Lexical acoustics:
Linking phonetic systems to the higher-order units they encode

By

Charles H. Redmon

Submitted to the graduate degree program in Department of Linguistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Committee members

Allard Jongman, Chairperson

Joan Sereno

Annie Tremblay

Jie Zhang

Michael Vitevitch

Date defended: October 30, 2020

The Thesis Committee for Charles H. Redmon certifies
that this is the approved version of the following thesis :

Lexical acoustics:
Linking phonetic systems to the higher-order units they encode

Allard Jongman, Chairperson

Date approved: October 30, 2020

Abstract

In the canonical analysis of the acoustic and perceptual structure of phonetic systems, experiments and modeling procedures are based on a balanced inventory of phones whose characteristics and relational structures are largely treated as independent of the wider system of higher-order distinctions in which they occur. The present study presents an alternative, *lexical* approach wherein the fundamental unit of phonetic analysis is rather a relation between contrastive words, thus shifting the state space of the system from a small, largely symmetric inventory of phones, to a large, heterogeneous ensemble of real-word contrasts in the lexicon. The ultimate goal in developing such an approach is twofold. First, by directly linking the study of phonetic systems to the higher-order units they encode, phonological and acoustic asymmetries, as well as interactions with top-down information, are implicit in the model, thus reducing the dependence of phonetic analyses on phonological inventory definitions, and providing for more scalable estimates of phonetic characteristics to real-word production. Second, the comparison of a lexical approach to the canonical inventory approach provides an explicit test of the consequences of the inventory assumption for models of cue integration and the acoustic/perceptual structure of the phonetic system.

Using the English obstruent system as a test case, several discrepancies in the acoustics and perception of obstruent contrasts in the two systems, as well as the link between the two in the form of models of cue integration, were uncovered. The data used as the basis for acoustic analyses and cue-integration models in the lexicon and inventory was based on a large sample of real words and controlled syllables, respectively, produced by a single native English speaker. Perception data came from two experiments

measuring listener word recognition of minimal-pair contrasts in noise.

For example, the presence of voicing in the signal is a more robust indicator of word-final voicing contrasts in controlled syllable data than in real words, whereas coarticulatory cues to obstruent place of articulation, such as in F2 at vowel onset/offset, are more robust among lexical contrasts. Perceptually, due to asymmetries in phonological distributions in the lexicon, the overall accuracy of listeners in word recognition depends primarily on their recognition of the plosives [p, t, k, b, d] and the fricatives [f, v, s, z], where the sibilants are generally robust in perception, while the plosives and labial fricatives are a common source of errors.

When directly linked in statistical models of cue integration in the inventory and lexicon—both of *ideal* discrimination and recognition patterns by human listeners—points of disagreement in cue weights can be classified into *distributional*, *acoustic*, and *composite* sources. Examples of distributional disagreement include the downweighting of relative F3 amplitude in the inventory model due to the reduced prevalence of sibilance contrasts, while acoustic disagreements include cases such as the poor scaling of vowel-offset F2 due to coarticulatory differences between controlled syllables and real words. Composite disagreement arises from a combination of these two sources.

Finally, the structure of the system of obstruent contrasts in the lexicon was studied by measuring the response of several measures, including minimal pair count, functional load, edge disjoint strength, and average path length, to noise- and cue-perturbation. Overall, the impact of background noise on plosive and sibilance contrasts has the greatest impact on the system, meaning cues such as noise duration, spectral tilt, F1, and relative F3 amplitude are the most influential in maintaining lexical distinctions.

The combination of these results challenges the viability of the independence assumption implied by the study of phonetic inventories, and provides empirical motivation for the study of phonetic systems as linked to the higher-order structures they encode.

Acknowledgements

This work would not have been possible without the generous support of many people over the past several years. First and foremost, I must thank my advisor, Allard Jongman, who encouraged me from the very beginning, supported me through successes and failures, and who has been an unwavering inspiration in the pursuit of hard problems and unasked questions in the field. To my committee members, Joan Sereno, Annie Tremblay, Jie Zhang, and Michael Vitevitch, thank you for generously offering your time and attention to this project, both in reading and critiquing the hundreds of pages below, and in providing feedback as the work has evolved over the years.

Several members of the University of Kansas and University of Alberta research communities provided direct support on this dissertation. Anna Holmes, Elizabeth Zollner, Sandy Kroeker, and Jarod Mariska spent painstaking hours staring at hundreds of waveforms and spectrograms in order to provide a robust measure of inter-annotater reliability. Shirley May, Ivy Mok, and Matthew Kelley helped me recruit and run participants in the pilot perception experiment at the University of Alberta. Finally, Benjamin Tucker was the person who initially introduced me to the MALD corpus, provided access to audio and force-aligned segmentation data, and supported my visit to Edmonton to collect the initial pilot data for the study. This project could not have been done without this early help.

To all of my teachers and colleagues in the Linguistics Department, thank you for creating a positive environment in which to learn and grow, and for inspiring me with your own work, both in the lab and in the classroom. To my students, thank you for your enthusiasm and engagement with subjects I love so much, particularly phonetics

and statistics, which are not helped by my monotone delivery but which I can now introduce with much greater clarity and purpose thanks to your involvement. In particular, I have benefited greatly from sharing ideas with the members of the LING 850 seminar, my colleagues in the Center for Research Methods and Data Analysis, and the members of the Network Science Working Group at KU. From this group I especially wish to thank Seulgi, Maite, Quentin, Yufu, Mingxing, Katrina, Rustle, Lucy, PJ, Ben, Zack, Chong, Eric, Trevor, and John.

Finally, to my family and friends, who endured many long-winded discussions over the years that have at least taught me two things: (1) that making research more broadly accessible is a neverending pursuit, and (2) that I have a possible future as a non-habit-forming sleep aid. Nick and Lena, you have been there through it all, and always helped me keep the work in perspective. Kathleen and Peter, thank you for never letting me get too big of a head, and Mom, thank you for being the rock that has given me the confidence to travel across the globe for the past decade in pursuit of this career. Lastly, Seulgi, you are my first sounding board for new ideas, my last refuge in times of crisis, and the reason any of this thesis made it to print. I could not have done it without you.

Financial support for this work was provided by the National Science Foundation (Dissertation Improvement Grant #BCS1918404).

Contents

List of Figures

List of Tables

Chapter 1

Introduction and literature review

1.1 Motivation and problem statement

The past two decades have seen a surge across the speech sciences in the development of high-dimensional statistical and computational models of the uptake of phonetic information from the acoustic signal, be it by human or computer. Much of this work is connected to wider efforts in the speech sciences—linguistics, psychology, speech pathology, and computer science/engineering, among others—to explicitly integrate linguistic structure at multiple levels (phonetics, morphology, semantics, etc.) and from multiple sources (auditory, visual, contextual, etc.) in model and theory building. Notable achievements of such integrative approaches include advances in automatic speech recognition (syntactic and semantic information now regularly augment traditional *n*-gram-based language models; Chelba & Jelinek, 2000; Tan et al., 2012), general models of psycholinguistic behavior (see Davis & Johnsrude, 2007, for review), and clinical speech intervention (Başkent, 2012, shows, for example, the use of top-down information to resolve information loss from signal distortion in cochlear implants).

And yet present models of phonetic contrast, and the acoustic information underlying its detection in the signal, remain largely unaffected by such developments. Cues are identified and weighted according to their ability to predict patterns in listener perception of carefully controlled, balanced sets of nonword syllable stimuli (or real-word exemplars of such syllables), where the constituents of those stimuli are drawn from the *sound inventory* of the language in question.[1]

Such an approach therefore assumes that the phonetic system, at least as far as its acoustic

[1]The term *sound inventory*, sometimes referred to as the *phone inventory*, or simply the *inventory*, is used throughout this work to refer to either a set of phonemes or a set of allophones in a particular language. Where this distinction is of theoretical consequence, the precise nature of the inventory (phonemic or allophonic) will be specified.

structure is concerned, is fully described by a sound inventory which is independent of the relative usage of those sounds in the composition of words in the language. The present work, adopting English obstruents as a case system, scrutinizes the effect of this assumption on the estimation of contrastive acoustic information in the speech signal, and ultimately presents a new approach to the study of phonetic contrast and the acoustic characteristics that realize such distinctions. Further details on the analytical form and logic behind this approach will be presented in Sections 1.4-1.6. But first we review the literature informing the approach, as well as introducing the primary research questions in Section 1.3.

1.2 Background

This project is informed by a wide body of research from several fields, but given the considerable overlap between fields, the review below has been structured as an outline of the layered assumptions behind, and information required for, navigation of the distinction between the canonical approach to the study of linguistic sound systems, and the approach developed in this dissertation. At each layer perspectives and contributions from several fields are incorporated.

1.2.1 Speech signal encoding, contrast, and phonetic inventory definitions

From the outset, the working assumption among linguists has been that the speech signal is encoded in a sequence of indivisible phonetic units; that is, words/morphemes are formally *strings*: finite sequences of symbols drawn from a finite alphabet Σ. In early philological work, including that of Sanskrit grammarians such as Patanjali in the 2nd century BCE,[2] and continuing through the work of European linguists of the 19th century CE such as Alexander Melville Bell, Baudouin de Courtenay, Henry Sweet, Paul Passy, and Mikołaj Kruszewski (Krámský, 1974), the most critical defining characteristic of elements of Σ was that their substitution could result in a distinct

[2]Most sources place the origin of the concept of phonemic contrast in the Sanskrit grammarians' use of the term *varṇa-sphoṭa* to describe a minimal speech unit that is itself devoid of meaning but when substituted with another such unit may result in a new word (Krámský, 1974; Fischer-Jørgensen, 1975).

word/meaning, and that for such a purpose they need not be decomposed further.

Bloch (1948) synthesizes much of this work into a more exhaustive statement of phonetic contrast as a relation between two segments which are phonetically distinct and occur in shared segmental environments but are not in free variation, where free variation is defined as the occurrence of phonetically distinct segments in the same phrase (i.e., yielding the same linguistic meaning). Later formulations include the distinctive features of Jakobson et al. (1951), whose binary values directly implement phonetic contrast as a kind of switch between opposing values, though at a feature level rather than segmental level, and the definition of the speech code in information theory (Shannon, 1948), where the elements of the code do not matter, just the number and distribution of distinct units in the code (formally, all codes of the same size and distribution are isomorphic). Here *distinct* implies a change in symbol changes the transmitted message.

In what we will refer to as the *canonical* approach to the study of sound systems in phonetics, the above definitions of contrast, which all reference the higher-order units whose meanings phonetic categories serve to distinguish, are typically simplified to category distinctions within the phoneme inventory. This simplification then carries over into general descriptions of the system (e.g., that English has a three-way contrast in plosive place of articulation), the assumption being that by reducing a complex ensemble of phonetic distinctions in context (i.e., in particular words and positional/prosodic environments) to an inventory of phonemes/allophones, critical phonetic information has not been lost. Such a result may in fact be the case, but without directly testing this assumption such a position is theoretically untenable.

1.2.2 Speech perception and acoustic cue weighting

The study of acoustic properties of speech sounds has its modern origins in the establishment of Bell Laboratories in 1925, and Haskins Laboratories in 1935. From this origin well into the 1980s, the general paradigm was to search for singular, invariant acoustic cues to a given phonetic contrast (see Jongman & McMurray, 2017, for review), such as VOT for onset plosive voicing (Lisker & Abramson, 1964), spectral shape in plosive place of articulation contrasts (Stevens &

Blumstein, 1978), and noise duration for fricative-affricate manner distinctions (Jongman, 1989). However, with Repp (1982), and related work in Whalen (1984) and Parker et al. (1986), the focus shifted to the analysis of multivariate cue structure, in particular applying more probabilistic frameworks of cue integration to the problem of parsing the acoustic signal (Massaro & Oden, 1980; Nearey, 1997; McMurray & Jongman, 2011). Nevertheless, the object of study remained controlled, balanced sets of nonword syllables, or arbitrarily chosen real-word minimal pairs, again balanced for the presence of target contrasts in controlled segmental/prosodic contexts.

From this framework several important results were obtained, including categorical perception (Liberman et al., 1957), coarticulatory dynamics (Öhman, 1966), and the nature of audio-visual integration in speech perception (McGurk & MacDonald, 1976); however, these results treat the phonetic system as independent of the higher-order structures it implements, meaning that the cues used in higher-order processes like word recognition are assumed to exhibit a constant structure, at least for equivalent phonetic contexts. In the next section we review the architectures of several models of spoken word recognition, focusing in particular on how they structure the speech input.

1.2.3 Models of spoken word recognition

Assuming a reliable link between the acoustic signal and a set of lower-dimensional phonetic units, the next task for the listener is to map this set onto meaningful units such as morphemes and words. Early models conceived of the lexicon as a dictionary, where words might be 'looked up' by narrowing down the set with each successive phoneme (Forster, 1976). The initial formulation of the Cohort model (Marslen-Wilson, 1987) largely adopts this perspective, though with greater clarity regarding the form of competition between candidate words as the signal unfolds. Both the Shortlist model (Norris, 1994) and the TRACE model (McClelland & Elman, 1986) similarly operate over a discrete string representation of the signal, though not as a series of phonemes, as in Cohort, but a sequence of changing distinctive feature bundles. Both TRACE and Shortlist employ connectionist architectures to relate the input to the target word, though the two differ critically in both the nature of feedback (TRACE allows word activation to reinforce an input phoneme layer; Shortlist

has no such mechanism) and competition (Shortlist, as the name implies, limits competition to a reduced set of candidate words, whereas TRACE assumes every word in the lexicon competes for recognition at all times). Shortlist B (Norris & McQueen, 2008) operates over an input closer to the acoustics, where just like the neighborhood activation model (NAM) of Luce & Pisoni (1998), similarities between input, target, and competitors are computed from behavioral data (the former uses gating tasks, the latter recognition in noise). Finally, there is the family of models described under Adaptive Resonance Theory (ART Grossberg, 2003), in particular the ARTPHONE model (Boardman et al., 1993; Grossberg et al., 1997), which attempts to model speech perception in a manner which is more compatible with existing data on the form and function of human memory systems. This model, while generally utilizing phonemes or phoneme-like units in its simulations, formally operates on a unit called the 'list chunk', which represents a span of the speech signal that can be held in working memory and serves to further activate higher-order units, including words in the lexicon.

All of the above cases assume an abstract encoding of the lexicon, however, but there are alternative approaches, such as the Lexical Access from Spectra (LAFS) model (Klatt, 1980) and a variety of Exemplar-theoretic models (see Johnson, 1997, for example), which assume the lexicon is not abstractly encoded, and that lexical access arises from a kind of acoustic distance matching between the input and stored acoustic instances of previously heard words. Yet even for such cases, the input cues are largely held to be the same; namely, a set of acoustic parameters is defined and weighted largely on the basis of inventory contrasts independent of their wider distribution in the language.

An approach which is perhaps intermediate between the two – abstractly encoded lexicon versus acoustically defined lexicon – is the FUL model (Lahiri, 1999; Reetz, 1999; Lahiri & Reetz, 2002), where the lexicon is assumed to consist of words encoded in sparse feature matrices that exclude all those features such as [CORONAL] that are considered underspecified in phonological theory (usually based on a combination of evidence from phonological typology and speech perception/production). Because the FUL model attempts a continuous matching of acoustic signal

properties (generally based on LPC decompositions of spectra computed from overlapping 20 ms windows) to acoustically defined features, which are then mapped onto the lexicon, FUL is less dependent on phonemic or allophonic segmentations of the signal. However, as with all of the above models, in practice the definition of the feature system (and its mapping onto the acoustic signal), particularly evident in cross-linguistic comparisons, is ultimately inventory-based. That is, attributes of [LABIAL], [CORONAL], and [DORSAL] features are studied independent of the words they serve to distinguish, despite the architecture of the recognition system being non-phonemic.

These perspectives treat the sound system as relatively simple: a set of symmetric relations between homogeneous units. However, unless speech sound categories/features are assumed to be innate (e.g., given in Universal Grammar, as proposed in Generative theory Chomsky & Halle, 1968) they must be learned from their occurrence in higher-order linguistic structures, in which case the acoustic cues latched onto by listeners in learning the sound system of their language becomes a complex function of the distribution of sounds in the words of the language, where the relative functional role of a given cue in facilitating successful message transmission is of particular importance (Boersma, 2012). Next, we provide background on one such alternative: a complex-systems approach.

1.2.4 Speech as a complex system

Broadly, a complex system is "a system composed of many interacting parts, such that the collective behavior of those parts together is more than the sum of their individual behaviors," (Newman, 2011). For the present system, the interacting parts are considered to be the words communicated, with an interaction defined as an acoustic distinction between words that is minimal enough to be confused (and thereby transmit information about the critical acoustic information listeners must attend to). The 'collective', or global behavior of the system is simply that in aggregate it serves to maintain sufficient acoustic distinctions between words such that successful communication is not compromised. For example, many phonetic phenomena, such as tonogenesis (Matisoff, 1973), chain shifts in vowel systems (Labov et al., 1972), and general trading relations in acoustic cues

(Repp, 1982), depend on this global characteristic of the system for their explanation; namely, the system must to a certain degree resist homophony so as to avoid communicative ambiguity. However, it should be clear that in focusing on the lexicon the present approach assumes homophony resistance is ultimately about preserving distinctions in meaning in the transmission of speech.[3]

There are many candidate models of such systems, such as agent-based modeling and genetic algorithms, two of several related simulation-based approaches Wedel (2004; 2007; 2012; 2017) and Blevins (2006) have most prominently applied to speech, but an analytic account that is necessarily prior to such computational approaches is that of modeling the topology of the system, typically as a graph/network (see Crane, 2018, for elaboration of this point). Briefly, a graph is a set comprised of a vertex/node set V, which defines the units of the system being represented, and an edge set E containing (un)ordered pairs of distinct vertices, where such pairs are defined for all instances of an interaction or relation between units (vertices) that is fundamental to some aspect the organization, behavior, and evolution of the system.

Vitevitch (2008) was the first to apply graph theory to the phonological encoding of the lexicon, emphasizing in particular the role of complex phonological similarity structures in accounting for both speech perception behavior and the general organization of speech systems (see Dautriche et al., 2017, for a recent extension of this approach to the cross-linguistic comparison of phonological systems). Figure 1.1 illustrates Vitevitch's approach, which while retaining the core assumption in phonological theory that minimal distinctions between words are marked by oppositions between phonemes (even if their encoding in the lexicon is a matter of debate), illustrates and makes mathematically precise the emergence of complex heterogeneous structure as one 'zooms out', as it were, from a single minimal pair to the remainder of minimal lexical contrasts with that set, and the contrasts with the subsequently larger set, and so on.

Note again that no new relations are being proposed with respect to existing linguistic theory; the graph merely serves to simultaneously represent (in a dependent fashion) all minimal pairs in

[3]The resistance of homophony, or the preservation of contrast, can also be formulated purely in an inventory sense, yet to do so is to ignore both the origin of the units that comprise the inventory (recall that phonemes are defined on the basis of minimal phonetic distinctions between words) and the ultimate purpose of avoiding homophony (to reduce ambiguity in communication).

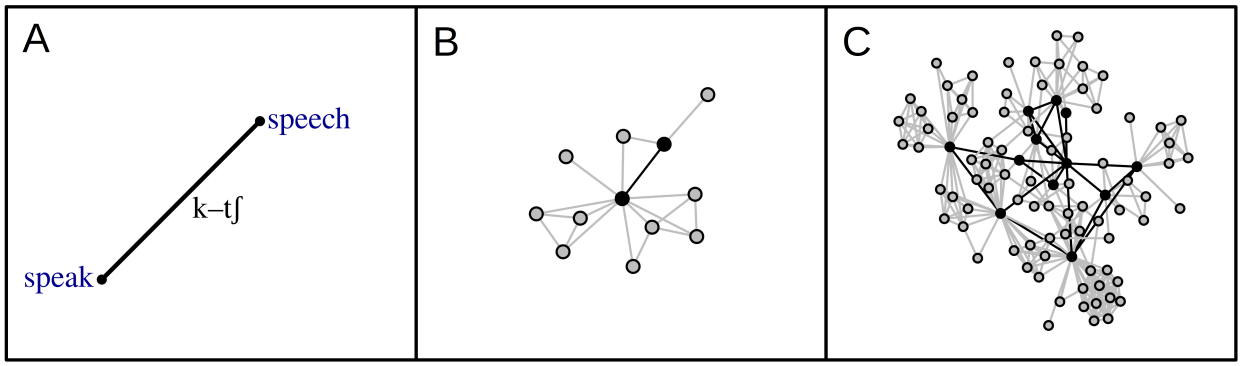

Figure 1.1: Illustration of the composition of a representation of the lexicon as a graph by expanding out from a single minimal contrast such as *speak-speech* (Panel A), which contrasts the obstruents [k] and [tʃ] in coda position, to the full set of minimal contrasts with each member of the pair (Panel B), and further to minimal pairs of that set (Panel C).

the lexicon. By examining the topology of this set of relations, we are able to identify the relative role different sets of contrasts play in the maintenance of acoustic distinctiveness in the lexicon as a whole. As such, this perspective is potentially informative for all models listed in the previous section, because even for the Exemplar-theoretic approaches, the general characteristics of oppositions between singular tokens should mirror to some extent the denser array of stored instances of those tokens. For more canonical approaches, where the encoding of words is ultimately assumed to be of an abstract, discrete form, the implications are more straightforward. By considering the phonetic system as a complex function of lexical structure, derived cues and dimensional weights remain fundamentally tied to the words whose meanings they serve to transmit, thus providing for models such as Shortlist or NAM acoustic information which is potentially more scalable to the word recognition problem.

Finally, we must clarify that we do not assume that the above approach precludes the generalization of phoneme-like categories, or even an inventory, from this complex ensemble of lexical interactions. What this approach does imply is that such generalizations cannot be held to be independent of the lexicon from which they derive, and as such the inventory does not represent a complete description of the system.

1.3 Research questions

Two primary research questions are posed in service of the goal stated in Section 1.1; i.e., to define the acoustic structure of the English obstruent system as a function of the ensemble of words in the lexicon distinguished by such sounds. First, we ask whether and to what degree canonical estimates of acoustic cue parsing—both their relative weight and multivariate structure—scale to the lexicon as a whole. This scaling assessment is measured by the predictability of listener word recognition from a model whose architecture upholds the aforementioned inventory assumptions of homogeneity, symmetry, and independence (discussed in more detail in Chapter 4), as compared with a model derived directly from the lexicon.

Second, we ask what a lexically-dependent complex speech system implies for the stability and resilience of information encoded in the signal. That is, we seek to assess whether the distribution of cue weights implied by the lexical model, as compared with an inventory model, more closely coheres with two core assumptions of any theory of language: (1) that information transmission through speech production and perception is resilient to various perturbations, including variation in background noise; and (2) that the system that broadly meets the communicative needs of a given community is relatively stable, meaning that while as a rule it will be subject to change over time, such change is expected to be slow in its progression.

Regarding the first question, we hypothesize that parameters that are heavily influenced by coarticulation (primarily those cuing place distinctions), such as formant frequencies or broad spectral characteristics at segment boundaries, will scale more poorly than those primarily indicative of changes in manner of articulation or voicing. This is because the balance in consonant-vowel combinations, as well as the reduced naturalness of nonword productions, skews the distribution of coarticulatory relations away from that observed in the real words which comprise the lexicon. Further, given the considerable asymmetry in contrast occurrence and phonotactic distribution in the lexicon, we expect relative cue weights to differ substantially between the two models, especially for more phonotactically constrained subsets of the system.

In answering the second question, we apply the general perspective that a fundamental driver

of organization in speech systems is perception (Martinet, 1952; Lindblom, 1990; Ohala, 1993), as successes and failures in a listener's receipt of the message, and thereby in communication more broadly, provide information about whether and how the speaker must adapt in speech production, and thus what aspects of the system must change and what can stay the same. This perspective leads to the hypothesis that lexical structure will complement fundamental acoustic characteristics of the speech sounds employed in a given language (in our case, English), meaning that more acoustically salient cues should receive more frequent and less phonotactically restricted use, while less salient cues should be sparser and more variable in both production (how they are realized in the signal) and perception (how they are utilized by listeners). Along these lines, we expect higher-order characteristics such as lexical frequency to further complement the information structure of the system; e.g., we do not expect two words of great acoustic similarity and low cue salience to also be of comparable lexical frequency and part of speech, because such configurations would be highly confusable, and thus more subject to change, and ultimately less likely to constitute significant portions of the system.

1.4 Overview of the analytical approach

1.4.1 Acoustic cue weighting and contrast distribution asymmetry

In considering the role of the lexicon in estimating the acoustic information underlying the obstruent contrast system in English, the first question to ask is *how does consideration of the lexical distribution of phonemic contrasts, which is marked by substantial asymmetries in both category and context frequency, change the weighting of acoustic cues relative to symmetric assumptions?* Baseline expectations are derived both from the large body of research on acoustics and perception of obstruents in controlled nonsense syllables, and from new simulations on both controlled nonword CV, VC, and VCV syllable data.

Next, the lexical distribution of obstruent consonants in English is incorporated as a category bias in the above syllable models. That is, in modeling category prediction from acoustic cues,

items in the training data for the model are drawn in proportion to their relative frequency in the lexicon, forcing the model to attend more to cues that have greater utility for more frequent contrasts. For example, cues to [s, ʃ] and [s, f] distinctions, such as spectral peak frequency and relative F3 amplitude, will be more heavily weighted than those to [f, θ].

Relative differences in cue rankings between the two classes of models serve to answer the above question, where assignment of greater weight to a given cue in the lexically biased model, relative to the baseline model, suggests that the broader role of that cue in distinguishing items in the lexicon has been underestimated by the canonical approach. Similarly, the baseline model which assigns equal importance to each contrast might overestimate the utility of certain cues that only serve to distinguish contrasts playing a limited role in the lexicon (e.g., Rise Time as a cue to the [ʃ, ʧ] contrast). Yet, the biased model remains limited to acoustic characteristics of controlled nonsense syllable productions, assuming such characteristics remain relatively constant across the lexicon. The next stage of analysis addresses this assumption by looking at real word data from a model lexicon of English.

1.4.2 Modeling phonetic contrasts under relaxed inventory assumptions

Given the theoretical source of phonemes in lexical oppositions—i.e., through the identification of minimal pairs—it follows that the same information relevant to the independent recognition of consonants must be present directly in acoustic differences between the minimally distinct words from which those categories derive. By directly predicting patterns in listener word recognition over a model lexicon of over 26,000 words (Tucker et al., 2018), we are able to do two things. First, by studying the acoustic and perceptual characteristics of obstruents produced in real words, we can test the extent to which acoustic measurements on controlled syllables are sufficient to predict recognition patterns on real words. Here we are not simply asking whether the acoustic properties of the two data sources are comparable (e.g., is the vowel length difference between /bɑt/ and /bɑd/ similar to the typical vowel length difference observed for word-final /t, d/ contrasts), but rather whether the same relative cue rankings emerge for models based on nonsense syllable or

real word acoustics.

Second, while minimal pairs are defined phonemically, our focus on obstruents as a class, and word pairs minimally distinguished by elements of that class, allows us to relax any assumptions about the internal organization of the obstruent system. That is, we simply observe that two words differ in some interval attributable to an obstruent speech sound (turbulence in airflow, noise in the acoustic signal), and aim to predict the likelihood that listeners will detect that difference based on the acoustic features parsed from that interval of the signal.

1.4.3 The role of higher-order information in signal parsing

The role of higher-order information in speech perception has been considered for over half a century—for instance, Howes demonstrated in 1957 a direct relationship between word frequency and intelligibility in noise—yet the role of such information in modulating the relative weight of acoustic cues to phonetic contrasts remains largely unexplored. For example, while various sources of top-down information have been shown to apply in the perception of stop consonant voicing in word recognition (Davis & Johnsrude, 2007), the role of voice onset time (VOT), fundamental frequency (f0), and first formant (F1) transition in delineating contrasts between /p, t, k/ and /b, d, g/, respectively, is generally assumed to be independent of such information. Following the initial prediction of word recognition errors from acoustic cues alone, we supplement the model with parameters reflecting the stimulus word frequency (both absolute, and relative to the frequencies of competitor words).[4] To the extent that such information reduces model error on listener response patterns, the resulting acoustic contribution should represent a more ecologically valid estimate of the role of the purely acoustic component in speech transmission.

[4]The simplicity of our present measure of higher-order information has to do partly with our experimental design in investigating only word production and recognition in isolation. Were we to study word recognition in larger utterance contexts, the contextual predictability of the target relative to competitor words (e.g., via *n*-gram/cloze probability, or syntactic/semantic felicity) must also be considered.

1.4.4 Distributed acoustic information and system resilience

Given a model of the system of obstruent contrasts as distributed over the English lexicon and modulated by higher-order sources of information, we are then in a position to ask: *how resilient is the system to perturbations of the acoustic signal?* In other words, given uncertainty in either specific cues parsed from the signal, or entire frequency bands, to what degree is the system able to retain its functional role in preventing lexical homophony. Further, by simulating the response of the system to acoustic perturbations we also arrive at estimates of what might be called the *global* information contributed by certain characteristics of the signal. This information may prove to be a valuable supplement to our present knowledge of what listeners attend to in the speech signal, particularly in cross-linguistic comparisons where lexical structure differs to a greater degree than do category inventories.

1.5 Logical challenges to the general approach

The approach outlined above takes important steps toward moving away from an inventory-centric view of the structure of phonetic systems. However, in the process of incrementally relaxing canonical assumptions, two key logical challenges appear to have been introduced. These challenges, and our approach to their solution, are addressed here so that certain anticipated points of confusion may be avoided in the interpretation of the analysis presented in Chapters 2-5.

The first challenge to the paradigm developed in this dissertation is an apparent tautology in the manner in which acoustic cues are identified and related to acoustic-perceptual distinctions between words in the lexicon. The acoustic cues from the literature have been defined precisely to distinguish between phonemes, which in turn are defined based on 'minimal' distinctions between lexical items. Thus, any relation between such cues and the lexicon remains fundamentally tied to the initial definition of phonological contrasts, especially when lexical distinctions are similarly restricted to phonological relations such as the minimal pair. We address this problem in two ways.

First, by studying English obstruent contrasts we assert that while such a distinction initially

rests on a phonological formulation, we are ultimately accounting for oppositions between semantically distinct lexical items where the primary acoustic distinction is localized to a shared interval of turbulence/noise in the signal. Thus, the aerodynamic and acoustic salience of the obstruent-sonorant distinction allows us to adopt the assumption that the signal can at least be parsed into approximate regions of turbulent or laminar flow, the remaining task being to distinguish acoustically two regions that are perceptually distinct and whose substitution yields a change in meaning.[5]

Second, even in the absence of a complete decoupling of lexical acoustics from formal phonological relations, this study provides information on how cue weights and the general perceptual structure of the phonetic system change when the broader distribution of speech sounds is considered. That is, even for perspectives which uphold the existing phoneme inventory, this work offers an alternative account of the acoustic information which underlies that system: one that is more compatible with perspectives incorporating *functional load* into phonological theory.

The second challenge to be addressed in a lexical approach to phonetic systems is the apparent circularity in (1) fitting a model to identification and confusion patterns between words, (2) identifying the necessary and sufficient acoustic cues (and their relative differences in weight/priority) for model fit, and finally (3) inferring critical characteristics of the structure of the cue system within the lexicon based on model output from (1) that necessarily has been optimized for that very task. In other words, under an argument restricted to the procedure in (1-3), there is no external validation of the cue system. A parallel situation may be found in the canonical approach to phonetic experimentation, in cases where cue weights are determined purely from a model fit to listener identification/discrimination patterns to controlled sets of phonemic contrasts. For this reason, among others, it has become standard to augment such work with the measurement of listener responses to synthetic manipulation of those cues, thereby validating the relative role of the different cues in a manner that is external to the model-fitting procedure. Experiment 2 (Chapter 4) adopts exactly this approach to validating cue weights, and uses a cross-splicing design to study both cue enhancements and reductions in lexical contrasts.

[5]There are of course certain contrasts like [b, w] that are acoustically and perceptually motivated, but which span the obstruent-sonorant distinction. These issues will be addressed in greater detail in Chapter 2.

1.6 Summary of thesis argument structure

Integrating the motivating literature and analytical approach with the issues outlined above, we arrive at the following general structure of the argument presented in this thesis.

1. The primary basis of phonetic analysis is phonemic. The premise that phonetics presumes phonology may appear counter to expectations based on their reversed ordering in levels of representation in linguistic theory. However, as much of the focus of phonetic analysis is on the acoustic and articulatory features critical to communication in a given language, phonemes are a natural unit of manipulation in the design of stimuli for production experiments, and in the analysis of identification and discrimination data from perception experiments (Pike, 1972).[6] Further evidence for the primacy of phonemic description in phonetic analysis comes from the *Principles of the International Phonetic Association*, which state that "The sounds that are represented by the [IPA] symbols are primarily those that serve to distinguish one word from another in a language." (Roach, 1989). The International Phonetic Alphabet is not strictly composed of symbols representing phonemic sounds (the 1989 report specifically calls attention to the introduction of diacritics for the narrow transcription necessary for many clinical applications), but phonemicity is the primary basis on which most symbols are justified for inclusion in the alphabet.

2. Phonemes by definition perform a lexically discriminative function. Whether mentalist or physicalist definitions of the phoneme are adopted, a necessary condition for any phonemic encoding is that it preserves, in symbolic form, distinctions between words that are perceived by native speakers of the language to be neither auditorily nor semantically equivalent (Krámský, 1974; Mugdan, 1985).[7] That is, the phoneme inventory of a language cannot be derived without some reference to the lexicon, and the points of equivalence and distinction between lexical items that allow the encoding to successfully transmit meaning through the speech signal.

[6] Some attempts at phonetic parameterizations that are not dependent on segment boundaries have been made (see for instance Pike, 1972; Catford, 1977); however, the majority of analytical approaches remain linked to phonemic segmentation in some way.

[7] The *auditory* condition is necessary to allow for homophones to share a common phonemic transcription.

3. The distribution of phonemic contrasts in the lexicon is non-uniform, with contrasts differing in their functional load. The non-uniformity of phoneme usage in the lexicon has been documented for nearly a century (Dewey, 1923), and this fact has received formal phonological attention since at least 1952, when Martinet argued the relative frequency of a given contrast was a relevant consideration (alongside articulatory and perceptual factors) in the prediction of the likelihood of that contrast to merge over time. This consideration of the relative impact of a merger as a function of the relative frequency of a contrast is referred to as the *functional load* (FL) of the contrast, and can be extended to distinctive features by considering classes of contrasts; e.g., place and manner of articulation have a higher functional load than voicing in American English (Surendran & Niyogi, 2003).

⇒ ***If the speech system is optimized for transmission of phonologically encoded messages, then perceptual weighting of acoustic information in the signal must reflect this distribution.*** Premises 1–3 imply the conclusion that the identification and relative weighting of acoustic cues in the speech signal must reflect to some extent the broader distribution of *lexical* contrasts in the language, provided that the speech system has been optimized to transmit phonologically encoded messages (taken here to be words, for simplicity). This is certainly not the only valid position—one can imagine a counter-position that the phonological inventory, though derived from lexical distinctions, exists as an independent set, and that acoustic signal parsing (and signal encoding, on the production end) operates exclusively on that set, with information flow between all higher-order units being purely phonological. But if some degree of optimization is held to apply at the level of message transmission, then any cue system structured to give equal weight to each contrast in the inventory (a position implied by the canonical analysis of speech sounds as an independent entity) must be sub-optimal, and therefore cue weights/rankings derived in this manner would be invalid.

1.7 Organization of the dissertation

This dissertation is organized as follows. Chapter 2 examines the acoustic parametrization of obstruent contrasts in English, presenting results of measurement distributions and simulations of contrast discriminability on both nonword-syllable and real-word databases. Chapter 3 presents results of a closed-class perception experiment designed to reveal the general discriminability of a wide range of obstruent contrasts in the lexicon, as well as to provide for a cue-integration model that tracks listener behavior. Chapter 4 assesses the relative cue weights in both the aforementioned model tracking listener word recognition in noise, and an ideal perceiver model of perfect lexical contrast discrimination. Further, results of a second perception experiment utilizing cross-splicing to verify predictions from the cue-integration model are presented. The primary focus of Chapter 4 is on the relative agreement between cue weights derived from models operating over an independent inventory of speech sounds and models directly linked to real-word contrasts in the lexicon. Chapter 5 examines the structure of the system of phonological contrasts in the lexicon from the standpoint of their resilience to global and cue-level acoustic perturbation, as well as considering the more general mathematical problem of estimating the lexically discriminative information contributed by acoustic features of obstruent consonants in English. Finally, Chapter 6 discusses conclusions that may be drawn from the experimental results and model simulations, as well as future directions for research within this general paradigm.

Chapter 2

English obstruent acoustics

Outline

This chapter provides a detailed description of the acoustic characteristics of obstruent contrasts in the lexicon, and an analysis of how such properties compare with measurements from controlled syllables that are the basis of much of the literature on speech acoustics and perception. After each parameter is introduced, including key findings in the literature, measurement details, and the physiological basis for the parameter's viability as a cue to obstruent consonant distinctions, cue distributions are presented for obstruents in both word and syllable databases. Finally, within each class of acoustic parameters—*temporal*, *amplitudinal*, and *spectral*—the relative discriminative power of each cue across the range of contrasts in the inventory and lexicon is assessed in preparation for their later inclusion in statistical models of cue integration in Chapter 4.

2.1 Introduction

2.2 Acoustic parameterization of English obstruents
The general methodology for identifying physiologically based acoustic parameters is presented along with a list of the parameters used in the analysis of obstruent contrasts in the present study.

2.3 Acoustic data sources
The item structure and recording characteristics of three databases are described: a model lexicon (real words) and model inventory (controlled syllables) produced by the same speaker, and reference syllable data from two previous studies.

2.4 Temporal parameters
Acoustic characteristics of parameters that are primarily temporal in nature—consonant duration, vowel duration, closure duration, noise duration, voice onset time, voice cessation time, and consonant voicing percentage—are presented.

2.5 Amplitudinal parameters
Acoustic characteristics of parameters that are primarily amplitudinal in nature—burst presence, noise amplitude, and vowel amplitude—are presented.

2.6 Spectral parameters
Acoustic characteristics of parameters that are primarily spectral in nature—spectral peak frequency/amplitude, dynamic amplitude, spectral tilt, spectral shape, spectral dispersion, low-frequency energy, relative F3/F5 amplitude, f0, F1, F2, and F3—are presented.

2.7 Discussion

2.1 Introduction

One of the central challenges introduced by the formulation of the *lexical* rather than *inventory* discrimination problem is the greater dimensionality and heterogeneity of contrasts involving obstruent consonants in the former. For instance, under a model where acoustic features probabilistically determine a choice among a candidate set of phonetic categories/phonemes (such as in the fuzzy-logical model of Massaro & Oden, 1980, for instance), any consideration of the demands of discriminating items in the lexicon forces a more expansive set of output categories to enter the model than are typically considered in controlled studies of single manner, voicing, or place classes. This is because if the model is to estimate, for example, the likelihood of misperceiving 'bird' [bɝd] as 'third' [θɝd], it needs to have been trained on such [b-θ] distinctions. Similarly, factors such as position (word onset), context (preceding the [ɝ] vowel), and higher-order information (the greater overall likelihood of 'third' introducing a bias that allows listeners greater tolerance of uncertainty in the acoustic input) must be considered in such a model.

With the exception of the final point on higher-order information, which will largely be addressed in Chapter 3, the above factors introduce unique challenges in parsing and integrating properties of the acoustic signal for obstruent identification in real-word recognition. The present chapter focuses on these challenges and describes the general acoustic properties of obstruent contrasts in the lexicon. In particular, we review a large set of acoustic parameters that have been proposed in the literature on English obstruent contrasts, discussing the descriptive and/or physiological motivations behind each parameter's introduction, addressing issues in parameter definition and measurement, and ultimately assessing each parameter's discriminative power in both controlled syllable and real word data used in the present study.

2.2 Acoustic parameterization of English obstruents

As noted in Chapter 1, a wide array of acoustic measurements have been proposed to distinguish various subsets of the English obstruent system, typically, though not exclusively, focused on de-

lineating categories along a single featural dimension (e.g., voicing, place, sibilance, etc.). This choice, while often for experimental control, clarity of argument, or simply convenience,[1] has in some instances constrained parameter definitions to be undefined for certain classes of sounds (e.g., voice onset time is canonically defined exclusively for plosives). Further, such constraints have led to inconsistencies between the theoretical coverage of a given parameter (the range of contexts where that parameter may be identified in the signal, based on its acoustic definition) and its experimental coverage (the contexts tested in the literature). For example, continuing with the example of VOT, existing acoustic definitions should include affricates and fricatives—they both have points of noise onset (the proper starting point for measuring VOT given that many plosives, particularly those in word-medial position, do not exhibit release bursts) and the timing of voice onset relative to this point can be used to distinguish voiced from voiceless in a manner parallel to that found for plosives.[2] However, studies of VOT remain restricted to plosives, which means that when all obstruents are viable candidates for identification or discrimination, it is unclear how the entire set will be distributed and ultimately perceived on the VOT dimension; e.g., must affricates and fricatives first be separated out by some other mechanism to prevent [ʤ] from being grouped with the voiceless plosives?

In addition to defining and reviewing the motivating literature behind each parameter used in the present study, the sections below address issues such as definition consistency and empirical coverage, particularly as they relate to challenges in multivariate cue integration in the generalized problem of spoken word recognition. The parameters in Sections 2.4-2.6 have been organized into three general classes—*temporal*, *amplitudinal*, and *spectral*—which characterize the primary signal dimension (time, amplitude, and frequency, respectively) along which a given parameter is measured. A complete list of parameters is provided in Table 2.1, including the features they have been shown or proposed to index, and citations for the source where they were first proposed, as

[1] Though as discussed in Chapter 1, the independence of features or cue dimensions was in some cases part of the motivating theory and constituted an assumption about the auditory processing and cognitive representation of speech (Liberman, 1993).

[2] The emergence of this pattern could also be due to inherent differences in noise duration between voiced and voiceless obstruents, commonly reflected in fricative distinctions, but this is beside the point as VOT remains defined, and therefore operative, for such cases.

well as subsequent sources presenting notable methodological revisions.

2.3 Acoustic data sources

Several data sources are used for the present analysis, including both controlled syllable and real-word data, which can be divided into two classes: *target* and *reference* data. The target data, which is used as stimuli in the perception experiments (Chapters 3 and 4) and as the basis for models of global cue information and distributed acoustic structure across the lexicon (Chapter 5), comprises approximately 27,000 words and 1,000 nonword syllables produced by a single speaker: a 28-year-old male native speaker of Western Canadian English. Thus, target data focuses on within-speaker characteristics of the phonetic system. Reference data integrates multiple existing databases of controlled syllables that have served as stimuli in perception experiments, and may therefore be used as a baseline for the cue weighting models of the inventory system presented in Section 4.4 of Chapter 4.

2.3.1 Target data

All target data were recorded from a 28-year-old male native speaker of Western Canadian English. The speaker had some phonetic training (he had taken an undergraduate phonetics class), but was instructed to read all items as naturally as possible. Items were recorded in isolation by presenting the speaker with a single word at a time on a computer monitor. Recordings were done at the Alberta Phonetics Lab in a sound-attenuated booth, using a Countryman E6 head-worn microphone with a 0 dB flat cap, powered by an Alesis MultiMix8 mixer and digitized on a Korg MR-2000S studio recorder at 44.1 kHz with 16 bit resolution. Finally, while the speaker and recording setup remained constant for both syllable and word data, the words forming the model lexicon were originally recorded in 2011–2012 as part of the Massive Auditory Lexical Decision project (Tucker et al., 2018), while the syllables were recorded in 2019 for use in the present study. For verification of minimal change in speaker production characteristics in the intervening years, the same speaker

Parameter	**Features**	**Literature**
Temporal Parameters		
1. Consonant Duration (DUR_C)	M, P, V	D55, PL60, U77
2. Vowel Duration ($DUR_{V1/V2}$)	V, M	L36, HF53, D55
3. Closure Duration (CD)	V, P	L57
4. Noise Duration (ND)	M, P, V	D55, HH56, G57, Y79
5. Voice Onset Time (VOT)	V	LA64, LA67
6. Voice Cessation Time (VCT)	V, M	D92
7. Voicing Percentage (VOI%)	V	D92
Amplitudinal Parameters		
8. Burst Presence (BURST)	M, P	F60, DSR77
9. Noise Amplitude (AMP_N)	M, P, S, V	S60, HS61
10. Vowel Amplitude ($AMP_{V1/V2}$)	V, M	LP59
Spectral Parameters		
11. Spectral Peak Frequency ($FREQ_{PK}$)	P	LDC52, HH56, S60, HS61
12. Spectral Peak Amplitude (AMP_{PK})	P	F54, OS83, STC96
13. Dynamic Amplitude (AMP_{DYN})	S, P	SM96
14. Spectral Tilt ($TILT_{C/V1/V2}$)	P, S	LGB84, STC96
15. Spectral Shape (SHAPE)	P, S	ERL98
16. Spectral Dispersion ($DISP_{C/VC/CV}$)	P, S	GM74, SB78, JWW00
17. Low Frequency Energy (LF)	V	HH56, MJ11
18. Relative Amplitude in F3 Region (AMP_{F3})	P	S85, HO93
19. Relative Amplitude in F5 Region (AMP_{F5})	S	S85, HO93
20. Fundamental Frequency ($f0_{CV/VC}$)	V	HF53, HAC70
21. First Formant Frequency ($F1_{CV/VC}$)	V	LDCG54, DLC55
22. Second Formant Frequency ($F2_{CV/VC/V1/V2}$)	P	LDCG54, DLC55, K89
23. Third Formant Frequency ($F3_{CV/VC}$)	P	O66, F73, STC96

Table 2.1: Acoustic parameter set, organized into *temporal*, *amplitudinal*, and *spectral* parameters, alongside the features each parameter serves to index (listed in order of relative importance), and the sources in the literature responsible for their original proposal and important later methodological contributions. The following codes were used for features and citations, respectively. M = manner of articulation, P = place of articulation, S = sibilance, and V = voicing. D55 = Denes (1955), D92 = Docherty (1992), DLC55 = Delattre et al. (1955), DSR77 = Dorman et al. (1977), ERL98 = Evers et al. (1998), F54 = Fischer-Jørgensen (1954), F60 = Fant (1960), F73 = Fant (1973), G57 = Gerstman (1957), GM74 = Gray & Markel (1974), HAC70 = Haggard et al. (1970), HF53 = House & Fairbanks (1953), HH56 = Hughes & Halle (1956), HO93 = Hedrick & Ohde (1993), HS61 = Heinz & Stevens (1961), JWW00 = Jongman et al. (2000), K89 = Krull (1989), L36 = Lloyd (1936), L57 = Lisker (1957), LA64 = Lisker & Abramson (1964), LA67 = Lisker & Abramson (1967), LDC52 = Liberman et al. (1952), LDCG54 = Liberman et al. (1954), LGB84 = Lahiri et al. (1984), LP59 = Lehiste & Peterson (1959), MJ11 = McMurray & Jongman (2011), O66 = Öhman (1966), OS83 = Ohde & Stevens (1983), PL60 = Peterson & Lehiste (1960), S60 = Strevens (1960), S85 = Stevens (1985), SB78 = Stevens & Blumstein (1978), SM96 = Shadle & Mair (1996), STC96 = Smits et al. (1996), U77 = Umeda (1977), Y79 = You (1979).

produced 300 words drawn from the item set Experiments 1 (Chapter 3). No notable deviations were observed between the 2011–2012 recordings and the 2019 recordings.

2.3.1.1 Model lexicon

A sample of 26,793 words was compiled by Tucker and colleagues (2018) from several sources: all unique word types in the Buckeye corpus (∼8,000; Pitt et al., 2007), an additional 9,000+ words from the English Lexicon Project (Balota et al., 2007), the 10,000 next highest frequency words in the Corpus of Contemporary American English (COCA; Davies, 2009), and 1,252 compound words from CELEX (Baayen et al., 1995). In addition to the particular requirements of the lexical decision project that motivated recording of these stimuli, for the present study this sample was deemed adequate to serve as a representative set of items sufficient to model the core phonetic structure of the mental lexicon shared by native speakers of English. That is, the items are considered sufficiently numerous and diverse to capture the range and frequency of phonetic contrasts listeners must distinguish to accurately perceive English speech.

In addition to the recording specifications detailed above, for the purposes of the lexical decision experiment, all words were scaled to have a 70 dB mean amplitude. Items were then force-aligned using the Penn Forced Aligner (Yuan & Liberman, 2008), and hand-checked by Tucker and colleagues. These annotations were then further edited manually in Praat (Boersma & Weenink, 2016) by the present author, with acoustic landmarks, such as burst onset, added for a complete parametric description of English obstruents according to the literature reviewed in Section 2.2. A subset of the data (10% of all items eligible for inclusion in the perception experiments; see Chapter 3 for details) was then independently annotated by two research assistants to assess the reliability of the measurements. Overall inter-rater reliability (IRR) was under 5 ms.

2.3.1.2 Controlled syllables

A total of 1,020 controlled syllables exhibiting obstruents in initial (CV), final (VC), and medial (VCV) positions were recorded for the present study for comparison with the real word data in

(Tucker et al., 2018). CV items were of the form: [*CV*b], where *C* is the set of all permissible obstruents in word-onset position—namely, [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, h, v, ð, z, ʒ]—and *V* is the set of monophthongal vowels [i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u], yielding 170 items × 2 repetitions = 340 CV syllables. VC items were similarly constructed with [b] as the onset consonant, 10 monophthongal nucleus vowels, and 16 offset consonants ([h] is excluded from the above set as it is phonotactically illicit), yielding 320 VC syllables (160 items × 2 repetitions). Finally, VCV sequences were constructed with the form [bV_1CV_2b], were V_1 is the same monophthongal vowel set, *C* is the 17-obstruent consonant set, and V_2 was constrained to match V_1 itemwise (i.e., only symmetric vowel contexts were recorded). In addition to the 340 items generated from this template (17 consonants × 10 vowels × 2 repetitions), 20 items were added to elicit the alveolar flap, [ɾ]. Such items were of the form: [ˈb*V*ɾəb], with *V* varied between the 10 monophthongs and stress always on the first syllable, and were repeated twice.

Acoustic signals were manually edited in Praat following the same procedures adopted for the real word data in the model lexicon. Further, to match the real word data and approximate stimulus pre-processing characteristics from the literature on controlled syllable perception, all items were normalized to 70 dB mean intensity.

2.3.2 Reference data

The following corpora were used in the construction of a reference data set. For controlled syllables, data from the California Syllable Test (CaST; Woods et al., 2010) were used. All controlled syllables were recorded from phonetically trained speakers. For the CaST, four speakers of Midwestern American English (2 female, 2 male) were recorded producing 20 onset consonants (including 16 obstruents, [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, h, v, ð, z]) in 3 vowel contexts [i, a, u] with 20 coda consonants (including 15 obstruents, [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, v, ð, z]). Each speaker produced 4 repetitions of the 1200 unique syllables, out of which the 2 best exemplars were chosen based on intelligibility. All recorded were made in a sound-attenuated booth with an AKG C-410 head-mounted microphone sampled at 44.1 kHz with 16 bit resolution. Finally, for purposes of the

perception experiment in Woods et al. (2010), stimuli were randomly normalized to between 70 and 75 dB mean intensity.

This 9600-syllable database was then reduced to a reference set for the present study consisting of 360 items that closely match the form of the CV and VC sets in the target inventory data: 192 [*CV*b] items (16 onset obstruents × 3 vowels contexts × 4 speakers × 1 repetition) and 180 [b*VC*] items (15 coda obstruents × 3 vowels × 4 speakers × 1 repetition).[3]

For the VCV reference data, we utilize the test stimuli from the 2008 Consonant Challenge organized at *Interspeech 2008* (CC08; Cooke & Scharenborg, 2008), which comprises 384 items chosen for presentation to listeners from much larger set of 12,096 recordings (24 consonants × 9 vowel contexts × 2 stress types × 28 speakers). The 24 consonants used to compose this database were the following: [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, h, v, ð, z, ʒ, m, n, ŋ, ɹ, l, w, j]. The 9 vowel contexts were formed from every possible combination of the corner vowels [i, a, u].[4] That is, unlike in the target data, vowel contexts were not symmetric in the reference data. Further, Cooke & Scharenborg (2008) did not record the alveolar flap [ɾ] as part of their study, which is both a part of the inventory data recorded from the target speaker and a critical part of the lexical contrast system in English. These characteristics make the CC08 database a less compatible reference for intervocalic obstruent contrasts than is the CaST for CV and VC contrasts, but given the shortage of VCV databases with accompanying perception data we must simply acknowledge this discrepancy as a possible confound. Recordings in Cooke & Scharenborg (2008) were made in a sound-attenuated booth at the University of Sheffield with a Bruel & Kjaer type 4190 microphone placed 30 cm in front of the talker, and were sampled at 50 kHz, following which all signals were high-pass filtered at 50 Hz, downsampled to 25 kHz, and intensity-normalized.[5]

From the test set, which contains 2 productions of each consonant by each of 8 speakers (4

[3]The total item count is 360 rather than 372 due to the 12-item overlap between [b*V*b] items in CV and VC sets.

[4]Cooke & Scharenborg (2008) describe the low vowel in this set as [æ], as in 'bat', but from the present author's auditory impressions of the available recordings, all speakers produce this vowel as [a].

[5]Cooke and Scharenborg do not specify what amplitude signals were normalized to, just that they were equated in "RMS levels." From our inspection of the data, mean intensities across items were not equal, but ranged between 65 and 75 dB. This could mean that larger sound files containing a set of syllables were first normalized before extracting individual items from the signal, but no further details are provided in the article to clarify this discrepancy.

female) for a total of 384 items (though items are not controlled to match in vowel contexts), we arrive at 272 items with obstruent consonants (17 obstruents $\times$ 2 items $\times$ 8 speakers), which were further reduced to a final reference set of 254 items after removing repetitions.[6]

2.4 Temporal parameters

Several acoustic features of potential discriminating value in obstruent identification are primarily temporal in nature, and generally serve to quantify in the acoustics the relative timing of the component gestural events of a given consonant. Such events include the relative timing of the onset/offset of the consonant constriction (reflected in consonant and preceding/following vowel duration), the duration over which the constriction is maintained (manifest as closure duration or noise duration, depending on the manner of articulation), and the relative timing of laryngeal gestures governing the initiation and cessation of voicing in the signal (reflected in the voice onset time and voice cessation time). In the subsections below, each of these parameters is reviewed, including discussion of previous literature on the use of a given parameter in distinguishing English obstruents, the physiological basis for such acoustic differences, the precise definition adopted in the present study, and the distribution of parameter values by category and contrast in both lexicon and syllable databases.

2.4.1 Consonant Duration (DUR_C)

2.4.1.1 Background and physiological basis

The duration of the consonantal interval of an obstruent (DUR_C), an interval which often comprises multiple distinct articulatory and acoustic phases (Docherty, 1992), has received less attention in the literature as a potential distinguishing feature in the acoustics of obstruent phones in English (far less than vowel duration as a cue to voicing, for example). One possible reason for this is that

[6]Because not all items were repeated—i.e., for some speakers the 2 items for a given consonant contained separate vowel contexts, while for others they were repetitions of the same context—the reference set was fixed to have no repetitions so as not to bias estimates of acoustic or perceptual variance on particular consonants.

unlike languages like Hindi and Italian, English does not have phonological length distinctions in its consonant system. Nevertheless, there are a number of reasons to consider the total duration of a consonant as a potentially informative acoustic feature given known characteristics of the mechanics and aerodynamics of speech production, most notably regarding obstruent voicing and manner of articulation, but also to some extent place of articulation may show systematic variation in duration due to the relative speed with which the initiation and cessation of a constriction between different articulators may be achieved (e.g., tongue tip gestures have long been known to be faster than those of the tongue dorsum or lip; Kuehn & Moll, 1976).

One of the first thorough studies on consonant length in English was Umeda (1977), which expanded on earlier findings in Denes (1955), Peterson & Lehiste (1960), Sharf (1962), Cole & Cooper (1975), Lisker & Abramson (1967), and Klatt (1975, 1976), among others. Umeda (1977) is further one of the most thorough studies of durational patterns in the consonant system as a whole, adopting a similar emphasis as that in the present work on breadth in item sampling and the search for *within-speaker* regularities. Approximately 20 minutes of read speech from a single male native speaker of American English was analyzed, with durations of all consonants—covering a wide range of segmental, syllabic, and prosodic environments—measured and analyzed for consistency with the parametric model proposed in Klatt (1976). Among the major patterns uncovered for obstruent consonants in Umeda (1977) are: (1) place of articulation shows the relation *labials* > *coronals* > *velars*; (2) voiced fricatives tend to be shorter than voiceless fricatives; (3) nonsibilant fricatives are generally shorter than sibilants, though results for voiced fricatives are more variable; (4) stops generally are less variable than fricatives, and consonants in word-medial contexts are less variable in duration than in word-initial or word-final positions; (5) word-initial consonants are shorter in function words than in content words, and in high-frequency words relative to low-frequency words.

Crucially, however, Umeda did not include voiceless stop aspiration intervals in her measurement of consonant duration. This choice was consistent with Peterson & Lehiste (1960) and Klatt (1976), among others, and was done primarily to reduce measurement error on vowel durations,

but as a consequence many later findings were obscured, and inconsistencies with earlier results were introduced, such as the place of articulation pattern in (1). Among the notable findings in this other work are the following. Given patterns in aspiration duration (voice onset time, VOT), where labials tend to be shorter than alveolars and velars, with velars typically the longest numerically (Lisker & Abramson, 1964, 1967), though not always significantly so (Docherty, 1992), we would expect either an equalizing of consonant durations based on the reverse pattern found in Umeda (1977), or a reversal toward the VOT pattern, depending on the relative sizes of the corresponding place effects on closure and aspiration durations (the results of Crystal & House, 1988, show some evidence of such a reversal, with the following relation found: labial/alveolar $<$ velar).

Regarding voicing, both earlier findings on duration as a cue to fricative and affricate voicing (voiced $<$ voiceless; Denes, 1955; Cole & Cooper, 1975), and combined findings on voiceless stop consonants exchibiting longer closure durations (Lisker, 1957; Sharf, 1962; Stathopoulos & Weismer, 1983) and longer VOTs (Lisker & Abramson, 1964, 1967; Docherty, 1992),[7] suggest the composite measure of consonant duration used in the present study should show robust differences between voiced and voiceless obstruents. Finally, with a few exceptions,[8] manner has generally been disregarded in studies of consonant duration. This is perhaps due to the clear phonological implications of the [continuant] and [delayed release] features for durational distinctions between plosives, affricates, and fricatives. Nevertheless, this theoretical architecture, though a possible reason for the lack of interest in such studies, lends further support to the consideration of consonant duration as a relevant cue in the perception of manner contrasts.

Given the diversity of effects of voicing, manner, and place on consonant length, the physiological basis for consonant duration as a whole is not singular, but rather comprises a range of biomechanical, aerodynamic, and motor-coordinative factors. The control of voicing is assumed to be primarily aerodynamic. Increases in supralaryngeal air pressure due to complete closure or

[7]Note, however that the former effect is more variable, with some studies, such as Crystal & House (1988), finding no reliable difference between voiced and voiceless closure durations.

[8]For example, Kluender & Walsh (1992) observed a durational confound in experiments on *rise time* in the perception of fricative-affricate distinctions, Byrd (1993) describes in passing duration differences between plosives and affricates in the TIMIT corpus.

narrow constriction of the vocal tract speed up the cessation of voicing, which occurs when air pressure equalizes across the glottis. As a result, voiced closures in the case of stops, and voiced frication in the case of affricates and fricatives, have a constraint on the duration over which they can be held; this constraint is not present for voiceless obstruents, and therefore there is a physiological motivation for the two classes to differ in duration. Further, the speed with which voicing for the following vowel can be initiated after the production of a preceding consonant, a quantity reflected in the VOT measure, is constrained by the configuration of the vocal folds during that consonant, with wider apertures for voiceless consonants taking a longer time to adduct and initiate voicing (Docherty, 1992). Thus, with differences in both VOT and closure duration attributable to voicing aerodynamics, the composite measure of consonant duration must also reflect such factors.

Regarding place of articulation, differences in the biomechanics of each articulator—both their speed, which depends on articulator mass and musculature, and the travel distance required to make a constriction, which depends on the configuration of the vocal tract—should impact both the timing of the initiation and release of consonant closure, and in the case of stop consonants the duration of noise produced at closure release. This latter expectation derives from the observation that a more slowly released constriction will spend slightly more time in a narrow-channel configuration sufficient for the production of turbulence. For these reasons, velars are expected to be longer than alveolars in both closure duration and noise duration (Kuehn & Moll, 1976), though as noted earlier, empirical work has shown the latter effect to be much more robust than the former.

Finally, manner of articulation effects on consonant duration have a physiological basis primarily in differences in the dynamics of gestural constraints on the different constriction types. With no complete obstruction of airflow in the production of fricatives, they can be maintained for longer durations while still transmitting information in the signal. Voiceless stop closures can of course be similarly maintained for an indefinite period, but such an extension of silence is much less informative; it can indicate some information about voicing, but little to nothing about place of articulation. Within the stops, differentiating plosives and affricates in terms of consonant duration is less clear, and likely has more to do with motor programming and the perceptual organization of

the speech code than in lower-level physiological regularities.

2.4.1.2 Definition and measurement

Consonant duration is defined simply as: $\mathrm{DUR_C} = t_r - t_c$, where t_r is the time of consonant release, coincident with vowel onset in CV and VCV positions, and the transition from noise to silence in VC position, and t_c is the point of consonant constriction, coincident with vowel offset in VCV and VC positions, and the onset of noise or prevoicing in CV position. In the latter case of measuring voiceless consonant onset in CV position, because we have no way to determine, in isolated productions, when the constriction gesture was initiated, we must use the first evidence of a consonant constriction in the signal, which in the case of voiceless stops happens to be the point of closure release. Thus, for such items we expect consonant durations to be underestimated. However, from the standpoint of perception, this later point of closure release will also be the first point when it is possible to detect the presence of a consonant in the input, meaning that such estimates should be accurate measures of consonant duration *as perceived*, even if they underestimate the actual duration of the phone *as produced*.

When either landmark is at a vowel boundary, the onset/offset of the vowel is defined as the point of rapid change in upper formant amplitudes (primarily F2) in the spectrogram, as well as in the relative presence of noise in the spectrogram/waveform, and the dampening of amplitude and decrease in complexity of the waveform. At utterance boundaries (t_c in CV position, t_r in VC position) landmarks are defined as the offset and onset of silence, respectively, in the waveform. See Figure 2.1 for sample measurements of these landmarks and of consonant duration in voiceless and voiced plosives, affricates, and fricatives in VCV sequences.

2.4.1.3 Category and contrast distributions

Here we review the distributions of consonant durations in the target inventory and lexicon databases; i.e., in the target speaker's productions of controlled syllables balanced by obstruent category and vowel context, and of the 960 minimal obstruent contrasts in the model lexicon which form the item

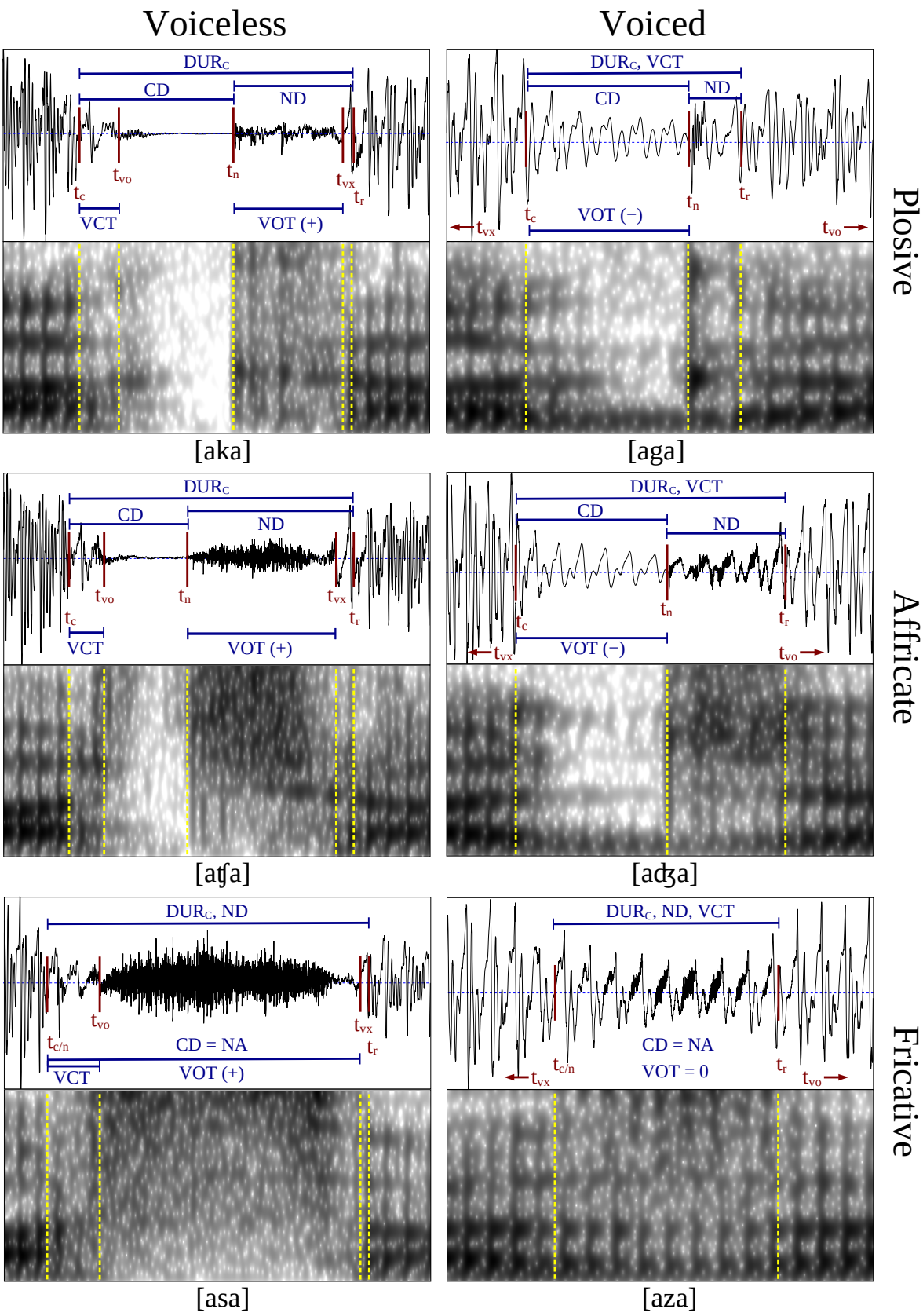

Figure 2.1: Sample measurements of temporal parameters: Consonant Duration (DUR_C), Noise Duration (ND), Closure Duration (CD), Voice Onset Time (VOT), and Voice Cessation Time (VCT).

set in the perception experiments in Chapters 3 and 4. Further, measurements from the reference syllable data are provided for comparison with the target speaker.[9] For category distributions, medians and inter-quartile ranges (IQRs) for each phone are presented, while for contrasts we present distributions of the absolute difference in consonant durations for each item pair, both overall (averaged across all obstruent contrasts), and by the minimal feature distinguishing the contrast (e.g., voicing contrasts are only considered for items which already share manner, place, and sibilance features). Results are presented separately for CV, VCV, and VC positions.

Word-initial position (CV). Figure 2.2 shows consonant duration distributions in CV position by obstruent phone and minimal feature contrast, where for the latter, estimates of the median and IQR of within-item differences in the inventory data (i.e., the difference between repetitions 1 and 2 of each syllable) are shown for reference. These values serve as an indicator of potential thresholds for chance distinctions that may arise due to random variation in production that is of limited utility in models of phonological and lexical contrast discrimination. In analyzing such contrast distributions, we are interested primarily in the degree to which consonant duration may signal abstract featural distinctions and the underlying physiological mechanisms they reflect.

From Figure 2.2 we find durations are generally shorter in the lexical database than in controlled syllables, consistent both general expectations that words tend to be produced more naturally (and therefore more rapidly) than nonwords, and with previous findings that consonant duration is significantly negatively correlated with word frequency (Umeda, 1977). The reference data also generally exhibits longer consonant durations than those of the target speaker, though this difference may arise as an artifact of laboratory recording conditions and inconsistencies in how participants interpret instructions on the production of controlled syllables. It is also worth noting that as part of the MALD project, the target speaker produced approximately 10,000 nonwords, all before producing the syllable data for the present study, and so this speaker may be able to produce such items more fluidly than are speakers producing nonword syllables for the first time.

[9]Distributions for female and male speakers are separated and indicated with ‘F’ and ‘M’, respectively, in parentheses; e.g., California Syllable Test, CaST (F), CaST (M).

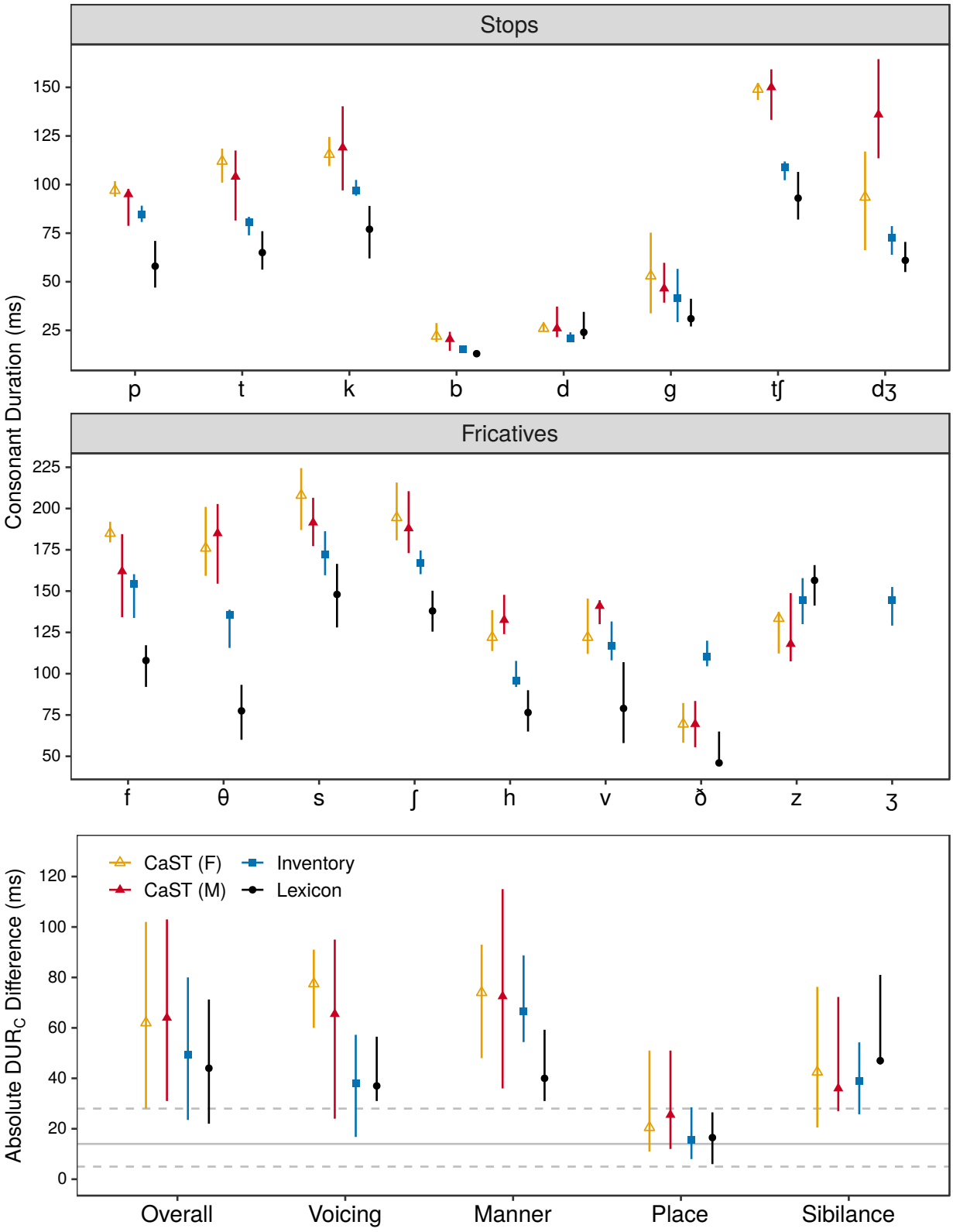

Figure 2.2: Consonant Duration (DUR_C) distributions in CV position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in DUR_C by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

However, in most other respects the relative differences between phones within each database are remarkably similar. Voiceless obstruents are generally greater in duration than their voiced counterparts, the alveolar sibilants being the one exception where this pattern does not hold in the lexicon, though this result could be due to the relative rarity of [z]-onset words in English. Regarding manner of articulation, plosives are shorter than affricates, which in turn are shorter than fricatives, the one exception being [ð], which tends to pattern more closely with the plosives. Place of articulation effects are also relatively consistent between the two data sources. With the exception of [p, t] in the inventory database, plosive duration generally increases with more posterior constrictions within a given manner and voicing class. Finally, sibilants tend to be longer than nonsibilants, particularly in real-word productions, though this distinction is somewhat weakened in the reference syllable data.

Turning next to contrast distributions, we see from the bottom panel of Figure 2.2 that obstruent contrasts in general tend to show notable differences in consonant duration. This is not surprising given previous literature on the range of physiological mechanisms responsible for inherent differences in obstruent duration, as well as the distinct duration profiles of most of the obstruent phones in Figure 2.2. Manner of articulation appears to be the most robust in terms of absolute differences in consonant duration, in the inventory data. Voicing and sibilance are less robust but remain generally above chance levels based on within-item variance, while place of articulation shows no clear contrast effect for consonant duration.

Word-medial position (VCV). Intervocalically, we observe the same trend in shorter durations for real word data relative to nonword syllables (Figure 2.3), as well as the generally greater durations for voiceless than voiced obstruents. Otherwise, compared to CV position there is a narrower separation of phones based on consonant duration, with manner distinctions notably reduced in both word and nonword data (the one exception being the introduction of a clear duration difference between flaps and non-flaps), and effects of place of articulation and sibilance also less clear. Examining the contrast distributions, consistent with the above description, only voicing contrasts

show duration differences greater than reference levels. All other features show substantial overlap with the reference distribution, indicating such distinctions are not reliably different from chance.

These results reflect in part the fact that consonant duration masks the durations of many distinct phases in consonant production (e.g., stop closure interval, noise interval, voicing intervals), phases which are demarcated in Figure 2.1 and captured in several parameters below. Further, the fact that many such intervals are either absent or not measurable in word-initial position means that the more robust results in the previous section could rather be a reflection of other coincident parameters such as noise duration, a hypothesis we will test when noise duration is formally introduced in Section 2.4.4.

Word-final position (VC). Distributions of consonant duration by category and featural contrast in VC position are shown in Figure 2.4. Word-finally, the utility of consonant duration as an index of phonetic contrast is even further diminished from VCV position. Voicing remains one dimension along which obstruents are differentiated according to this parameter, but the plosive distinction is much narrower. The lower overall durations of plosives, however, do introduce a manner distinction between plosives and non-plosives, and there appears to be a narrow effect of sibilance in the voiceless fricatives, though fricative durations in VC position are also more variable than other manners of articulation.

Finally, regarding featural contrasts, as in VCV position voicing is the most robust feature delineated by consonant duration, with the remaining three lower in absolute duration differences in all four databases, though the relative ranking among the three features is otherwise inconsistent between words and syllables (Inventory, CaST). The variances in word-final position, however, are larger than in VCV, a result partly driven by the greater amount of missing data word-finally due to the measurement problem of determining the point of consonant offset in unreleased plosives.

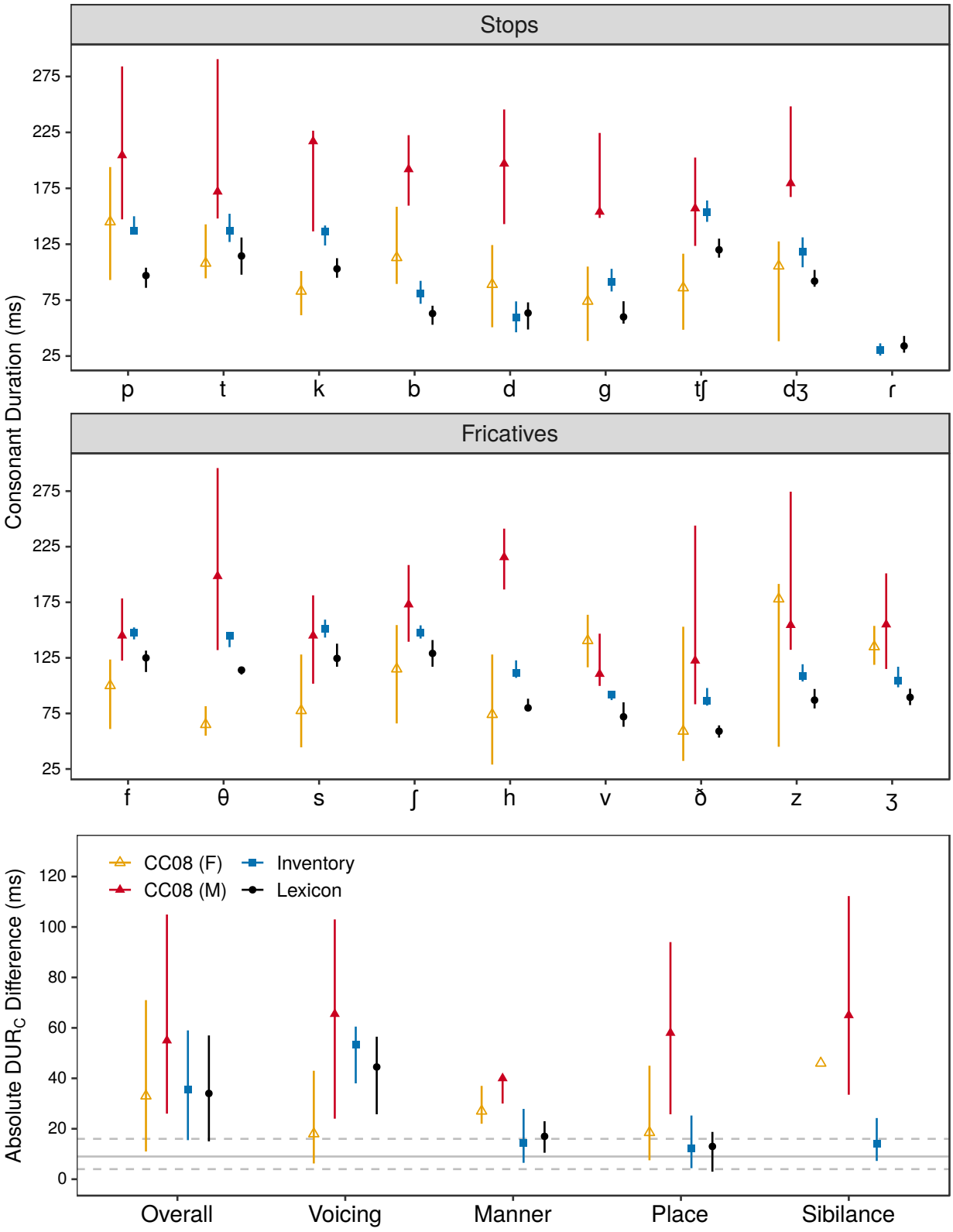

Figure 2.3: Consonant Duration (DUR_C) distributions in VCV position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in DUR_C by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

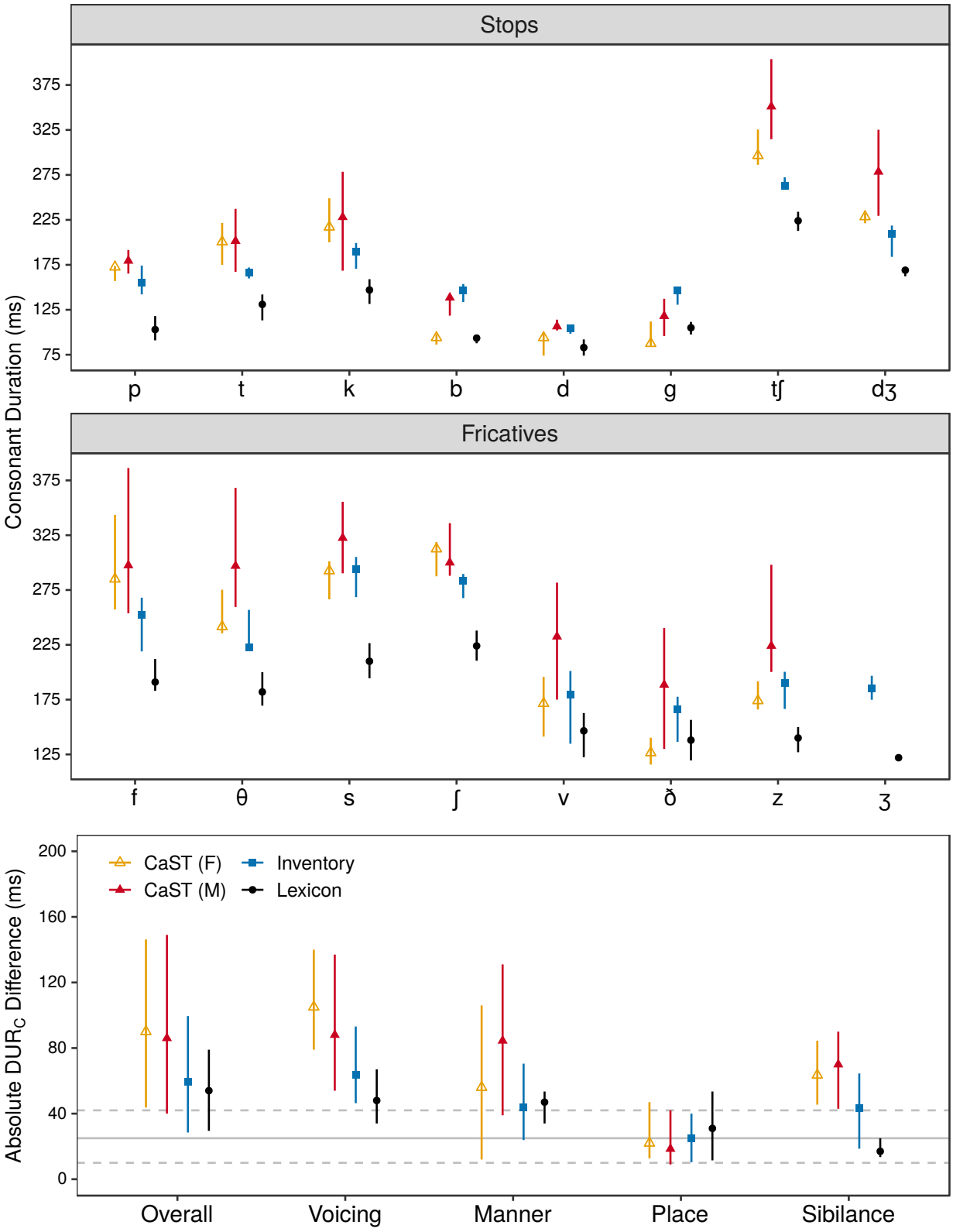

Figure 2.4: Consonant Duration (DUR_C) distributions in VC position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in DUR_C by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

2.4.1.4 Summary

Consonant duration is an acoustic parameter that is both of potential utility as a cue to the discrimination of obstruent phones, and of explanatory value given the various physiological factors with predictable timing effects (e.g., articulator movement velocity, aerodynamic constraints on voicing in a closed/contrained tube). However, as discussed in the introduction, and as the VCV and VC data show, consonant duration, particularly as applied to obstruents, is also subject to considerable variability. This variability stems from both the complexity of obstruents as sounds comprised of multiple articulatory events, and measurement constraints at word boundaries.

2.4.2 Vowel Duration ($DUR_{V1/V2}$)

2.4.2.1 Background and physiological basis

Consonantal effects on preceding vowel duration have been observed acoustically since at least the 1930s (Lloyd, 1936; Heffner, 1937; Locke & Heffner, 1940; Lehmann & Heffner, 1940, 1943), with much of this early work inspired by observations from auditory impressions presented in general phonetic documentation on American English (e.g., Kenyon, 1924). House & Fairbanks (1953) present perhaps the most comprehensive description of such effects in this early period, including accounting for the multivariate distribution of consonant categories along dimensions of vowel duration, fundamental frequency, and vowel amplitude. Denes (1955) was the first to examine such acoustic differences perceptually, finding evidence in cross-splicing and synthetic stimulus designs for listener weighting of vowel duration in the identification of voiced [z] in 'use (V)' versus voiceless [s] in 'use (N)'. In this work, due to the lack of independence of consonant and vowel durations, a further parameter was derived—the ratio of consonant to vowel duration—that proved an accurate predictor of listener responses. These results were later replicated and extended to the perception of plosive voicing contrasts (Raphael, 1972; Port, 1976, 1981), and informed a growing literature on timing cues in general, with a particular focus on their relation to the production and perception of voicing (Peterson & Lehiste, 1960).

While conditioned differences in vowel duration as a function of following consonant voicing (and to a lesser extent place and manner of articulation) have been thoroughly documented in production data, explanations of the ultimate source of this regularity—physiological, phonological, or some combination of the two—have received considerable debate in the decades following initial reports in House & Fairbanks (1953) and Denes (1955). Some authors have pursued general auditory explanations (Kluender et al., 1988), citing changes in vowel duration as intentional modifications for the purpose of auditorily enhancing differences in closure duration presumed to more directly reflect consonant voicing distinctions. However, far greater attention has been paid to potential articulatory motivations. Such work includes examinations of EMG (Raphael, 1975; Bell-Berti, 1975), aerodynamic (Malécot, 1966; Lisker, 1970; Harris, 1974), and kinematic (Gracco, 1994; Löfqvist & Gracco, 1994) data, much of which finds greater velocity and displacement of consonant articulators in closing gestures for voiceless consonants, as well as earlier onset of intraoral pressure buildup in vowels preceding voiceless consonants.

Gracco (1994) integrates the kinematic, aerodynamic, and EMG data into the following explanation of the vowel duration effect: due to the greater intraoral pressure required to cease voicing during voiceless consonants, an earlier onset of constriction is required to achieve this increase, therefore leading to shorter vowels on average before voiceless consonants. We should note, however, that Gracco exhibits some reservations about this explanation, notably due to the lack of evidence for relatively greater stiffness and force of contact in voiceless consonant closures (Lubker & Parris, 1970), which he argues would be expected under this explanation. Finally, because much of these effects have parallels in the cessation of the consonant gesture in CV transitions, similar effects may be predicted for the duration of the following vowel, though this context has received little attention.

2.4.2.2 Definition and measurement

Given the above discussion we formulate the definition of the vowel duration cue as follows: $\mathrm{DUR}_{\mathrm{V1/V2}} = t_c - t_r$, where t_c and t_r are the times of consonant constriction/onset and release/offset,

which, given the previous definition of such parameters in the measurement of consonant duration, correspond to points of vowel offset and onset, respectively. Here vowel duration is defined for both preceding (DUR_{V1}) and following (DUR_{V2}) vowels, as the available articulatory data suggests consonant voicing effects on V2 duration are at least consistent with the physiological explanation for changes in V1 duration, though the latter has received the majority of attention in the literature.

Measurement of vowel duration then depends on the identification of the same landmarks that were used in the measurement of consonant duration, except in the case of vowel-onset/offset words, or vowels adjacent to non-obstruent phones, where previous definitions based on formant amplitudes, noise in the spectrogram/waveform, and waveform amplitude and complexity may not apply. In the former case of vowel-onset/offset words, vowel onset/offset is defined as the point of appearance or disappearance of formant structure in the spectrogram. In the latter case of vowels adjacent to nasals, liquids, or other vowels, inflection points in second and third formant trajectories, as well as discontinuities in upper formant amplitudes (in the case of nasals and laterals), are used to define vowel boundaries. Figure 2.1 shows sample measurements of the primary landmarks in the definition above for different voicing and manner classes, and though it does not indicate a full vowel segmentation, the criteria used to define consonant onset and offset times (t_c, t_r) may be used to derive vowel offset and onset times, respectively.

2.4.2.3 Category and contrast distributions

Next we present vowel duration distributions in both databases by category (medians and IQRs for each phone) and contrast (distributions of absolute vowel duration differences for each item pair as a function of the minimal feature distinguishing the contrast). Because the availability of preceding and following vowel duration as potential acoustic cues is necessarily constrained by context, rather than presenting separate results for each position (CV, VCV, VC), we present results by parameter, beginning with preceding vowel duration (DUR_{V1}) in VCV and VC contexts, followed by DUR_{V2} in CV and VCV contexts.

Preceding vowel duration ($\mathbf{DUR_{V1}}$). Distributions of V1 durations in VCV and VC positions are shown in Figures 2.5 and 2.6, respectively. In both word-medial and word-final contexts, there is a clear separation in DUR_{V1} distributions between voiced and voiceless obstruents, particularly in word-final position, where vowels preceding voiceless obstruents can be over 100 ms shorter than their voiced counterparts. This pattern extends across manners of articulation, and appears to be similarly robust for each class. Further, this pattern is consistent in both word and syllable data, as well as from the present speaker and from the reference speakers in the CaST and CC08 databases. The only other notable pattern, which is restricted to VC position, is the moderately longer vowel duration preceding fricatives than preceding stops. However, this pattern in category means does not extend to vowel duration differences in specific minimal pairs, as the manner effect in the bottom panel of Figure 2.6 shows. All other featural distinctions do not appear to vary notably by preceding vowel duration.

Finally, it bears mentioning that there is an apparent place effect in the lexicon data ([t, d] $<$ [p, b, k, g]), but this is an artifact of the stress contraints on the allophonic distribution of the alveolar plosives and the alveolar flap, the former occurring intervocalically at the onset of a stressed syllable, and the latter preceding unstressed syllables. Thus, instances of VCV [t, d] in the lexicon almost exclusively follow unstressed vowels, which is why this pattern does not show up in the syllable data, and why such differences are not expected to be reliable cues in the discrimination of plosive minimal pairs, as once the stress pattern is controlled we would expect similar drops in preceding vowel duration for labials and velars. Similarly, vowels preceding [h] are substantially lower in the lexicon than in the syllable data. This result is also due to stress patterns, as the glottal fricative only occurs intervocalically as the onset of a stressed syllable, a distribution that is likely due to perceptual constraints on [h] that similarly govern its absence from coda position in English.

Following vowel duration ($\mathbf{DUR_{V2}}$). Figures 2.7 and 2.8 show following vowel duration distributions in CV and VCV positions, respectively. In both figures, results by category and contrast

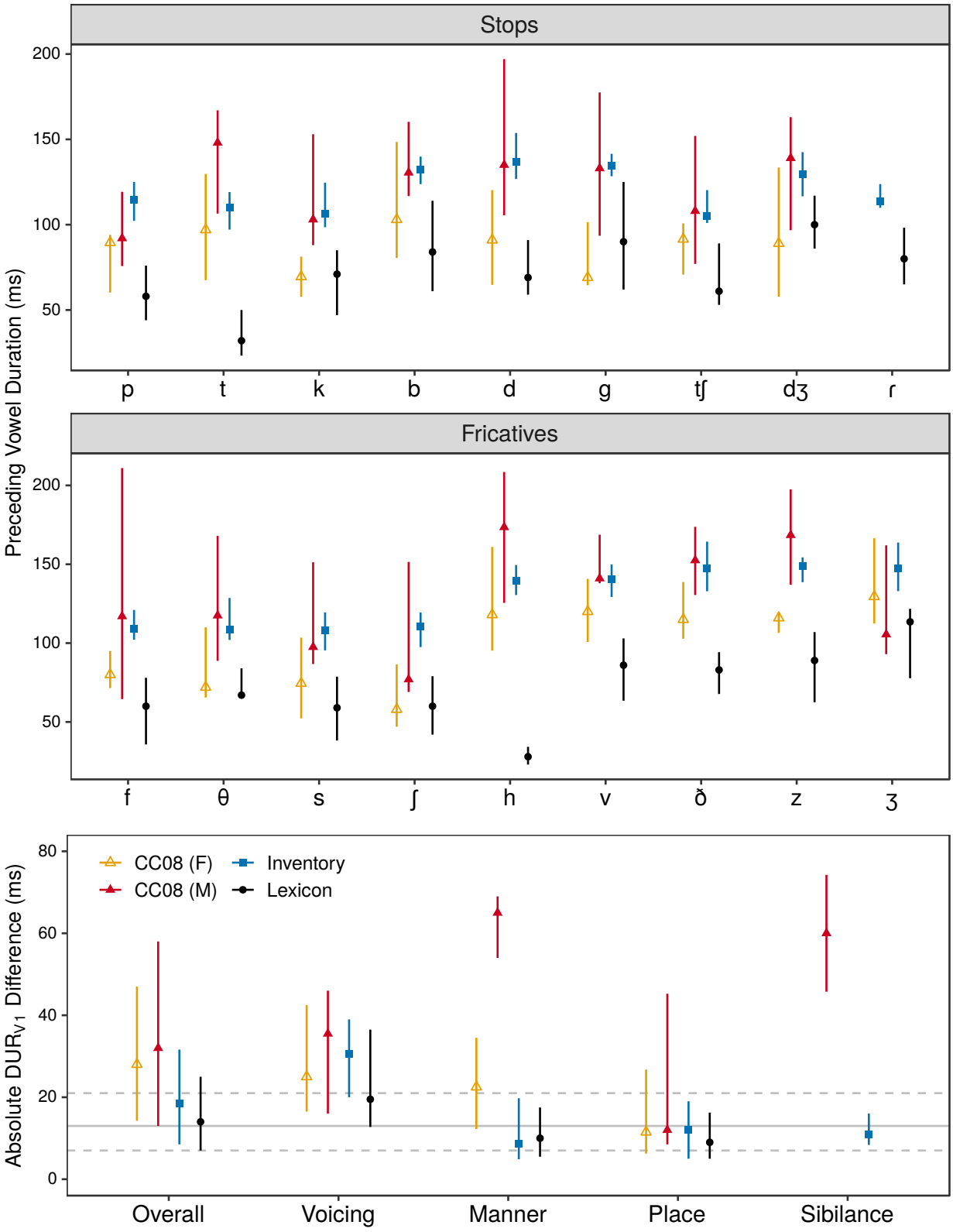

Figure 2.5: Preceding Vowel Duration (DUR_{V1}) distributions in VCV position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in DUR_{V1} by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

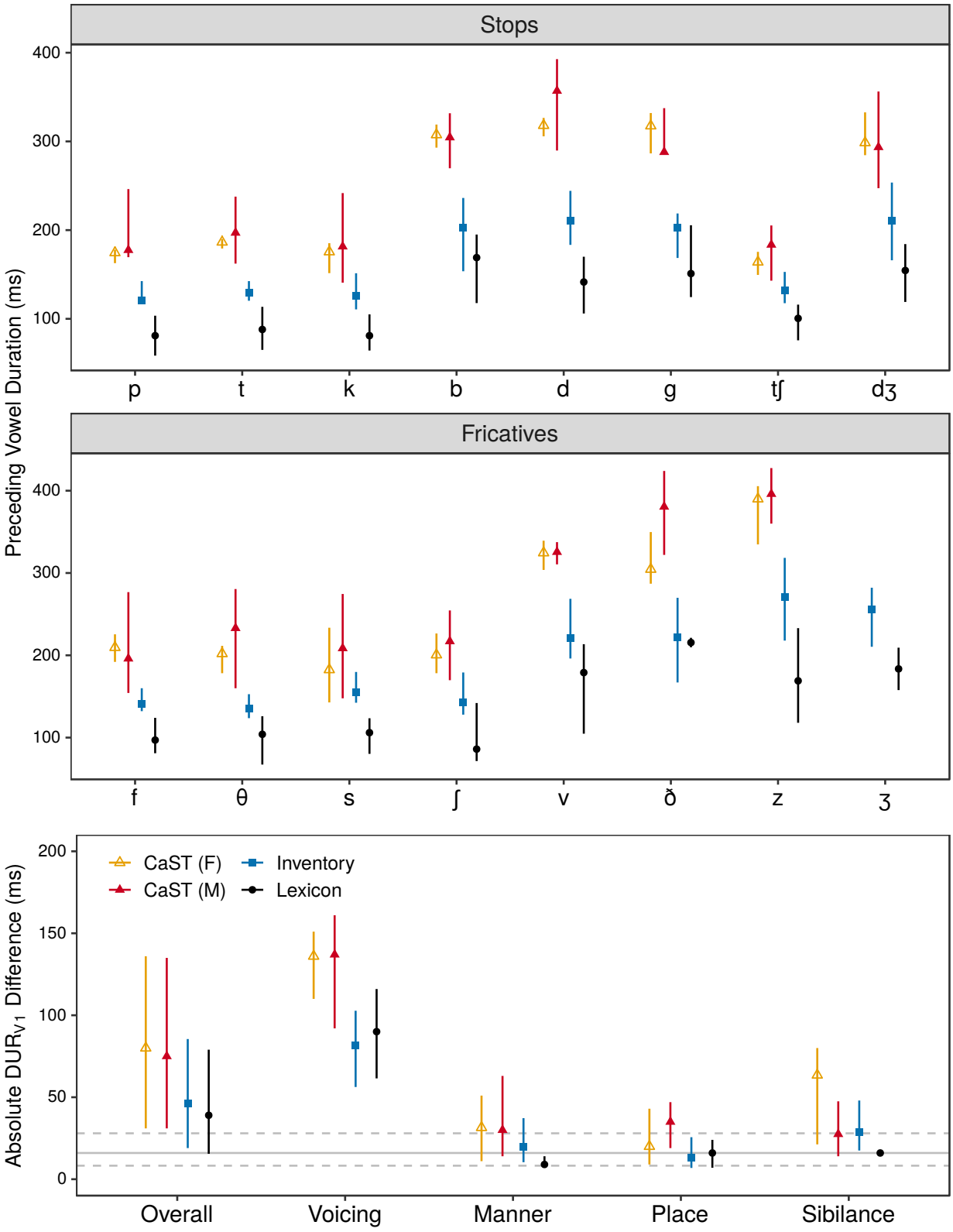

Figure 2.6: Preceding Vowel Duration (DUR_{V1}) distributions in VC position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in DUR_{V1} by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

show few clear patterns. Durations are shorter in the real-word data, and for the inventory data there is a significant drop in V2 duration following the alveolar flap [ɾ], which simply reflects the difference in vowel environment (reduced [ə] vs. full vowel contexts for the other obstruents). Reference V2 durations from the CaST and CC08 databases are again longer overall, but with the same general category trends, or lack thereof. When broken down by featural contrast, most conditions overlap substantially with the chance region between 10 and 30 ms, particularly the inventory and lexicon data from the target speaker whose within-item variance was used to derive estimates of chance effects. Overall, both lexical and inventory-based phonological contrasts exhibit approximately the same differences between items as within. The one distinction of note, though still within the estimated chance range, is in the effect of voicing in CV contrasts, where there does appear to be a relatively consistent decrease in duration from voiced to voiceless that is observed primarily in plosives and affricates. Again, this result could be due to differences in gestural timing between voiced-voiced and voiceless-voiced sequences, or it could simply reflect measurement effects, with vowel onsets delayed in voiceless contexts due to the perturbation of higher formants by noise. Yet, even for the latter case, if listeners are tracking higher formant amplitudes for vowel identification, then this result would lead them to perceive shorter vowels following voiceless obstruents, making $\mathrm{DUR}_{\mathrm{V2}}$ a viable cue in the perception of word-initial obstruent voicing.

2.4.2.4 Summary

Vowel duration, particularly the duration of vowels preceding word-final contrasts, reliably differs between voiceless and voiced obstruents, and this distinction is robust across data types (controlled syllable, real word), and across manner (plosive, affricate, fricative) and place (labial, coronal, dorsal). However, unlike consonant duration, vowel duration is otherwise unproductive at distinguishing other features. Therefore, in terms of the wider utility of a given parameter in distinguishing items in the lexicon, vowel duration may be more limited in scope.

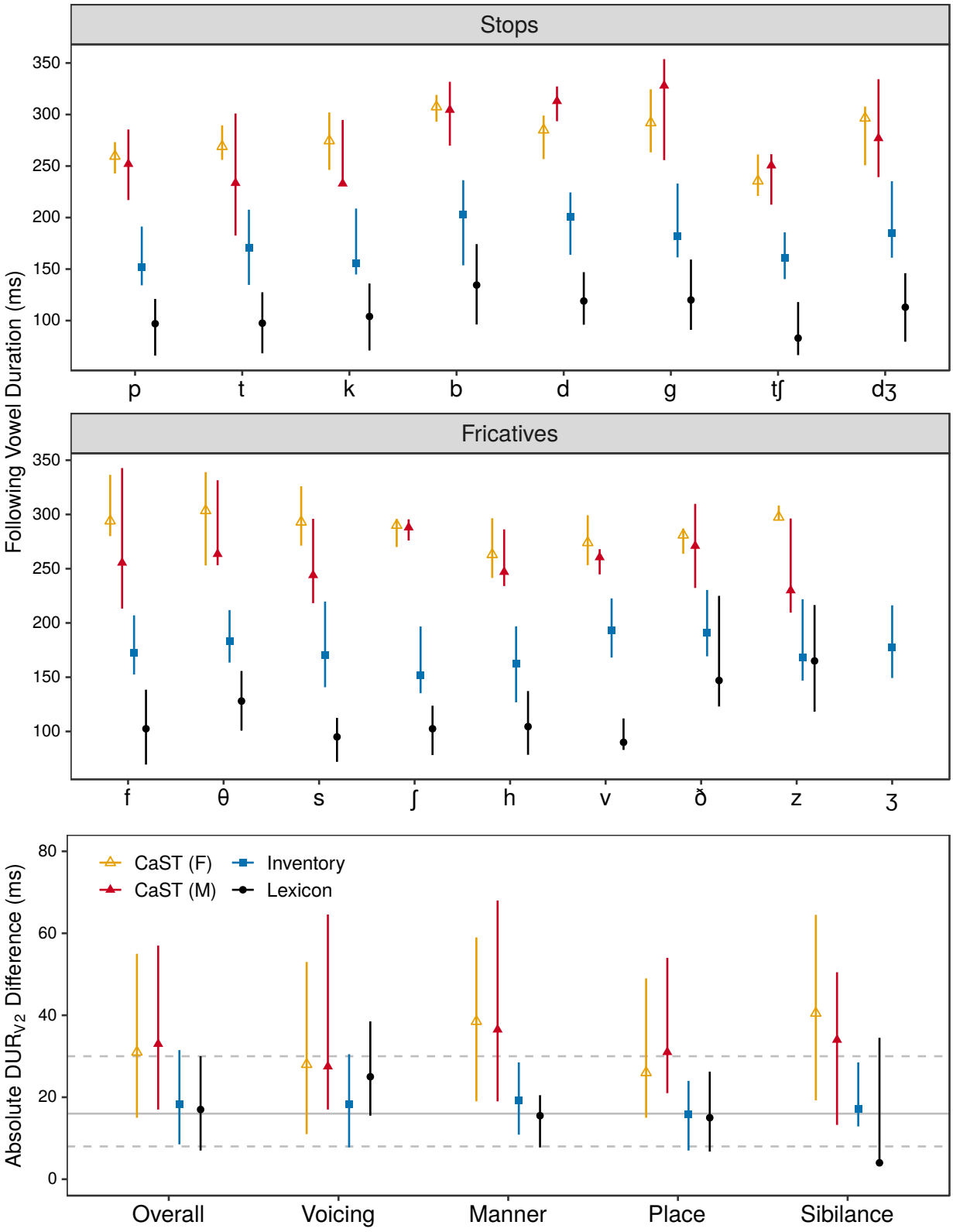

Figure 2.7: Following Vowel Duration (DUR_{V2}) distributions in CV position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in DUR_{V2} by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

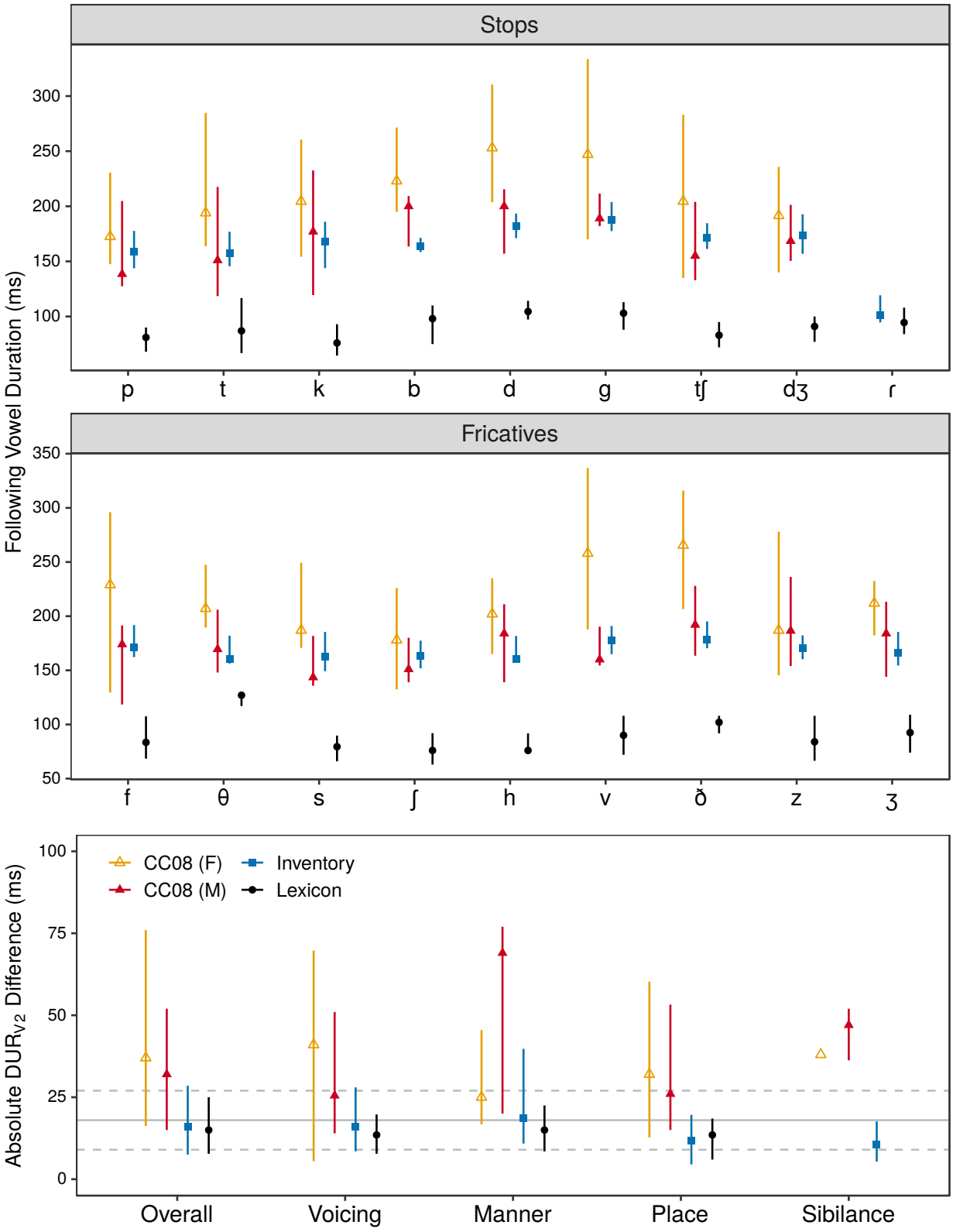

Figure 2.8: Following Vowel Duration (DUR_{V2}) distributions in VCV position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in DUR_{V2} by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

2.4.3 Closure Duration (CD)

2.4.3.1 Background and physiological basis

The duration of stop consonant closure, or *closure duration* (CD), was introduced by Lisker (1957) as an acoustic feature of potential utility in distinguishing English voiced and voiceless stops in intervocalic position. The measurement was partly inspired by a previous finding in Denes (1955) that not only is the word-final voicing contrast in English partly cued by preceding vowel duration, but also that such effects vary as a function of the duration of the consonant (operationalized as the ratio between vowel and consonant durations). Lisker (1957) found not only spectrographic support for the /p, b/ distinction, but also demonstrated perceptual shifts in voicing identification with changes in closure duration in a tape-splicing design.

The physiological explanation for the regularity in the distinction between voiced and voiceless stop closure durations is well motivated on aerodynamic grounds. Due to the trans-glottal pressure differential required to sustain voicing, with greater closure duration there is greater pressure buildup in the oral cavity and therefore greater likelihood the pressures above and below the glottis will equalize and cause vocal fold vibration to cease (Malécot, 1966; Lisker, 1970).Therefore, in order to sustain voicing, speakers must initiate consonant release earlier in voiced stops than in voiceless (Gracco, 1994; Löfqvist & Gracco, 1994).

Identification patterns in Lisker (1957) were also shown to be modulated by preceding and following vowel formant cues (tested via cross-splicing from preceding/following vowel intervals from the opposing VCV), consistent with later findings of cue interactions in stop place identification (Harris et al., 1958; Hoffman, 1958). Notable later uses include investigations of trading relations and multivariate cue integration (Repp, 1978; Fitch et al., 1980; Lisker, 1986; Parker et al., 1986; Castleman & Diehl, 1996), effects of speaking rate on cue perception (Summerfield, 1981; Port, 1976; Miller & Grosjean, 1981; Kidd, 1989), effects of context (Luce & Charles-Luce, 1985), prosodic effects (Cole et al., 2007), developmental changes in the use of closure duration in voicing perception (Kuhl, 1979; Cohen et al., 1992) and production (Smith, 1978; Mack &

Lieberman, 1985; Tauberer, 2010), and variability in read speech (Byrd, 1993).

Finally, though closure duration has primarily been used in the literature as a voicing cue, as noted above, the parameter has also been shown to influence perception of stop place of articulation (Repp, 1984a), as well as being used in the detection of stop consonants at syllable boundaries (Raphael & Dorman, 1980) and in consonant clusters (Marcus, 1978), and in distinguishing manner contrasts between fricatives and affricates (Rakerd et al., 1982).

2.4.3.2 Definition and measurement

Given identified acoustic landmarks of the point of consonant closure (complete contact between passive and active articulators, obstructing airflow in the vocal tract), t_c, and the point of release/noise-onset, t_n, closure duration is defined simply as: $\text{CD} = t_n - t_c$. Detecting these landmarks, however, is a non-trivial matter. Common signatures of the point of closure (assuming the consonant is in a post-vocalic environment) include the loss of energy in the higher formants (F2, F3, etc.) and the loss of signal amplitude in the waveform. In the case of voiced stops, rather than silence, a voice signal from the larynx may remain in the signal as a quasi-sinusoid, in which case formant characteristics play a critical role in defining the point of closure. Signatures of consonant release generally mirror consonant closure, but also vary according to consonant manner and place of articulation. Figure 2.1 illustrates the measurement of closure duration for a sample of voiced and voiceless intervocalic stops.

2.4.3.3 Category and contrast distributions

Because closure duration can only be reliably measured in post-vocalic context, and is undefined for fricative consonants, CD distributions by category and featural contrast will only be presented for stops in VCV and VC position. Further, because closure duration cannot be measured for unreleased stops, a common occurrence in real-word productions in English, the VCV results should be held as primary in the presentation below.

Word-medial position (VCV). Figure 2.9 shows closure duration results for intervocalic stops in inventory and lexicon data sets. For both unit types (words and nonword syllables) the following pattern in plosive place of articulation was observed: coronals $<$ velars $\leq$ labials. POA differences are larger overall in the controlled syllable data than in real words,[10] but the general trends do not appear to differ between the inventory and lexicon. There is also an apparent interaction between place and voicing, where the distinction between labials and velars is greater in voiceless ([p] $>$ [k]) than in voiced ([b] $\approx$ [g]), and the distinction between coronals and both labials and velars ([t, d] $<$ [p, b, k, g]) is greater in voiced than in voiceless plosives. These effects of place are broadly consistent with expectations based on articulator velocity (Kuehn & Moll, 1976), though the difference among voiced plosives conflicts with expectations based on aerodynamic constraints on voicing as a function of the point of supralaryngeal occlusion. However, the latter expectation is not generally borne out in the data (Stathopoulos & Weismer, 1983).

In general, voiceless plosives have slightly longer closure durations than voiced plosives, consistent with the original theory and demonstration in Lisker (1957), but this effect does not extend to affricates. In fact, in the real-word data the voiceless $>$ voiced relation even reverses ([ʧ] $<$ [ʤ]). Finally, regarding manner of articulation, affricates appear to exhibit slightly shorter closure durations on average than plosives, at least within a given voicing class. In the case of voiced stops, this result could simply be a side-effect of articulator differences, with the tongue blade being relatively faster than the lips or tongue dorsum, making it easier for sounds like [d] and [ʤ] to be achieved in a narrower time window. However, for voiceless stops this pattern breaks down, as [ʧ] is substantially shorter in closure duration than [t], an effect which may be attributed to the need to release the constriction earlier in an affricate in order allow sufficient time for frication before the onset of the vowel (assuming rhythmic factors place some constraints on overall syllable length). Given the inconsistencies between the voiced and voiceless affricates, and the general sparsity of alveolar plosives intervocalically (due to the flapping rule), we cannot draw any firm conclusions at this point on the factors underlying manner effects on closure duration.

[10] Effects in the reference data, however, are reduced, particularly among female speakers.

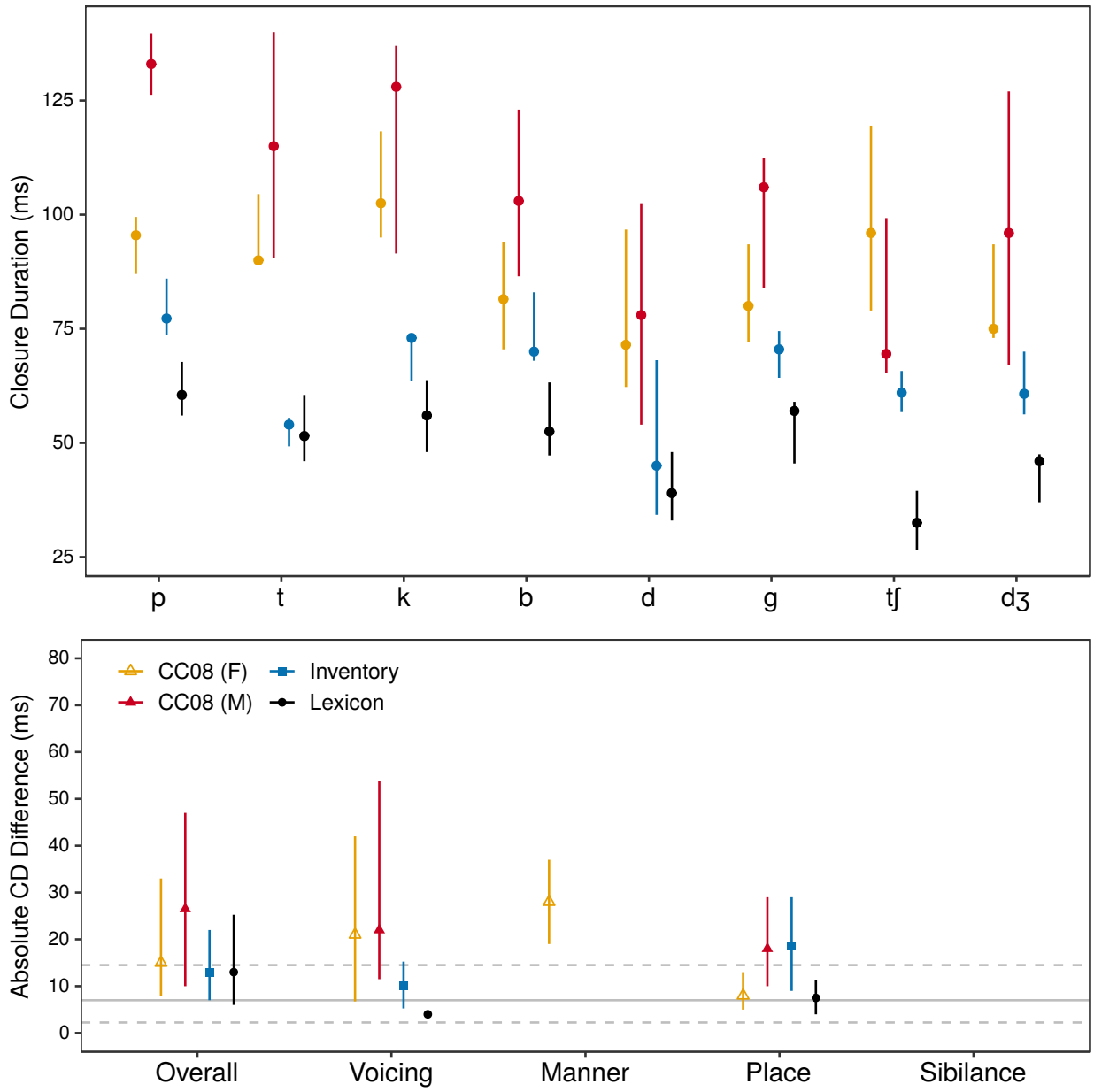

Figure 2.9: Closure Duration (CD) distributions in VCV position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in CD by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

Regarding featural contrasts, many distinctions are sparse or entirely absent in the lexicon because of the pervasive role of the alveolar flap [ɾ] in VCV contrasts, as the present work has left CD undefined for flaps.[11] The one feature of note in Figure 2.9 is place of articulation, which shows much greater itemwise contrast differences in the inventory than in the lexicon, though the variance of both distributions is high.

Word-final position (VC). Closure durations of obstruents in VC contrasts are shown in Figure 2.10, and are generally consistent with the intervocalic patterns. That is, alveolar plosives are consistently shorter than labials and velars, while affricates show less consistency, [ʧ] being shorter than [ʤ] in the inventory data, but longer than [ʤ] in the lexicon and in the reference data. Further, word-final contrasts in general are not robustly distinguished by closure duration, but among the two features where CD may play a role (manner and sibilance are restricted due to their reliance on fricatives for minimal contrasts), place contrasts show the greatest distinction. In fact, despite the greater measurement reliability in VCV position, contrast effects for place of articulation are more robust word-finally. This result is likely due to the general sparsity of [t, d] intervocalically, as both positions show similar distributions for labial, alveolar, and velar plosives.

2.4.3.4 Summary

The consistent place of articulation effect in VCV and VC positions, particularly word-finally where plosive POA is more evenly distributed, does lend support to the physiological explanations introduced at the beginning of this section. However, it is important to emphasize the limited applicability of CD to obstruent contrasts, being defined only for postvocalic stop consonants. This limitation restricts the potential utility of closure duration as a cue in perception, but given the frequency of word-final voiceless plosive contrasts in English, CD may remain informative in the lexicon. This hypothesis will be tested directly in the cue integration models in Chapter 4.

[11] This choice was made in part because while we know flap consonants exhibit a very brief closure, the acoustic output of flap productions does not match the criteria used to identify consonant closures in all other stop consonants. Acoustically, flaps are much more similar to some voiced fricatives, showing a depression but not cessation of formant amplitudes at flap boundaries, and lacking any trace of a release burst.

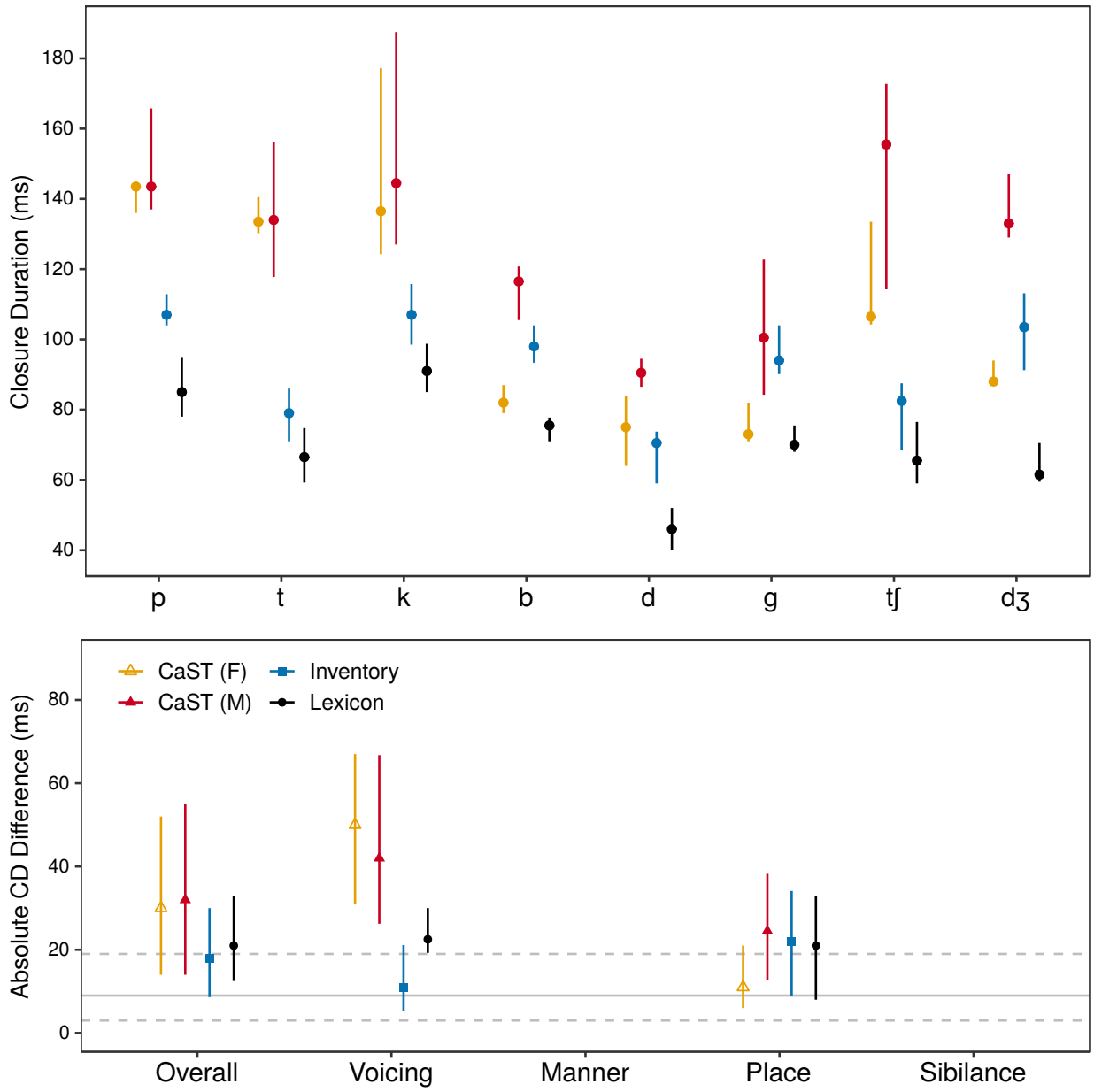

Figure 2.10: Closure Duration (CD) distributions in VC position. The top two panels show category medians and IQRs. The bottom panel shows distributions of absolute differences in CD by feature contrast. The dashed gray lines indicate medians and IQRs of within-item differences in consonant duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

2.4.4 Noise Duration (ND)

2.4.4.1 Background and physiological basis

Related to closure duration is the duration of noise due to turbulence in the vocal tract, or the *noise duration* (ND), primarily applied in the analysis of fricative contrasts (Cole & Cooper, 1975; You, 1979; Baum & Blumstein, 1987; Behrens & Blumstein, 1988; Jongman et al., 2000), but also relevant to affricate voicing (Cole & Cooper, 1975) and the fricative-affricate manner distinction (Kluender & Walsh, 1992). Further, as the plosive burst is also a source of noise, burst duration may be considered as within the same cue class, though the two are not typically directly compared in studies of obstruent acoustics.

There are two main physiological bases of distinctions in noise duration that may define obstruent phone and feature contrasts in English, one of which may be viewed as *active* or controlled, and the other as *passive*. The active regulation of noise duration is what we see in the manner distinctions between plosives, affricates, and fricatives, where there are major differences in noise duration between the three classes that are the result of differences in articulatory control over the degree of consonantal constriction and the timing of consonant release. Of course, there are some place dependencies worth noting, where in English there is no strict plosive-affricate distinction that does not also coincide with differences in place of articulation (e.g., there are gestural motivations for the difficulty of engaging a rapid, plosive-like release between the tongue blade and palate without the initiation of significant frication); however, it is reasonable to assume that in the present system English speakers exert some control over affricate noise duration.

The second physiological property that leads to systematic variation in noise duration is passive and has both articulatory and aerodynamic components. On the articulatory end, obstruents may vary in noise duration due to differences in constraints on the maintenance of a constriction (e.g., alveolar/postalveolar sibilant fricatives versus labial/dental nonsibilant fricatives) or on the speed of articulator movement (as in the difference between labial and velar articulations discussed above). On the aerodynamic end we have primarily constraints on the maintenance of voicing dur-

ing the production of noise from a narrow supralaryngeal constriction, because such constrictions are too narrow to prevent the pressure buildup behind the constriction from equalizing with subglottal pressure and thereby halting vocal fold vibration. Of course, this constraint could also manifest in differences in the percentage of the noise interval that is voiced, but if complete obstruent voicing is required there is reason to expect concomitant reductions in noise duration.

Therefore, noise duration serves as a potential cue to all four primary obstruent features—voicing, manner, place, and sibilance—due to the numerous physiological constraints imposed on constriction timing and airflow regulation in the production of obstruent consonants.

2.4.4.2 Definition and measurement

Unlike closure duration, noise duration is more directly defined on acoustic grounds, though points of noise onset and offset do correspond to changes in airflow from the consonant constriction. Here, noise duration is defined as follows: ND $= t_r - t_n$, where t_n is the point of noise onset—i.e., when noise is first introduced in the signal, either from constriction onset in the case of fricatives, or stop closure release in the case of plosives and affricates—and t_r is the point of consonant release, defined earlier in Section 2.4.1.2. These landmarks are illustrated in VCV context in Figure 2.1.

2.4.4.3 Category and contrast distributions

In the sections below, we review noise duration distributions by phonetic category and featural contrast in inventory, lexicon, and reference databases. Results are presented separately for CV, VCV, and VC positions.

Word-initial position (CV). With the exception of prevoiced stops, which are relatively infrequent word-initially in English, *noise duration* as an acoustic parameter coincides with the *consonant duration* measure presented earlier. This can be seen in the remarkable similarity between Figures 2.2 and 2.11, which present category and contrast distributions for consonant duration and noise duration, respectively, in CV position. That is, the same general patterns reported for con-

sonant duration—voiced > voiceless; plosives < affricates < fricatives; [p, b] < [t, d] < [k, g]; nonsibilants < sibilants—are present for noise duration, the primary difference being that noise duration exhibits consistently lower within-category variance than consonant duration. The one other distinction of note between ND and DUR_C is in the patterning of the voiced nonsibilant fricatives, [v, ð], whose noise durations more closely align with those of the voiced plosives than do the consonant durations of the two series (particularly for [v]; [ð] already shows similarities to the plosives in the DUR_C results in Figure 2.2). The mechanism behind this outcome, which is confirmed in the VOT and Burst parameter results below, is the occasional fortition of [v] and [ð] word-initially into their plosive counterparts, [b̪] and [d̪]. That is, such items exhibit an initial interval of periodicity consistent with closure voicing (i.e., quasi-sinusoidal, with no clear noise or formant structure), followed by a release burst and short noise interval before vowel onset.

Contrast distributions follow the same trend for noise duration as for consonant duration; i.e., in terms of robustness, manner > voicing, sibilance > place. These distinctions between features are reduced in the lexicon data relative to the inventory, while the reference data shows a different pattern from the enhancement of ND distinctions in voicing contrasts: voicing, manner > sibilance > place. However all such effects, as well as the overall distributions for contrasts irrespective of featural composition, are moderately stronger for noise duration than consonant duration, and only place distinctions are within the chance range.

Word-medial position (VCV). Figure 2.12 shows noise durations of obstruents and obstruent contrasts in VCV position. Effects of voicing and manner, though reduced relative to those observed in CV position, remain robust word-medially. However, there is no clear difference between sibilant and nonsibilant noise durations in VCV position, and place effects are only consistent for voiced plosives. Among voiceless plosives, velars and alveolars remain longer than labials, but this difference is much reduced in the inventory relative to the lexicon. Further, [t] exhibits a longer noise duration than [k] in the target speaker's data (the lexicon and inventory sets), a pattern which is inconsistent with word-initial plosives and is not shared in the reference VCV data

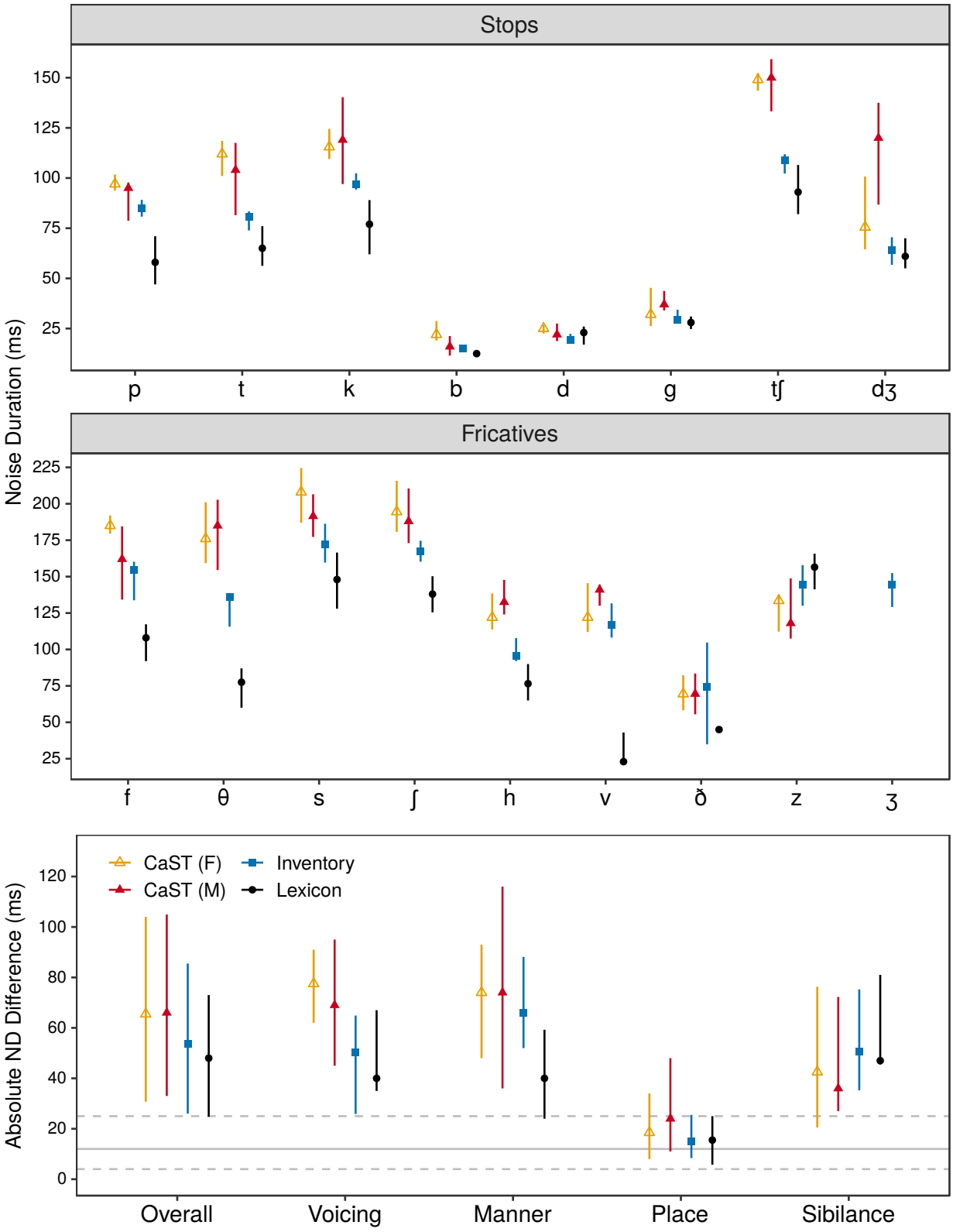

Figure 2.11: Noise Duration (ND) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in noise duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

from the CC08 study, but this pair is also more variable and shows greater distributional overlap than the labial < alveolar/velar distinction. As a result of these inconsistencies, only voicing and manner features show clear separation from chance ND differences intervocalically. These effects, however, are substantial—greater than 50 ms on average—and drive an overall distinction in noise duration between contrasting items in all four data sets.

Word-final position (VC). The robust voicing effects presented above for CV and VCV positions are notably reduced word-finally, though as Figure 2.13 illustrates, with the exception of the labial plosives, all other places and manners show the voiced < voiceless relation observed above, and this effect is not restricted to a particular speaker or item set. As in VCV position, ND differences in place and sibilance contrasts are largely at chance levels, though here the reference data diverge from the target data in showing a sibilance effect. Yet ultimately, manner of articulation is responsible for the great majority of variation in noise duration word-finally, as both the category and contrast distributions show. This effect would be further pronounced if unreleased plosives were taken into consideration as instances of null noise durations; however, we have retained the convention that unmeasurable quantities are treated as missing (i.e., NA's) and thus do not contribute to category or contrast averages.

2.4.4.4 Summary

Noise duration has the advantage over consonant duration in being simpler acoustically and articulatorily, while retaining many of the same phone and feature effects, and that too with generally less variance. Plosives generally are shorter in ND and DUR_C than affricates, which are shorter still than fricatives, and within each manner class voiced obstruents tend to be shorter than voiceless. The one notable distinction between the two in terms of their role in the system comes in the postvocalic context. In VCV and VC positions, consonant duration shifts to a parameter that primarily indexes voicing, whereas noise duration primarily distinguishes manner classes. Word-initially, ND and DUR_C are largely the same measurement, and so their role is conflated, but more

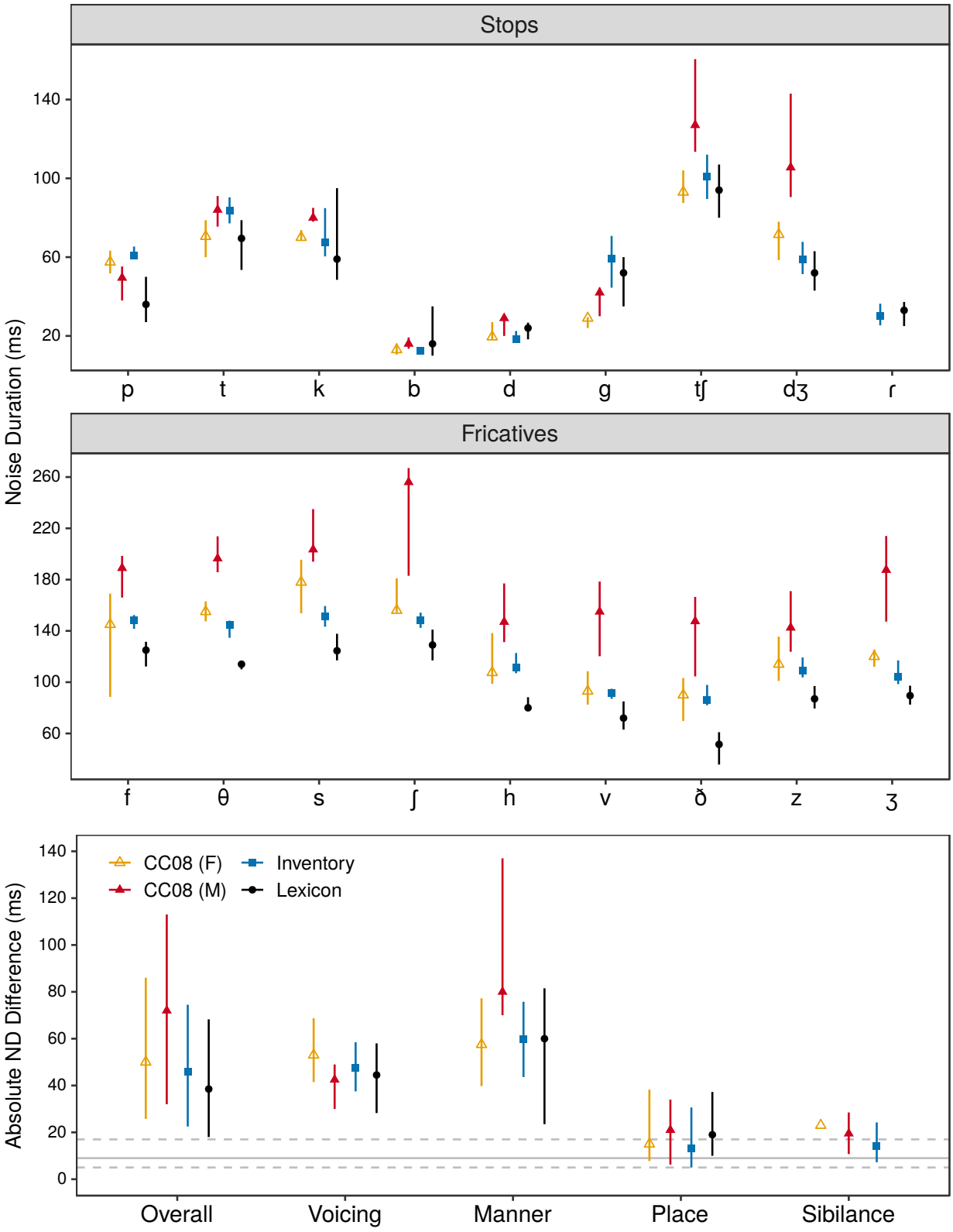

Figure 2.12: Noise Duration (ND) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in noise duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

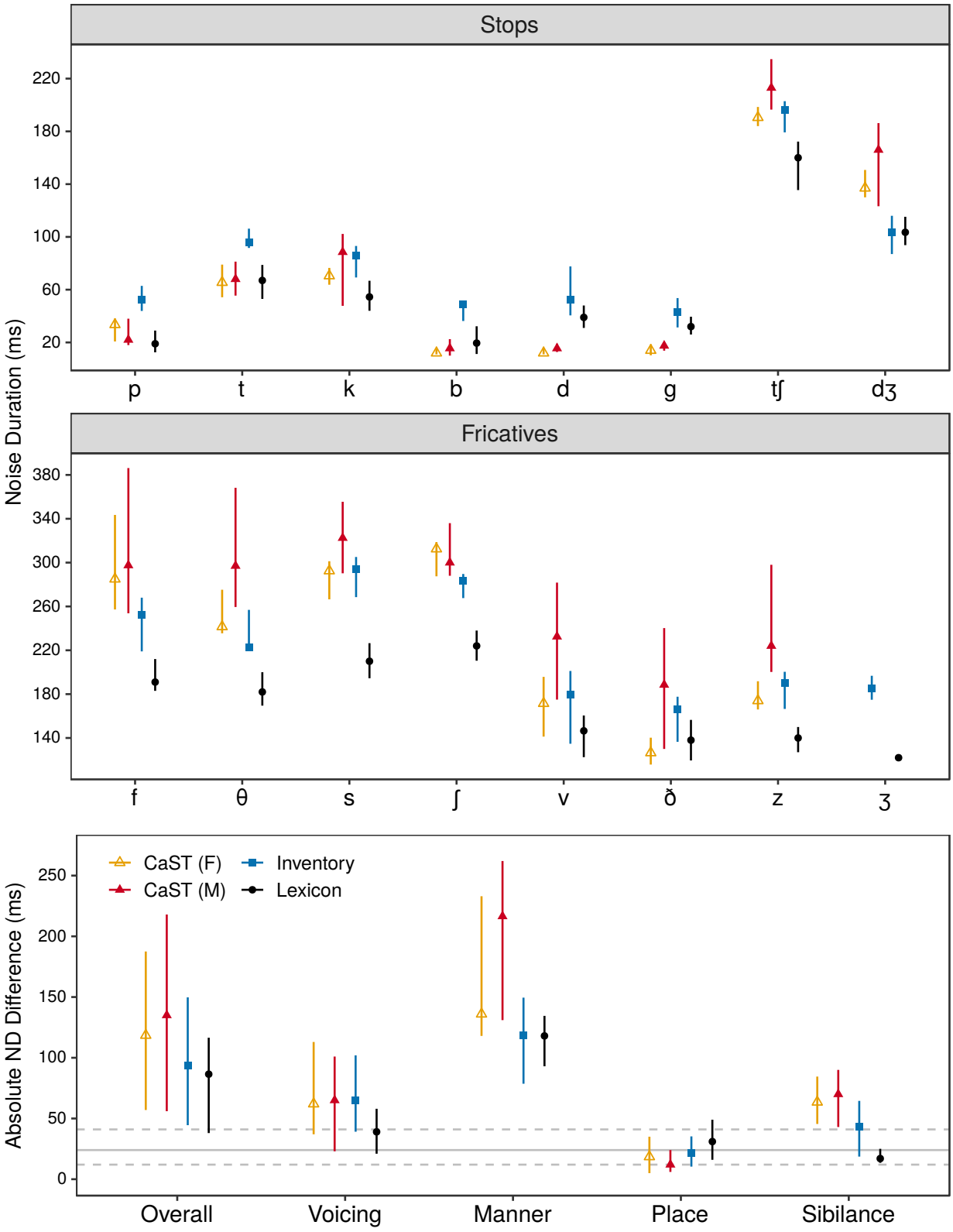

Figure 2.13: Noise Duration (ND) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in noise duration in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

broadly noise duration reflects manner of articulation, and consonant duration reflects voicing.

2.4.5 Voice Onset Time (VOT)

2.4.5.1 Background and physiological basis

The complete definition and demonstration of *voice onset time* (VOT) was presented in the seminal 1964 paper by Leigh Lisker and Arthur Abramson, where VOT—the difference between the point of voicing onset (for the consonant or following vowel in a CV sequence) and the point of release of the consonant closure release—was shown to reliably index both two-way (Cantonese, Dutch, English, Hungarian, Spanish, Tamil) and three-way (Eastern Armenian, Korean, Thai) voicing distinctions, though for two languages with four-way contrasts—Hindi and Marathi—VOT was not able to distinguish breathy-voiced stops ('voiced aspirated' stops, in Lisker and Abramson's terminology) from plain-voiced ('voiced unaspirated') stops. VOT has received further study acoustically, articulatorily, perceptually, developmentally, and cross-linguistically in thousands of studies since Lisker and Abramson's initial work (see Abramson & Whalen, 2017, for review).

VOT was anticipated in earlier work at Haskins on the delay in F1 onset relative to F2 (also known as the *F1 cutback time*), where aspiration noise was added in speech synthesis from the Pattern Playback to fill the interval between F2 and F1 onset, effectively generating a VOT distinction alongside the formant cue of interest (Liberman, 1993). This relationship between VOT and F1 establishes an important acoustic and physiological relationship that will be revisited in the section on F1 onset/offset in the discussion of spectral parameters in Section 2.6. Finally, the physiological basis of voice onset time is relatively straightforward, and is one of the reasons it has been the focus of such active exploration in the years since Lisker & Abramson (1964): VOT captures the relative timing in voicing and consonantal noise onset, which can also be seen as the degree of laryngeal coarticulation between consonant and vowel, which depends in part on the phonological voicing status of the consonant. Traditional restrictions to plosive series have also defined this anchor point as the point of consonant release, but extensions to fricatives, such as in Massaro & Cohen (1976), require that the more general *noise onset* point be used. Further, this modification

allows for differences in laryngeal timing among fricatives and affricates to be captured as well, which all should vary in voice timing to a parallel, though not necessarily identical, degree.

2.4.5.2 Definition and measurement

Voice onset time is defined as the relative temporal lag between the onset of obstruent noise and the onset of voicing in a consonant-vowel sequence. Here, *noise onset* is used as a generalization of the more common definition based on the point of consonant release, as the release burst is only applicable as a landmark in stops, whereas noise onset applies to all obstruents. That is, VOT $= t_{vx} - t_n$, where t_{vx} is the time of voice onset, and t_n is the time of noise onset. From this definition, cases of *prevoicing*, where voicing occurs prior to noise onset (e.g., when stop consonants exhibit voicing during the closure) have negative VOTs; however, across the obstruent class positive VOTs are expected to be more common. This is because definitionally, fricatives cannot have a negative VOT, as the point of noise onset (t_n) defines consonant onset as well, meaning that the theoretical minimum VOT for a fricative is 0 ms, indicating a fully voiced token where voicing begins at or prior to noise onset. Such cases are rare word-initially, but quite frequent intervocalically.

Demonstrations of the measurement of the two critical VOT landmarks—noise onset (t_n) and voice onset (t_{vx})—are shown in Figure 2.1 for several obstruents of different voicing and manner classes. For stop consonants (the first four panels, illustrating intervocalic [k, g, ʧ, ʤ], in Figure 2.1), noise onset is generally marked by a release burst; i.e., an acoustic impulse defined by the sudden introduction of noise in the waveform (in contrast to the more gradual onset of noise in fricative consonants) and broad distribution of energy across the entire frequency range of the spectrogram, a consequence of the flat spectra characteristic of impulses. For fricatives, noise onset is identified through the combined observation of noise in the waveform and in the spectrogram. Here, as with consonant release (t_r), we do not distinguish between frication noise (noise due to turbulence at a supralaryngeal constriction, or due to the deflection of airflow by a downstream surface such as the palate or the teeth) and aspiration noise (noise generated in the larynx under non-modal phonation), and so many points of noise onset, such as in the bottom-left panel of Figure

2.1, correspond to the initial states of an interval of pre-aspiration at the transition between modal voicing during a vowel and the voicelessness of a following fricative consonant. This interval defines what is known as the *voice cessation time* (VCT), which is reviewed in Section 2.4.6.

The onset of voicing, t_{vx}, is measured primarily from the identification of the initiation of periodicity in the waveform, though the appearance of low-frequency energy in the spectrogram (referred to sometimes as the 'voice bar') is also an indicator of voicing onset. In order to distinguish periodicity due to vocal fold vibration from non-laryngeal periodic sources (e.g., uvular vibration in some dorsal obstruent releases, or periodic sources or line noise in the recording environment), voicing cycles are tracked backward from the vowel, and the period duration of the first cycle at voicing onset is compared with that in the vowel to ensure that the two are part of the same laryngeal gesture. To be clear, period durations are expected to change during different phases of the same gesture (consonant closure, aspiration at CV transition, modal voicing during the vowel), but such differences should be within a restricted range and should change gradually.

2.4.5.3 Category and contrast distributions

VOT distributions by category and featural contrast are reviewed below for word-initial and word-medial positions. Voice onset time is undefined for VC position, and therefore such contrasts are excluded from the analysis below; however, the relative timing of laryngeal gestures between consonant and vowel in word-final position will be captured in a related measure, *voice cessation time* (VCT), which is addressed in the subsequent section (§ 2.4.6).

Word-initial position (CV). Figure 2.14 shows VOT distributions for obstruents and obstruent contrasts in CV position. As the featural breakdown in the bottom panel of Figure 2.14 indicates, the primary dimension captured by VOT is unsurprisingly voicing, though VOT distinctions are much larger in controlled syllables, both in the inventory and reference data, than in the lexicon. The voiceless $>$ voiced relation is observed in all three manner classes in every data set, with affricates showing the narrowest difference, though still quite robust at around 50 ms, and fricatives

the greatest, between 75 and 200 ms. Plosives show an average distinction of 50 and 100 ms between voiced and voiceless. This interaction between manner and voicing also implies that when voicing is controlled, VOT may also serve as an indicator of manner of articulation; i.e., for voiceless obstruents, we have the relation *plosive* < *affricate* < *fricative*, whereas for voiced obstruents we find fricatives exhibit the lowest VOTs, followed by plosives, and then affricates. These relations between category distributions are further grounded in the slight but robust manner contrast effect in Figure 2.14, particularly in the lexicon data.

Within a given voicing and manner class, obstruents are relatively similar in VOT. There are slight place distinctions between voiceless plosives, with velars showing a slightly longer voicing lag than coronals or labials, though the distinction between the latter two is less consistent across data sets. Among voiced plosives, the only notable place effect is the relatively greater occurrence of prevoicing in [g] than in [b] or [d], resulting in IQRs that extend into the negative range of VOT. In general, place contrasts are not well indexed by VOT, as the bottom panel of Figure 2.14 shows.

Among the voiceless fricatives, excepting [h], there are no sizeable distinctions in VOT for the reference data; however, for the target speaker [θ] is notably lower in voice onset time than [f, s, ʃ] in both lexicon and inventory data; this increase in the distinction between [θ] and [s] is responsible for the sibilance distinction shown in the bottom panel of Figure 2.14. The other major factor in generating a sibilance effect in the lexicon is the frequent occurrence of negative VOTs for the voiced labiodental, [v]. As noted earlier, this distribution (along with that of [ð] in the inventory data), which diverges substantially from the consistent 0 ms VOTs of the other voiced fricatives and the [v] tokens in the other databases (indicative of fully voiced fricatives), derives from the common fortition of [v] into a plosive [b̪] with a long prevoicing interval.

Finally, all data sets show generally lower VOTs for the glottal fricative than for the remainder of the voiceless fricatives, though in the lexicon this difference is negligible. However, despite this distinction, [h] remains more clearly aligned with the voiceless series than with the voiced.

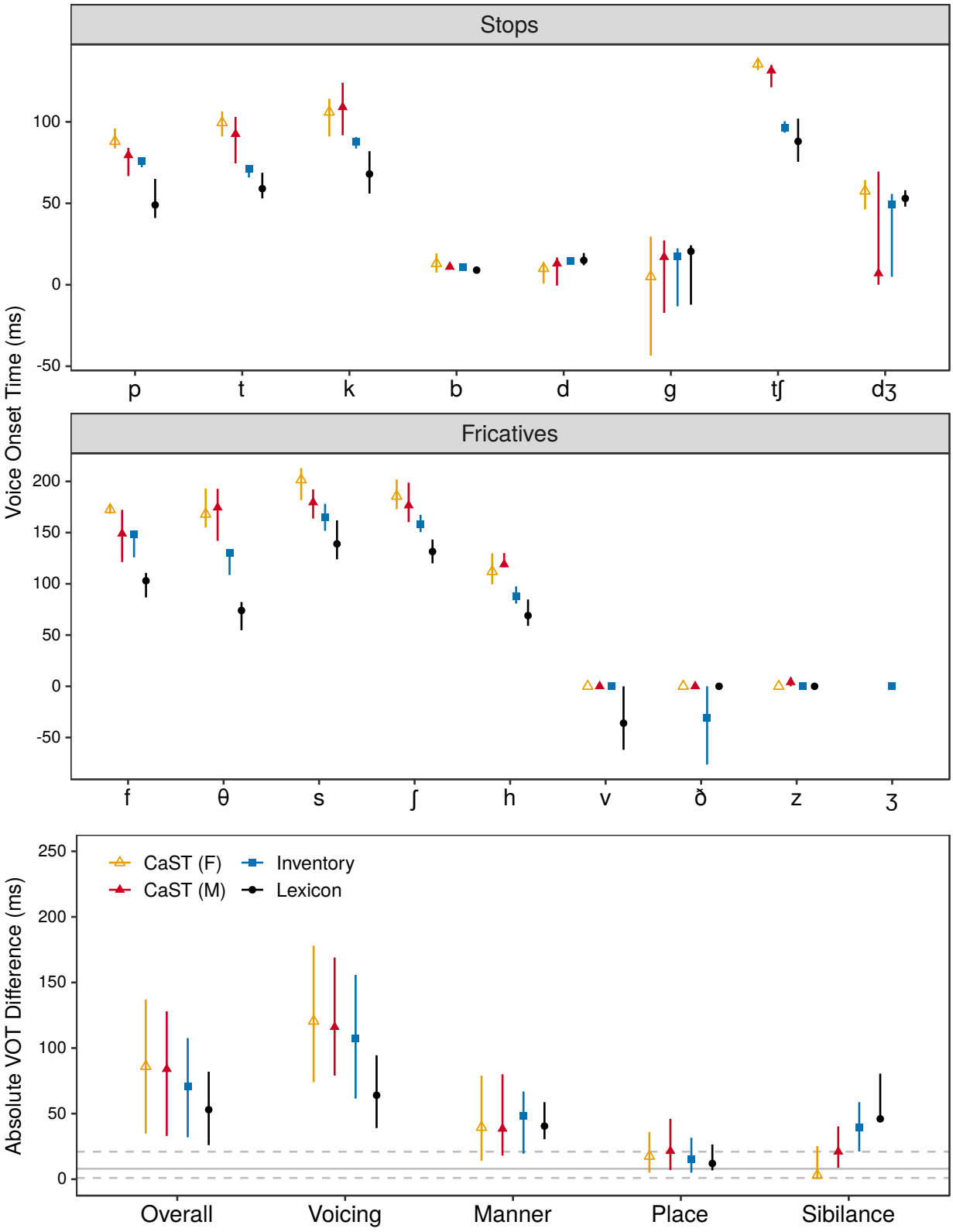

Figure 2.14: Voice Onset Time (VOT) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in VOT in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

Word-medial position (VCV). Turning next to intervocalic obstruents, Figure 2.15 indicates a much more robust role of VOT in distinguishing voicing contrasts than in CV position, and in fact here we see a major difference between the target speaker's data and the reference data from Cooke & Scharenborg (2008), where VOT differences in voicing contrasts are notably more distinct in the target data. VOT distinctions in manner contrasts are also more robust in VCV position than in CV, though the sibilance effect disappears intervocalically. Here we should note that no pure sibilance distinctions are present in the lexicon in VCV position, and such distinctions are in general quite rare given that sibilance and place overlap to such a great extent; the only sibilance contrasts that are otherwise matched for place, manner, and voicing are [s, θ] and [z, ð].

Examining the VOT distributions in individual obstruent categories, we see that VCV position is where prevoicing begins to play a critical role, especially in the lexicon, where all stops have median VOTs well below zero. Voiceless plosives exhibit relatively shorter VOTs intervocalically, in ther range of 25-75 ms, but given the considerable prevoicing of voiced plosives the end result is a wider voicing gap of 100-150 ms. Further, prevoicing is also robust in the target speaker's productions of [ʤ], leading to a VOT distinction between voiced and voiceless affricates that is larger than in the other two manners.

Regarding place of articulation, intervocalic obstruents show the same negative correlation between VOT and the anteriority of constriction as in CV position; namely, VOT decreases with more posterior articulations. This distinction is most robust between labials and non-labials, though voiced plosives, particularly in the lexicon, show a further distinction between alveolars and velars. This increase in the effect of place of articulation among voiced plosives in VCV position relative to CV can be linked to the aerodynamic constraint on maintaining voicing when the size of the supralaryngeal cavity is reduced. Given that nearly all intervocalic stops are prevoiced, voicing can be maintained for a longer duration during labial closures before the pressure equalizes above and below the glottis than it can when the closure is at the alveolar ridge or velum, resulting in [b] having the longest negative VOT, followed by [d], and finally [g]. In this respect, the prevoicing duration of postalveolar [ʤ] is exactly what we expect: shorter (i.e., yielding a greater VOT) than

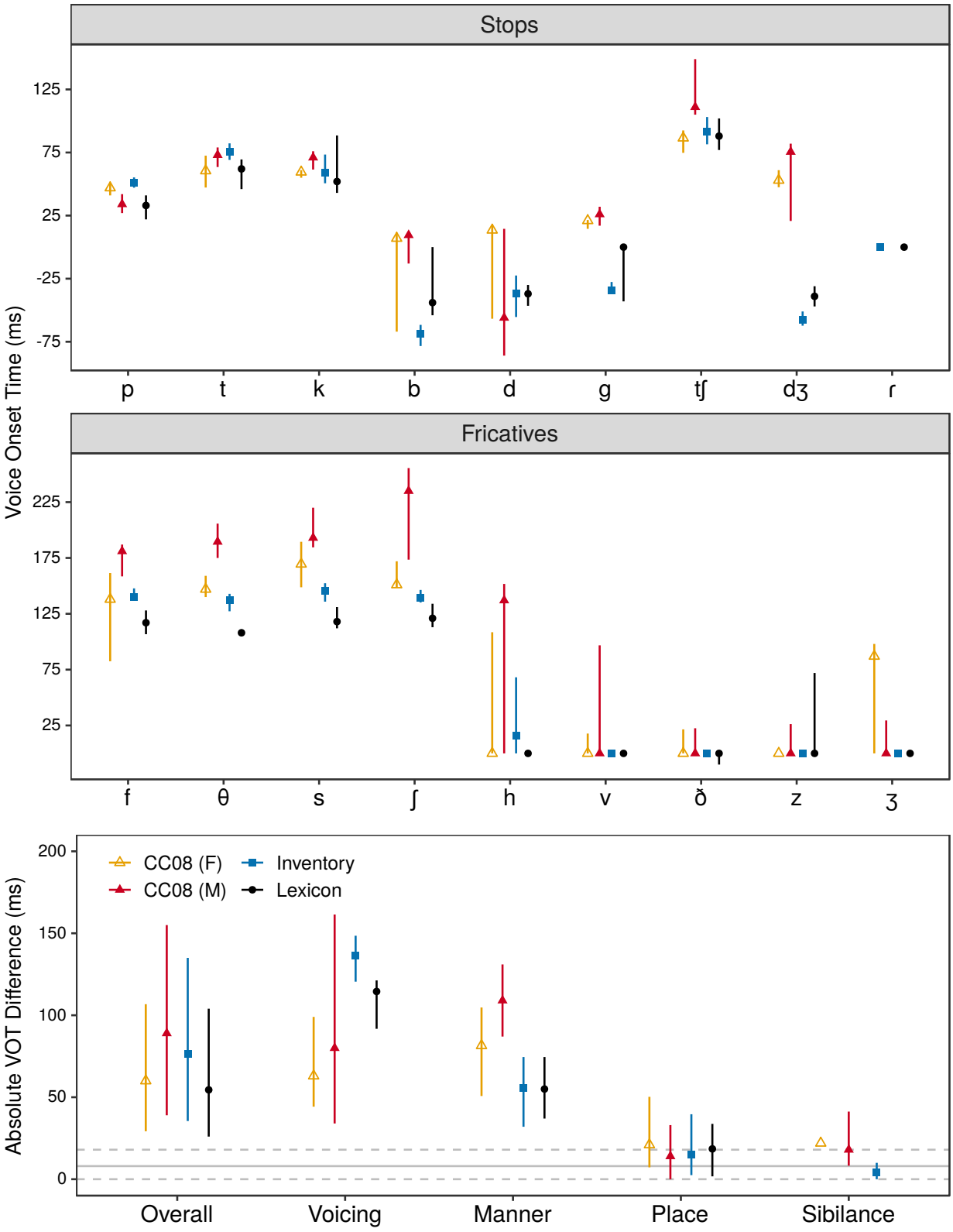

Figure 2.15: Voice Onset Time (VOT) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in VOT in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

labial [b] but longer than velar [g], and relatively consistent with alveolar [d]. Finally, we see that intervocalically, [h] is consistently voiced and diverges from the CV result in patterning much more closely in VOT with voiced fricatives than voiceless.

2.4.5.4 Summary

Both word-initial and word-medial VOT distributions are consistent in providing a robust acoustic index of voicing contrasts. VOT distinctions in such contrasts, particularly in the target data, are well above estimated chance levels in both positions. Further interactions between manner, place, and voicing yield regularities in VOT that are promising for the discrimination of obstruent contrasts. In addition to the robust voiceless $>$ voiced relation, there are manner distinctions within voicing classes, with plosives $<$ affricates $<$ fricatives in both word-initial and word-medial voiceless obstruents, fricatives $<$ plosives $<$ affricates in voiced CV obstruents, and plosives $<$ affricates $<$ fricatives in voiced VCV obstruents. Finally, there are minor but consistent place effects within a given voicing and manner class, where in stop consonants, VOT generally increases with more posterior articulations.

2.4.6 Voice Cessation Time (VCT)

2.4.6.1 Background and physiological basis

The study of voice cessation time first received thorough analysis in Docherty (1992), who examined laryngeal timing over all portions of the vowel-consonant-vowel interval. Theoretically it is a simple counterpart to voice onset time, and captures the relative lag in laryngeal control over the separation of vocal folds and the initiation of voicelessness. This parameter is also closely related to measures of the presence or absence of closure voicing, but is more precise in distinguishing perseveratory voicing coarticulation (in VCT) from anticipatory coarticulation in VOT. Voice cessation time is conceived of as a cue to voicing distinctions, but due to its dependence on laryngeal and supralaryngeal timing, VCT may also vary as a function of manner and place of articulation, as

different locations and degrees of obstruent constriction have different aerodynamic consequences for the maintenance of vocal fold vibration.

2.4.6.2 Definition and measurement

Voice Cessation Time (VCT) is defined as the duration of the interval between the point of consonant constriction onset, t_c, and voicing offset, t_{vo}; i.e., VCT $= t_{vo} - t_c$. See Figure 2.1 for sample measurements of VCT for a range of voicing and manner classes. The identification of consonant constriction onset has already been discussed in the context of DUR_{C}, $\text{DUR}_{\text{V1/V2}}$, and CD measurement, while the identification of voicing offset mirrors that of voicing onset in VOT measurement. Namely, t_{vo} is measured at the point where periodicity ceases in the waveform and there is a notable drop in low-frequency energy in the spectrogram. In another important respect, VCT is the mirror of VOT in the behavior of boundary cases such as when voicing continues throughout the consonant. For VOT, we saw that such cases either result in a VOT of 0, such as in fricative consonants when noise and voice onset are coterminus in the consonant interval, or a negative VOT that at maximum can be as long as the closure duration, despite the fact that in VCV and VC contexts voicing was initiated much earlier for the vowel gesture. In the definition of VCT, when voicing continues throughout the consonant, VCT becomes equivalent to consonant duration.

Both solutions to cases where landmarks such as t_{vx} and t_{vo} fall outside of the consonant interval are not ideal from the standpoint of directly reflecting the timing of laryngeal gestures, and represent the inevitable conflict between attempting to map continuous gestures onto discrete, bounded segments. Nevertheless, the solutions presented above retain the general goal of establishing acoustic parameters that are consistent and monotonic in their delineation of obstruent categories and phonological features: voiced obstruents should have longer VCTs and shorter/negative VOTs relative to voiceless obstruents. And thus, we retain them for use in the present study while noting important caveats in their interpretation.

2.4.6.3 Category and contrast distributions

Distributions of voice cessation times by category and featural contrast are reviewed below for word-medial and word-final positions. Voice cessation time is undefined for CV position in isolated words where word onset coincides with utterance onset.

Word-medial position (VCV). Figure 2.16 shows VCT distributions for intervocalic obstruents. As with VOT, there are clear voicing distinctions that are robust across manner classes. In plosives, VCT differences between voiced and voiceless average around 40-60 ms, whereas for fricatives and affricates they are moderately greater, at between 80 and 100 ms. However, unlike VOT, there is no significant manner effect for voice cessation time, a result that can be seen in Figure 2.16 from the relatively constant VCT in voiceless obstruents. That is, the manner distinctions described above are due to effects of manner on VCT in voiced obstruents. This constraint, though responsible for the reduction in overall manner effects relative to those observed for VOT, does mean that VCT remains a potential cue to manner of articulation in voiced obstruents.

Place effects on voice cessation time are generally absent, the one exception being the consistently lower VCT in the voiced alveolar plosive [d] relative to [b, g]. This result is consistent with the consonant duration pattern, and likely reflects continuous voicing throughout the consonant closure, in which case the two measures are equivalent. Therefore, this place distinction may reflect effects of place of articulation on supralaryngeal rather than laryngeal gesture timing, in which case VCT does not provide any independent place information.

Word-final position (VC). Voice cessation times word-finally are generally shorter than those word-medially because there is no following-vowel trigger for the continuation of voicing throughout the consonant. As Figure 2.17 shows, however, this result has little effect on the VCT differences between voiced and voiceless stops, as both voiced plosives and the voiced affricate [ʤ] exhibit notably greater voicing into the closure (on the order of 50-100 ms) than their voiceless counterparts. Further, the place distinction among voiced plosives—alveolar $<$ labial/velar—remains

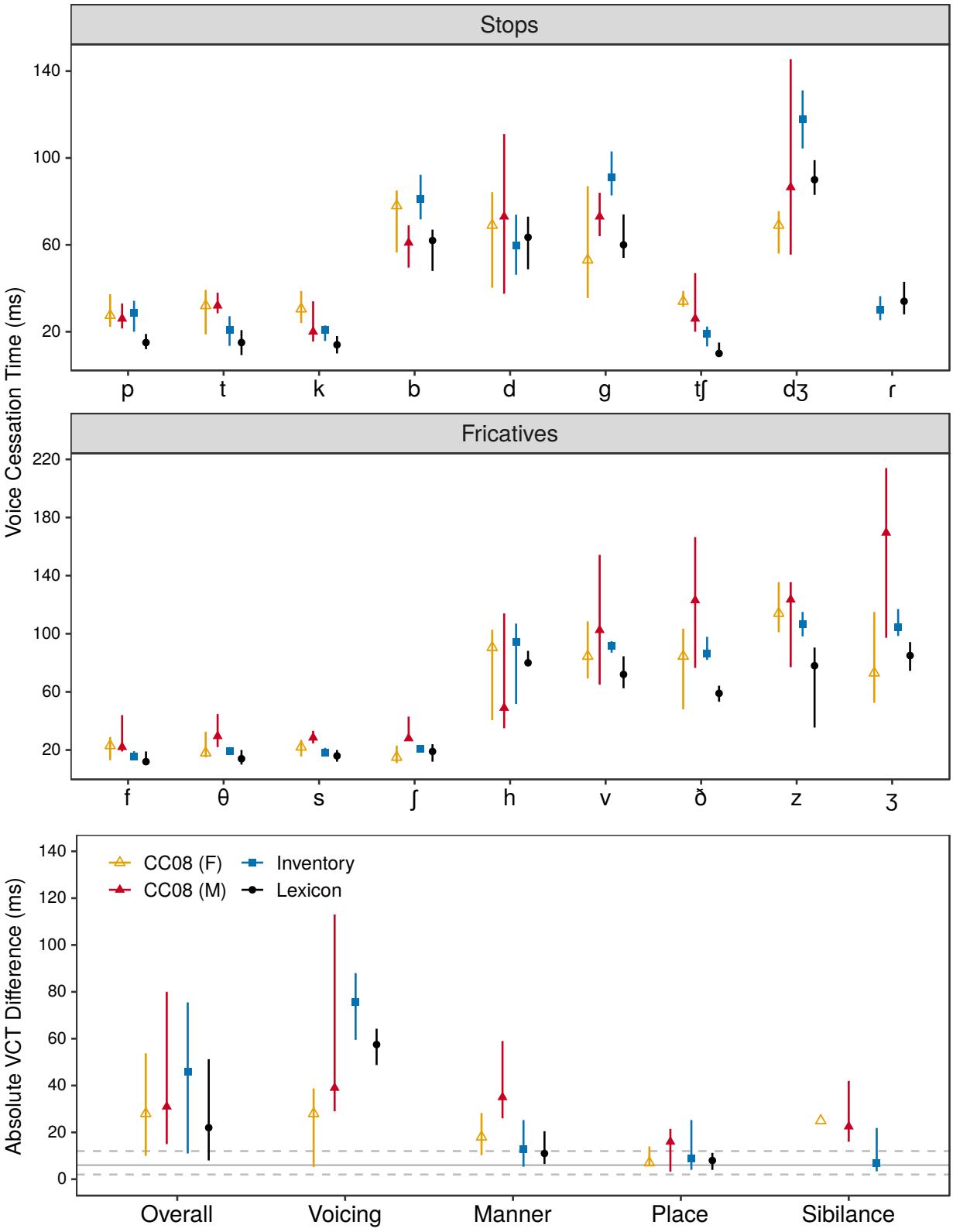

Figure 2.16: Voice Cessation Time (VCT) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in VCT in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

and is more consistent with the closure duration pattern than with consonant duration, indicating that this may indeed reflect an articulatory difference based on the relative timing with which different articulators can achieve consonant closure and release (alveolars, utilizing a tongue tip gesture, should be faster in this regard than labials or velars). These results are generally preserved in both word and syllable data, and in both target and reference speakers.

Fricatives, on the other hand, show a notable departure of the lexical pattern from the syllable pattern. In controlled syllables, there remains considerable voicing in word-final voiced fricatives, leading to a robust VCT difference of 100-150 ms. However, in the lexicon this set is almost entirely devoiced, showing voice cessation times that are comparable to those in the voiceless fricatives and well below what we observe in voiced stops. Thus, for fricatives in real word productions, listeners may need to rely more on other cues such as the preceding vowel duration to distinguish voiced from voiceless. Finally, this neutralization of fricative VCT distinctions results in an overall voicing contrast effect that is absent from the lexicon though it is present and robust in the syllable data.

2.4.6.4 Summary

For stop consonants, the time it takes for voicing to cease from the preceding vowel is a reliable indicator of the voicing distinction in both VCV and VC obstruent contrasts. Voiced stops exhibit longer VCTs than their voiceless counterparts in a manner that is consistent with both consonant and closure duration patterns reported above. Further, among voiced plosives there is a place distinction between alveolars labials/velars that is preserved in both positions. However, for fricatives VCT is only a reliable voicing cue in word-medial position, where it is much more common for voicing to continue throughout the consonant as a way of minimizing laryngeal effort.

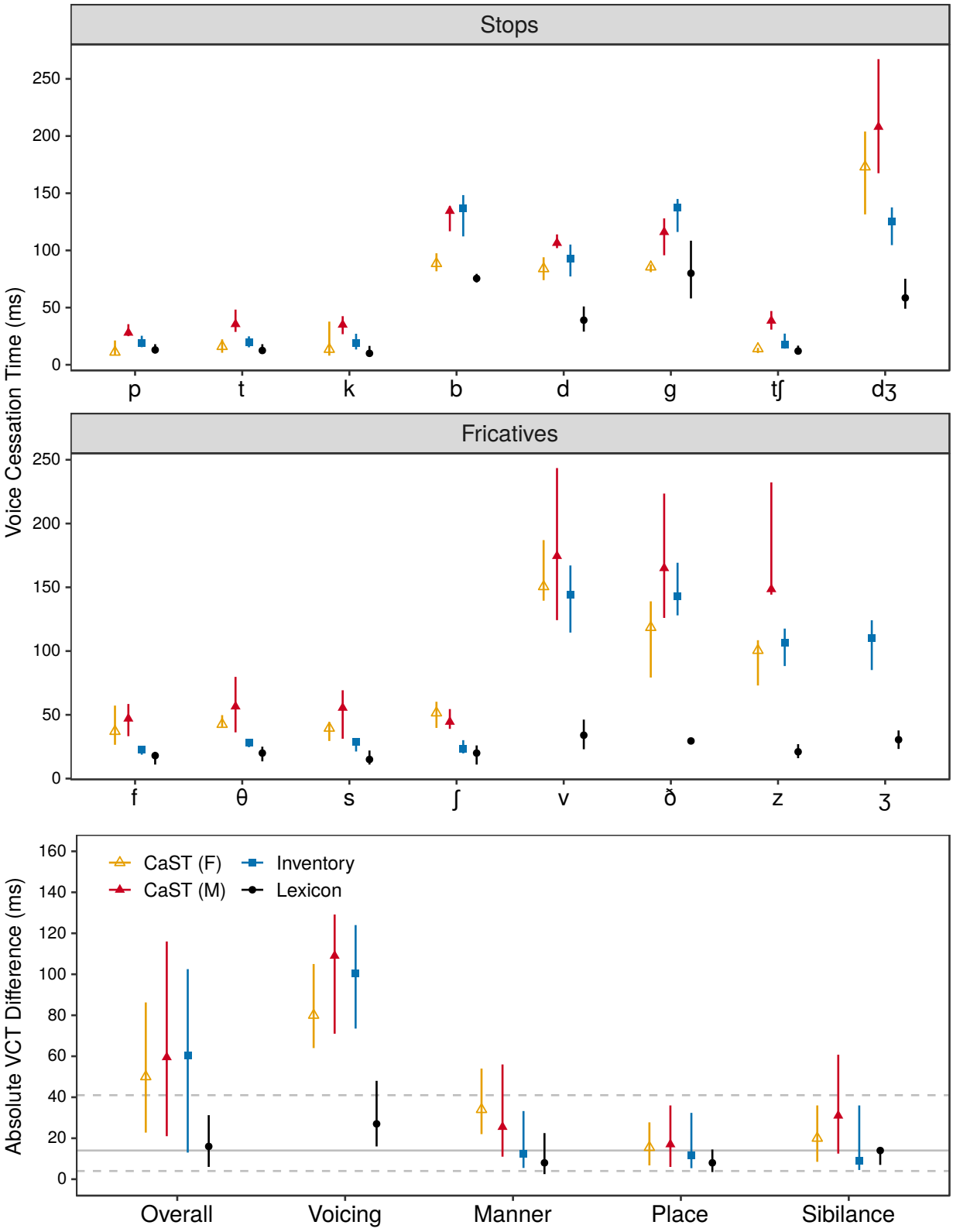

Figure 2.17: Voice Cessation Time (VCT) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in VCT in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

2.4.7 Consonant Voicing Percentage (VOI%)

2.4.7.1 Background and physiological basis

The percentage of the consonant interval that is voiced is a derived parameter, combining several measures discussed above into a simple metric for the discrimination of voiced from voiceless obstruents. This measurement has primarily been used in phonological and typological work (cf. Myers, 2002; Colantoni & Steele, 2007), but has physiological and acoustic precursors in the study of voice timing in Docherty (1992). Briefly, we expect phonologically voiced obstruents to be voiced over a greater percentage of the consonant interval than their voiceless counterparts because unlike with voiceless obstruents, voicing from adjacent vowels does not need to be terminated in order to cue the category. This means that even in devoicing environments in English, such as word-initial and word-final position, there should be greater voicing coarticulation, and therefore greater voicing percentages, in voiced obstruents relative to voiceless ones. Therefore, as a physiologically grounded measure, *consonant voicing percentage* captures both active and passive differences in laryngeal control in voiced and voiceless obstruents, the former referring to active control of the larynx to produce voicing during the closure or noise interval, and the latter to the passive spread of voicing due to laryngeal coarticulation and the lack of active control over voice suppression.

In this latter respect we expect differences in voicing percentage may also arise as a function of manner of articulation, as different manner classes have both different coarticulatory propensities and different contrast constraints in terms of the sounds they need to remain distinct from (e.g., glottal fricatives regularly undergo intervocalic voicing, which is not an impediment to speech transmission as there is no voiced glottal counterpart to [h] in English).

2.4.7.2 Definition and measurement

The definition of Consonant Voicing Percentage (VOI%) can be derived from the previous three parameters—Noise Duration (ND), Voice Onset Time (VOT), and Voice Cessation Time (VCT)—in

the following manner:

$$\text{VOI\%} = \min\left\{100 \cdot \frac{\text{VCT} + (\text{ND} - \text{VOT})}{\text{DUR}_\text{C}}\ ;\ 100\right\},$$

where ND and VOT are put in parentheses to indicate that if VOT is not measurable (i.e., is a missing value, NA), such as in word-final position, the entire (ND − VOT) term is removed from the sum in the numerator. Similarly, missing values for VCT, which occur in word-initial position, are dropped from this sum. The fact that VOI% can be derived from the other temporal parameters introduces some redundancy into our analysis, just as with consonant duration as the summation of closure and noise duration. However, given that the cue integration model adopted in Chapter 4 is not a linear model (it uses a Bayesian variant of the bagged decision tree model) this dependency does not introduce collinearity problems. Further, the composition of VOI% from VOT and VCT, among other parameters, may provide a more efficient voicing cue than the use of VOT and VCT separately. The relative utility of VOT, VCT, and VOI% will be assessed in the cue integration results later in this chapter and in Chapter 3.

2.4.7.3 Category and contrast distributions

Word-initial position (CV). Figure 2.18 shows consonant voicing percentages (VOI%) in word-initial obstruent phones and contrasts. Among stops, percentages are much higher in voiced plosives than in the voiced affricate [ʤ], and are further highly variable, ranging up to 100% in some cases. This leads to a robust voicing distinction among plosives that is not preserved in the affricates, where word-initially [ʧ] and [ʤ] are both mostly voiceless, with the primarily differentiated by noise duration, [ʧ] being generally longer, consistent with patterns among plosives and fricatives. Further, there is a trend in the reference data from (Woods et al., 2010) for voiced plosives to show greater voicing percentages with more posterior places of articulation; however, this trend is much reduced in the inventory data and absent entirely from the lexicon, though [g] does show a longer VOI% range that extends up to 100%. Among fricatives, voicing percentage be-

comes a near-dichotomous variable separating voiced from voiceless, as the voiced fricatives are all at or near 100% voiced across the databases, and conversely, voiceless fricatives are near-zero, ranging between 0 and 10% on average. As a result, there is a robust voicing contrast effect for VOI% that is present in both word and syllable data, though the ranges for this effect are wide (20-90%). No other featural contrasts are reliably distinguished by consonant voicing percentage.

Word-medial position (VCV). In VCV position we see a further trend toward the dichotomous potential of VOI% as an index of obstruent voicing. In the target data, plosives and affricates extend the fricative pattern from CV position in showing consistent 100% voicing of [b, d, g, ʤ], while voiceless obstruents are all around 20% voiced. As Figure 2.19 shows, however, the reference data shows greater devoicing of voiced stops intervocalically, but remains robust at distinguishing voiced from voiceless by approximately 50%. This discrepancy between the two data sets leads to a much more robust voicing contrast effect in the target data than in the reference data, as well as a manner distinction in the reference data that is not replicated in the target inventory or lexicon. Finally, [h] is commonly fully voiced intervocalically, leading it to pattern with the voiced fricatives in VOI% in VCV position.

Word-final position (VC). Figure 2.20 shows consonant voicing percentages in word-final position, and diverges from the CV and VCV patterns above in a manner consistent with the VCT results reported in the previous section. Voiced fricatives in real words are mostly devoiced word-finally, though there is around a 10-15% difference between [f, θ, s, ʃ] and [v, ð, z, ʒ] that retains some potential for distinguishing the two sets. Voiced plosives, with the exception of [d] in the lexicon data, are near ceiling in VOI%, while [ʤ] is somewhat lower at between 40 and 80% voiced. This weakening of the distinction among stops, and near-neutralization of the distinction among fricatives in the lexicon, results in a voicing contrast effect that remains robust in the syllable data, but which shows some distributional overlap with the estimated chance range in the lexicon.

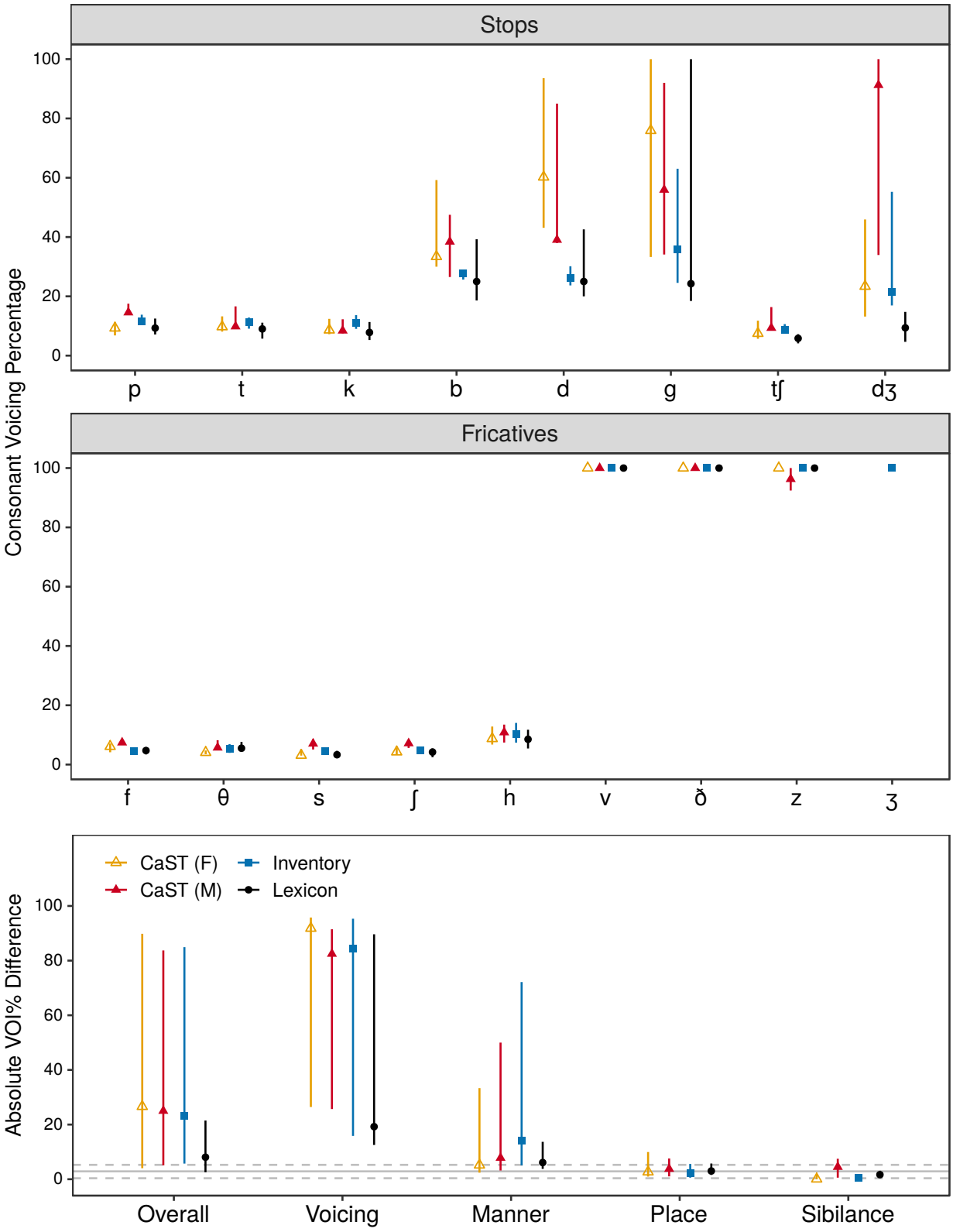

Figure 2.18: Consonant Voicing Percentage (VOI%) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in VOI% in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

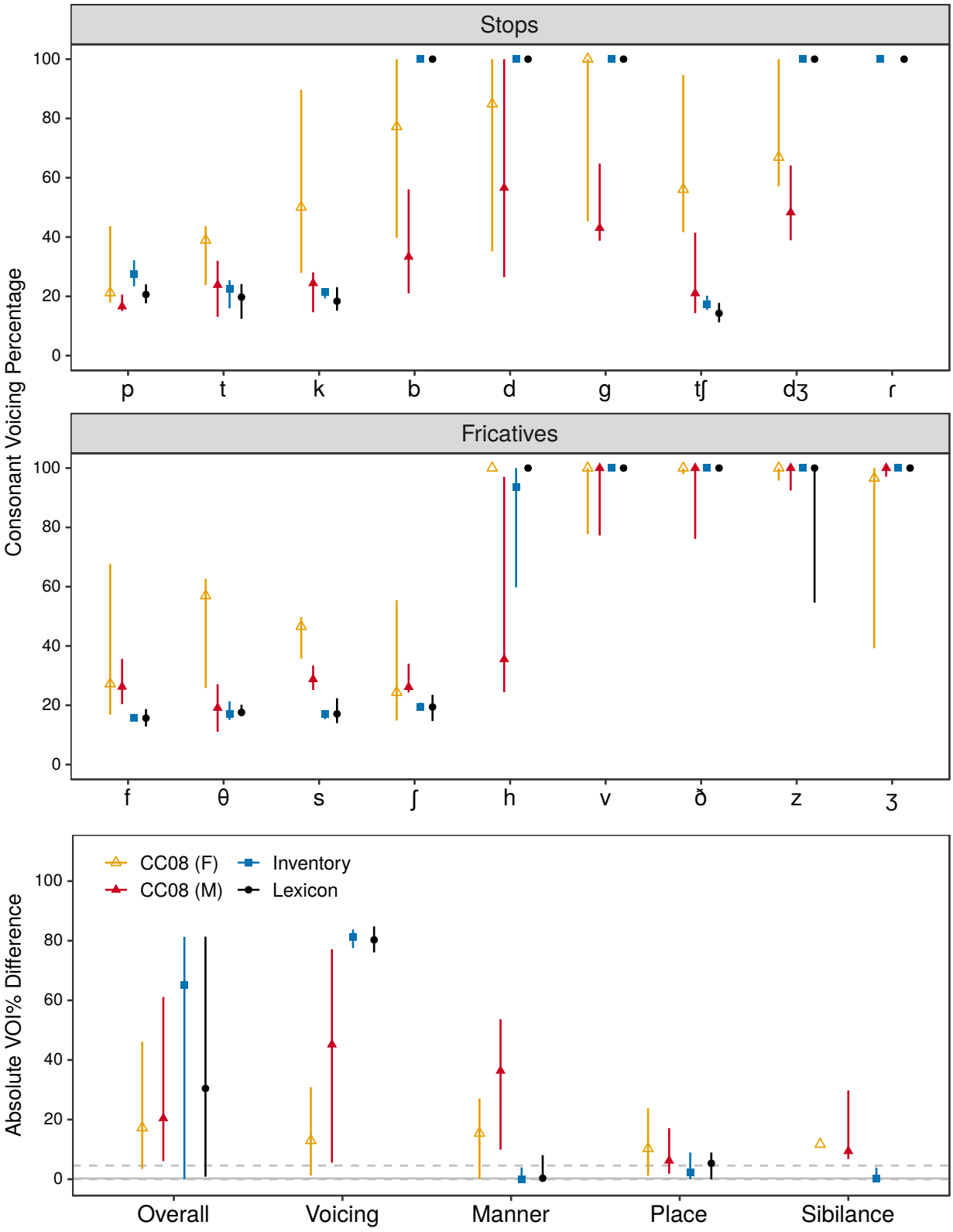

Figure 2.19: Consonant Voicing Percentage (VOI%) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in VOI% in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

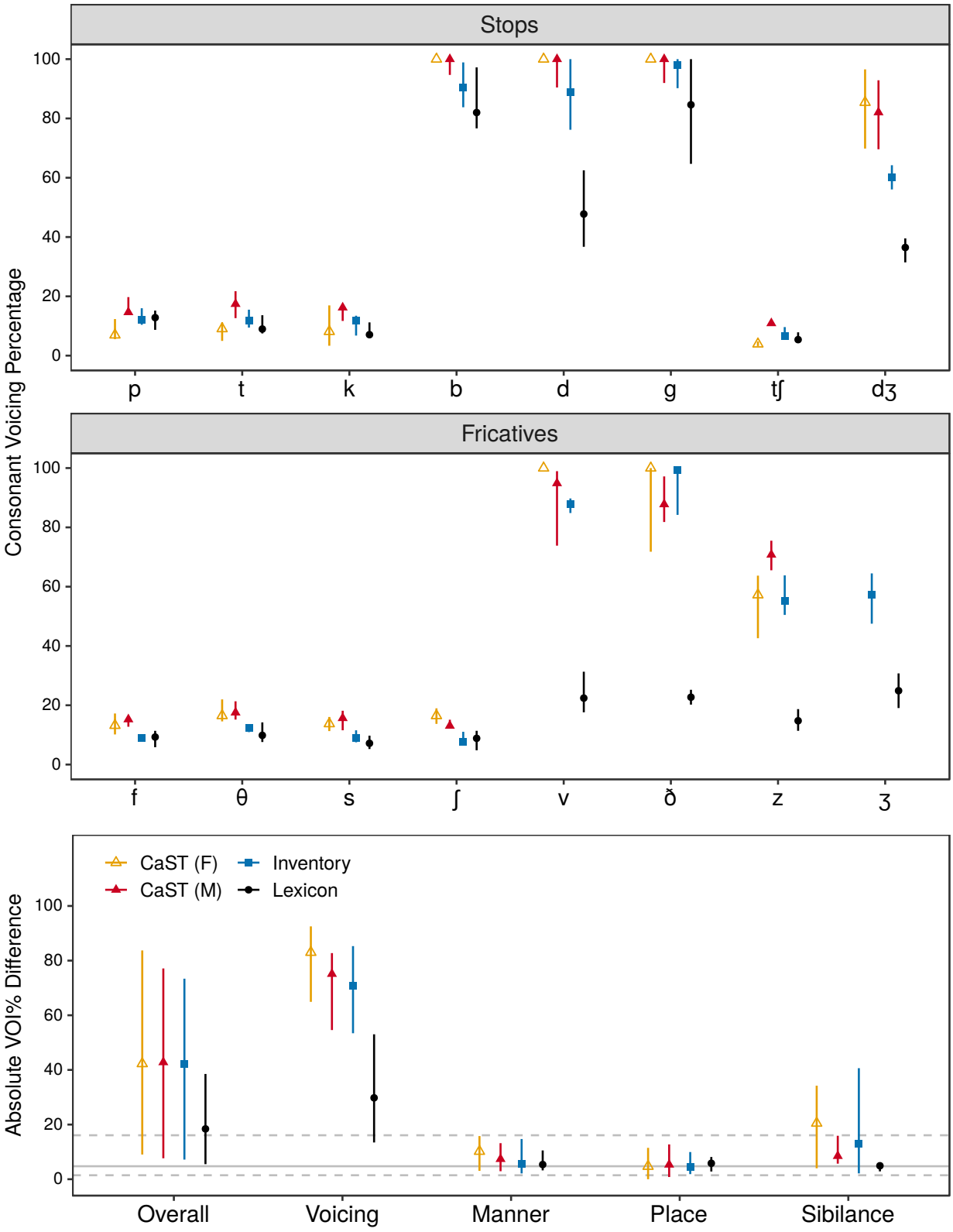

Figure 2.20: Consonant Voicing Percentage (VOI%) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in VOI% in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

2.4.7.4 Summary

As noted earlier, consonant voicing percentage as a derived parameter retains many of the indexical features of VOT, VCT, and DUR_C. However, by collapsing these independent features into a composite measure we gain some efficiency in distinguishing voiced from voiceless, as in many cases—fricatives in CV position, all manners in VCV position, and plosives in VC position—VOI% shows a dichotomous separation of the two classes into either entirely voiced or entirely voiceless. This result may be of particular value in the cue integration models of Chapter 4, which utilize a decision tree learning rule that more easily incorporates dichotomous cues than do linear models.

2.4.8 Comparative discriminative power of temporal parameters

Having reviewed the details of each temporal parameter above, including their definition, measurement, physiological basis, and category/contrast distributions, we now compare the overall discriminative power of each parameter in each database, where critical attention is paid to the distinction between the target *inventory* and *lexicon* databases. Figure 2.21 shows normalized contrast effects—the mean of the absolute value of itemwise contrast differences between each scaled (transformed to range between 0 and 1) parameter—of each temporal parameter in both target and reference data sets in CV, VCV, and VC positions.

Word-initially, the relative discriminative power of the five temporal parameters (DUR_{V1}, CD, and VCT are not measurable in CV position) is remarkably consistent across databases, with consonant voicing percentage (VOI%) the most discrimative in all but the lexicon, followed by consonant duration, noise duration, and voice onset time, and DUR_{V2} consistently the weakest parameter of the five. The considerably smaller contrast effect for VOI% in the lexicon is consistent with the VC pattern for VOI%, and reflects the fact that voiced obstruents at word edges are much more prone to devoicing than in controlled syllables where items are likely to be hyperarticulated. Nevertheless, the consistency between databases in CV position is notable, and reflects, among other factors, the greater balance in word-initial obstruent distributions in the lexicon relative to VCV or

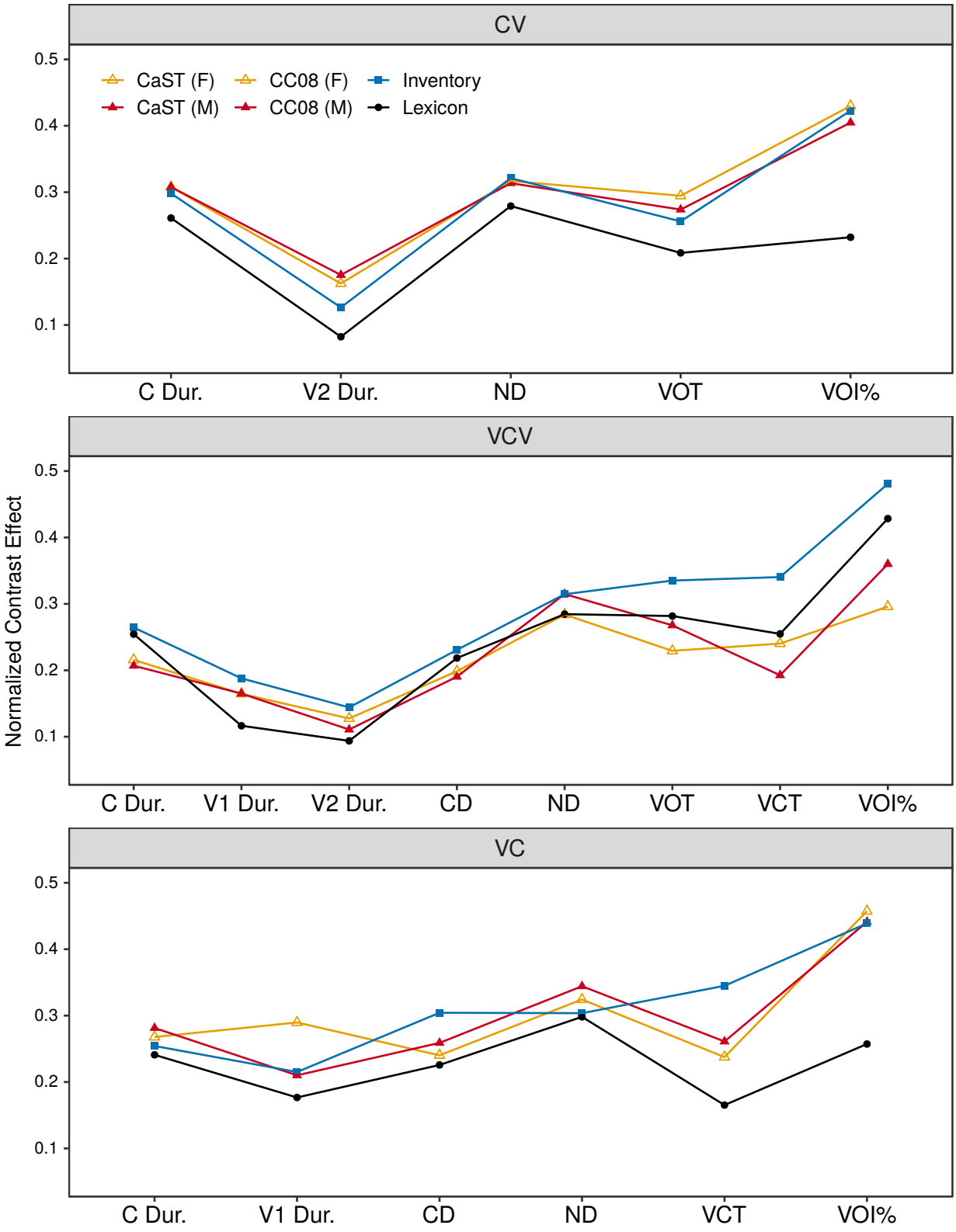

Figure 2.21: Comparative discriminative power of temporal parameters in CV, VCV, and VC positions, as measured by the *normalized contrast effect*, the mean of the absolute differences by contrast between scaled (transformed to range between 0 and 1) parameters. The CaST and CC08 reference data sets share the same color palette because their contexts (CV/VC and VCV, respectively) are mutually exclusive.

VC positions, allowing the set of lexical contrasts to more closely mirror the definitional balance in the inventory and reference databases. We will see later in the presentation of listener perception results in Chapter 3 that such distributional differences between CV, VCV, and VC contrasts in the lexicon are also reflected in listener recognition patterns.

The relative discriminative power of the eight temporal parameters in VCV position shows moderately less consistency across databases than in CV position, but the general patterns are largely preserved. Namely, VOI%, VCT, VOT, and ND are the most discriminative, followed by DUR_C, with CD, DUR_{V1}, and DUR_{V2} the least informative word-medially, particularly in the lexicon. The most notable differences between the real-word and inventory data are the lower relative discriminative power of preceding vowel duration and voice cessation time. The difference in lexicon and inventory DUR_{V1} effects likely reflects the greater prosodic variability in real-word contrasts relative to controlled syllables, while the reduced power of VCT is consistent with the greater voicing of intervocalic voiceless obstruents in real words than in controlled syllables, a pattern which is again consistent with differences in articulation of the two stimulus types—the lexicon exhibiting greater hypoarticulation, and the inventory greater hyperarticulation.

Word-finally, we find the least consistency between databases. VOI% and ND are generally the most discriminative in VC position, though CD and VCT are slightly more informative than ND in the inventory, and slightly less informative in the lexicon, with consonant and vowel duration among the less discriminative parameters overall. The greatest discrepancy between the lexicon and inventory data can be seen in the final voicing cues: VCT and VOI%. Both cues are considerably less discriminative in the lexicon, a result which again is partly a consequence of the greater prevalence of word-final partial devoicing in real words than in nonword syllables. Another important factor that will be addressed in greater detail in Chapter 3 is the distribution of featural contrasts in VC position in the lexicon as compared with the inventory. Because the inventory is completely symmetric in obstruent voicing word-finally (comprising 8 phones each of voiced and voiceless), the systematic pairing of each phone with each other in the inventory results in 53% of

contrasts involving a voicing distinction,[12], as compared with 42% in the lexicon. One of the reasons for this reduced proportion in the lexicon is the large influence of [d, z] contrasts word-finally (21% of VC contrasts in Exp. 1), the majority of which are morphological distinctions between present and past tense.

2.5 Amplitudinal parameters

The next major class of acoustic parameters for obstruent identification/discrimination tracks characteristics of the amplitude envelope of the signal. These parameters—Burst Presence (BURST), Noise Amplitude (AMP_N), and Vowel Amplitude ($AMP_{V1/V2}$)—primarily capture differences in the manner of obstruent production, though due to the effects of the point of constriction and the configuration of the larynx on the aerodynamics of the speech stream, amplitudinal parameters may also reflect voicing, place, and sibilance features.

2.5.1 Burst Presence (BURST)

2.5.1.1 Background and physiological basis

The presence/absence of a consonant release burst received early acoustic and perceptual attention in Fant (1960) and Dorman et al. (1977). Acoustically, a burst is realized as an impulse in the signal: a point of amplitude increase that is of infinite slope in the theoretical limit, and which exhibits a near-uniform distribution of energy in the spectrum. Articulatorily, bursts occur under conditions of rapid pressure and airflow change in the vocal tract due to the release of a complete oral obstruction. They can also be found as artifacts of brief intervals of contact between articulators in the production of the dental/labio-dental fricatives [f, v, θ, ð], or during incomplete phases of tongue dorsum release in the production of velar plosives, but we will be focused primarily in the present study on the first definition, and thus only bursts at the initiation of a noise interval (not after noise onset), will be considered in the measurement of burst presence. This choice provides a clearer

[12]The reason this number is 53% and not 50% is because of the exclusion of identity pairs (i.e., non-contrastive pairs of phones, such as [p, p] or [g, g]) from the measurement of the featural composition of phonetic contrasts.

physiological link to burst presence as an indicator of both the type of consonantal constriction and the timing of consonant release, and is discussed in greater detail in the section below.

2.5.1.2 Definition and measurement

The measurement of burst presence, a dichotomous variable (BURST = 1 if present, 0 if absent), is based on the joint observation of two properties, one in the waveform and one in the spectrogram. In the waveform, bursts appear as an acoustic impulse: an instantaneous jump in noise amplitude, or alternatively, a non-repeating spike in the sound wave. In the spectrogram, bursts exhibit a uniform distribution of energy in the spectrum, apparent as a dark vertical band across the complete frequency range. See Figures 2.1 and 2.22 for sample stop obstruent burst measurements.

2.5.1.3 Category and contrast distributions

Below we present burst presence percentages by category and contrast in the lexicon, inventory, and reference data. Results are considered separately in CV, VCV, and VC positions, but a review of the effect of contrast position on the frequency of occurrence of release bursts will be presented in the summary at the end of this section.

Word-initial position (CV). Figure 2.23 shows the percentage of each obstruent category exhibiting a release burst in word-initial position, as well as the percentages of minimal featural contrasts where the contrast is marked by a difference in burst presence. Word-initially, the vast majority of stops exhibit release bursts, with affricates in the lexicon moderately more likely to be produced without a burst than in the inventory, and the reference stops [p, ɖʒ] also exhibiting lower burst presence rates. However, in general from this data we expect word-initial stops to be produced with a release burst. Fricatives, on the other hand, predictably show few if any bursts, consistent with their classification as being initiated with a narrow constriction rather than a complete closure. Among the fricatives, however, nonsibilant dental and labio-dentals do on occasion form a closure that at noise onset produces a burst. This occurs primarily in the target speaker's

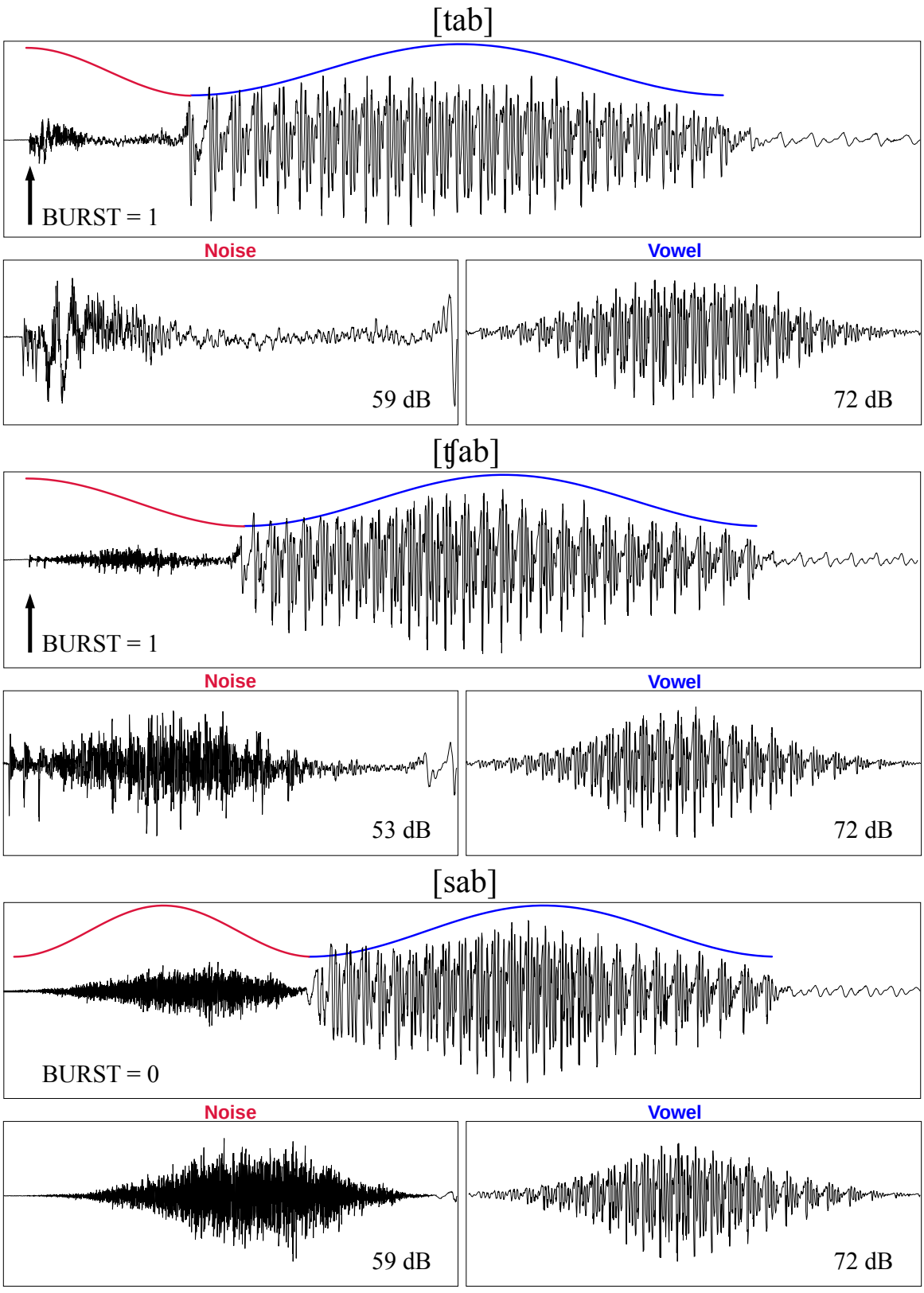

Figure 2.22: Sample measurement of amplitudinal parameters: Burst Presence (BURST), Noise Amplitude (AMP_N), and Vowel Amplitude (AMP_V). The half/full Hamming windows are shown in the top panels, with the windowed signals used to compute the mean amplitudes shown in the bottom two panels.

data, and particularly in the lexicon, with bursts more common in voiced nonsibilants than voiceless, especially in real-word productions of [v], where approximately 80% of CV tokens occur as [b̪]; that is, as a voiced labio-dental plosive with a notable release burst (see Figure 2.11 for noise durations consistent with this description). As a consequence of these results, there are substantial manner contrast effects across the lexicon, inventory, and reference databases, with the occurrence of release bursts in nonsibilants in the target data further yielding substantial sibilance effects in both the inventory and lexicon.

Word-medial position (VCV). Intervocalically, overall burst presence rates decline, particularly in the lexicon and in the more posterior post-alveolar and velar stops. This result is expected for affricates, which even in CV position are not consistently produced with release bursts (for this reason characteristics such as Rise Time and Noise Duration were explored as cues to the [ʃ, ʧ] distinction; Gerstman, 1957; Howell & Rosen, 1983; Kluender & Walsh, 1992), but the velar results suggest a quite frequent process of plosive spirantization that is not commonly described in English, but which occurs in over half of the VCV [k] and [g] tokens in the lexicon. From the inventory results, this lenition process does appear to be more typical of the target speaker's speech than of the reference speakers', particularly for [g], though we should also note that the reference data appears from auditory impression to exhibit greater hyper-articulation (overall consonant and vowel duration differences between target and reference data are consistent with this description). Given that the target speaker had experience producing thousands of words and nonwords in the laboratory prior to recording the inventory data, this result is not surprising. A case of lenition that is more consistent across databases is that of [b], which commonly is reduced to an approximant intervocalically, a result which is less common for its voiceless counterpart, [p].

Finally, as in CV position, the rare occurrence of bursts a fricative noise onset is limited to the nonsibilants, though in VCV position there are notably fewer such cases, occurring only for [f] and [ð], and not consistently across databases. These patterns in burst presence by obstruent category yield substantial manner contrast effects across databases, while there is also a notable place effect

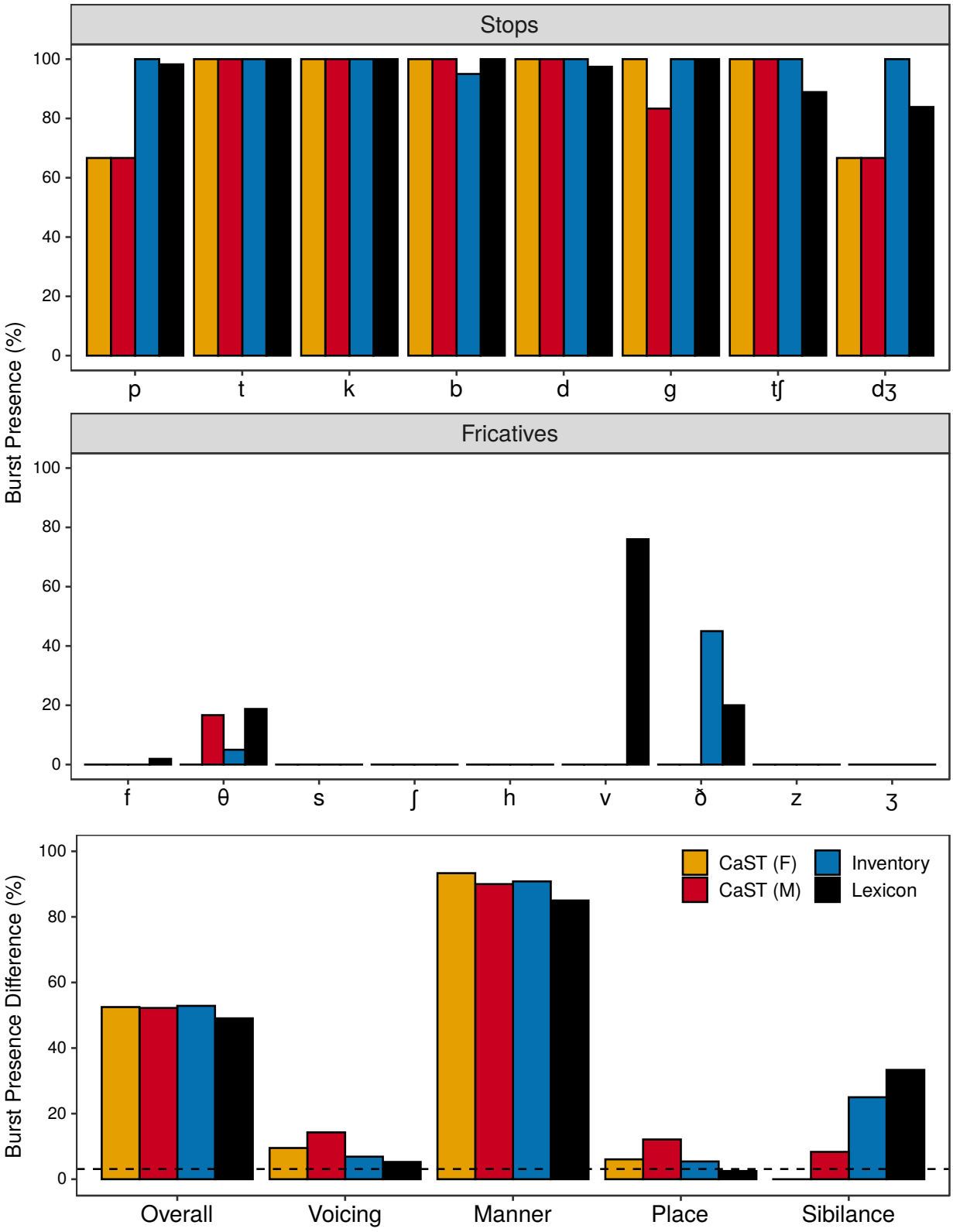

Figure 2.23: Burst Presence percentages in CV position. The top two panels show the percentage of bursts observed for each obstruent. The dashed line indicates mean within-item differences in burst presence in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

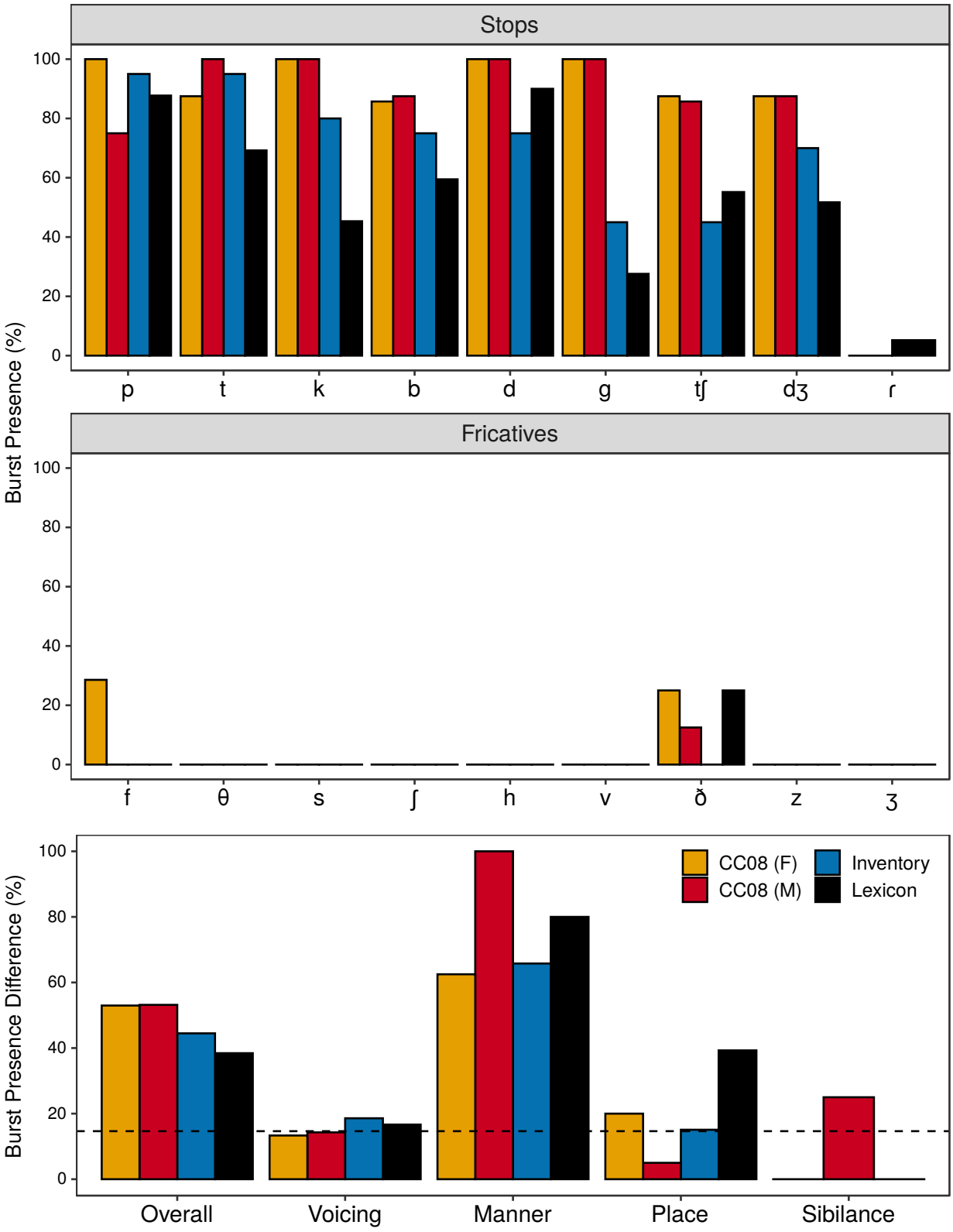

Figure 2.24: Burst Presence percentages in VCV position. The top two panels show the percentage of bursts observed for each obstruent. The dashed line indicates mean within-item differences in burst presence in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

in the lexicon due to the frequent lenition of velar plosives intervocalically.

Word-final position (VC). In VC position fricatives are never produced with a release burst, suggesting that such fortition effects are limited to syllable onset. Among the stops, bursts are again more common in the reference data than in the target data, with coronals—both plosives and affricates—showing the greatest lenition as burst percentages range between 40 and 80% in the lexicon and inventory. And while the majority of such cases remain released (just without a burst), the distribution of unreleased stops word-finally is primarily restricted to [t, d]: 8/12 instances, with the other 4 occurring with [p]. However, given that coronal plosives are far more prevalent in word-final contrasts than are labials or velars, this result is expected; further the rate of unreleased voiceless labial plosives is the highest at 9%, as compared with 6% and 3%, respectively, for [t, d].

Thus, word-finally the presence/absence of a release burst remains a clear discriminator of manner contrasts, and to some extent differentiates place contrasts in the lexicon, though the latter result is primarily due to the distinction between coronal and non-coronal plosives, which may in part be a consequence of the greater prevalence of coronals word-finally. Being more frequent, coronal plosives could be more likely to undergo lenition, a process that is motivated by characteristics of both production and perception, or the lowered burst presence percentages could simply be a case of regression to the mean.

2.5.1.4 Summary

The presence/absence of a release burst, and the relative rate at which this acoustic feature appears in obstruent productions, is a reliable indicator of the manner of articulation of the consonant, and as such appears to have great potential as a cue in discriminating manner contrasts, both in the inventory and the lexicon. Further, due to variable rates of fortition of nonsibilant fricatives word-initially, and place-dependent lenition word-medially and word-finally, burst presence may also cue sibilance and place contrasts in the lexicon, though these effects are much weaker than those for manner. Overall, when comparing CV, VCV, and VC positions we find that bursts are least

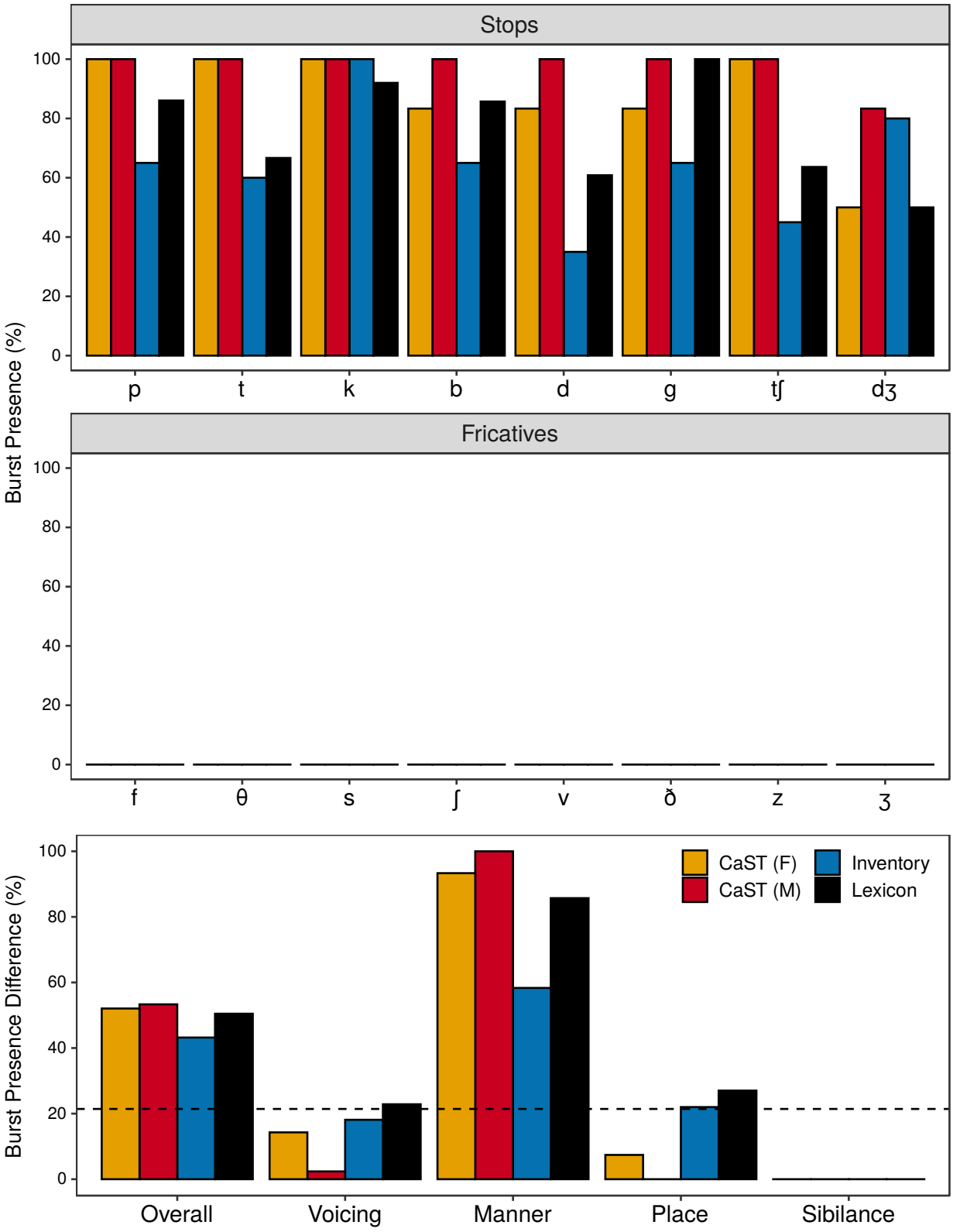

Figure 2.25: Burst Presence percentages in VC position. The top two panels show the percentage of bursts observed for each obstruent. The dashed line indicates mean within-item differences in burst presence in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

common intervocalically, at 33% (45% in the reference data, 32% in the inventory, and 28% in the lexicon), followed by VC position at 42% (49% in the reference data, 32% in the inventory, 42% in the lexicon), and CV position at 56% (46% in the reference data, 50% in the inventory, 60% in the lexicon). Thus, any use of burst presence/absence must be conditional on syllable context, as in the models in Chapter 4. Finally, though the presence of a release burst is descriptively powerful in the discrimination of obstruent contrasts, particularly those differing in manner of articulation, its utility in speech perception will ultimately depend on listeners' ability to detect bursts in the signal, a process which is determined in part by the amplitude of the burst. In the next section we examine a broader feature of noise amplitude as an obstruent cue, and while this cue does not isolate the amplitude of the burst (noise amplitudes are measured over windows including both friction and aspiration noise), the two should correlate.

2.5.2 Noise Amplitude (AMP_N)

2.5.2.1 Background and physiological basis

The amplitude of the noise interval in obstruent production was one of the earliest characteristics investigated for the discrimination of contrasts, with the primary focus being on sibilance distinctions among voiceless fricatives (Strevens, 1960; Heinz & Stevens, 1961). From both of these early studies the connection between noise amplitude and the regulation of airflow in the vocal tract has been at the forefront of the analysis. Briefly, more posterior fricatives tend to exhibit higher noise amplitudes because of the presence and increasing size of a resonating cavity downstream from the point of constriction; labial and dental fricatives, for instance, have little-to-no resonating cavity and therefore are relatively low in amplitude. Sibilants also tend to be higher in amplitude because of the presence of an additional noise source due to the impingement of the airstream on the teeth, generating an obstacle-turbulence sound source that is often much louder than the constriction/wall turbulence found in most other fricatives (Shadle, 1985). Finally, glottal fricatives, though nonsibilants, tend to be the loudest among the English fricatives, because unlike the other fricatives there is no supralaryngeal constriction to dampen the airflow and reduce the overall volume of the sound

emitted from the vocal tract. This result is similar to the vowel amplitude distinction between [i] and [a], where the more open cavity for the latter leads to higher amplitudes on average.

Regarding noise amplitude in plosives and affricates, research has primarily focused on the former through the study of *burst amplitude* as a cue to obstruent place of articulation (Repp, 1984b), where labials tend to be lower in burst amplitude than coronals and velars, a result which is consistent with the explanation of fricative noise amplitude differences on the basis of differences in anterior vocal tract size. The noise amplitude distinction between velars and coronals, however, is less clear, as the former benefits from a larger resonating cavity, but the latter benefits from additional turbulence at the teeth in many cases.

In summary, noise amplitude is broadly motivated as a cue to many place and sibilance distinctions that is firmly rooted in the aerodynamic and acoustic consequences of noise production at different points in the vocal tract. Next we review the exact definition and measurement procedures for noise amplitude adopted in the present study.

2.5.2.2 Definition and measurement

Noise Amplitude (AMP_N) is measured as the mean amplitude of the signal in a window covering the noise portion of the consonant interval. Figure 2.22 illustrates the different windows used on obstruents with and without a burst. For those exhibiting a release burst, a half Hamming window is applied to the noise interval such that the sample weights reduce with increasing temporal distance from the burst. That is, following the formula $w_a(n) = 0.54 + 0.46 \cdot \cos\left(2\pi \cdot \frac{n}{2N}\right),\ 0 \leq n \leq N$, where n is the sample number and N is the total number of samples in the noise interval. From this formula we see that the point of burst onset is given a weight of 1 (no amplitude reduction), while the point of noise offset is given a weight of 0.08, with all points in between reduced in amplitude by a factor equal to the raised cosine function (i.e., the range of y is converted from $[-1, +1]$ to $[0.8, 1]$) over the interval $[0, \pi]$. For obstruents lacking a release burst, a full Hamming window is applied to the noise interval following the formula: $w_b(n) = 0.54 - 0.46 \cdot \cos\left(2\pi \cdot \frac{n}{N}\right),\ 0 \leq n \leq N$. The full Hamming window, $w_b(n)$, is a symmetric window where the signal amplitude is reduced

with increasing temporal distance from the midpoint of the interval (i.e., noise onset and noise offset each receive weights of 0.08, while the center of the noise interval receives a weight of 1).

These windows were chosen primarily to maintain consistency with the spectral analysis (see Section 2.6), where the goal is to place the greatest weight on the point in the noise interval where the articulatory constriction is estimated to be at its maximum. For stops (operationalized as obstruents exhibiting release bursts), the point of maximal constriction is of course in the closure, with the burst representing the portion of noise most characteristic of that point. In the case of fricatives, speakers tend to reach the point of greatest constriction at the midpoint of the noise interval (some research has shown points of constriction maximum tend to be slightly offset from midpoint, at around the 2/3 point, but given the sparsity of data on this topic, the midpoint will be retained for simplicity and as a point of minimal influence from the surrounding vowel context; Shadle & Scully, 1995). As such, for fricatives and fricative-like obstruents lacking a release burst, a full Hamming window over the noise interval places the greatest weight on the center of the noise, whereas for stops and stop-like obstruents exhibiting release bursts, a half Hamming window places the greatest weight on the release burst at noise onset. Finally, we should note that the presence/absence of a burst is only an approximation to the consonant production characteristics that inform the choice of window, but this acoustic rather than phonological definition was adopted in order to account for cases of lenition and fortition that commonly arise in the lexicon.

2.5.2.3 Category and contrast distributions

Below we illustrate noise amplitude distributions in the lexicon, inventory, and reference data for obstruent categories and contrasts in word-initial, word-medial, and word-final positions.

Word-initial position (CV). Figure 2.26 shows noise amplitudes in CV position, and with the exception of the voiceless plosives the general patterns in the target and reference data are largely consistent. Voiced fricatives tend to be greater in amplitude than their voiceless counterparts, particularly among nonsibilants, while [ʧ] is slightly louder than [ʤ], though this distinction is

notably reduced in the lexicon. Plosives, on the other hand, do not show consistent effects of voicing. Voiceless labials tend to be lower in amplitude than voiced labials, but the opposite pattern is obtained for coronals (consistent with the affricate results), and velars show little difference in amplitude as a function of voicing. However, the reference data shows much greater variability in these patterns, with a clear [g] > [k] difference among both male and female speakers, while [d] > [t] difference among male speakers does not extend to the reference female data.

Regarding manner contrasts, fricatives are generally louder than plosives and affricates, though voiceless nonsibilants, including [h], exhibit the lowest amplitudes of all the obstruents, at between 43 and 53 dB. Among stops, there are no clear differences between affricates and plosives, though plosives are more variable and extend into lower amplitudes. These results, however, could be attributed to differences in place of articulation, as the amplitudes of the coronals [t, d, ʧ, ʤ] are fairly consistent.

From the patterns above, the two greatest contrasts effects that emerge, particularly in the lexicon, are for sibilance and manner. Largely due to the inconsistencies among plosives, noise amplitude is not a reliable cue to place or voicing across obstruents, though as discussed above it may be of some utility in fricative and affricate distinctions.

Word-medial position (VCV). Intervocalically, relative to CV position the distinctions between plosive noise amplitudes increase, showing a greater effect of voicing ([p, k] < [b, g]), though among affricates and fricatives the voicing effect reduces, as [ʤ] shows an increase in noise amplitude making it comparable to [ʧ], and [f, θ] increase in amplitude, becoming more similar to their voiced counterparts [v, ð], though there remains a notable distinction between voiceless and voiced nonsibilant fricatives of around 5 dB. This result further enhances the general distinction between sibilants and nonsibilants, though the bottom panel of Figure 2.27 illustrates that the AMP_N differences observed on specific sibilance contrasts reduces word-medially relative to word-initially, at least in the inventory data (no pure sibilance contrasts are present in the lexicon, though there are many multi-feature contrasts involving sibilance distinctions that where noise amplitude would

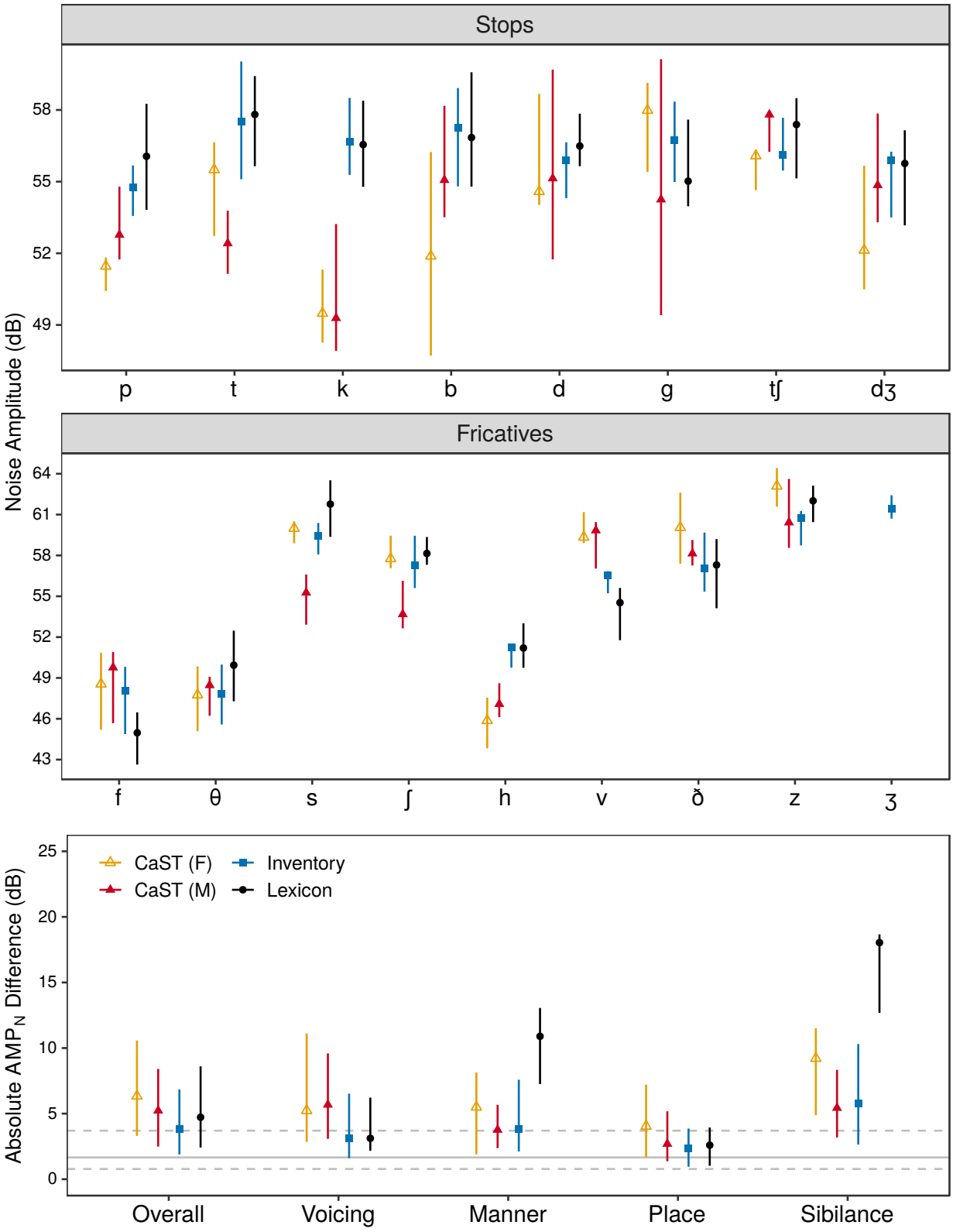

Figure 2.26: Noise Amplitude (AMP_N) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_N in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

likely play a role).

The most notable change from CV to VCV position is the reduction in manner effects on noise amplitude. Word-initially, the three manner classes occupied different amplitude ranges, with stops in the lexicon generally between 55 and 58 dB, nonsibilant voiceless fricatives below 52 dB, and sibilant fricatives above 58 dB, with voiced nonsibilants intermediate between the two. These different ranges led to a robust contrast effect for manner of articulation that is substantially reduced intervocalically, going from around 10 dB on average to around 5 dB, with much greater overlap with the chance range in VCV position. This result appears to be due partly to the increase in plosive amplitudes intervocalically, and partly to the role of the alveolar flap [ɾ] in VCV contrasts, as [ɾ] is generally of higher amplitude (around 60 dB in the lexicon and 63 dB in the inventory) due to its weakened constriction relative to the other obstruents, preserving more closely the amplitudes of the surrounding vowels. Overall, noise amplitude is less discriminative word-medially than it is word-initially, though it appears to play a role an a more diverse range of featural contrasts, as voicing, manner, and place all show effects that range above chance level.

Word-final position (VC). Figure 2.28 shows noise amplitude distributions word-finally that in many respects follow the same pattern as those word-initially. There are few distinctions of note between the plosives, while affricates show a notable voicing effect ([ʧ > ʤ]) similar to that observed between voiced and voiceless sibilant fricatives. The one major difference between the two positions which also represents a notable departure of the real-word data from the controlled syllables is the low amplitude of [v, ð], which are much more similar to their voiceless counterparts word-finally than word-medially or word-initially.

In terms of featural distinctions among observed contrasts in the lexicon, noise amplitude has the greatest effect word-finally, distinguishing voicing, manner, and sibilance contrasts well above chance levels. These patterns are mirrored, though generally reduced, in the inventory and reference data, a result which could be due to the lower amplitudinal variance induced by the hyperarticulation that is more characteristic of nonword syllable productions.

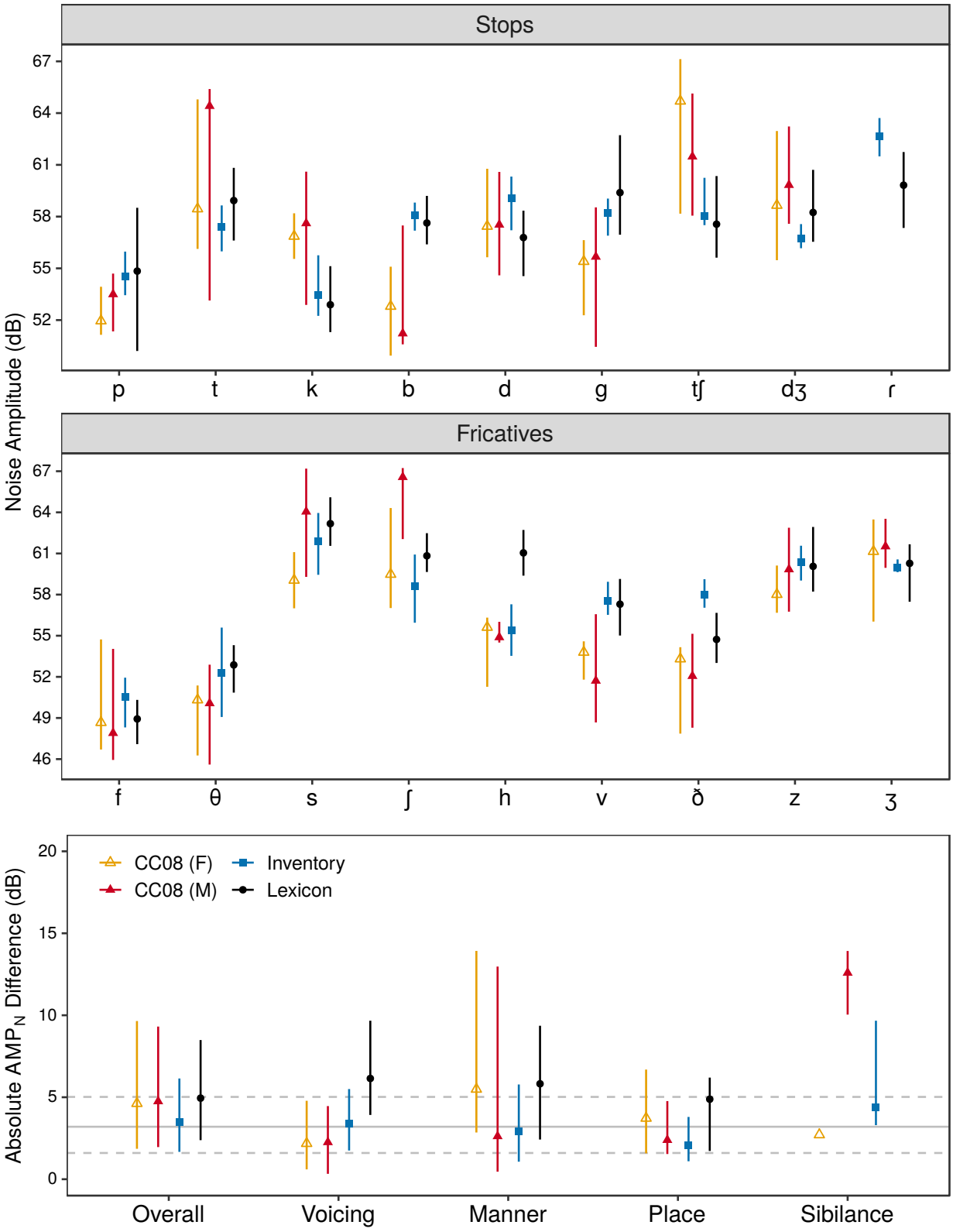

Figure 2.27: Noise Amplitude (AMP_N) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_N in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

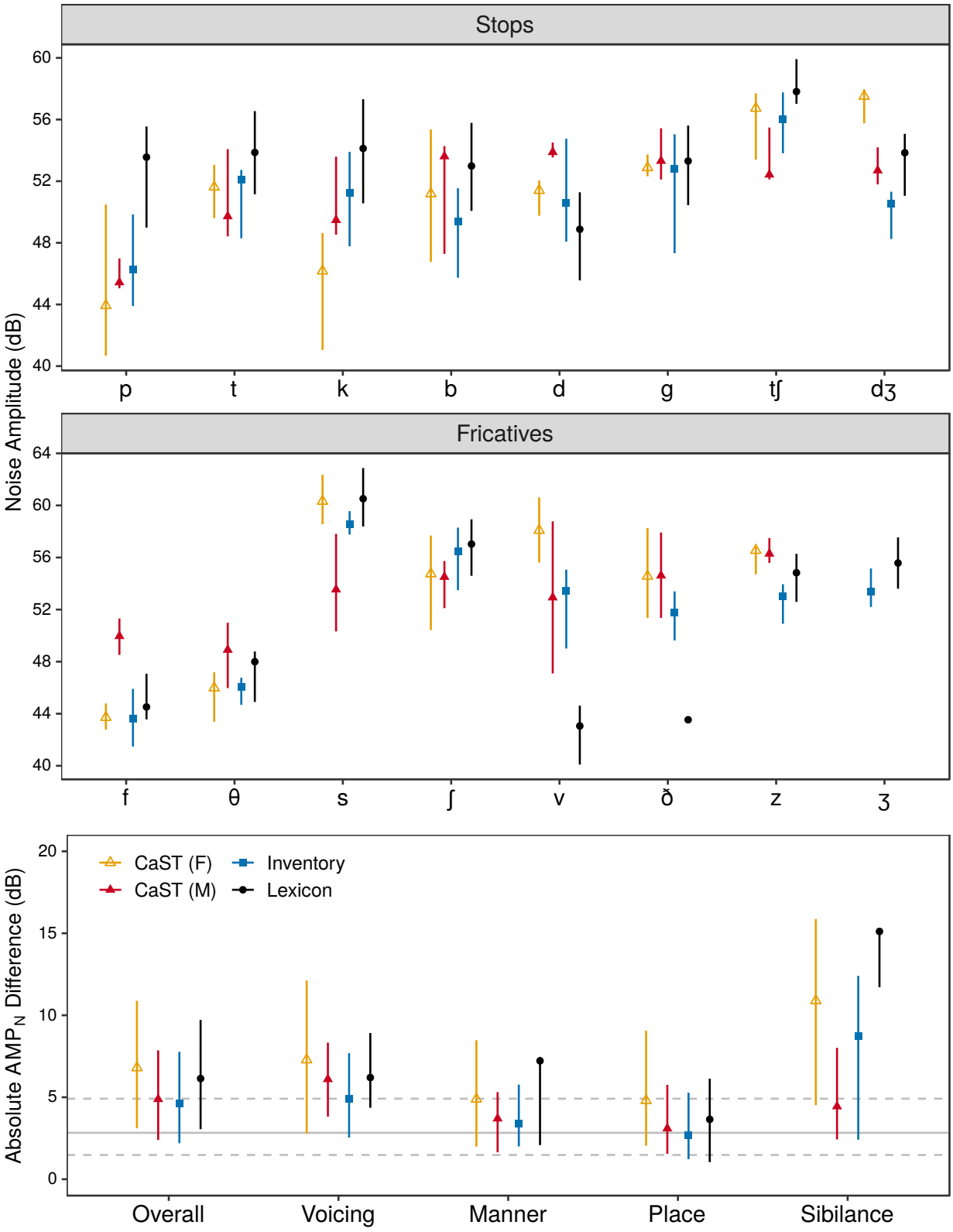

Figure 2.28: Noise Amplitude (AMP_N) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_N in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

2.5.2.4 Summary

As expected from the numerous physiological differences that may impact to some degree the amplitude of the noise produced by different obstruent articulations, AMP_N shows a fairly broad range of distinctions among different obstruent classes that promises to be of some utility in speech perception. Most notably, noise amplitude varies consistently as a function of manner of articulation, with manner contrast effects observed in all three positions, though the relative size of these effects does differ by position (CV $>$ VC $>$ VCV). Other effects of voicing and sibilance are generally robust among fricatives, and with the exception of VCV position, voiced and voiceless affricates are also well-differentiated by noise amplitude. Plosives, on the other hand, are generally poorly distinguished by noise amplitude and show both wide within-category variance and inconsistent patterns between the target and reference data. This is the primary reason for the general lack of place of articulation effects, appearing only moderately in VCV position in the lexicon data, a result which appears to be primarily due to the greater amplitude of [t] relative to [p, k] given that [t] only appears intervocalically as the onset of a stressed syllable, and thus likely reflects the influence of prosodic factors on noise amplitude rather than place of articulation.

2.5.3 Vowel Amplitude ($AMP_{V1/V2}$)

2.5.3.1 Background and physiological basis

The amplitude of the adjacent vowel in CV, VCV, and VC sequences has been used primarily as a reference for the normalization of noise amplitude in the consonant interval (Jongman et al., 2000; Behrens & Blumstein, 1988), and its inclusion in the present study is also motivated in part by this work, though concern over amplitude variation in the target data is reduced due to the constant speaker, style, and recording procedures adopted in (Tucker et al., 2018). However, there is some precedent for the study of obstruent effects on vowel amplitude, and the use of vowel amplitude as a cue in its own right. Lehiste & Peterson (1959), for instance, found in an early study of vowel amplitude and English stress that there were systematic differences in vowel

amplitude as a function of the voicing and manner of the preceding and following consonant. Vowels following/preceding voiced plosives exhibit the lowest amplitudes on average, followed by voiced fricatives, voiceless fricatives, and finally voiceless plosives with the highest average preceding/following vowel amplitude.

This result has a direct physiological basis in the aerodynamics of syllable production and the role of voicing and manner in regulating airflow through the vocal tract. At the lower end of the scale are voiced plosives, which require low air pressure in order to sustain voicing behind the constriction, a requirement which is lessened to some degree in voiced fricatives. Given this relatively low pressure compared to voiceless obstruents, a consequent reduction in airflow in the following vowel is expected, leading to the vowel amplitude distinction noted above. Voiceless fricatives and plosives exhibit the opposite relation, the latter yielding larger vowel amplitudes than the former due to the relatively larger pressure buildup behind a complete obstruction in the vocal tract, though both sets tend to produce louder vowels than their voiced counterparts.

2.5.3.2 Definition and measurement

Vowel Amplitude ($AMP_{V1/V2}$) is measured similarly to noise amplitude. As Figure 2.22 illustrates, a Hamming window is first applied to the vowel interval, and then the root-mean-squared amplitude of that window is taken as the vowel amplitude. Separate amplitudes are recorded for preceding (V1) and following (V2) vowels (noted as AMP_{V1} and AMP_{V2}, respectively), though this differentiation is only relevant for intervocalic contrasts.

2.5.3.3 Category and contrast distributions

Below we review vowel amplitude distributions by category and featural contrast. As in the previous analysis of vowel duration, results are separated by preceding (AMP_{V1}) and following (AMP_{V2}) vowel rather than by position.

Preceding vowel amplitude ($\mathbf{AMP_{V1}}$). Figure 2.29 shows V1 amplitude distributions in word-medial obstruent categories and contrasts. Overall the amplitude of the preceding vowel is relatively constant across obstruent contexts, though there does appear to be a slight effect of voicing where the amplitude preceding voiceless obstruents is marginally greater than that preceding voiced obstruents. However, given that all items have been amplitude-normalized to some extent in preparation for inclusion in perception experiments, this result could simply be a consequence of the preprocessing procedure, as voiceless obstruents tend to exhibit lower noise amplitudes word-medially. Of course, one could argue from the opposite direction that amplitude normalization is responsible for the differences in noise amplitude discussed above, but given that those effects are much greater and more articulatorily motivated, the initial explanation appears more plausible. Nevertheless, these are still patterns that listeners may use in perception.

There are a couple of distinctions that stand out as being notably divergent from the general pattern of relatively constant amplitudes in the 71-74 dB range: the low V1 amplitudes preceding [t] and [h]. Both results are likely a consequence of the restricted stress pattern required to elicit intervocalic [t] and [h] (i.e., unstressed V1, stressed V2), as unstressed vowels tend to be lower in amplitude relative to their stressed counterparts, while most other VCV contrasts should exhibit stress on the preceding vowel. Note that in Figure 2.32 both [t] and [h] raised V2 amplitudes relative to the other obstruents, consistent with prosodic expectations. This does leave unexplained the lack of such an effect on vowels preceding [d], though in many respects [d] is acoustically more similar to [ɾ] than to [t] (see, for instance, the consonant duration distributions in Figure 2.3). Finally, regarding specific contrasts and minimal featural distinctions, vowel amplitude is generally at or below chance levels, with only one or two decibels separating most contrasts. The male reference data is one exception to this trend, though it is not clear what is responsible for this discrepancy. The Cooke & Scharenborg (2008) data is different from the inventory data in varying stress patterns (trochaic versus iambic), but the male and female groups are balanced in both the relative frequency of each pattern, and the frequency with which such patterns occur with different obstruents. The vowel contexts do differ as a function of stress, however, with the male

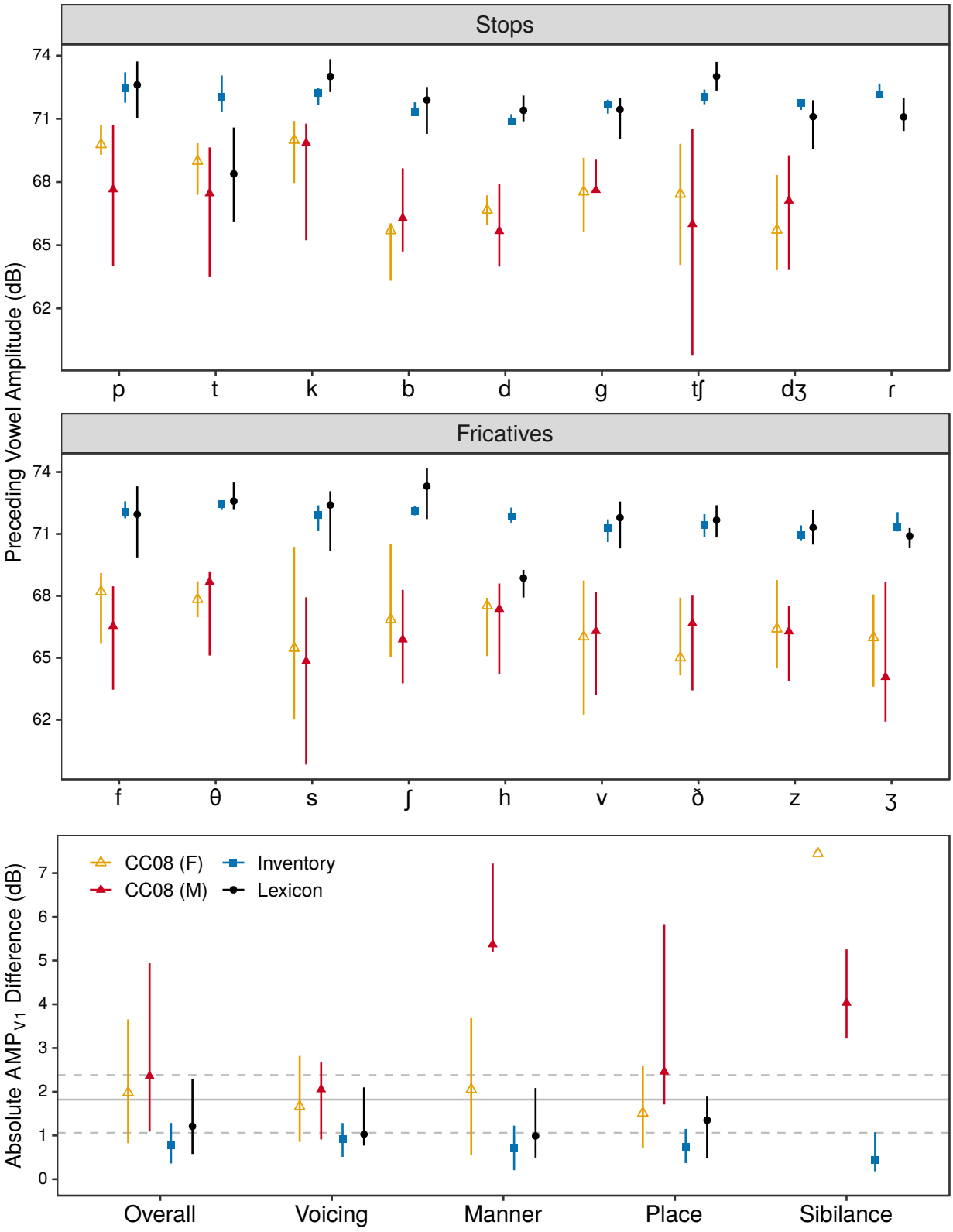

Figure 2.29: Preceding Vowel Amplitude (AMP_{V1}) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{V1} in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

reference data exhibiting a stressed low vowel [a] in V1 position much more often (approximately 2/3 of the time) than unstressed [a]. Given that the female reference data is far more symmetrical in this respect, the anomalous patterns in the male reference data could be due to some interaction between stress and vowel context.

Word-finally there is a greater differentiation of obstruents according to preceding vowel amplitude, primarily as a function of voicing, but also to some extent manner. In the case of voicing, vowels preceding voiceless obstruents tend to be louder than those preceding voiced obstruents. The manner effect, which is only present in the lexicon data, is around 0.5 dB lower than that for voicing and overlaps to a greater extent with the estimated chance range. Nevertheless, the distribution remains well above the ranges in other databases, and well above lexical and non-lexical effects of place and sibilance. The greatest contributors to this effect are the [ʧ, ʃ] contrast, which averages greater than 2 dB and the [p, f] contrast, which averages around 1.5 dB. In both contrasts the amplitude of the vowel preceding the stop is greater than that preceding the fricative. The alveolar contrasts [t, s] and [d, z] also present manner distinctions of a similar magnitude, but do not contribute to the exclusive manner effect in the bottom panel of Figure 2.30 because each contrast is also distinguished by sibilance. Finally, there is the question of whether these results are due to stimulus amplitude normalization rather than being articulatory or perceptual in nature. In general, the noise amplitude results are not consistent with this explanation, as word-final stops are also louder in noise amplitude than word-final fricatives. However, considering that the signal amplitude would be close to zero during the closure, the preprocessing explanation remains a possibility, especially given that all such effects are around 1-2 dB.

Following vowel amplitude ($\mathbf{AMP_{V2}}$). Figure 2.31 shows vowel amplitudes following obstruent categories and contrasts in CV position. The most notable distinctions that emerge word-initially in the lexicon are those between voiced and voiceless obstruents, with vowels following voiceless obstruents slightly louder than those following voiced obstruents, with the greatest distinctions observed between the contrasts [s, z] and [θ, ð]. These differences, however, are not consistently

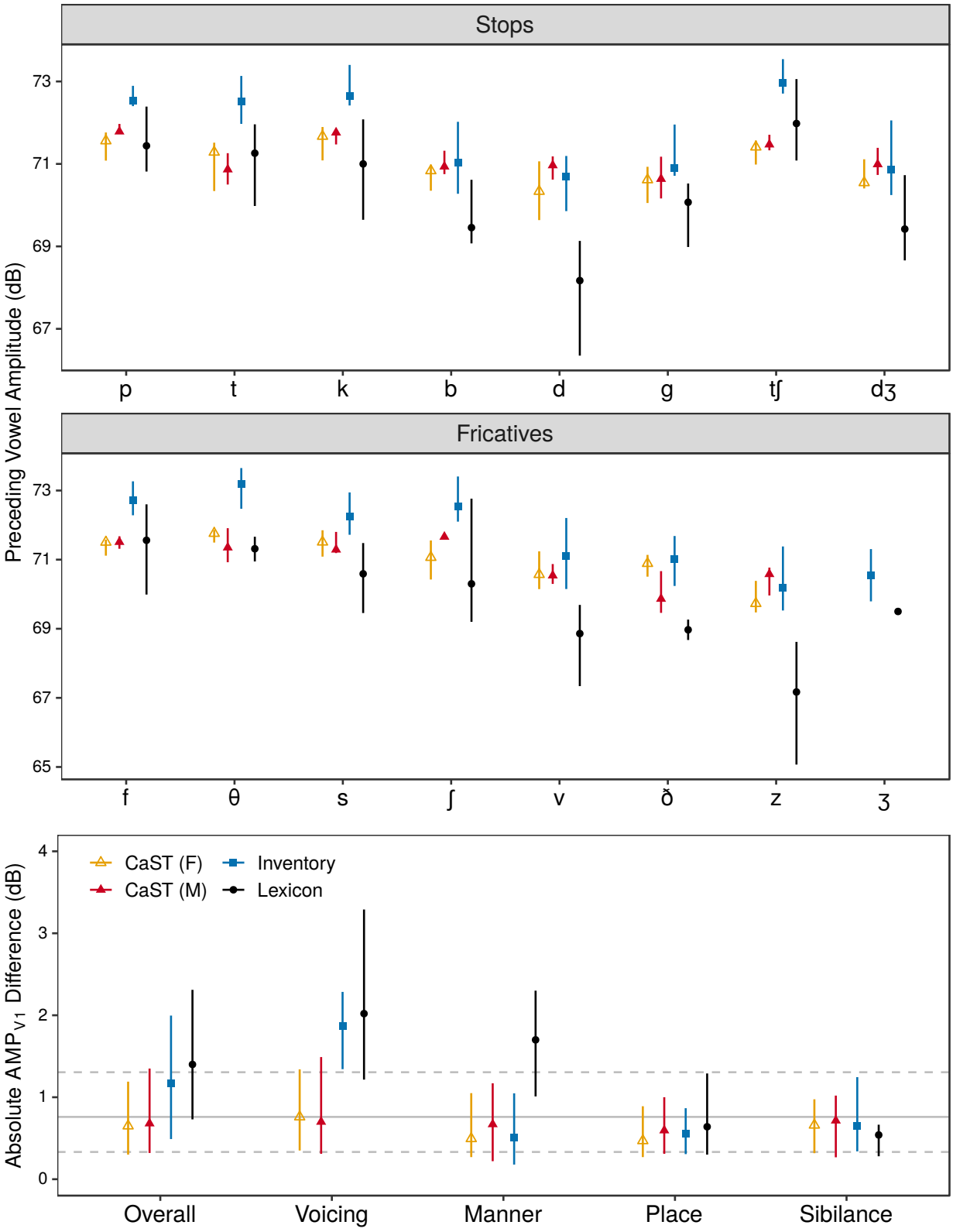

Figure 2.30: Preceding Vowel Amplitude (AMP_{V1}) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{V1} in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

present in the syllable data—either in the inventory or reference data—unlike the AMP_{V1} effects in word-final position which were present in both the lexicon and inventory. There is also a minor effect of manner, where vowels following fricatives tend to be louder than those following stops, though this difference only amounts to 1 dB on average and overlaps considerably with the chance range based on within-item variability in the inventory data. Thus, for word-initial contrasts, following vowel amplitude does not appear to be a robust cue.

Word-medially, V2 amplitudes are relatively constant at between 64 and 66 dB on average in the lexicon, 66-69 dB in the inventory, and 64-66 dB in the reference data. The few exceptions to this trend, all occurring in the lexicon, are the obstruents [t, h, θ, ð], each of which is relatively infrequent intervocalically and thus unlikely to contribute to robust contrast effects. The outlier behavior of [t, h] was previously discussed as likely a consequence of prosodic factors, as each occurs only as the onset of stressed syllables intervocalically, and the dental fricative results are consistent with this explanation given their exclusive occurrence preceding unstressed syllables, resulting in lower V2 amplitudes on average. However, the fact that [ɾ] does not follow this trend despite occurring in the same environment weakens the prosodic explanation for low vowel amplitudes following [θ, ð]. Ultimately, with the exception of a singular sibilance effect in the female reference data and a minor manner effect in the male reference data, no contrast effects emerge for AMP_{V2} intervocalically. Further, given the large within-item differences in following vowel amplitude (2.6 dB on average with an IQR covering 1.5-3.5 dB), AMP_{V2} does not appear to be a reliable cue for intervocalic obstruent contrasts.

2.5.3.4 Summary

Vowel amplitude was already considered of limited utility based on its sparse usage in the literature, and was included primarily as a means for the cue-integration model to relativize noise amplitude without directly coding relative amplitude into the cue set. The reason this latter option was not adopted, as discussed in the introduction to this section as well as in the introduction to noise amplitude (see also the discussion of relativized temporal cues in Section 2.4), was that composite

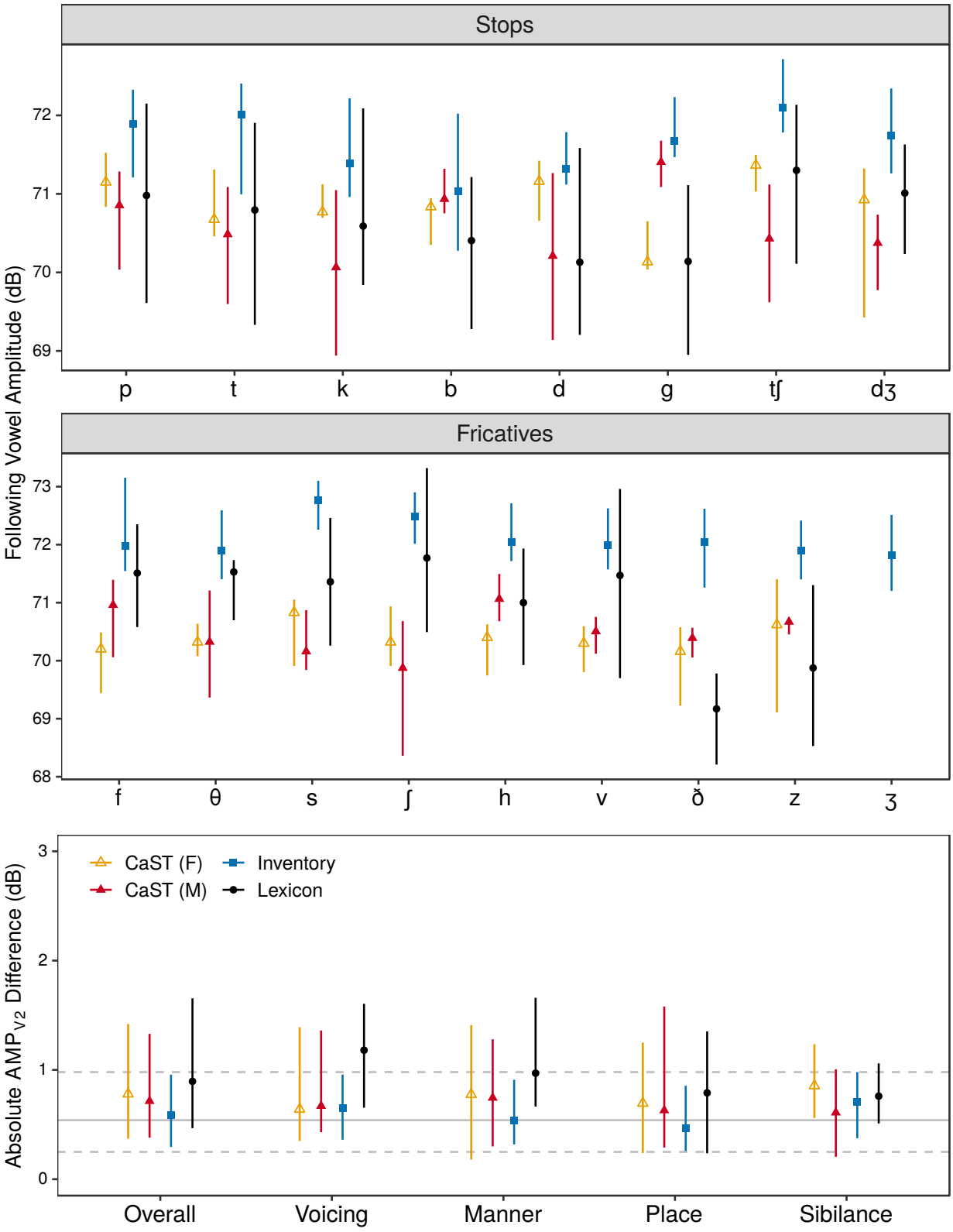

Figure 2.31: Following Vowel Amplitude (AMP_{V2}) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{V2} in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

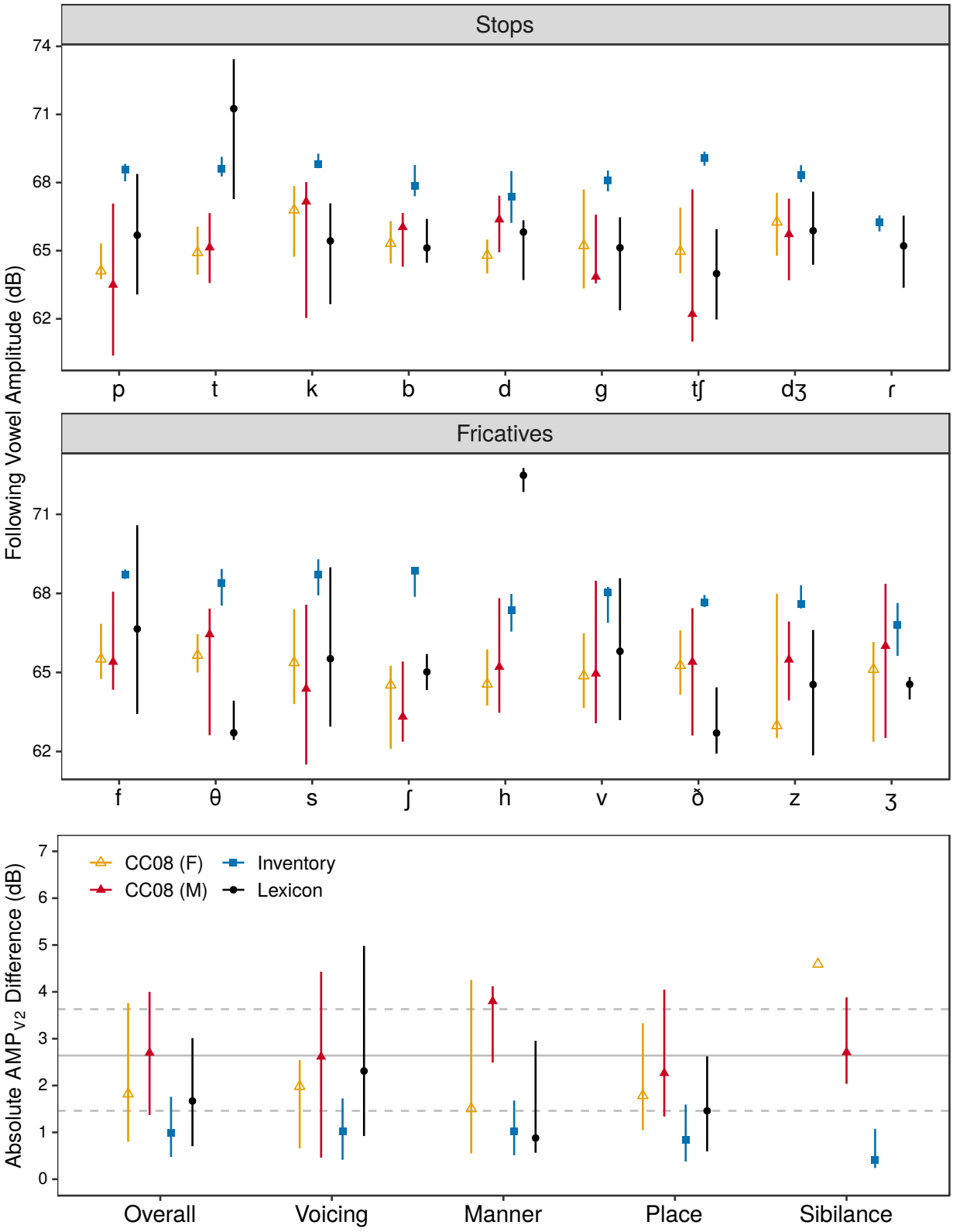

Figure 2.32: Following Vowel Amplitude (AMP_{V2}) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{V2} in the inventory data (between reps 1 and 2) and serves as a reference for potential chance effects.

parameters introduce an additional assumption about the way amplitude is perceived, as well as potentially introducing noise by linking variability in the consonant to that in the vowel, which may be further affected by its own segmental context in a manner that is not relevant for the perception of the target consonant. Ultimately, however, vowel amplitudes were relatively constant across obstruent contexts, making both V1 and V2 amplitudes of limited value as independent cues, but also suggesting the patterns observed for noise amplitude would remain consistent in both raw and relativized versions. Where vowel amplitude is of greatest potential as a cue is in its structured variation as a function of voicing for CV and VC contrasts, though it remains to be seen in the cue-integration models in Section 4.4 of Chapter 4 whether this subtle though consistent effect is at all predictive of listener behavior.

2.5.4 Comparative discriminative power of amplitudinal parameters

Comparing the four amplitudinal parameters discussed above in terms of their overall discriminative power, Figure 2.33 shows that burst presence far outweighs noise and vowel amplitude in its discriminative power. Word-medially there is little difference in discriminative power between AMP_N and $AMP_{V1/V2}$. For CV and VC contrasts, however, noise amplitude is relatively more robust in the target data, though still sizeably less discriminative than burst presence. Here it is important to emphasize that such distinctions are estimates of cue potential assuming both perfect reception of the signal, as well as assuming that listeners are indeed tracking these acoustic properties as we have defined them. This point is all the more critical for the analysis of amplitudinal parameters given that the assessment of listener recognition, both in the present study and in the majority of the literature, is on stimuli embedded in background noise. We will return to these results in Chapter 4 when cue-integration models of listener perception are presented.

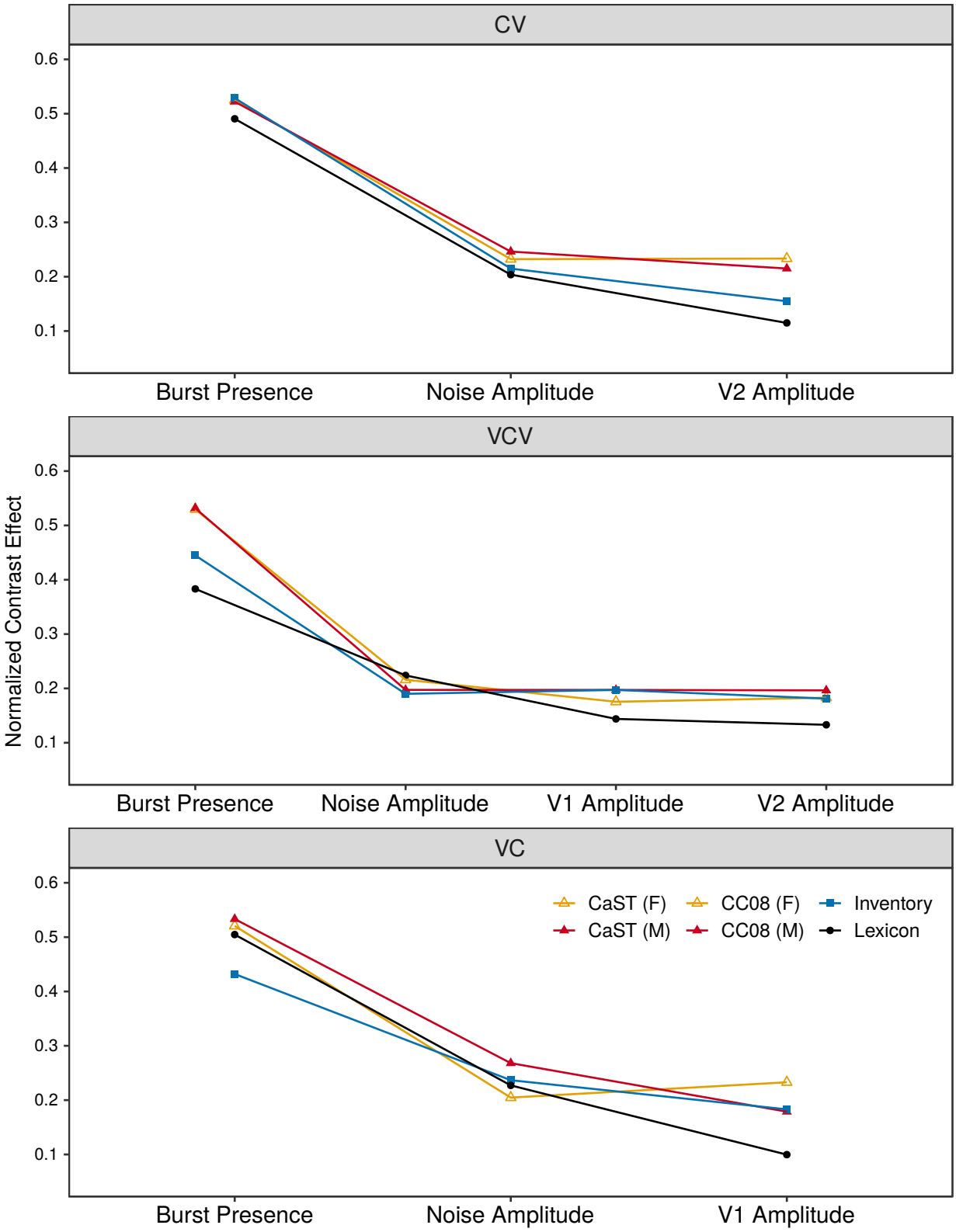

Figure 2.33: Comparative discriminative power of amplitudinal parameters in CV, VCV, and VC positions, as measured by the *normalized contrast effect*, the mean of the absolute differences by contrast between parameters that have been scaled between 0 and 1. The CaST and CC08 reference data sets share the same color palette because their contexts (CV/VC and VCV, respectively) are mutually exclusive.

2.6 Spectral parameters

The final class of parameters used to characterize English obstruent acoustics comprises those that are primarily spectral in nature; i.e., those characterizing the relative amplitude of the signal at different frequencies and the overall distribution of energy in the spectrum. This class is also the largest and the most closely related to the parameterization that forms the backbone of most automatic speech recognition (ASR) systems, though as in the previous sections the focus of the present study is on only those acoustic properties that can be directly linked to a gestural source and are explanatory in terms of the acoustics, aerodynamics, and mechanics of speech articulation. The following spectral parameters are used in the present study, each of which is assessed in the subsections below: Spectral Peak Frequency (FREQ_{PK}), Spectral Peak Amplitude (AMP_{PK}), Dynamic Amplitude (AMP_{DYN}), Spectral Tilt ($\text{TILT}_{\text{C/V1/V2}}$), Spectral Shape (SHAPE), Spectral Dispersion ($\text{DISP}_{\text{C/VC/CV}}$), Low-Frequency Energy (LF), Relative F3 Amplitude (AMP_{F3}), Relative F5 Amplitude (AMP_{F5}), Fundamental Frequency ($\text{f0}_{\text{VC/CV}}$), First Formant Frequency ($\text{F1}_{\text{VC/CV}}$), Second Formant Frequency ($\text{F2}_{\text{VC/CV}}$), and Third Formant Frequency ($\text{F1}_{\text{VC/CV}}$). Spectral information, though also reflective in part of *source* characteristics determined by voicing, manner, and sibilance, is most notable for what it reveals about the configuration of the vocal tract *filter*. Thus, unlike the sparse cues to place of articulation in the temporal and amplitudinal sets above, here we expect more robust discrimination of place contrasts, thereby completing the featural scope of the acoustic parameterization of obstruent consonants in the present study.

2.6.1 Spectral Peak Frequency (FREQ_{PK})

2.6.1.1 Background and physiological basis

Spectral peak frequency was one of the earliest obstruent cues explored in the phonetic literature, receiving attention in Liberman et al. (1952), Hughes & Halle (1956), Strevens (1960), and Heinz & Stevens (1961) as a cue to plosive and fricative place of articulation. The analysis of spectral peak frequency has a fundamental physiological and acoustic basis in the theory of resonating

tubes and vocal tract geometry (Fant, 1960), and reflects the fact that obstruent constrictions generally produce a decoupling of front and back cavities, and with a primary excitation source in the noise generated at or downstream of the constriction, the frequency of highest energy in the spectrum generally corresponds to the resonance frequency of the anterior cavity, making it a powerful acoustic measurement of place of articulation. Differences in spectral peak frequency can also arise due to differences in noise source, such as the distinction between obstacle and wall turbulence in sibilant and nonsibilant fricatives (Shadle, 1985), and thus spectral peak frequency also provides a direct and physiologically explanatory cue to obstruent sibilance.

2.6.1.2 Definition and measurement

Spectral peak frequency is defined as the frequency of maximum amplitude in the spectrum between 550 and 10,000 Hz. Spectra are computed by taking an FFT of the full/half Hamming windows shown in Figure 2.22, which are chosen based on the presence/absence of a release burst. To reiterate the justification of the choice of windows that was discussed in Section 2.5.2, where the measurement of noise amplitude was introduced, our aim is to characterize the distribution of energy in the spectrum at or near the point of maximum consonantal constriction. Thus, the greatest weight in the window chosen for stop consonants (the half Hamming window) is at noise onset, whereas for fricatives a full Hamming window ensures the greatest weight is placed on the midpoint of the noise interval. Affricates are treated either as stops or fricatives in this regard based on the presence or absence, respectively, of a release burst (the same is also true for plosives that are occasionally produced without bursts). This use of bursts as a criterion for distinguishing contriction types was adopted in order to provide an objective measure that does not rely on phonological classifications and can flexibly capture cases of stop lenition and fricative fortition.

The lower bound of the frequency range, 550 Hz, serves to prevent the identification of peak frequency with components of the source. This of course is not an exact point, but thresholds of 550 and 500 Hz have been used variously in the literature to separate source and filter components of the spectrum (Shadle & Mair, 1996; Koenig et al., 2013). Both 500 and 550 Hz thresholds were

tested on the present data, and 550 Hz was chosen as it exhibited far fewer peak frequencies at the lower bound compared to 500, suggesting a 500 Hz threshold more often sits at the right edge of a source harmonic and therefore is less adequate at separating source and filter components. The upper bound of 10 kHz is less critical and reflects the fact that most linguistically relevant information in the spectrum, including the spectral peak, lies below this point. See Figure 2.34 for sample measurements of spectral peak frequency in consonant noise intervals of [aʤa] and [afa], two items differing in voicing, manner, place, and sibilance.

2.6.1.3 Category and contrast distributions

Below we review spectral peak frequency distributions for obstruent categories and contrasts in the lexicon, inventory, and reference databases. Results are presented separately for word-initial, word-medial, and word-final positions.

Word-initial position (CV). At word/syllable onset, spectral peak frequency shows generally robust separation of obstruent phones along multiple dimensions. Beginning with the plosives, Figure 2.35 shows that for both voiceless and voiced plosives labials are consistently lower in peak frequency than coronals and velars. This result is partially consistent with theoretical expectations (cf. Fant, 1960; Stevens & Blumstein, 1978, and the discussion in the introduction to this section); namely, labials exhibit the lowest peak frequencies, corresponding to primarily to F1, while alveolars and velars show higher-frequency spectral peaks. The lack of separation between alveolars and velars in the lexicon occurs in part due to the fact that both places exhibit spectral peaks that are coupled with the higher formant frequencies (F2, F3, and F4) and will vary according to vowel context. For this reason [t, k, d, g] show much greater variances than [p, b]. No clear effect of voicing is present in peak frequency differences among plosives.

Turning next to the affricates, both [ʧ] and [ʤ] are highly consistent in exhibiting spectral peaks around 3 kHz, as are their homorganic fricatives [ʃ] and [ʒ]. Thus, we can see that as predicted, the frequency of the peak in the spectrum is a consistent indicator of place of articulation, both

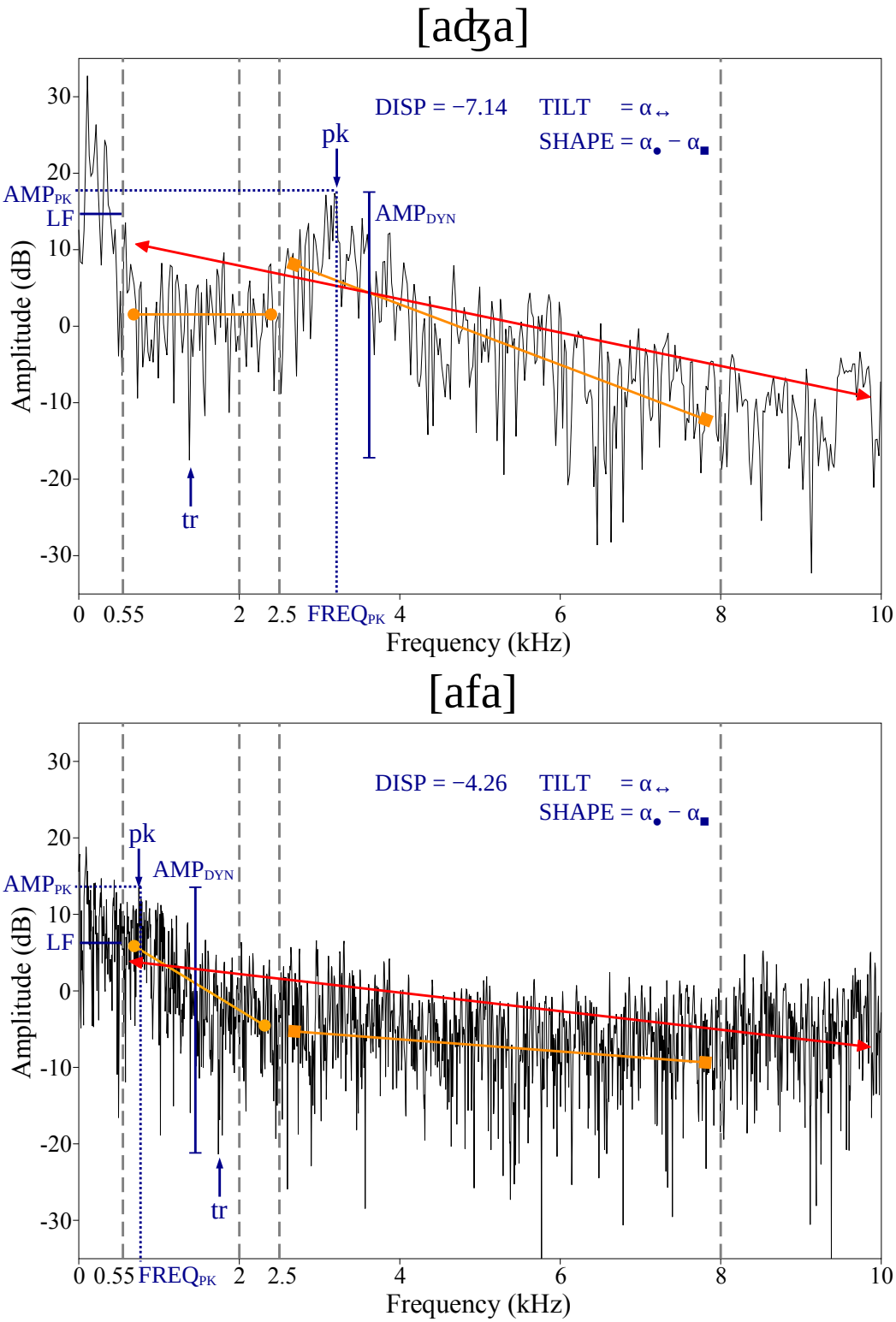

Figure 2.34: Sample measurement of spectral parameters (Set I): spectral peak frequency (FREQ_{PK}), spectral peak amplitude (AMP_{PK}), dynamic amplitude (AMP_{DYN}), spectral dispersion (DISP), spectral tilt (TILT), spectral shape (SHAPE), and low-frequency energy (LF). Landmarks for the measurement of FREQ_{PK}, AMP_{PK}, and AMP_{DYN} include the peak of the spectrum above 550 Hz (pk) and the trough of the spectrum below 2 kHz (tr). Spectral slopes (α) for TILT and SHAPE parameters are indexed by line endpoints.

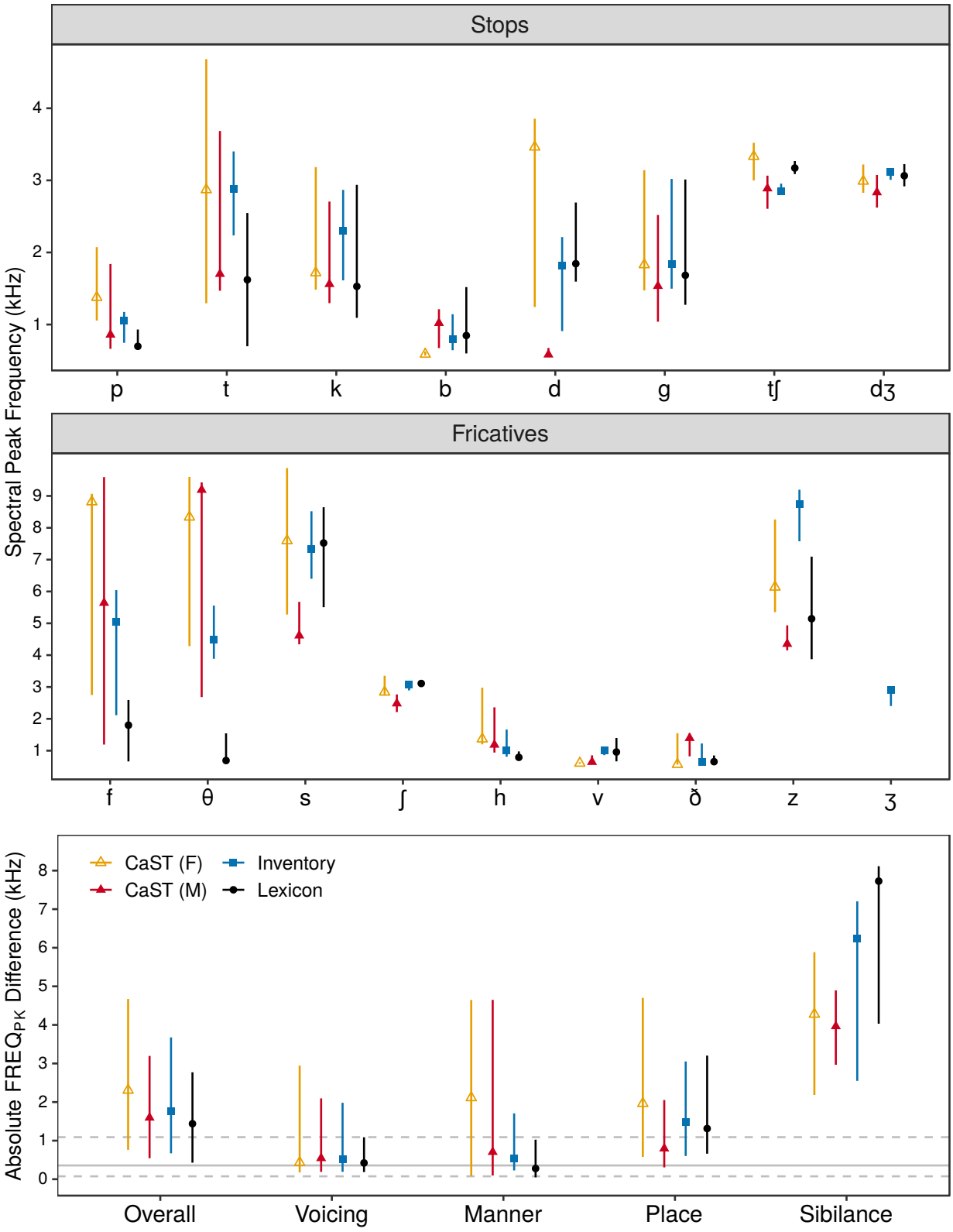

Figure 2.35: Spectral Peak Frequency ($FREQ_{PK}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $FREQ_{PK}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

in the lexicon and the inventory/reference data. Finally, among fricatives the two most notable distinctions indexed by FREQ_{PK} are place and sibilance. The most robust place distinction occurs between [s] and [ʃ], with [s] markedly higher in peak frequency than [ʃ]. Sibilance contrasts, particularly in the lexicon, are even more robust, with nonsibilants exhibiting low peak frequencies at or below 2 kHz in the lexicon,[13] well below the 3–9 kHz range observed for sibilants.

As the bottom panel of Figure 2.35 shows, sibilance by far exhibits the greatest contrast effects in all four data sets, with place of articulation the only other feature that is above estimated chance levels on average. This place effect remains notable, however, because it occurs more widely than the sibilance effect, which in its pure contrastive form only distinguishes [s] from [θ] and [z] from [ð], as the nonsibilant fricatives [f, v, h] also differ in place of articulation.

Word-medial position (VCV). Intervocalically, we see in Figure 2.36 similar place and sibilance effects, with nonsibilant fricatives lower in peak frequencies than sibilants (though [θ] is considerably more variable in this regard, both within and across databases), postalveolar fricatives and affricates again consistent around 3 kHz and well below peak frequencies for the alveolar sibilants [s, z], and labial plosives generally lower in FREQ_{PK} than alveolar/velar plosives. The one notable difference from CV position is in the relative patterning of alveolar plosives, as [t] shows higher peak frequencies than [k], consistent with expectations from Stevens & Blumstein (1978), while [d] patterns more closely with [b]. This latter relation is unexpected, but may be a consequence of [d] being produced more like the flap [ɾ] than like its voiceless plosive counterpart (recall similar results were obtained for consonant duration and preceding vowel amplitude), as [ɾ] generally exhibits low spectral peak frequencies as well, consistent with the negative spectral tilts typical of approximants and vowels.

In terms of contrast effects in Figure 2.36, the only notable featural effect in the lexicon is for place of articulation. No 'pure' sibilance contrasts are present intervocalically in the lexicon, though many multi-feature contrasts between sibilants and nonsibilants are expected to contribute

[13] Peak frequencies for [f, θ] are much higher in the controlled syllable data, a result which appears to be largely a consequence of the greater hyperarticulation observed in such items.

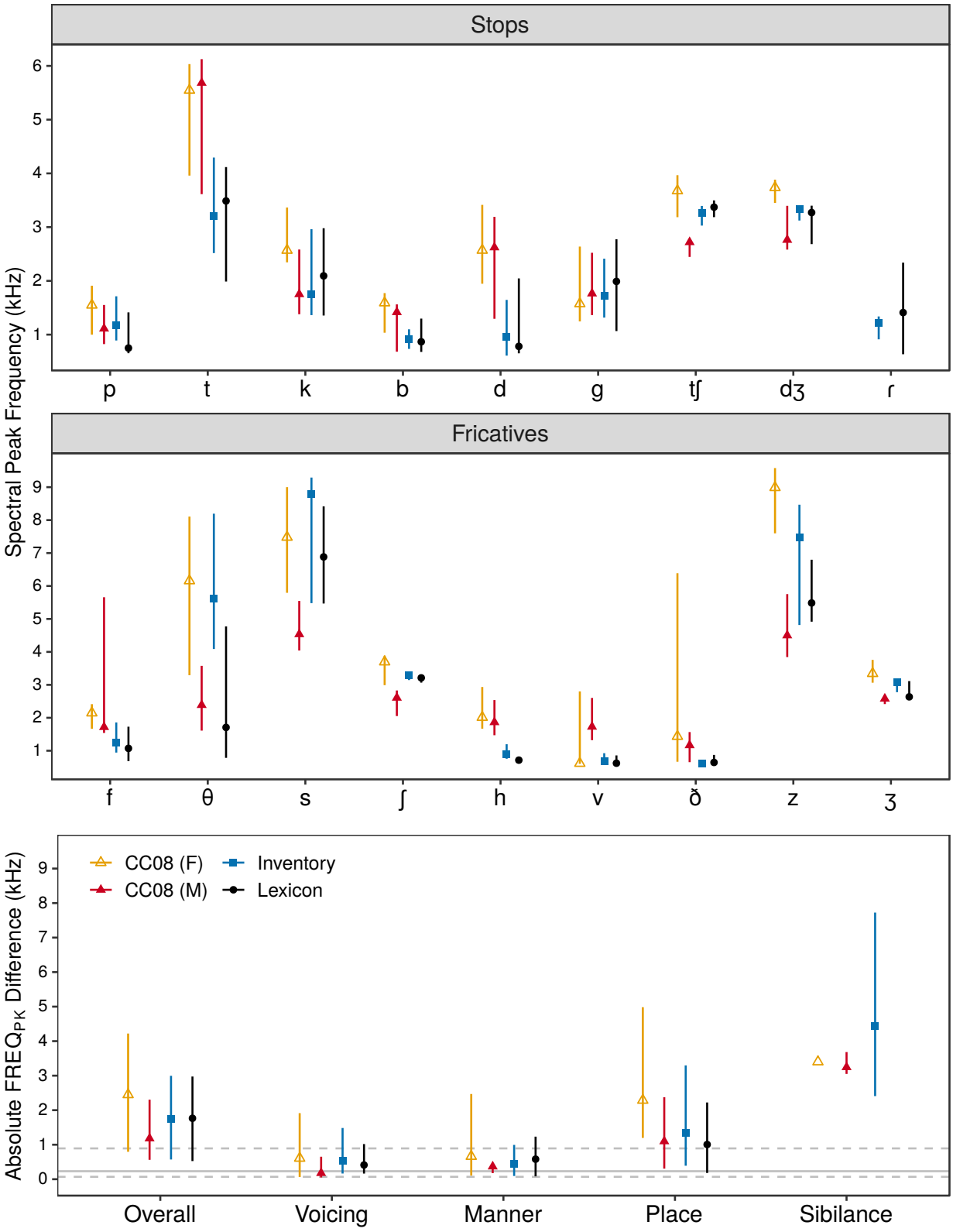

Figure 2.36: Spectral Peak Frequency ($FREQ_{PK}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $FREQ_{PK}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

to the overall contrast effect observed in lexicon. Place effects are consistently observed in the inventory and reference data as well, though sibilance effects are much more robust in the three controlled syllable databases.

Word-final position (VC). Figure 2.37 shows spectral peak frequency distributions among word-final obstruent contrasts in the lexicon, inventory, and reference data. Here we find the greatest contrast effects among the three positions, as FREQ_{PK} plays a role in discriminating sibilance, place, and voicing contrasts in the lexicon, where the latter voicing effect is exclusive to lexical contrasts. From the category distributions in Figure 2.37, we see that the voicing effect is largely due to the higher peak frequencies observed in the voiceless alveolar plosives [t] as compared with its voiced counterparts [d], though there is also a minor difference in the same direction between the velars [k, g]. All other category distributions by place and sibilance remain consistent with those observed in CV and VCV positions.

2.6.1.4 Summary

Spectral peak frequency is remarkably consistent as an index of place of articulation and sibilance across contrast positions and data types, and with the exception of the word-final distinction between [t] and [d], these effects also show little variation as a function of voicing and manner of articulation. This robustness of FREQ_{PK} as a cue for place of articulation serves as a first demonstration of the critical role spectral information plays in characterizing the filter characteristics of obstruent consonants, as the temporal and amplitudinal parameters discussed in previous sections largely reflected source characteristics, both in terms of laryngeal configurations and the regulation of airflow (and thereby noise source characteristics) in obstruent consonant production. The next two parameters, spectral peak amplitude (AMP_{PK}) and dynamic amplitude (AMP_{PK}), however, fall in this latter category as primarily reflecting source characteristics.

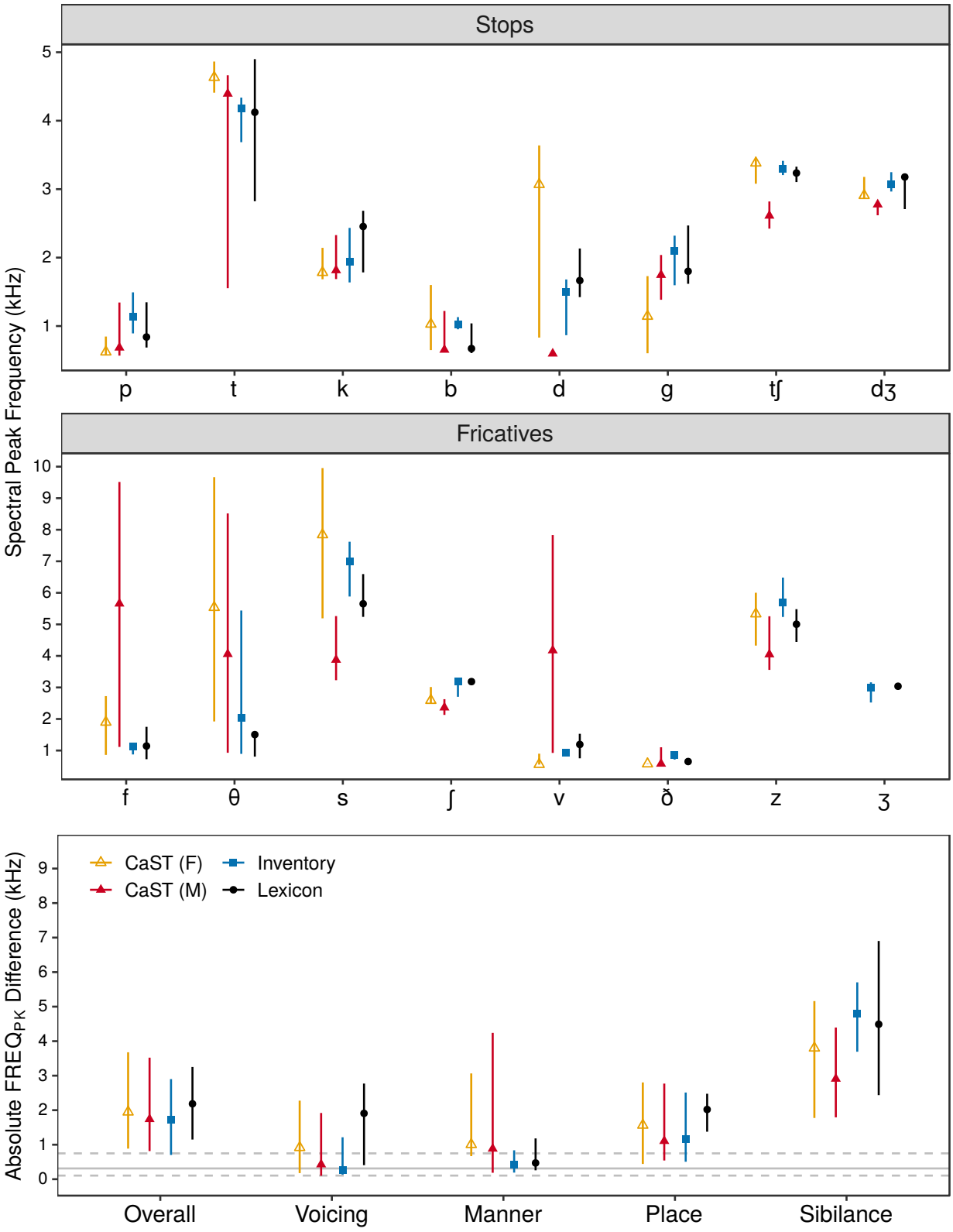

Figure 2.37: Spectral Peak Frequency ($FREQ_{PK}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $FREQ_{PK}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

2.6.2 Spectral Peak Amplitude (AMP_{PK})

2.6.2.1 Background and physiological basis

Spectral peak amplitude is similar to spectral peak frequency in being directly linked to variation in place of articulation and sibilance (Fischer-Jørgensen, 1954; Ohde & Stevens, 1983; Smits et al., 1996). The reason behind this link is that changes in anterior cavity size not only affect the frequency of the main resonance. They also affect the volume of air exciting that resonance, with larger cavities exhibiting higher peak amplitudes, on average, controlling for other factors such as noise source. Finally, regarding the impact of the noise source on the amplitude of the main resonance peak in the spectrum, the turbulence noise source at the teeth in sibilant production generally produces greater excitation of the spectral peak than friction generated at the point of constriction, such as in the production of labial and dental fricatives.

2.6.2.2 Definition and measurement

Spectral peak amplitude is defined as the maximum amplitude in the spectrum between 550 and 10,000 Hz, or alternatively as the *y* component of the spectral peak (i.e., the amplitudinal counterpart of spectral peak frequency). See Figure 2.34 for sample measurements of AMP_{PK} in the noise intervals of [aʧa] and [afa].

2.6.2.3 Category and contrast distributions

AMP_{PK} distributions for obstruent categories and contrasts in the lexicon, inventory, and reference data are presented below according to contrast position; i.e., word-initially (CV), word-medially (VCV), and word-finally (VC).

Word-initial position (CV). Figure 2.38 shows spectral peak amplitude distributions for obstruents in CV contrasts, and broadly illustrates that peak amplitudes are the highest among sibilants, with the wide variance among voiceless fricatives further contributing to a manner effect where [f, θ] exhibit spectral peaks of much lower amplitude than their plosive counterparts [p, t]. Among

plosives, the target data shows a consistent increase in peak amplitude with more posterior constrictions, however this pattern is only shared among voiced plosives in the female reference data, and is not robust enough to yield above-chance AMP_{PK} distinctions across place contrasts in either the lexicon or inventory. Finally, there is a slight but consistent voicing effect among stop consonants, with voiceless stops exhibiting slightly louder spectral peaks than their voiced counterparts, but again, these effects are minimal given their overlap with the range of AMP_{PK} differences observed for non-contrastive within-item comparisons.

Word-medial position (VCV). The sibilance and manner effects observed above for word-initial contrasts are greatly reduced in VCV position, though sibilants and nonsibilants remain largely distinct in their spectral peak amplitudes. Manner of articulation, though reduced intervocalically (primarily due to the greater variability in AMP_{PK} among plosives), is also generally above-chance in its contrast effects, a result which can again be largely attributed to the relatively lower amplitudes of [f, θ] in comparison with the plosives [p, t]. Voiced plosives and nonsibilant fricatives, on the other hand, are more similar in this regard. Finally, as in CV position, obstruent voicing does appear to impact the amplitude of the main prominence in the spectrum, as voiced stops and sibilant fricatives generally exhibit lower amplitudes than their voiceless counterparts. However, no robust contrast effect for voicing is evident in Figure 2.39. Nevertheless, the median absolute AMP_{PK} difference for voicing contrasts remains above that observed for different repetitions of the same item in the inventory, so peak amplitude is not entirely uninformative about obstruent voicing, just less informative in relation to sibilance and manner of articulation.

Word-final position (VC). Word-finally, the primary distinctions in peak amplitude among obstruents are between sibilant and nonsibilant fricatives and between voiceless and voiced stops, with manner effects largely reduced relative to CV and VCV position. It is not clear, however, where this reduction in manner distinctions derives from, as nonsibilant fricatives remain lower in peak amplitude than their plosive counterparts. This result must then reflect the greater similarity in peak amplitudes among postalveolar obstruents (i.e., [ʧ, ʃ] and [ʤ, ʒ]) word-finally than was ob-

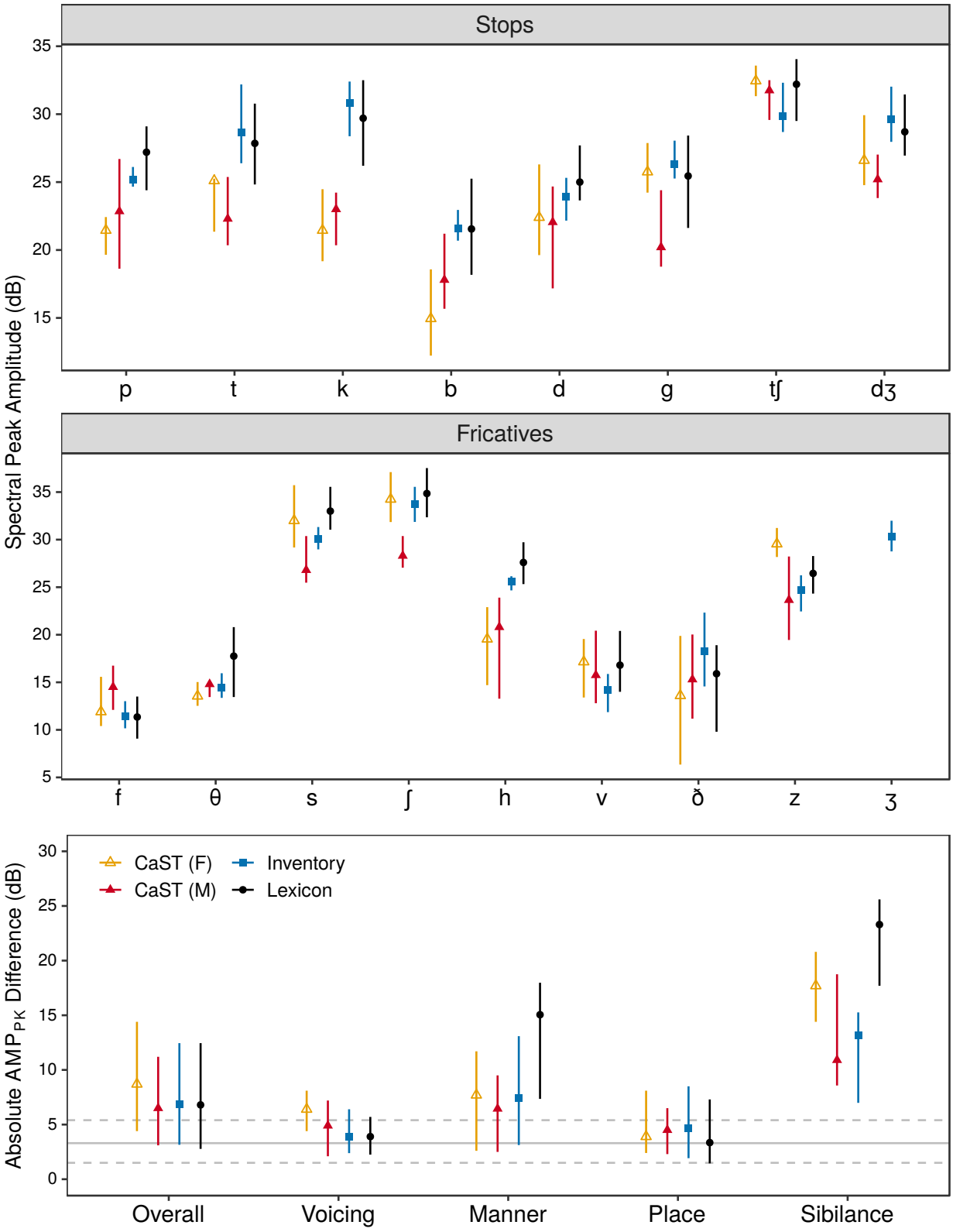

Figure 2.38: Spectral Peak Amplitude (AMP_{PK}) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{PK} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

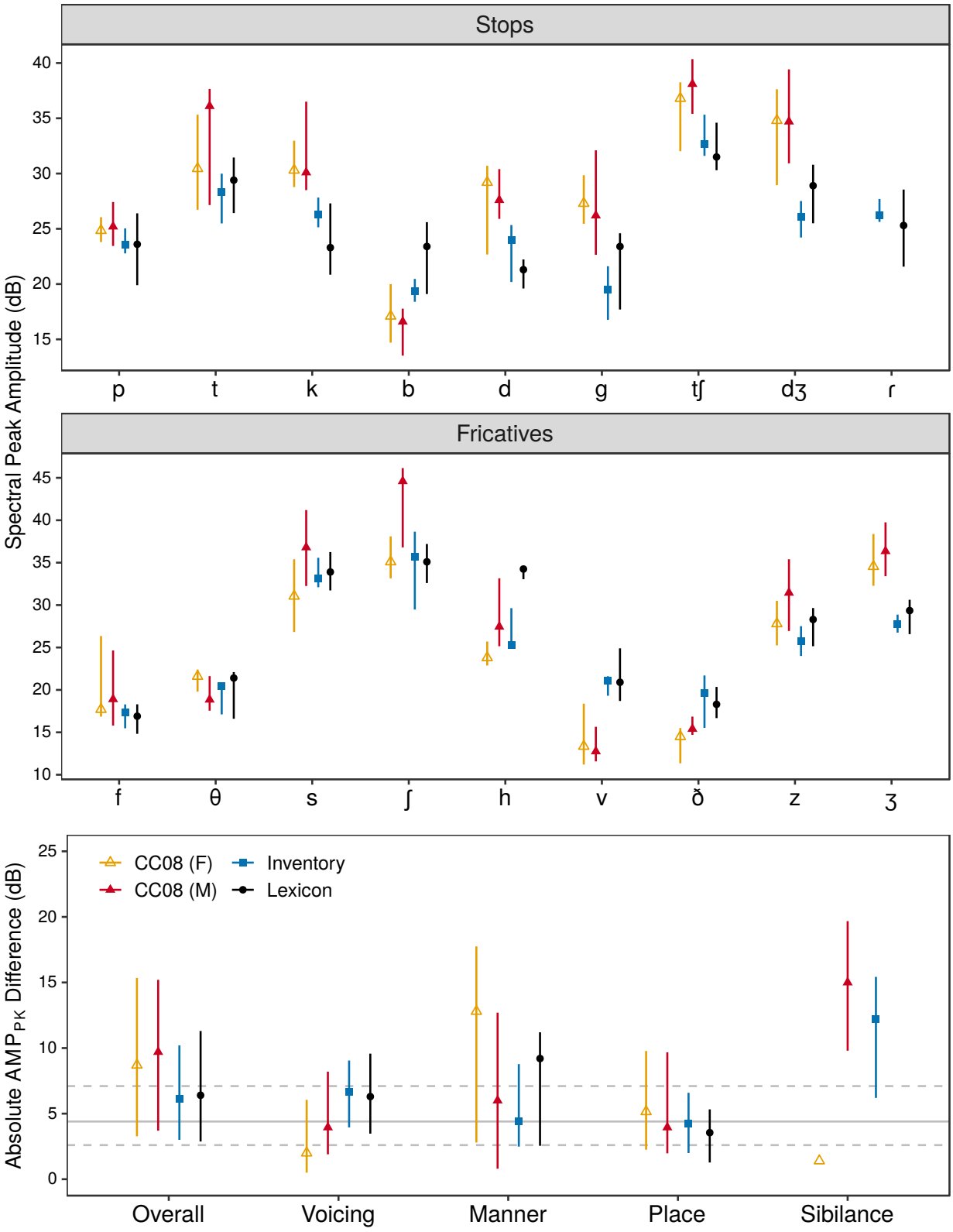

Figure 2.39: Spectral Peak Amplitude (AMP_{PK}) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{PK} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

served in word-initial or word-medial positions, as all other pure manner contrasts show absolute AMP_{PK} differences on the order of 5–10 dB.

2.6.2.4 Summary

In general, spectral peak amplitude retains much of the positional consistency of spectral peak frequency, primarily with regard to distinctions between sibilants and nonsibilants, but also to some extent as a function of voicing and manner of articulation, the former being more robust postvocalically and the latter more robust prevocalically. Regarding place of articulation, the picture is less clear. Though all three positions showed overall chance-level contrast effects for pure place distinctions, there were minor but consistent distinctions in peak amplitude among voiced and voiceless plosives, particularly between labial and lingual (alveolar/velar) places of articulation that may remain of some utility in obstruent contrast discrimination.

2.6.3 Dynamic Amplitude (AMP_{DYN})

2.6.3.1 Background and physiological basis

Dynamic amplitude was defined in Shadle & Mair (1996), with precursors in Shadle (1985), as a measure of the prominance of the main spectral peak with respect to the broader amplitude range in the spectrum. Physiologically, dynamic amplitude primarily reflects differences in sibilance, with sibilants exhibiting larger dynamic amplitudes than their nonsibilant counterparts. This distinction reflects differences in both airflow and anterior cavity size between the two classes. Dynamic amplitude also varies systematically with changes in vocal intensity (Koenig et al., 2013), but we are not concerned with this latter property at present.

2.6.3.2 Definition and measurement

Dynamic amplitude is defined as as the amplitude difference between the spectral peak amplitude (defined above as the maximum amplitude between 550 and 10,000 Hz) and the lowest ampli-

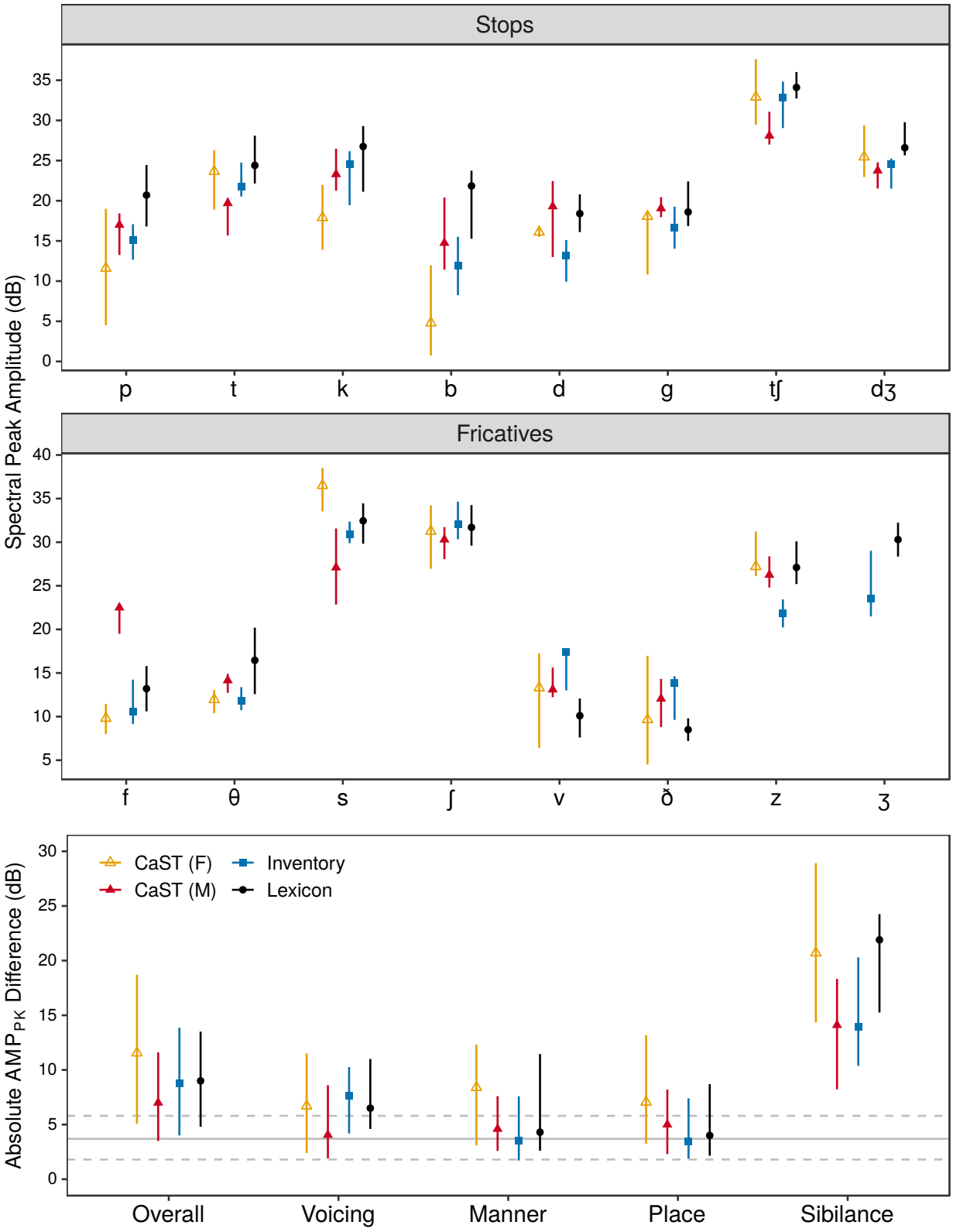

Figure 2.40: Spectral Peak Amplitude (AMP_{PK}) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{PK} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

tude—the amplitude *trough*—below 2 kHz. See Figure 2.34 for sample measurements of dynamic amplitude in noise spectra for [aʤa] and [afa]. Note from Figure 2.34 that in some cases, particularly for labials, the spectral peak is expected to occur before the low-frequency trough. This relation was not anticipated in the original definitions from Shadle (1985) and Shadle & Mair (1996), but follows from the definition and thus such cases are retained in the present study. We will see in the following section the extent to which dynamic amplitude, when applied to a diverse set of obstruents, positions, and databases, reflects the noise source distinctions—particularly between obstacle and non-obstacle turbulence generators—that it was originally designed to capture.

2.6.3.3 Category and contrast distributions

Here we present dynamic amplitude distributions for obstruent categories and contrasts in the lexicon, inventory and reference data. As before, results are presented separately for word-initial, word-medial, and word-final positions.

Word-initial position (CV). At word/syllable onset, dynamic amplitude appears to reflect characteristics of both laryngeal and noise sources, as the primary contrast effects in Figure 2.41 are for sibilance and voicing. Among plosives all four data sets show a marked separation between voiceless and voiced, with voiceless plosives approximately 10 dB higher in dynamic amplitude than their voiced counterparts, though this distinction weakens moderately at more posterior places of articulation. Affricates also show some reduction in dynamic amplitude from voiceless [ʧ] to voiced [ʤ], though this difference is narrower at around 5 dB. Finally, among fricatives the effect of voicing appears to be restricted primarily to sibilants, though the nonsibilants [f, θ] do show slightly higher dynamic amplitudes than their voiced counterparts [v, ð].

Regarding sibilance effects, both sibilant affricates and sibilant fricatives tend to show the highest dynamic amplitudes at between 50 and 70 dB. The one exception to this trend is the relatively high dynamic amplitude observed for [h], which based on its low spectral peak frequency (Figure 2.35) and steep negative spectral tilt (Figure 2.44 appears to be a characteristic of [h]'s more vowel-

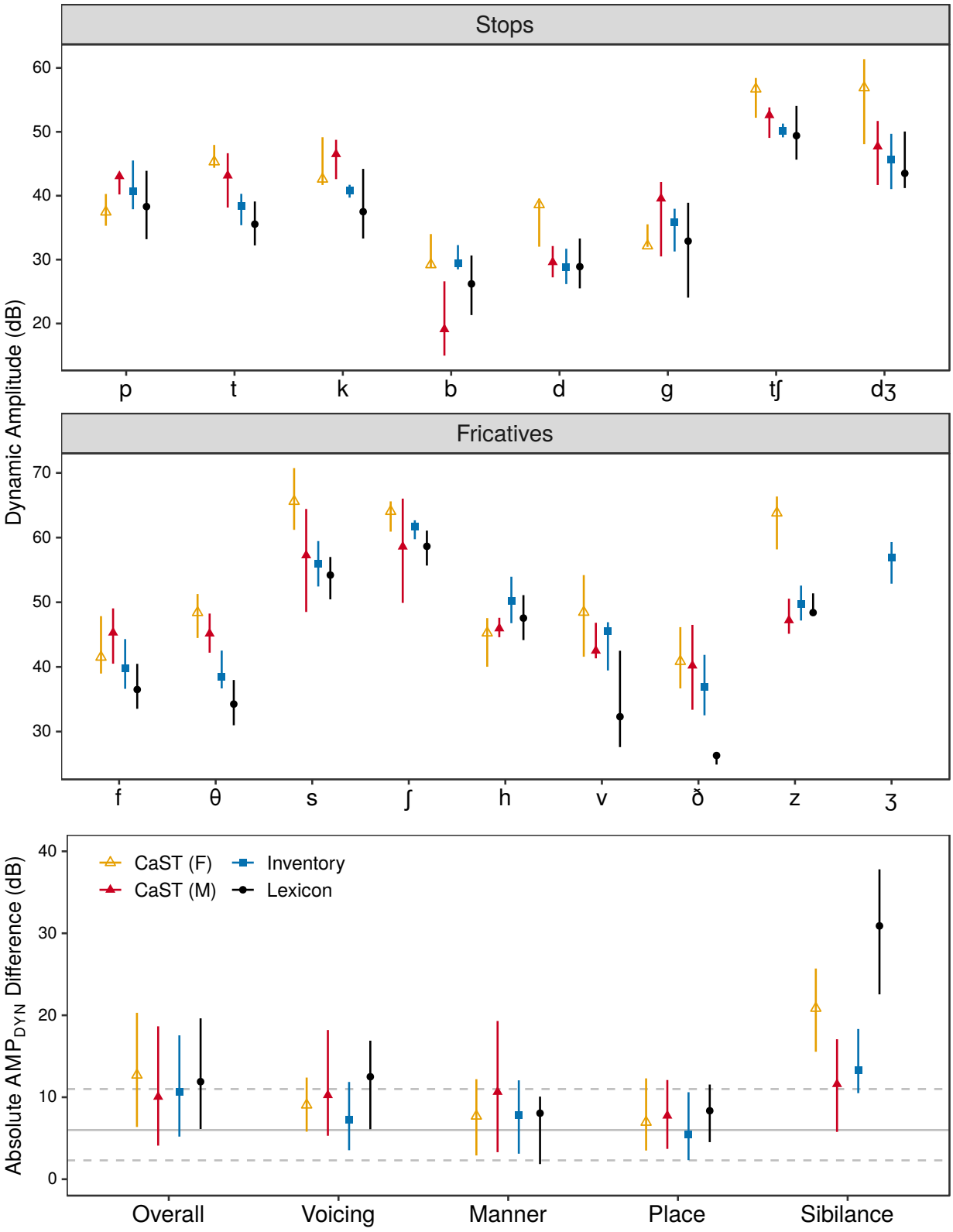

Figure 2.41: Dynamic Amplitude (AMP_{DYN}) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{DYN} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

like noise spectrum (see also the spectral tilts of vowel onset in Figure 2.45), consistent with many featural descriptions of [h] as [−consonantal] (Jakobson et al., 1951; Chomsky & Halle, 1968).

Word-medial position (VCV). Figure 2.42 shows dynamic amplitude distributions for intervocalic obstruent categories and contrasts, and generally agrees with the CV patterns in terms of sibilance, though voicing distinctions are notably reduced. The sibilance contrast effect in the inventory data, though apparent in the category distributions in Figure 2.42, is minimal and highly variable due to the fact that when the two contrasts, [s, θ] and [z, ð], are paired by vowel context, there are many pairs that are similar in dynamic amplitude, while a few contrasts show outlier differences of 20–30 dB. Thus, when aggregated the contrast effect for sibilance is notably reduced (around 4–5 dB) from the effect implied by the distributions for [s, z, θ, ð] (9–10 dB).[14] There is also a manner distinction that emerges, where fricatives are moderately higher in dynamic amplitude than their plosive counterparts, though as the bottom panel of Figure 2.42 shows, this effect varies widely, particularly in the lexicon. In general the absolute AMP_{DYN} differences observed for intervocalic obstruent contrasts are comparable to those in CV position, though there is greater overlap between the inventory and lexicon ranges and the chance range. This result partly reflects the greater within-category variance in dynamic amplitudes in VCV position, where IQRs average 9 dB as compared with 7 dB word-initially despite the similarity between the two positions in overall AMP_{DYN} range: 56 dB in CV (15–71 dB), 57 dB in VCV (17–74 dB).

Word-final position (VC). Among obstruent contrasts in VC position, the dynamic amplitude distributions in Figure 2.43 reflect a combination of effects present in CV and VCV positions. The one constant across contexts remains the sibilance effect, where sibilants exhibit higher dynamic amplitudes than nonsibilants, though the distinction is moderately reduced word-finally relative to differences observed in CV and VCV contrasts. Regarding voicing, there is the same voiceless >

[14] Here it is worth emphasizing that unlike the sample mean, the difference between sample medians is not equivalent to the median of the paired sample differences. This distinction is important for our understanding of the distribution of acoustic information in the lexicon and inventory, as our primary concern is the reliability of a given cue in distinguishing two items, and the median reflects the typical distinction more accurately than the mean.

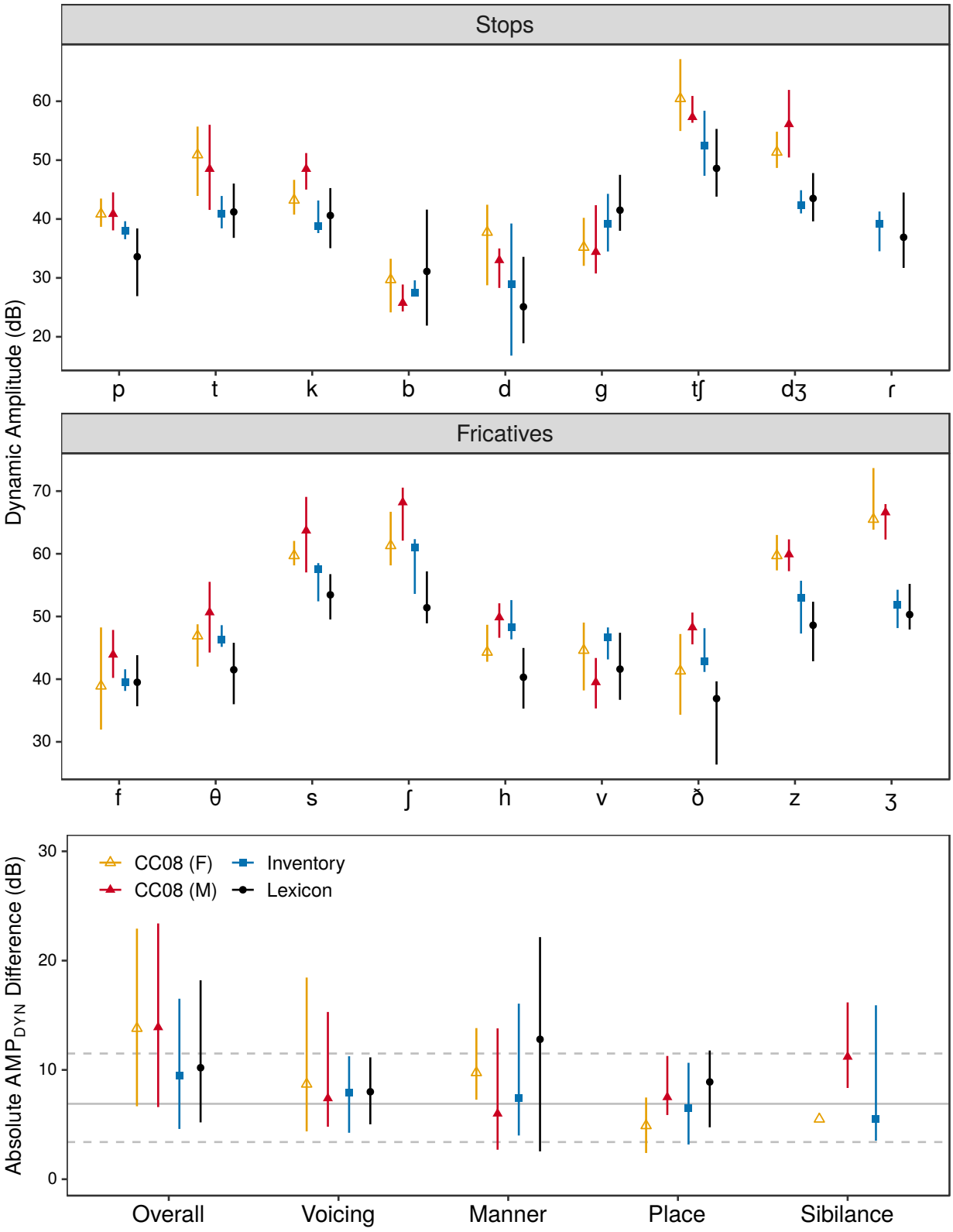

Figure 2.42: Dynamic Amplitude (AMP_{DYN}) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{DYN} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

voiced relation word-finally as that observed in word-initial position, though the word-final voicing contrast effects in the lexicon show greater overlap with the chance distribution.

Turning next to manner of articulation, there are fairly large absolute differences in dynamic amplitude for pure manner contrasts in the controlled syllable data, but in the lexicon distinctions are much reduced. The manner effect in the syllable data derives from the generally greater dynamic amplitudes of fricatives relative to their stop counterparts. Finally, as in CV and VCV position, there is a modest effect of place of articulation on dynamic amplitude, where alveolars and velars generally exhibit higher dynamic amplitudes than labials, though this pattern is more consistent in the lexicon than in the inventory or reference data. There is also a place distinction between the voiceless sibilant fricatives [s, ʃ], where [s] $<$ [ʃ], that is fairly robust in the inventory and male reference data of approximately 10 dB that is not present in the lexicon but which is interesting in that it stands in direct opposition to the results in Shadle & Mair (1996) and the theoretical predictions of Shadle (1985) based on noise source differences between alveolar and postalveolar sibilants. This relation is not in fact unique to VC position, but occurs word-initially and intervocalically as well in most databases, though only CV contrasts show this distinction in the lexicon. Given the generally higher spectral peak amplitudes observed for [ʃ], as well as the postalveolar's much narrower spectral peak frequency range, this result appears to derive from [s] exhibiting a much less defined and consistent spectral peak than [ʃ] due to its shorter and more variable front cavity (Tabain, 2001).

2.6.3.4 Summary

Dynamic amplitude is fairly consistent in its featural effects across different contrast positions, though all such effects are more variable than those observed for the frequency and amplitude of the spectral peak (FREQ_{PK}, AMP_{PK}). As in earlier work, the greatest distinction observed in dynamic amplitude was that between sibilants and nonsibilants. However, unlike in previous studies which have focused primarily on voiceless fricatives, we were also able to demonstrate modest but consistent effects of voicing (voiced $<$ voiceless) and place (labials $<$ alveolars, velars)

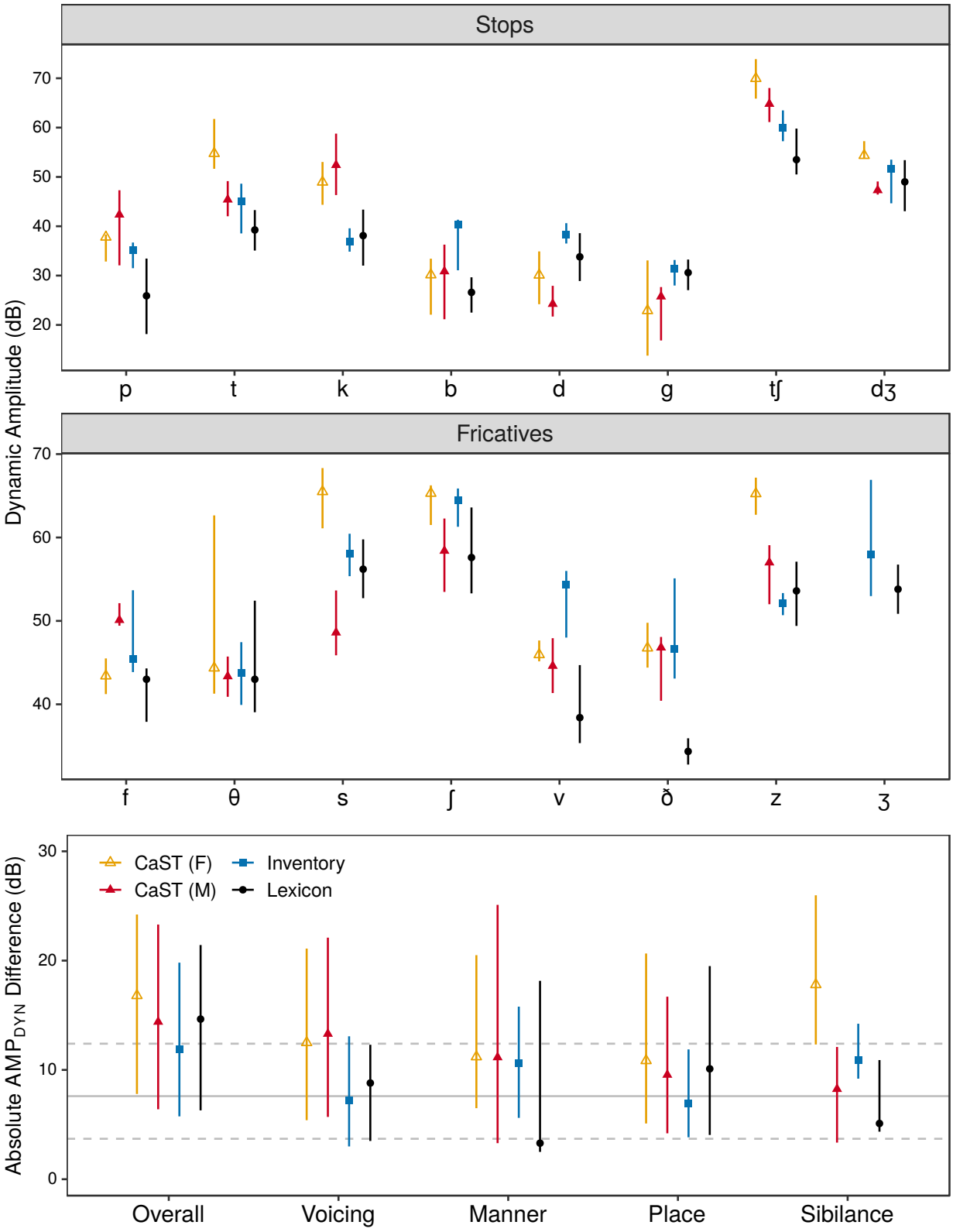

Figure 2.43: Dynamic Amplitude (AMP_{DYN}) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in AMP_{DYN} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

on the dynamic amplitude of stop spectra, as well as more general manner effects wherein fricatives tend to exhibit greater dynamic amplitudes than their stop counterparts.

2.6.4 Spectral Tilt ($\text{TILT}_{\text{C/V1/V2}}$)

2.6.4.1 Background and physiological basis

The spectral tilt of the consonant noise spectrum, and its tilt relative to that in the adjacent vowel, was first explored in detail in Lahiri et al. (1984), who found that labials exhibit steeper spectral tilts in the noise interval relative to alveolars, and consequently greater change in spectral tilt from alveolar bursts into the vowel onset relative to labial transitions. The variation in spectral tilt as a function of place of articulation is physiologically motivated in a manner similar to the previous analysis of spectral peak frequency; namely, excepting labials, which have no anterior resonating cavity, spectral tilt varies as a function of the resonance patterns in the vocal tract. This means that alveolars and velars, which tend to excite the upper formants (F2 and F3 for velars, F3–F5 for alveolars), exhibit the flattest, least negative spectral tilts, labials and glottals are more steeply negative tilting, and sibilant fricatives exhibit positive spectral tilts due to their relatively high spectral peak amplitudes. In terms of dynamics of the relationship between spectral tilt in the consonant and in the vowel, such relations are expected to reflect differences in coarticulation and noise dispersion at CV and VC transitions.

2.6.4.2 Definition and measurement

Spectral tilt is defined as the slope of the log-frequency spectrum between 550 and 10,000 Hz. It is measured by first taking the log-transform of the frequency domain of the spectrum, and then measuring the slope of a line fit to the 550–10,000 Hz interval via least-squares regression. Both spectral tilts of the consonant (TILT_{C}) and the onset/offset of the adjacent vowel ($\text{TILT}_{\text{V1/V2}}$) are measured, the former computed from the half/full Hamming windows on the consonant noise interval (as shown in Figure 2.22), and the latter derived from 20 ms half Hamming windows at

vowel onset/offset, where the greatest weight is placed on the CV/VC boundary, with window weights tapering off to zero as they move into the vowel.

2.6.4.3 Category and contrast distributions

In the sections below, which are organized according to contrast position, the spectral tilt distributions of both the consonant and the adjacent vowel(s) will be presented; i.e., TILT_{C} and TILT_{V2} in CV position; TILT_{C}, TILT_{V1}, and TILT_{V2} in VCV position; and TILT_{C} and TILT_{V1} in VC position. Here the consonant and vowel spectral tilts are grouped together because of their formal link in Lahiri et al. (1984) as the change in spectral tilt from the consonant to the vowel, though we do not use this composite parameter in the present study.

Word-initial position (CV). Figure 2.44 shows distributions of consonant spectral tilt for obstruent categories and contrasts in word-initial position, with stops exhibiting generally negative tilts between -30 and 0 dB, and fricatives spanning a much wider range from steep negative tilts around -30 dB for [h] to positive tilts of 10–30 dB for the alveolar fricatives [s, z]. Beginning with the plosives, while there is a consistent separation of the voiceless set [p, t, k] according to place of articulation, labials showing the most negative slopes, followed by velars, and then alveolars, the only place effect among voiced plosives is between [d] and [g], which though in the same direction as for [t, k]—i.e., velars steeper than alveolars, the distinction is much narrower. As a result of the flattening of spectral tilt distinctions between voiced plosives, a modest voicing distinction of voiced $<$ voiceless emerges between the lingual plosives, a pattern which extends to and widens among the affricates, with spectral tilts in the target data around -20 dB for [ʤ], and around -10 dB for [ʧ]. Finally, this difference of around 10 dB between voiceless and voiced affricates is also present in their fricative counterparts, with nonsibilant fricatives even further distinguished by voicing, particularly in the target data, where [v, ð] are around 20 dB steeper than [f, θ]. The lower panel of Figure 2.44 reflects these patterns in moderate voicing contrast effects across all four data sets, though effects are larger on average in the syllable data than in the lexicon.

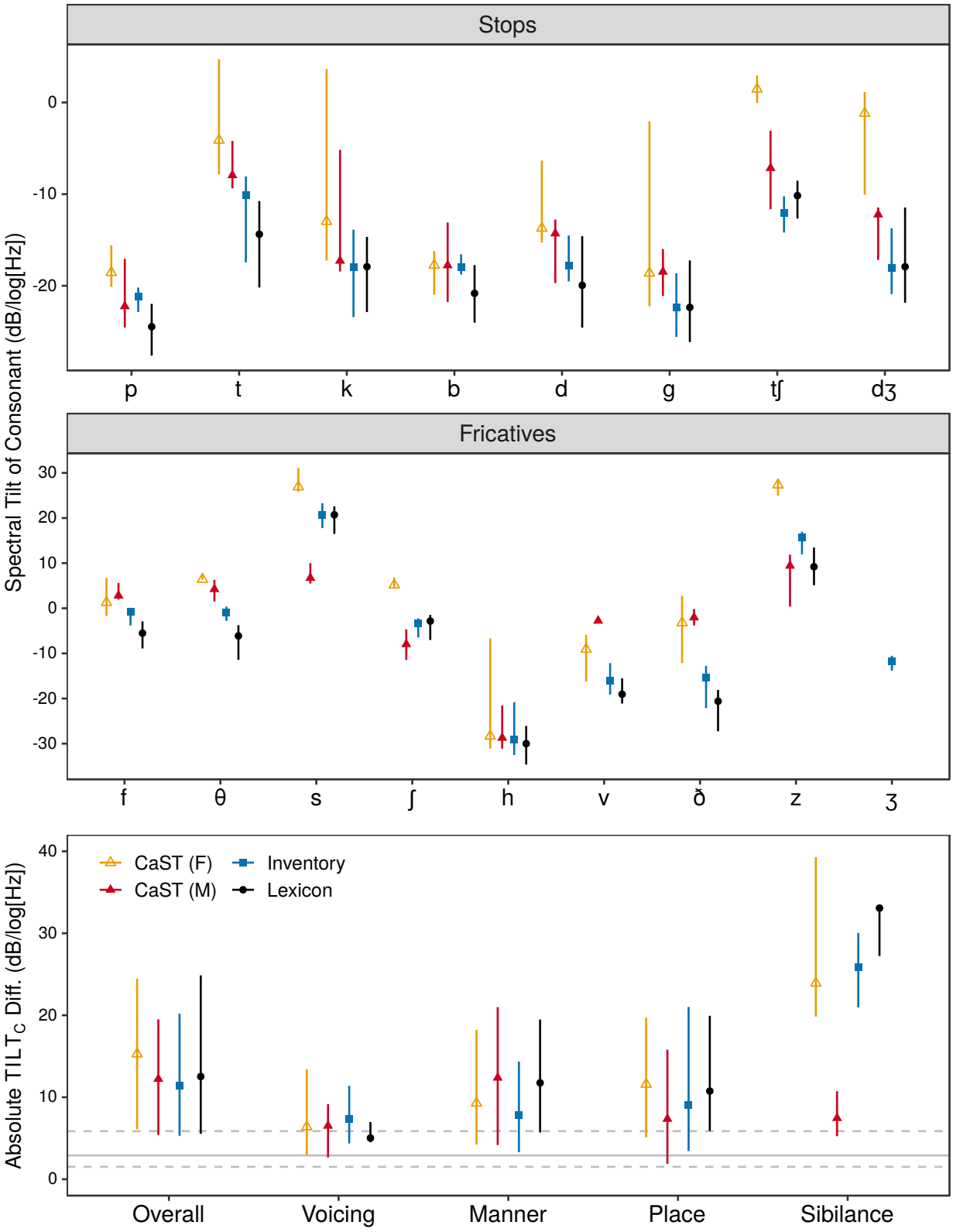

Figure 2.44: Spectral Tilt of Consonant Noise ($TILT_C$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $TILT_C$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

Regarding place of articulation effects among fricatives, spectral tilts among nonsibilants are constant across place, while sibilants show notable differences in tilt, with [ʃ, ʒ] relatively flat at −10–0 dB, and [s, z] more steeply positive at 15–25 dB. This 25–35 dB difference, combined with the place distinctions among plosives, contributes to robust place contrast effects, particularly in the lexicon. Manner effects are also robust, as noted earlier, with stops generally more negative than fricatives, particularly among voiceless obstruents. Absolute TILT_{C} differences range between 5 and 20 dB and show little to no overlap with the estimated chance distribution. Finally, the greatest distinction in spectral tilt is between sibilants and nonsibilants. Voiceless nonsibilant fricatives are generally flat, while voiced nonsibilants exhibit negative tilts similar to those observed in plosives. By comparison, [s, z] show large positive spectral tilts, resulting in sibilance effects of up to 40 dB.

Figure 2.45 shows spectral tilts at the onset of the vowel following obstruent contrasts in CV position. Given that the primary motivation for including TILT_{V2} in the present study is to serve as a reference for consonantal spectral tilt, the wide overlap in vowel spectral tilts as a function of preceding obstruent is not surprising. Nevertheless, there are a few consistent patterns of note. First, excluding the postalveolar and glottal obstruents, there is a moderate but consistent decrease in spectral tilt (i.e., steeper negative tilts) at vowel onset following more anterior places of articulation. That is, among plosives we find that overall, though primarily among the voiced series, labials are steeper than alveolars, which are further steeper than velars. Similarly, among fricatives we find the relation *labial* < *dental* < *alveolar*, where again the effect is more pronounced for voiced fricatives. This place effect reflects both differences in the excitation of higher formant frequencies as a function of the point of vocal tract constriction, and differences in the spread of aspiration noise into the vowel. Labials, for instance, due to the lack of a front resonating cavity show little amplification of the formants above F2, and thus have relatively steep negative spectral tilts. On the other hand, the noise generated at alveolar constrictions and at the teeth excites formants up to F5, while for velar constrictions the excitation of F2 and F3, combined with their proximity in frequency, serves to generate a strong mid-frequency prominence in the spectrum that also flattens out the spectral tilt, particularly when measured on a log-frequency scale.

This pattern, *labials* < *velars* < *alveolars*, is what was observed for consonantal spectral tilts, but it should theoretically extend into the vowel onset as well (Fant, 1960; Stevens & Blumstein, 1978). However, we find the opposite relation between alveolars and velars. From closer inspection of the spectrograms and vowel-onset spectra in [d]- and [g]-onset items, we find that the primary cause of this inversion is the greater persistence of high-frequency noise into vowels following [d] relative to vowels following [g]. In other words, the transition from plosive aspiration to modal voicing tends to be more gradual following alveolars than velars, meaning that at vowel onset (defined primarily based on the onset of F2) following [d] the upper formants are notably more damped by aspiration noise relative to the formants at [g]-offset. This relative increase in high-frequency energy in vowels following velars ultimately leads to a flatter spectral tilt relative to vowels following alveolars. While Lahiri and colleagues did not study [d, g] contrasts in their 1984 study where the $\Delta TILT$ measure was originally proposed, these results are consistent with the theory motivating their study: namely, that obstruents should differ both in the excitation characteristics of their noise spectra and in the dynamics of the spectral transition into the vowel.

Among lingual obstruents—that is, excepting labials and glottals—there is a slight voicing effect for TILT_{V2}, where the spectral tilt at vowel onset following voiced obstruents tends to be flatter than that at corresponding voiceless CV transitions. This pattern is consistent with the place effect discussed above, and derives from the greater spread of high-frequency noise from voiceless obstruents into the vowel, and the upper formant dampening that results from such diffusion. Voicing contrast effects, however, are not as robust as the place effects, and are further weaker than the manner effect that derives primarily from the steeper negative spectral tilt following nonsibilant fricatives relative to their plosive counterparts. Finally, there is a notable sibilance effect on V2 spectral tilts in the lexicon that is consistent with the place description above, given that pure sibilance contrasts based on the present feature set are restricted to the [s, θ] and [z, ð] distinctions, though we expect such effects to extend to multi-feature sibilance contrasts as well.

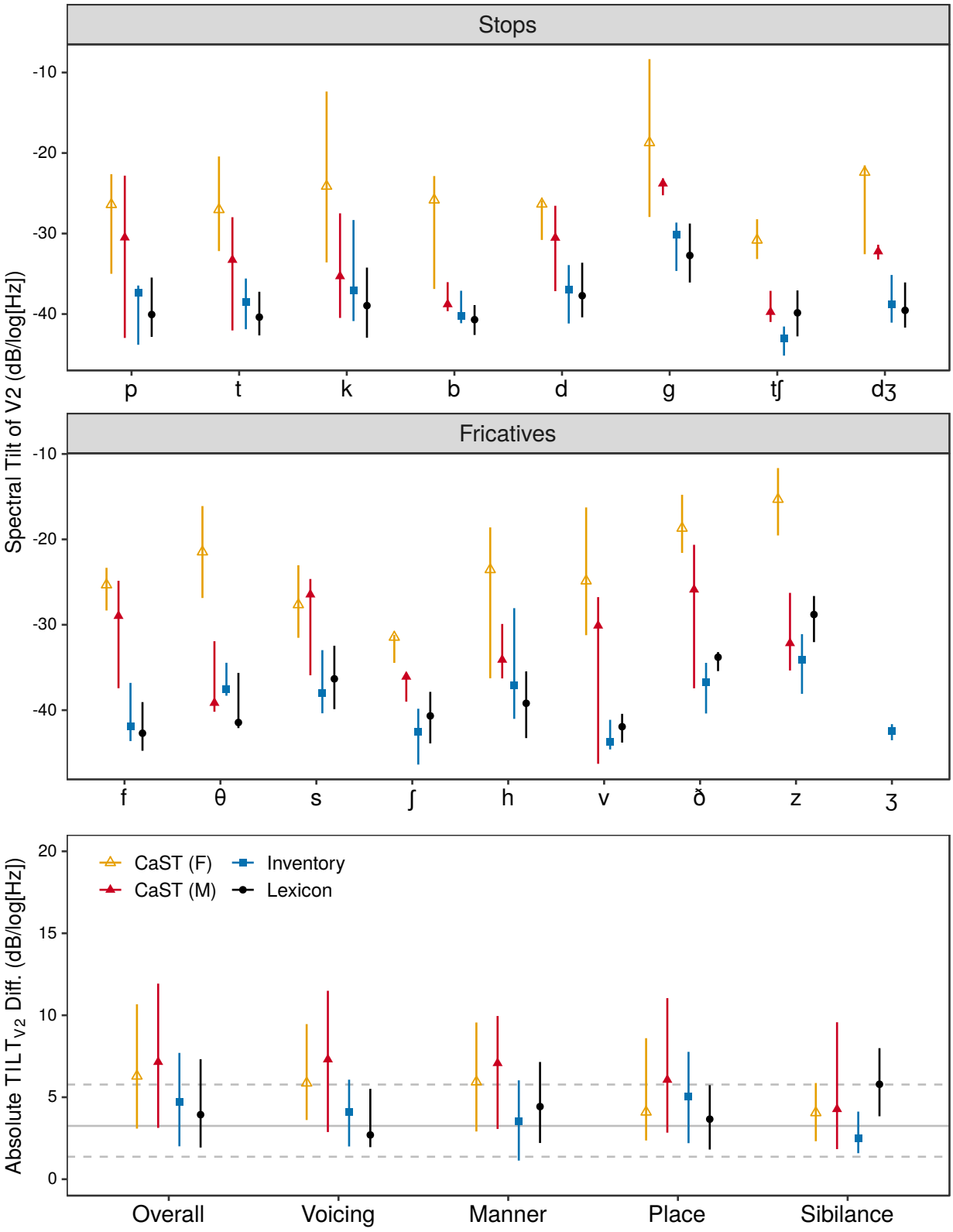

Figure 2.45: Spectral Tilt at Vowel Onset (TILT$_{V2}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in TILT$_{V2}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

Word-medial position (VCV). Intervocalically, many of the same patterns observed for consonant spectral tilt in CV position are preserved. As Figure 2.46 shows, sibilant fricatives remain the highest in positive spectral tilt, while the glottal fricative [h] and the alveolar flap [ɾ] exhibit the most negative spectral tilts, consistent with their more vowel-like quality in comparison to the other obstruents. Sibilants are also notably distinct from nonsibilants in this regard, particularly among voiced obstruents where [v, ð] exhibit steep negative spectral tilts consistent with their more approximant-like character intervocalically. Regarding voicing, the *voiceless* > *voiced* pattern is more robust in VCV position, at least in the target data, as is the manner distinction in the lexicon, which in addition to the manner effects discussed for CV contrasts is aided by the frequent occurrence of contrasts with the alveolar flap, whose spectral tilt as noted earlier is atypically negative. The place effects shown in CV position are even greater intervocalically, as there is a robust *labial* < *velar* < *alveolar* relation among the plosives in all four data sets, as well as a distinct *alveolar* > *postalveolar* effect and a slight but consistent *labial* < *dental* effect among the fricatives.

Finally, as in CV position, the effect of obstruent sibilance is the greatest at around 25 dB on average in the inventory, and though no pure sibilance contrasts are present intervocalically in the lexicon, the substantial positive spectral tilts of the alveolars [s, z] and the nearly flat tilts of the affricates [ʧ, ʤ] in comparison with all other stops except [t], suggest obstruent sibilance is well indexed by TILT_{C} in the lexicon as well. Thus, overall contrast effects from consonant spectral tilt are greater in VCV position than in CV position. Next we consider whether these results extend to spectral tilt at the preceding and following vowel transitions.

Beginning with spectral tilt at vowel offset, TILT_{V1}, Figure 2.47 shows little continuity with the TILT_{V2} results for CV contrasts. There are slight effects of voicing and manner in the lexicon, but both contrast effects are well within the estimated chance range and from the category distributions. Further, manner effects appear to be restricted primarily to the [p, f] and [t, θ] contrasts and contrasts with the alveolar flap, while voicing effects are only apparent for labial and alveolar plosives. Neither place nor sibilance exhibit above-chance TILT_{V1} differences in any of the databases. These results, though less promising for the discrimination of intervocalic con-

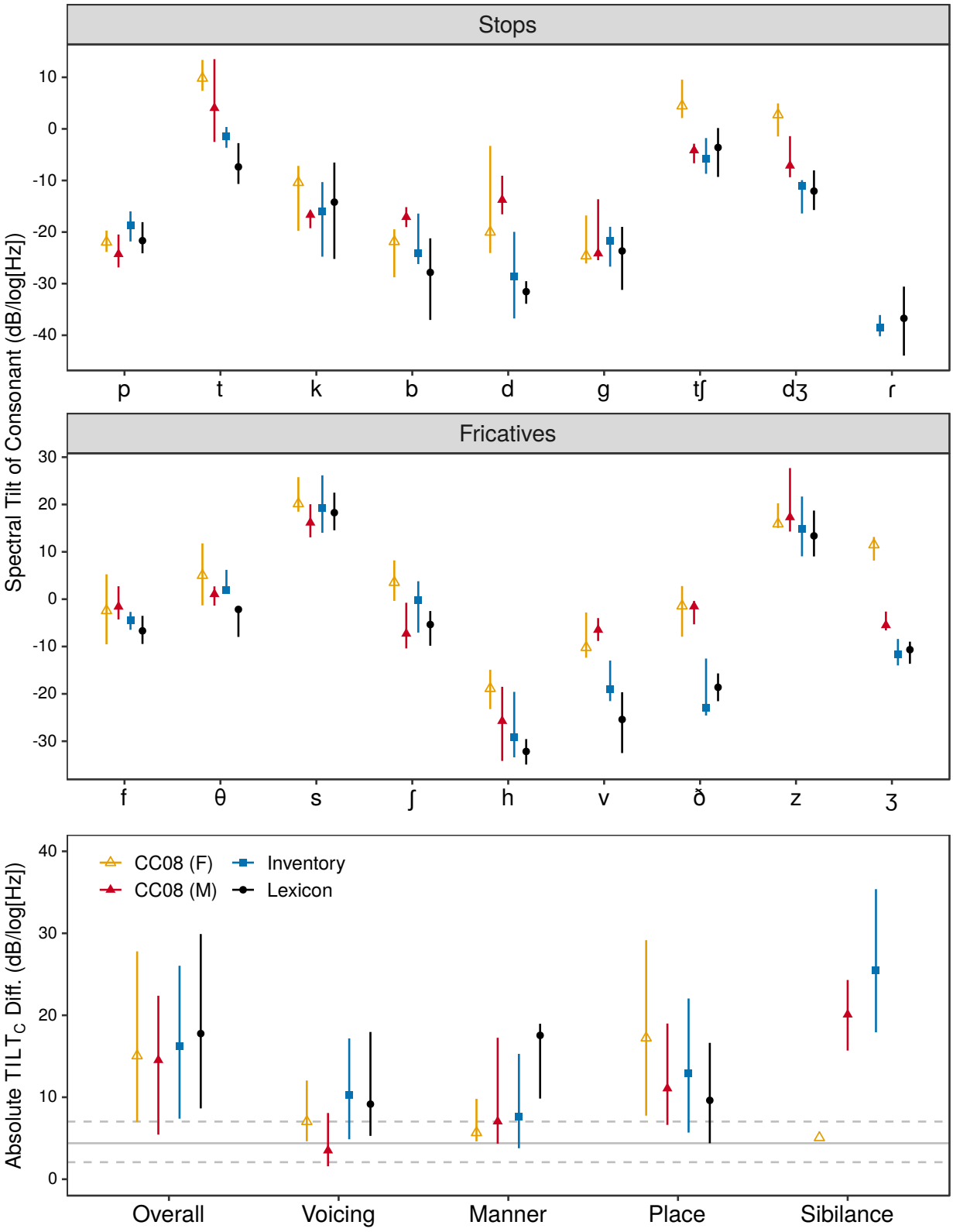

Figure 2.46: Spectral Tilt of Consonant Noise ($TILT_C$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $TILT_C$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

trasts, are not surprising given that both explanations for the vowel-onset spectral tilt patterns in CV position—differences in upper formant excitation as a function of place of articulation and differences in noise diffusion characteristics at the CV transition as a function of voicing, manner, and sibilance—rely on the obstruent preceding the vowel and occurring at syllable onset. When word-final contrasts are examined in the next section we will be able to test the extent to which coda obstruents may impact the overall spectral tilt at vowel offset.

Spectral tilt distributions at V2 onset in VCV contrasts are shown in Figure 2.48, and mirror the TILT_{V2} patterns in CV position, though voicing and manner effects are generally reduced, as are sibilance effects due the greater similarity in vowel transitions between [s, z] and [θ, ð] intervocalically. Place of articulation, however, retains its impact on vowel-onset spectral tilt, as transitions from labials exhibit more negative tilts than coronals and velars, and among voiced obstruents, particularly voiced plosives, alveolar transitions are further steeper than velars. Further, the [s, ʃ] and [z, ʒ] distinctions in CV position—i.e., *alveolar* > *postalveolar*—are also present in VCV position. Thus, TILT_{V2} remains a useful parameter intervocalically, though its featural contribution is somewhat narrower than its role in word-initial contrasts.

Word-final position (VC). Figure 2.49 shows consonant spectral tilt distributions for obstruent categories and contrasts in VC position. Overall, the place and sibilance effects in the other two positions are retained word-finally, as is the voicing effect that was present word-initially. However, manner effects are much reduced word-finally, though the category distributions show the same general patterns as before: i.e., excepting [t], plosives exhibit steeper spectral tilts than their nonsibilant fricative counterparts, while the stop–fricative distinction among postalveolars is much narrower. Regarding voicing, voiceless stops, particularly coronals, exhibit flatter tilts than their voiced counterparts, while the place effect is largely consistent with that observed in CV and VCV contrasts. Alveolar fricatives show spectral tilts approximately 30 dB greater than postalveolars, while the [p] < [t] distinction is similarly robust at around 15–20 dB. The noise spectra of word-final obstruents do not, however, exhibit the [k] < [t] relation that is present in CV

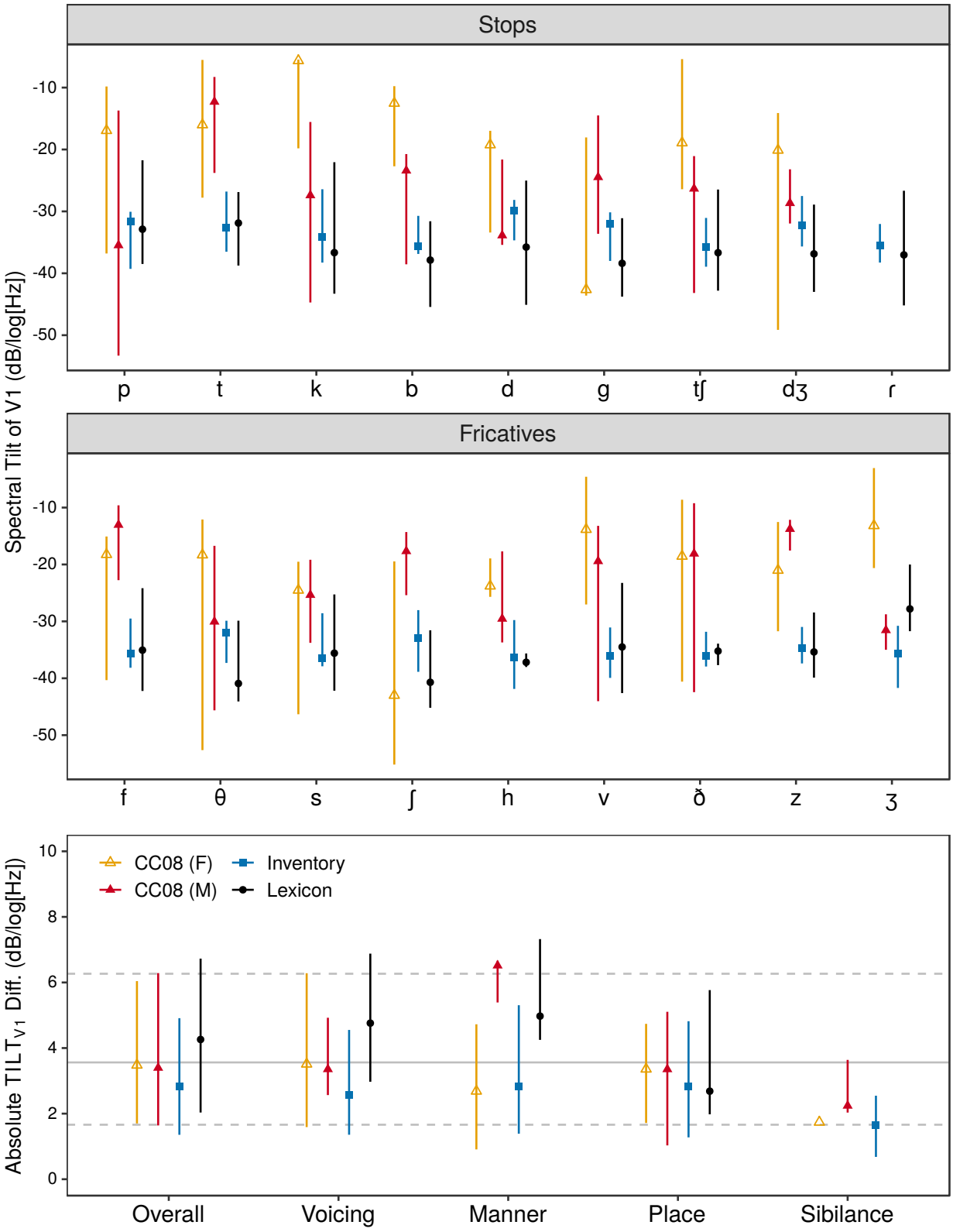

Figure 2.47: Spectral Tilt at Vowel Onset ($TILT_{V1}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $TILT_{V1}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

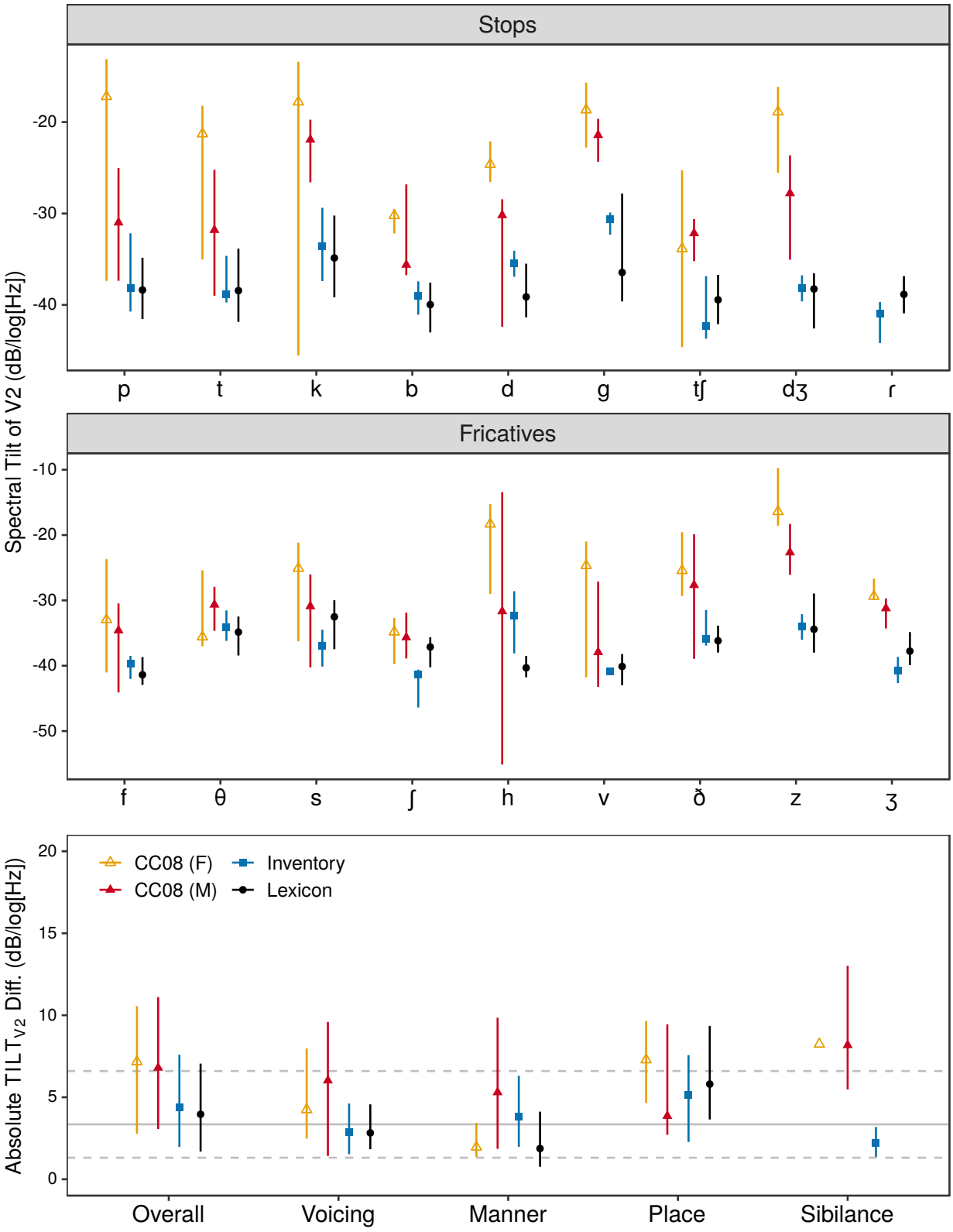

Figure 2.48: Spectral Tilt at Vowel Onset ($TILT_{V2}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $TILT_{V2}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

and VCV position. Finally, the *labial* < *dental* relation among intervocalic fricatives is notably reduced word-finally, bringing VC contrasts more in line with those in CV position. Considering word-final contrast effects overall, $TILT_C$ is the most discriminative in the lexicon of any of the three positions, a result which reflects in part the greater asymmetry in VC contrast distributions, as over 20% of word-final lexical contrasts are between [d] and [z], whose spectral tilt difference ranges between 30 and 50 dB.

Spectral tilts at vowel offset (Figure 2.50), though still less informative than $TILT_{V2}$, are more discriminative of word-final contrasts in the lexicon than word-medial lexical contrasts, primarily along the voicing dimension. Figure 2.50 shows that among plosives, the spectral tilt of the vowel preceding voiceless plosives, particularly in labial and velar contexts, tends to be shallower than that preceding voiced plosives, while affricates and fricatives show the opposite pattern. Examination of word-final plosive and fricative transitions revealed that among plosives this effect derives largely from the relative amplification of F1 preceding voiced plosives, as well as the relatively higher amplitudes of upper formants such as F3 and F4 preceding voiceless plosives. As for fricatives and affricates, the *voiceless* < *voiced* relation appears to be due to the more gradual VC transitions of voiced fricatives and affricates relative to voiceless, resulting in vowel spectra that are generally more diffuse and therefore less steep. Ultimately, however, these results are restricted to obstruent contrasts in the lexicon. Vowel-offset spectral tilt remains relatively constant as a function of voicing in both the inventory and reference syllable data.

2.6.4.4 Summary

The results above reveal substantial differences between obstruent consonants in terms of the global tilt of their noise spectra, and to a lesser extent the spectra at vowel onset, while vowel-offset spectral tilt is only of limited utility in word-final lexical contrasts. The most consistent patterns in $TILT_C$ distributions are those reflecting obstruent sibilance and place of articulation, where alveolar sibilants exhibit large positive spectral tilts in comparison to the flat and sometimes negative spectral tilts of nonsibilants, and place effects can be seen primarily in the distinction between

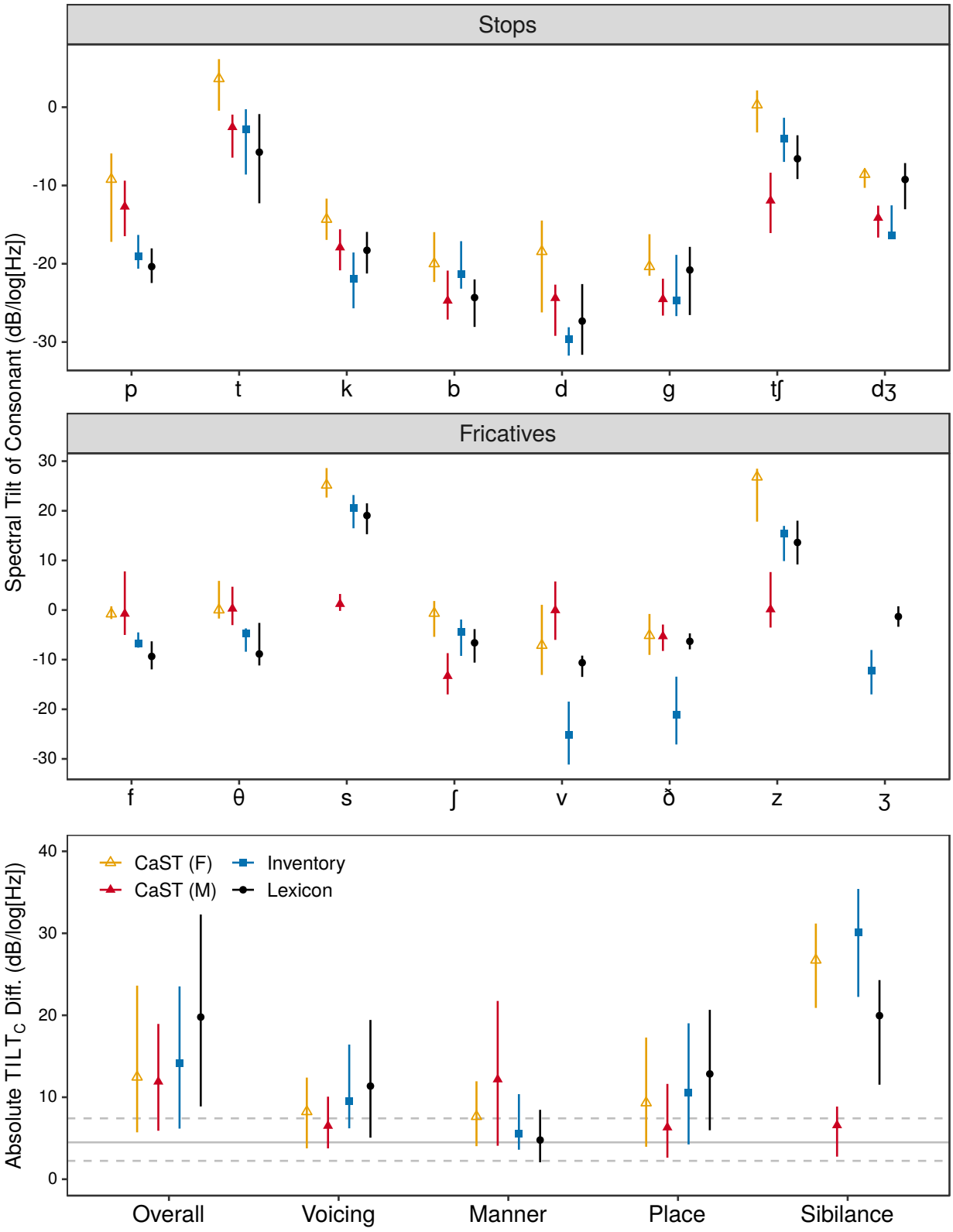

Figure 2.49: Spectral Tilt of Consonant Noise (TILT$_C$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in TILT$_C$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

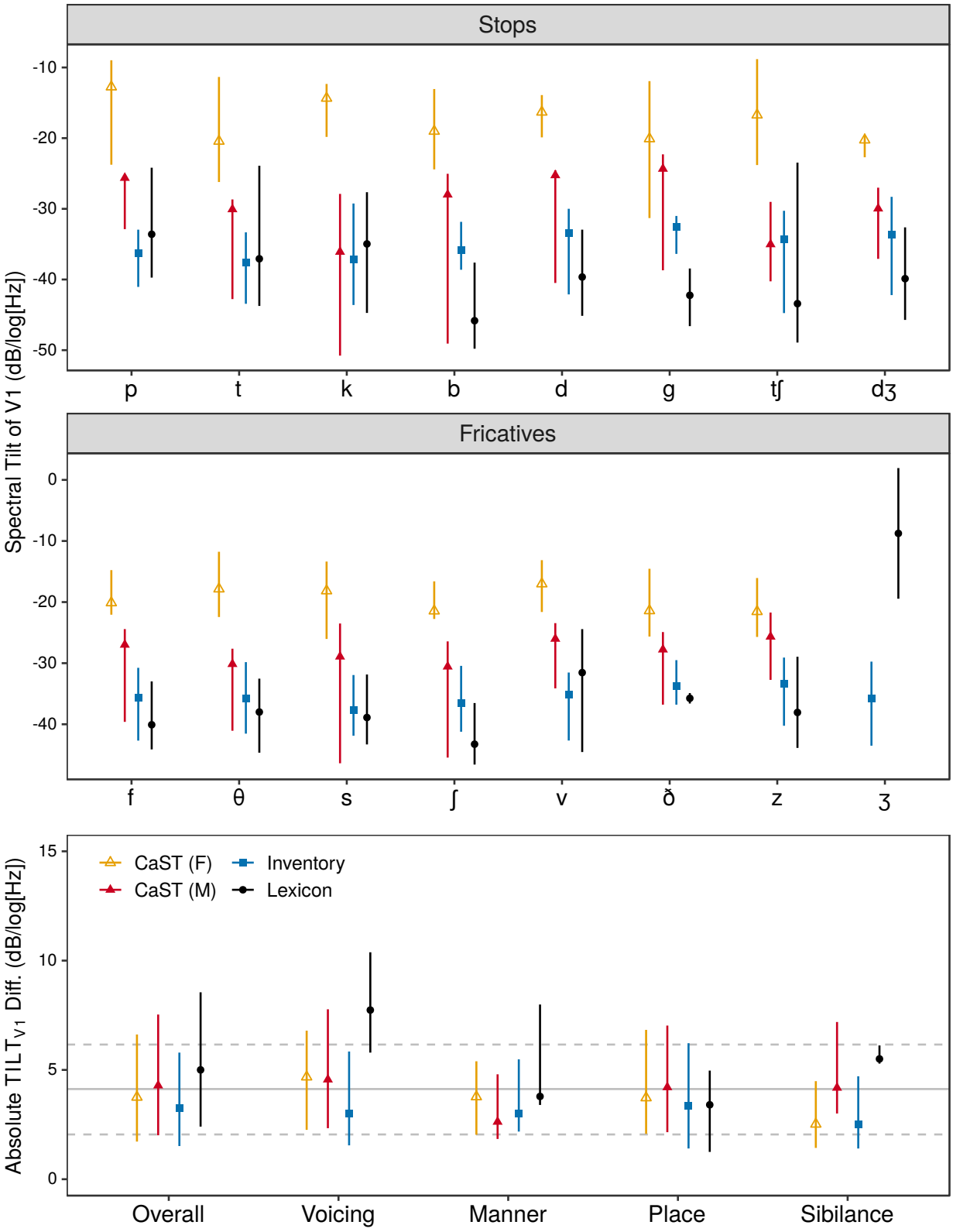

Figure 2.50: Spectral Tilt at Vowel Offset (TILT$_{V1}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in TILT$_{V1}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

labials and non-labials (lab. $<$ dent./alv./vel.), and to a lesser extent between voiceless alveolar and velar plosives ([t] $<$ [k]). Voicing and manner of articulation also affect consonant spectral tilt, though less consistently across CV, VCV, and VC positions. Generally, voiceless obstruents, particularly affricates and fricatives, exhibit shallower spectral tilts than their voiced counterparts, while the voicing distinction among plosives is less uniform and depends on place of articulation. Regarding manner, stop noise spectra tend to exhibit more negative spectral tilts than their fricative counterparts, though these effects are notably reduced word-finally.

Spectral tilt at vowel onset is much less discriminative, but does provide information about obstruent place of articulation that is consistent across CV and VCV positions: namely, CV transitions tend to exhibit greater spectral tilt (more negative) following more anterior constrictions. This pattern results from differences in both the excitation of higher formant frequencies as a function of anterior cavity size, and in the spread of high-frequency noise into the vowel. Taken together, the TILT_{C} and TILT_{V2} results suggest that there is value in retaining both parameters for use in models of acoustic cue integration, as opposed to combining them into a single spectral tilt change parameter, as they reflect subtle but meaningful differences in articulatory causes, particularly with regard to consonant-vowel coarticulation in the lexicon, where the spectral tilt of the CV/VC transition reflects in part the relative phasing of the consonant and vowel gestures and the degree to which information from one spreads into the other.

2.6.5 Spectral Shape (SHAPE)

2.6.5.1 Background and physiological basis

The parameter referred to as *spectral shape* in the present study, is the *steepness difference* metric proposed in Evers et al. (1998) for the cross-linguistic discrimination of sibilant place distinctions. Broadly, spectral shape aims to capture differences in the broad distribution of energy in the spectrum by measuring the spectral tilt above and below the mid-frequency region around 2500 Hz. This measure was favored over statistical measures such as spectral kurtosis (Forrest et al., 1988; Jongman et al., 2000) because it is less sensitive to variation in spectral peak frequency and low-

frequency energy, and has a more direct physiological link in capturing the distinction between noise generated at postalveolar constrictions and noise generated at other places of articulation.

2.6.5.2 Definition and measurement

Spectral shape is defined as the difference in spectral tilt between a low-frequency interval of 550–2500 Hz and a high-frequency interval of 2500–8000 Hz; i.e., SHAPE = $\alpha_L - \alpha_H$, where α_L is the spectral tilt between 550 and 2500 Hz, and α_H is the tilt between 2500 and 8000 Hz. Figure 2.34 illustrates the measurement of spectral shape via separate least-squares line fits over the aforementioned low- and high-frequency regions of the spectrum. Note that here, unlike in the measurement of spectral tilt in the previous spectrum, raw frequencies are used rather than log-transformed frequencies, meaning spectral shape is measured in dB/kHz rather than dB/log(Hz).

2.6.5.3 Category and contrast distributions

Below we review spectral shape distributions in the lexicon, inventory, and reference data, with results presented separately for word-initial, word-medial, and word-final obstruent contrasts.

Word-initial position (CV). From Figure 2.51 we see that word-initally, spectral shape is notably distinct for postalveolars in comparison to all other obstruents. Most obstruents show spectral shapes ranging between −5 and 0 dB, while postalveolars exhibit large positive shapes on the order of 5–15 dB, reflecting their main prominence around 2.5–3 kHz which results in a steep positive slope in the low-frequency range and a steep negative slope in the high-frequency range. That is, the results in Figure 2.51 are consistent with the robust [s, ʃ] distinctions reported in Evers et al. (1998), but otherwise spectral shape shows little differentiation of other word-initial obstruent contrasts. Among the other featural effects in Figure 2.51 are moderate effects of voicing and manner, where the former are restricted primarily to the lexicon and reflect the fact that voiced stops exhibit larger spectral shape values than their voiceless counterparts, and the latter derive from larger spectral shape values in [ʃ, ʒ] than in [ʧ, ʤ], and to a lesser extent between voiceless nonsibilant

fricatives and plosives: [p] < [f], [θ] < [t].

However, one problem that is immediately apparent in the interpretation of spectral shape is in its interpretability, as without examining the spectrum itself or the relation between spectral shape and other spectral parameters, it is unclear whether a given effect is due to differences in the distribution of energy in the high- or low-frequency regions, or some combination of the two as in the robust distinction between [s] and [ʃ]. Closer inspection of low- and high-frequency spectral tilts for word-initial obstruents reveals that the [p, f] distinction is due to the relatively flat [f] spectra as compared with [p] spectra where there is a steeper tilt in the low-frequency range, followed by a shallower tilt over the high-frequency range. Similarly, the [θ] < [t] result derives from [t] exhibiting a relatively flat spectrum as compared with [θ], which exhibits a steep fall between 550 and 2500 Hz and a relatively flat spectrum between 2500 and 8000 Hz. Regarding voicing, the lower spectral shape values among voiceless plosives reflect negative spectral tilts in the low-frequency region that are generally steeper than the corresponding tilts in their voiced counterparts. These results are inconsistent with the spectral shapes anticipated in Evers et al. (1998), where the location of the spectral peak either in the mid-frequency region for [ʃ], or in the high-frequency region for [s], results in radically different spectral shapes. Much of the voicing and manner effects reported above can be captured in the overall spectral tilt of the consonant noise spectra (see Figure 2.44, for example), with the 2500 Hz threshold between low- and high-frequency regions exhibiting little to no impact on most obstruent distinctions.

Word-medial position (VCV). Intervocalically, the distinction between postalveolar and non-postalveolar obstruents narrows, though the [s, ʃ] and [z, ʒ] contrasts remain robust. Figure 2.52 shows, however, that voicing and manner effects in the lexicon widen in VCV position relative to CV position, with [d, g] notably higher in spectral shape values than [t, k], though labials show the opposite pattern: [b] < [p]. Among fricatives, voiced fricatives (including the mostly voiced intervocalic [h]) tend to exhibit lower spectral shape values than their voiceless counterparts, though postalveolar obstruents tend to show the opposite pattern, as in CV position; i.e., [ʤ] >

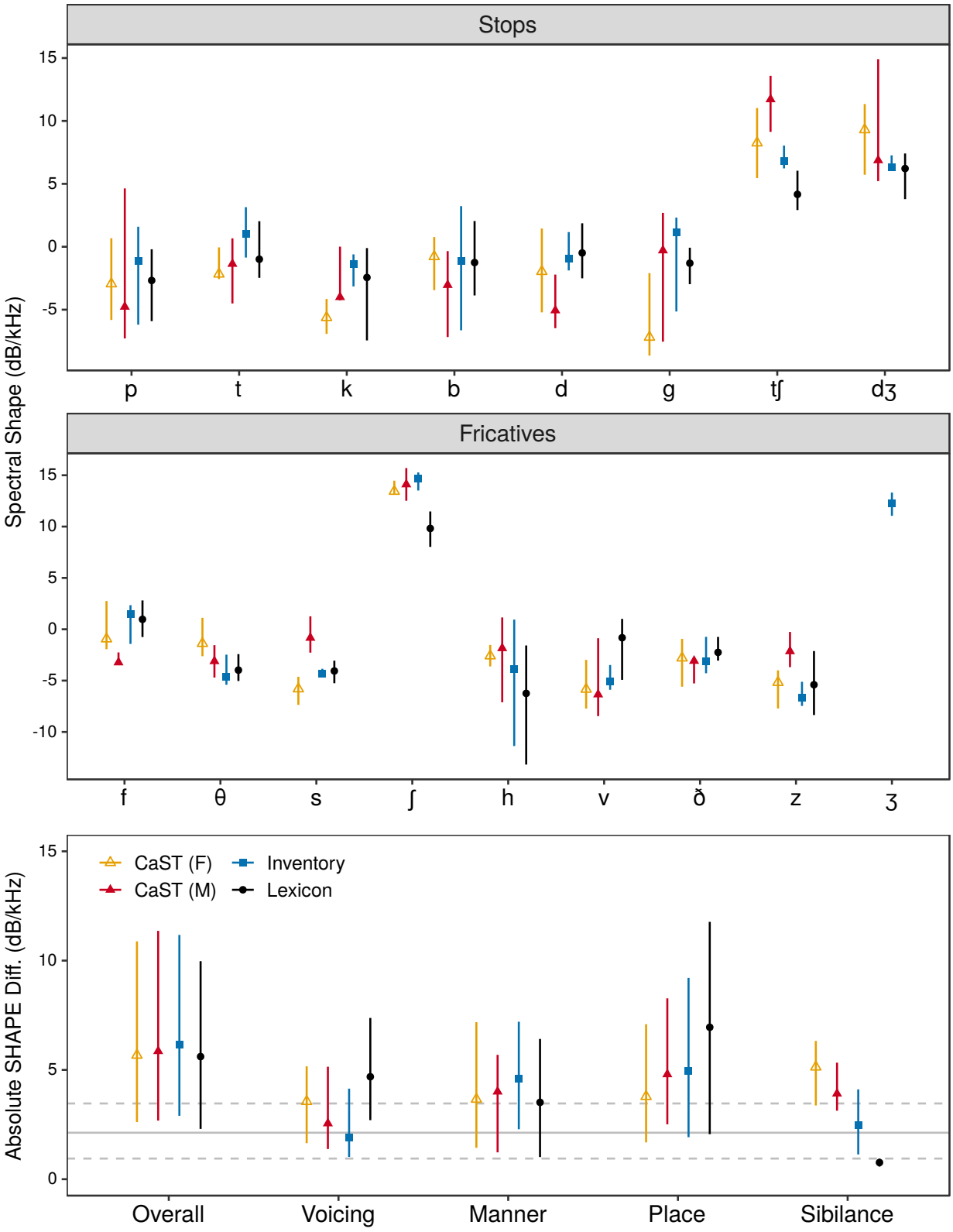

Figure 2.51: Spectral Shape (SHAPE) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in SHAPE in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

[ʧ], [ʒ] > [ʃ]. Here, unlike in CV position, the separation of low- and high-frequency spectral tilts does result in distinctions not captured in the global spectral tilt, TILT_C, as voiced plosives exhibit overall steeper negative spectral tilts than their voiced counterparts, but amplitudes drop more rapidly with increasing frequency in voiceless plosives than in voiced plosives, leading to more negative spectral shapes for [t, k] than [d, g]. Labials do not follow this pattern due to the enhanced low-frequency F1 and F2 amplitudes of [b] relative to [p], resulting in steeper spectral tilts between 550 and 2500 Hz, and thereby lower spectral shape values. Voicing effects among fricatives, on the other hand, are consistent with the global spectral tilt patterns, wherein voiceless nonsibilant fricatives exhibit generally flat spectra, unlike the negative spectral tilts observed for voiced nonsibilants. Ultimately, spectral shape is similarly discriminative in VCV position as in CV position, though its effects are more broadly distributed among different contrasts and features.

Word-final position (VC). Figure 2.53 shows spectral shape distributions in word-final obstruent contrasts. As in CV and VCV positions, postalveolars, particularly the fricatives [ʃ, ʒ], exhibit the highest spectral shape values and contribute to a fairly robust place contrast effect in the lexicon, inventory, and reference data. Other place distinctions evident in Figure 2.53 are the higher spectral shapes of velar plosives relative to labials and alveolars in the voiceless series, and the higher shape values of alveolars and velars relative to labials in the voiced series, all of which derive primarily from differences in spectral tilt over the low-frequency range. Voicing and manner effects are less robust than in word-initial or word-medial position, contributing to a lower overall contrast effect for spectral shape in word-final obstruents.

2.6.5.4 Summary

In general, spectral shape is less broadly discriminative than the global spectral tilt, but it does provide a stark discrimination of postalveolar fricatives and affricates from the remainder of the obstruent phones that is consistent across contrast positions. Further, there are effects of voicing and manner of articulation captured in spectral shape that derive from differences in the distribution

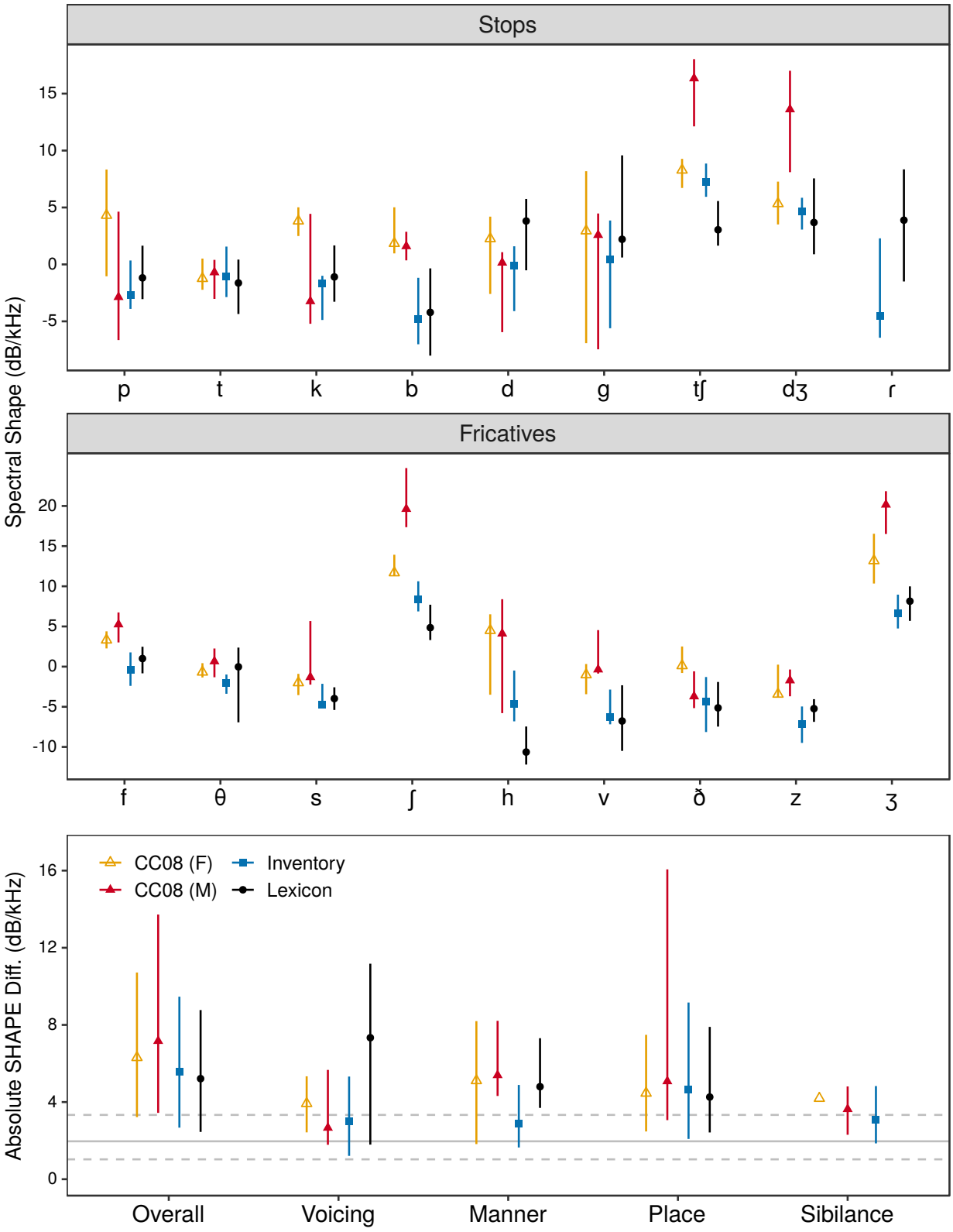

Figure 2.52: Spectral Shape (SHAPE) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in SHAPE in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

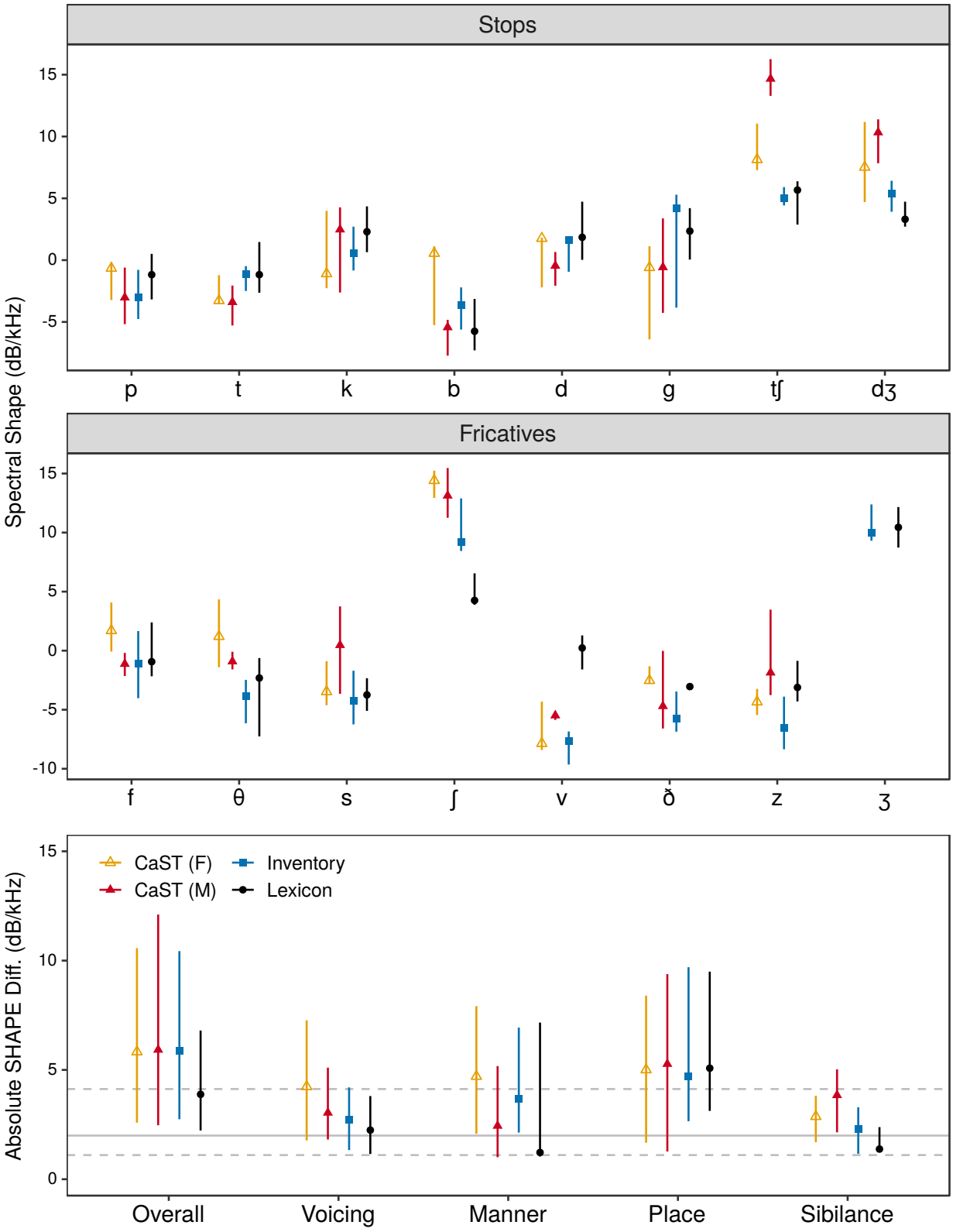

Figure 2.53: Spectral Shape (SHAPE) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in SHAPE in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

of low-frequency energy that provide a useful physiologically motivated alternative to the more general statistical approaches of spectral moment decomposition and the discrete cosine transform (DCT), both of which are further sensitive to distributional assumptions that are rarely met by consonant noise spectra. The primary question that remains for future work to address is how thresholds such as the 2500 Hz boundary between low- and high-frequency regions should be determined, and to what extent are such thresholds generalizable across a range of consonants and segmental/prosodic contexts, all while retaining physiological interpretability.

2.6.6 Spectral Dispersion ($\text{DISP}_{\text{C/VC/CV}}$)

2.6.6.1 Background and physiological basis

Early work on the spectral characteristics of fricative and stop articulations noted a distinction between relatively more *dispersed* spectra, and spectra where energy is concentrated in a narrower frequency band (Stevens & Blumstein, 1978). Initial work by Stevens and Blumstein categorized stop place of articulation by visually fitting different templates meant to capture broad differences in spectral shape. Later work utilized statistical measures of dispersion in the *spectral variance* measure of Forrest et al. (1988), Jongman et al. (2000), and others, but the present work adopts the Wiener entropy, or spectral *flatness* measure of (Gray & Markel, 1974), which has more stable acoustic properties and broadly reflects differences in the concentration of energy in the noise spectrum. Nonsibilant fricatives are expected to yield the highest dispersion values, while sibilants and labial/velar plosives are expected to exhibit much more concentrated spectral energy distributions and therefore lower dispersion values. Thus, physiologically, spectral dispersion reflects differences in both the nature of the noise source (nonsibilant versus sibilant) and the place of articulation of the consonant, as different points of constriction yield different excitation profiles that impact the distribution of energy in the spectrum.

2.6.6.2 Definition and measurement

Spectral dispersion is defined as the Wiener entropy, also known as the *spectral flatness*, of the spectrum between 550 and 10,000 Hz, which is measured by taking the ratio between the geometric mean and the arithmetic mean of the spectrum over the above interval (Gray & Markel, 1974; Niyogi & Sondhi, 2002; Agus et al., 2018). That is, spectral dispersion is calculated via the following equation:

$$DISP = \frac{\exp\left(\frac{1}{N}\sum_{n=0}^{N-1}\ln x(n)\right)}{\frac{1}{N}\sum_{n=0}^{N-1}x(n)},$$

where n is the bin number, N is the number of bins in the spectrum, and $x(n)$ is the amplitude of the bin, measured in Pascals. The output of this equation yields spectral dispersion values within the [0, 1] range, where a maximally flat spectrum, equivalent to the uniform distribution and similar to the spectrum of white noise, has a value of 1, and less dispersed spectra with greater energy concentration in a narrow range of frequencies approach DISP values of 0. Dispersion values in the [0, 1] range are then log-transformed to provide greater separation at the lower end of the range where most speech spectra are concentrated.

Spectral dispersion is measured for both the consonant noise spectrum (DISP_C) and the VC/CV transitions ($\text{DISP}_\text{VC/CV}$). The former is computed from the half/full Hamming windows on the consonant noise interval (as shown in Figure 2.22), and the latter is measured from 20 ms full Hamming windows centered at vowel onset/offset, where the greatest weight is placed on the CV/VC boundary, with window weights tapering off to zero as they move away from the boundary into the vowel and into the consonant. These consonant-vowel transition windows are used in order to mirror the windows adopted in Jongman et al. (2000), where spectral variance was particularly discriminative of English fricatives according to place of articulation.

2.6.6.3 Category and contrast distributions

Below we present spectral dispersion distributions separately for CV, VCV, and VC contrasts, where results for both consonant and transition windows are shown for a given contrast position,

similar to the organization of spectral tilt results in Section 2.6.4. That is, $DISP_C$ and $DISP_{CV}$ are presented for word-initial contrasts; $DISP_C$, $DISP_{VC}$, and $DISP_{CV}$ for word-medial contrasts; and $DISP_C$ and $DISP_{VC}$ for word-final contrasts.

Word-initial position (CV). Figure 2.54 shows spectral dispersion distributions for obstruent categories and contrasts in word-initial position in the lexicon, inventory, and reference data. From the bottom panel of Figure 2.54 we see that in the lexicon, spectral dispersion is most robust as a cue to sibilance and manner contrasts, while across databases sibilance and place are the most robust effects. The sibilance effect derives from the generally greater spectral dispersion in non-sibilant fricatives than in sibilants, though this effect is primarily restricted to voiceless fricatives. Regarding manner, most homorganic stop-fricative pairs exhibit similar spectral dispersion values, the most notable exception being [p, f], where the [p] is less dispersed than [f] by approximately 8 dB. Similar differences can be seen for [t, θ] and [b, v], though to a much lesser degree. This pattern reflects in part relative differences in the contribution of different noise sources in plosives and fricatives, where a greater proportion of the noise interval is occupied by aspiration noise in the former than in the latter, with aspiration exhibiting greater spectral tilt and thereby less dispersion.

At the CV transition, the primary distinction that emerges in Figure 2.55 is between voiced and voiceless obstruents, the latter being typically less dispersed, particularly among the stop consonants. This pattern reflects the greater spread of high-frequency noise into the vowel following voiceless obstruents than following voiced obstruents, as this noise dampens the energy in upper formants, causing the spectrum to exhibit relatively greater low-frequency energy and therefore a lower dispersion index. Figure 2.55 also shows minor effects of place and manner, where the former is due to relatively greater dispersion in transitions following more posterior places of articulation, and the latter reflects lower dispersion at fricative-vowel transitions than at plosive-vowel transitions. Both results are consistent with the voicing pattern described above; namely, greater excitation of upper formants at the transition yields more dispersed spectra, which occurs both for more posterior obstruents relative to anterior ones, and for aspiration noise relative to frication.

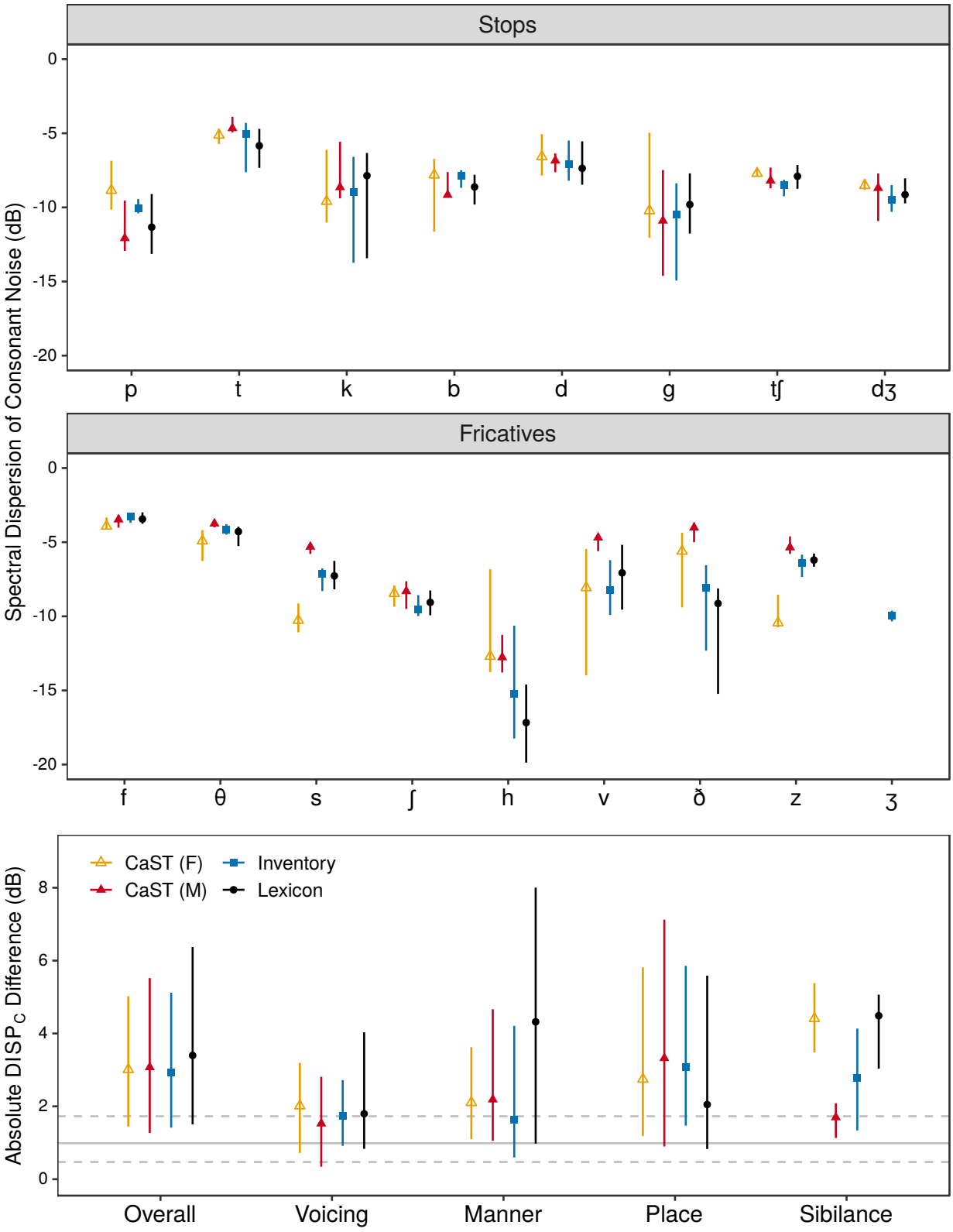

Figure 2.54: Spectral Dispersion of Consonant Noise ($DISP_C$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $DISP_C$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

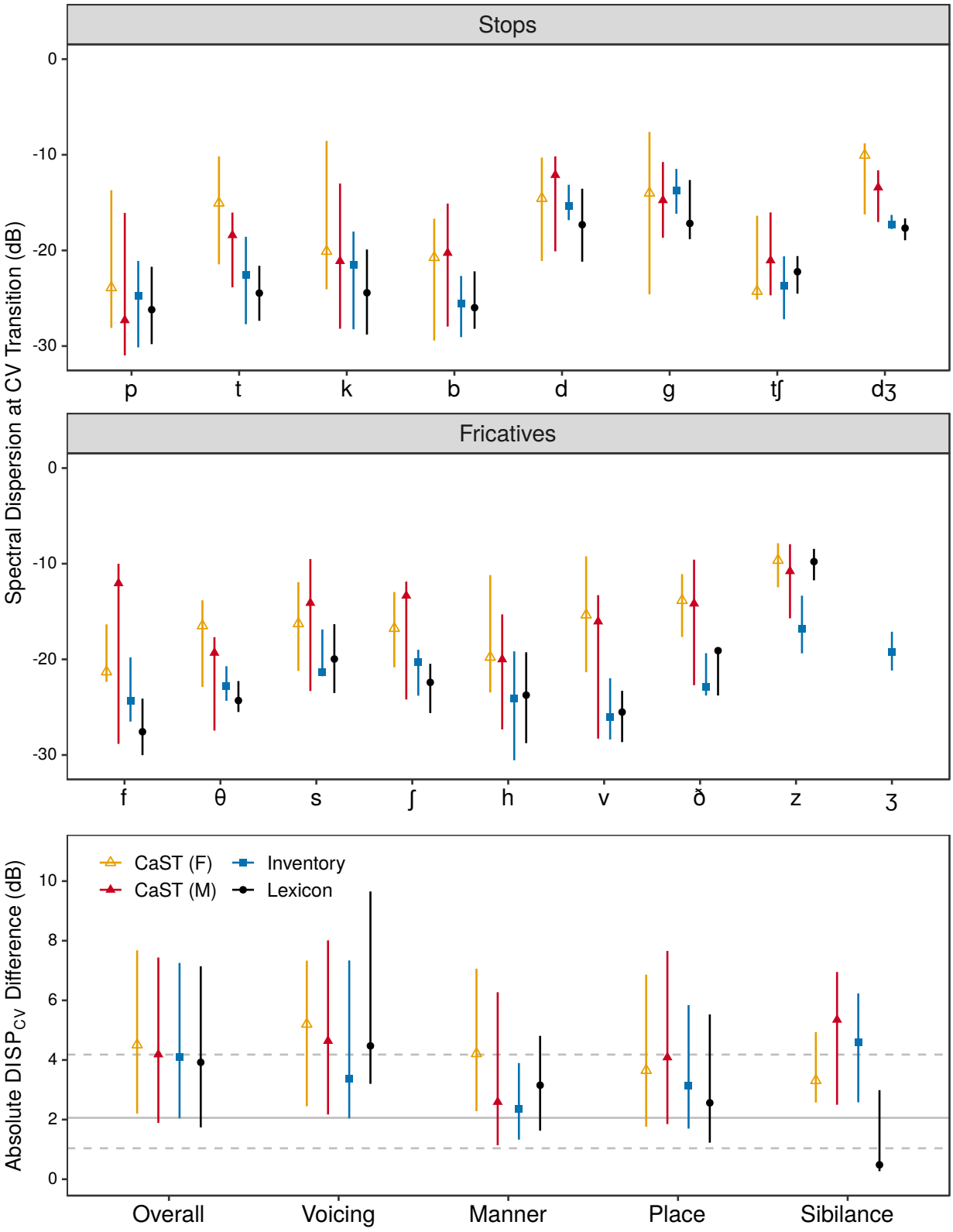

Figure 2.55: Spectral Dispersion at CV Transition ($DISP_{CV}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $DISP_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

Word-medial position (VCV). Dispersion patterns for the noise spectra of intervocalic obstruents are similar to those in word-initial position, though with greater overall effects of voicing. From Figure 2.56 we see that for all three plosive places of articulation, and for nonsibilant fricatives, the noise spectra of voiceless obstruents exhibit greater spectral dispersion than their voiced counterparts. This notable change in the size of the voicing contrast effect in VCV position relative to CV position reflects the greater high-frequency energy in voiceless noise spectra than in voiced spectra (see Figure 2.46 for spectral tilt values consistent with this result), a difference which is reduced word-initially due to the gradient but consistent process of obstruent devoicing at word boundaries. Therefore, from this result we expect voicing contrast effects for spectral dispersion to be similarly reduced word-finally (see Figure 2.49 for spectral dispersion results in VC position, results which are partially consistent with this prediction).

The other major featural contrast effect for consonant spectral dispersion intervocalically is that corresponding to distinctions in manner of articulation, where just as in CV position, the largest difference observed is that between [p] and [f], the plosive being approximately 8 dB lower in spectral dispersion than the fricative. Similar patterns are present among the other nonsibilants, though to a considerably lesser degree. Finally, the remaining effects of place and sibilance are consistent with those in CV position, where for place the *labial* $<$ *velar* $<$ *alveolar* relation is obtained for voiceless plosives, and the *postalveolar* $<$ *alveolar* relation is obtained among sibilant fricatives, while for sibilance, voiceless nonsibilant spectra are notably more dispersed than voiceless sibilants. Overall, consonant spectral dispersion differences ($\Delta\text{DISP}_\text{C}$) are robust between intervocalic obstruent contrasts, particularly in the target data where they average between 5 and 6 dB and range up to 10 dB for the 75th percentile in the lexicon.

Results are similar in size for spectral dispersion at the VC transition, though the within-item variance of transition spectra is larger than for noise spectra. Further, DISP_VC exhibits different featural patterns than DISP_C, where the most robust effect is due to place of articulation, followed by voicing and sibilance, with manner effects the smallest at around chance levels in all four data sets. The place effect derives primarily from a distinction between labial and non-labial plosives,

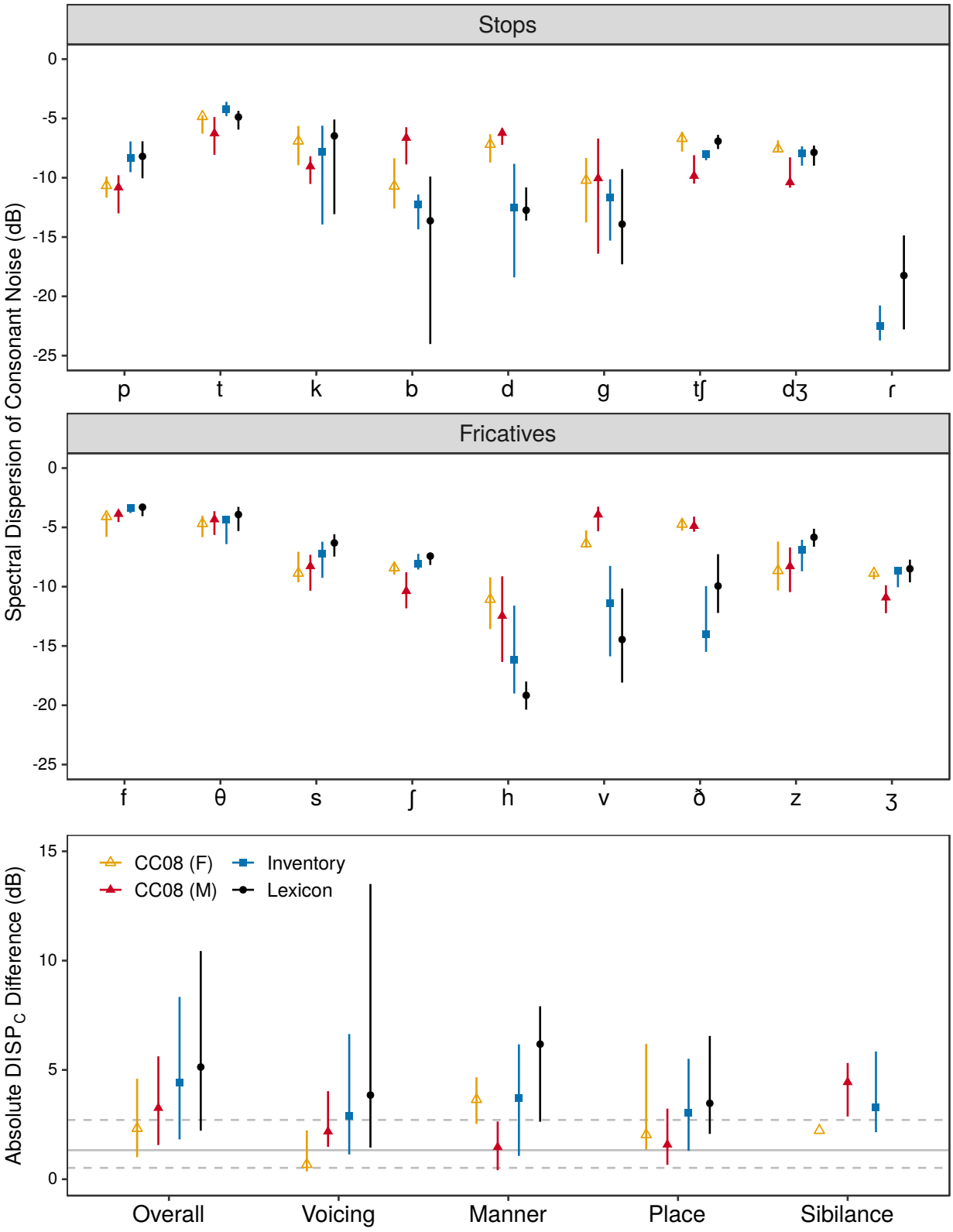

Figure 2.56: Spectral Dispersion of Consonant Noise ($DISP_C$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $DISP_C$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

the former being less dispersed at the VC transition than the latter due to the greater amplification of low-frequency resonances by labial constrictions (the fricatives [f, v] are similar in this regard). Similarly, transitions into sibilant fricatives are notably more dispersed than transitions into their nonsibilant counterparts, a pattern which is in direct opposition to the sibilance effect in consonant noise spectra where nonsibilants are more dispersed than sibilants. Finally, regarding voicing the major difference between voiced and voiceless obstruents in Figure 2.57 occurs among sibilants, where [z, ʒ, ʤ] are consistently more dispersed by around 10 dB than their voiceless counterparts. All of the above results are most robust in the lexicon, though all four data sets show notable contrast effects for spectral dispersion at the VC transition.

Spectral dispersion distributions at the CV transition for VCV contrasts are shown in Figure 2.58, and generally exhibit the same patterns as in CV position, though with a notably more robust place distinction and a reduced voicing effect. That is, as with word-initial contrasts, intervocalic obstruents exhibit greater dispersion at the CV transition with more posterior articulations within a given manner class, the only exceptions being the fricatives [ʃ, ʒ, h], where the postalveolars are less dispersed than their alveolar counterparts due to their greater concentration of energy in the mid-frequency region, and the glottal fricative exhibits among the lowest dispersion values given its great concentration of energy at lower frequencies and its notably damping of higher formant frequencies. The voicing effect in Figure 2.58 is consistent with the VC transition results in deriving primarily from distinctions among sibilants, as nonsibilant fricatives and plosives are much more consistent between the voiced and voiceless series in terms of spectral dispersion at the consonant-vowel transition.

These results indicate that part of the distinction between voiced and voiceless obstruents in CV position reflects prosodic differences in the phasing of consonant, vowel, and laryngeal gestures, where consonant-vowel transitions are more distinct word-initially than word-medially. This is because the greater difference in noise duration word-initially (see Figures 2.11 and 2.12) means there is a correspondingly greater difference between voiced and voiceless plosives in terms of the acoustic onset of the vowel (based largely on the appearance of F2 in the spectrogram) relative to

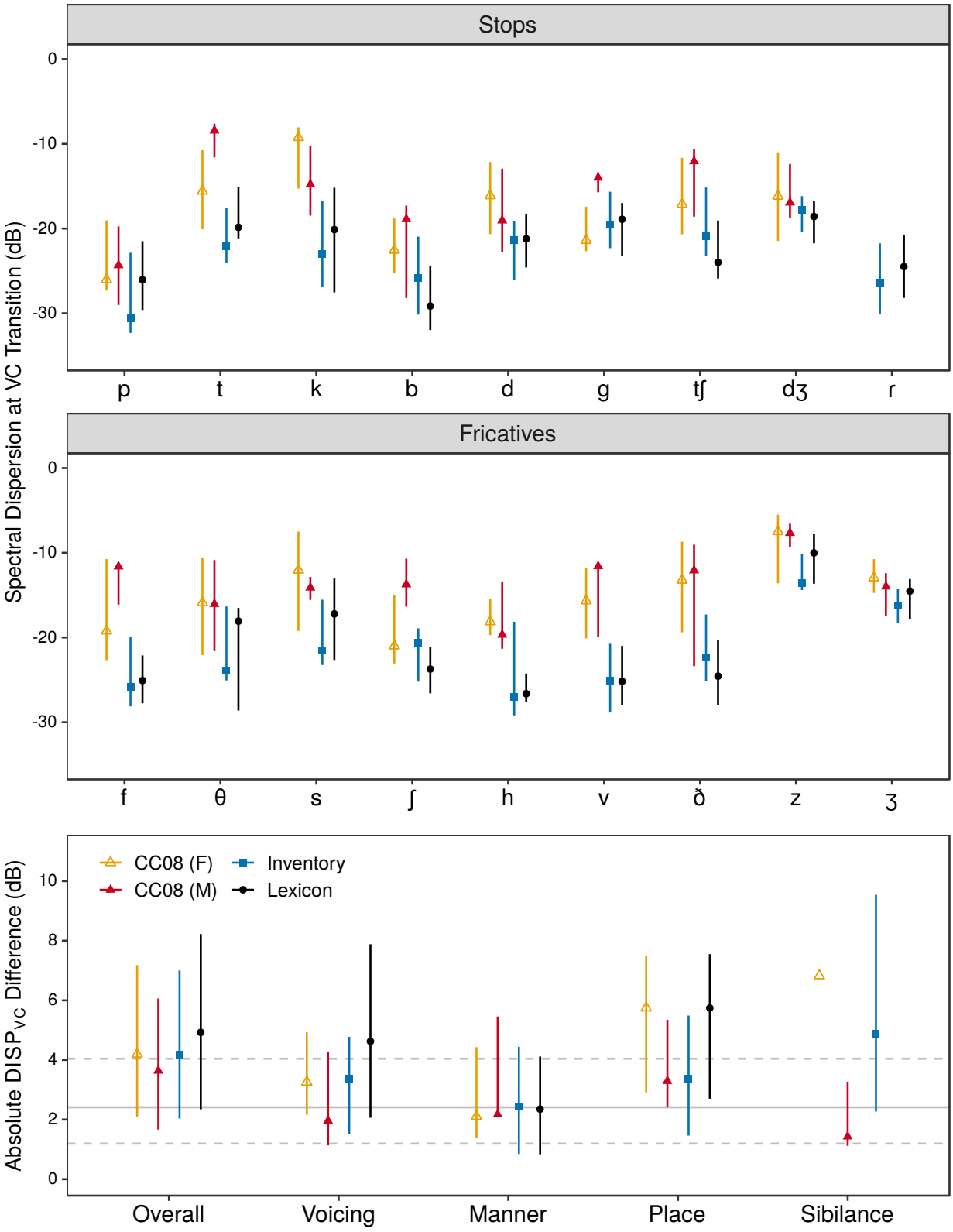

Figure 2.57: Spectral Dispersion at VC Transition ($DISP_{VC}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $DISP_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

the articulatory transition from consonant to vowel (see also Figures 2.79 and 2.83 for differences in F2 onset as a function of voicing in CV versus VCV position). As a result of the greater similarity between voiced and voiceless CV transitions word-medially relative to word-initially, their spectral characteristics, including DISP_{CV}, should exhibit greater similarity and thus a lesser voicing contrast effect in VCV position than in CV position.

Word-final position (VC). Figure 2.59 shows spectral dispersion distributions from consonant noise spectra in word-final obstruent contrasts. Among fricatives, the general patterns word-finally are consistent with those in CV and VCV positions; namely, nonsibilants are more dispersed than sibilants, and alveolars are more dispersed than postalveolars. The one effect missing from VC position is the voicing distinction among nonsibilant fricatives, where contrary to the *voiced* > *voiceless* relation in CV and VCV positions, voiced and voiceless nonsibilants are comparable in spectral dispersion. The one exception to this trend is the low spectral dispersion values for [v, ð] in the inventory data, a result which is due to the frequent approximant-like production of voiced nonsibilants by the target speaker in controlled syllable stimuli. As we can see from the lexicon distributions in Figure 2.59, this outlier behavior does not extend to the same speaker's pronunciation of real words.

Among stops, the most notable effects are the effect of place of articulation among voiceless plosives ([p] < [k] < [t]) and the effect of voicing among plosives, where with the exception of velars, voiceless plosives are more dispersed than their voiced counterparts by approximately 5 dB. This voicing effect is much reduced among affricates, which pattern more closely with their fricative counterparts in exhibiting minimal differences between voiced and voiceless. Finally, examining stops and fricatives for manner distinctions reveals a similar result to that in CV and VCV positions; namely, nonsibilant fricatives tend to be more dispersed than their plosive counterparts. However, unlike in other positions, the word-final manner effect in the lexicon is more consistent across nonsibilants; i.e., [p] < [f], [t] < [θ], [b] < [v], [d] < [ð]. Yet because distinctions between sibilant affricates and fricatives are much narrower word-finally, the overall contrast effect

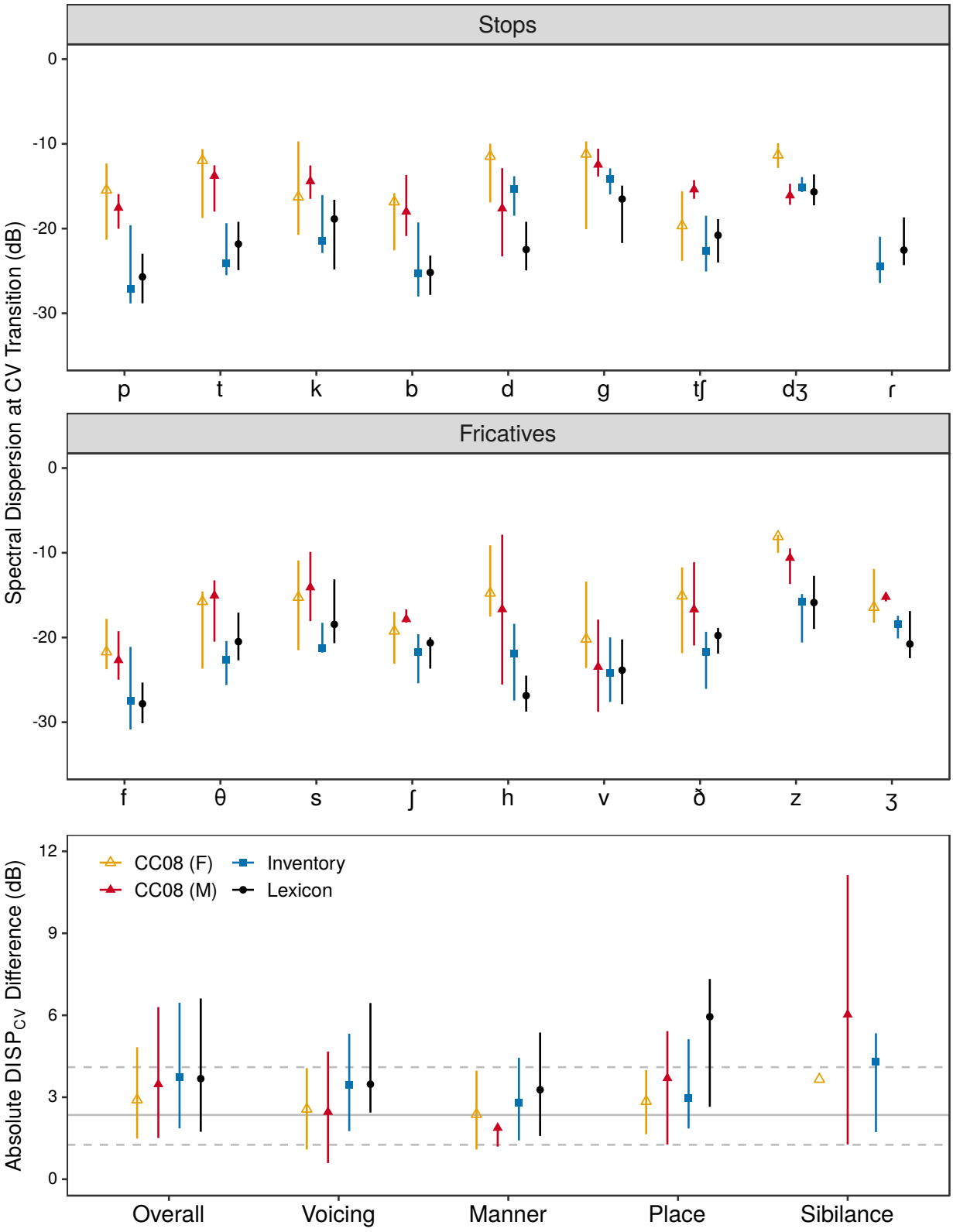

Figure 2.58: Spectral Dispersion at CV Transition ($DISP_{CV}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $DISP_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

for manner of articulation is reduced in VC position relative to CV and VCV positions.

Considering next the spectral dispersion of VC transitions, the discriminability of such transitions is much less robust word-finally than word-medially, with the largest effect occurring for manner contrasts in the lexicon. This result largely derives from the lower dispersion of labial plosives relative to their fricative counterparts [f, v], though similar but reduced effects are present for the contrasts [d, ð] and [ʤ, ʒ]. The more consistent featural effect word-finally, though reduced in overall size relative to obstruent manner, is place of articulation. Figure 2.60 shows that with the exception of post-alveolars, there are gradual increases in spectral dispersion with more posterior articulations in both stops and fricatives, just as in VCV position, a result that again reflects the fact that larger resonating cavities exhibit greater excitation of higher formants, yielding a greater spread of energy across the frequency range and thus larger spectral dispersion values. Finally, though voicing and sibilance distinctions among VC transitions are present word-finally—*nonsibilants* < *sibilants*, *voiceless* < *voiced* among fricatives and affricates—these effects are relatively minor and show near-complete overlap with estimated chance distinctions based on within-item DISP_{VC} variance in the inventory data.

2.6.6.4 Summary

The spectral dispersion of both the consonant noise interval and the CV/VC transitions is broadly discriminative of obstruent contrasts along an array of featural dimensions, particularly place and sibilance for DISP_{C}, place for DISP_{VC}, and place, manner and voicing for DISP_{CV}. When applied to the noise spectra of sibilant and nonsibilant fricatives, spectral dispersion reliably captures the distinction between diffuse spectra with little structure, as in [f, θ], and more complex spectra with clear regions of energy concentration, as in [s, ʃ] (Stevens & Blumstein, 1978; Shadle & Mair, 1996; Jongman et al., 2000). However, at the VC and CV transitions, given that all spectra exhibit notable low-frequency energy from the vowel offset, spectral dispersion primarily reflects the degree of high-frequency energy in the spectrum, either from the excitation of upper formants from more posterior consonant constrictions or the spread of high-frequency noise into the vowel

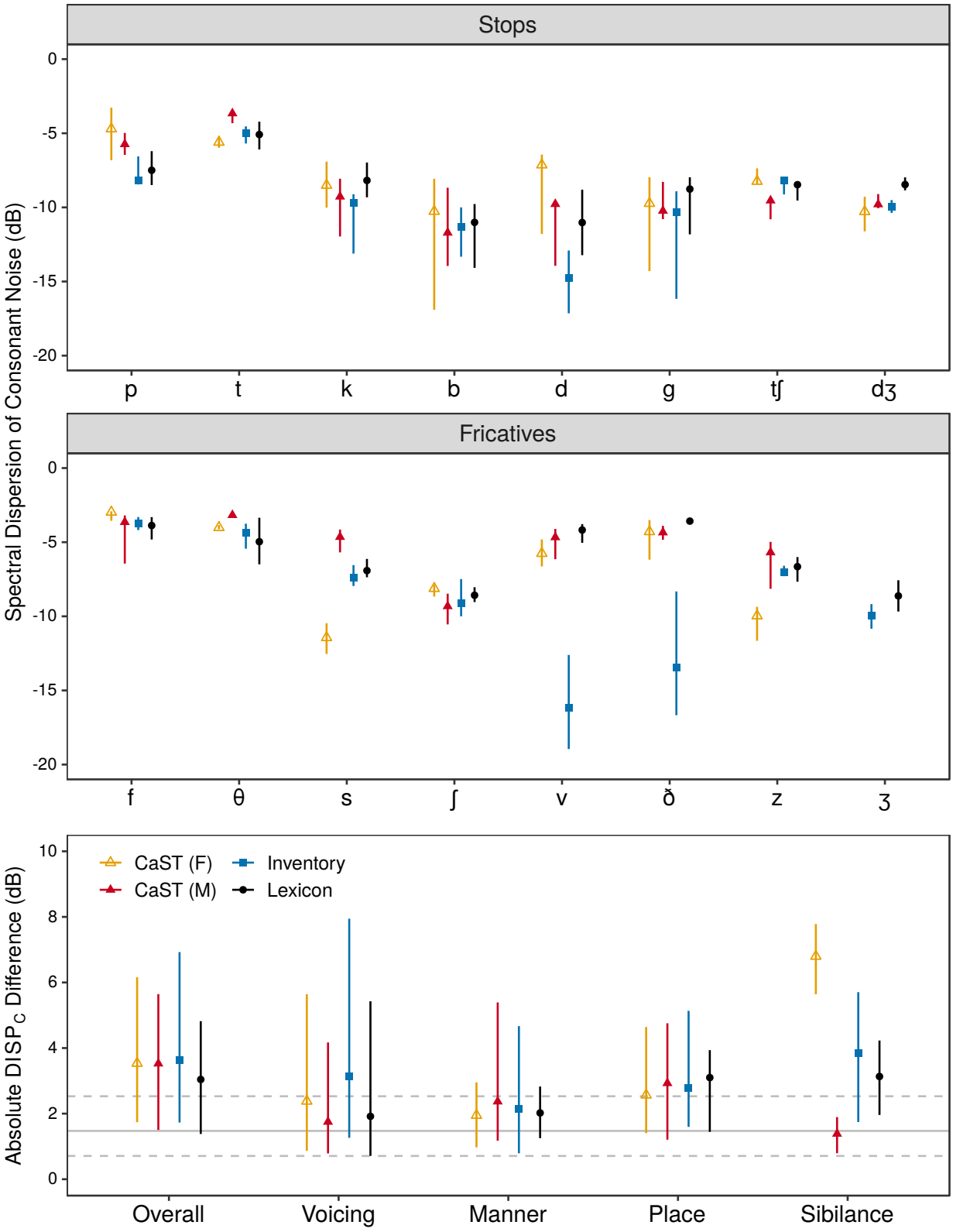

Figure 2.59: Spectral Dispersion of Consonant Noise ($DISP_C$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $DISP_C$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

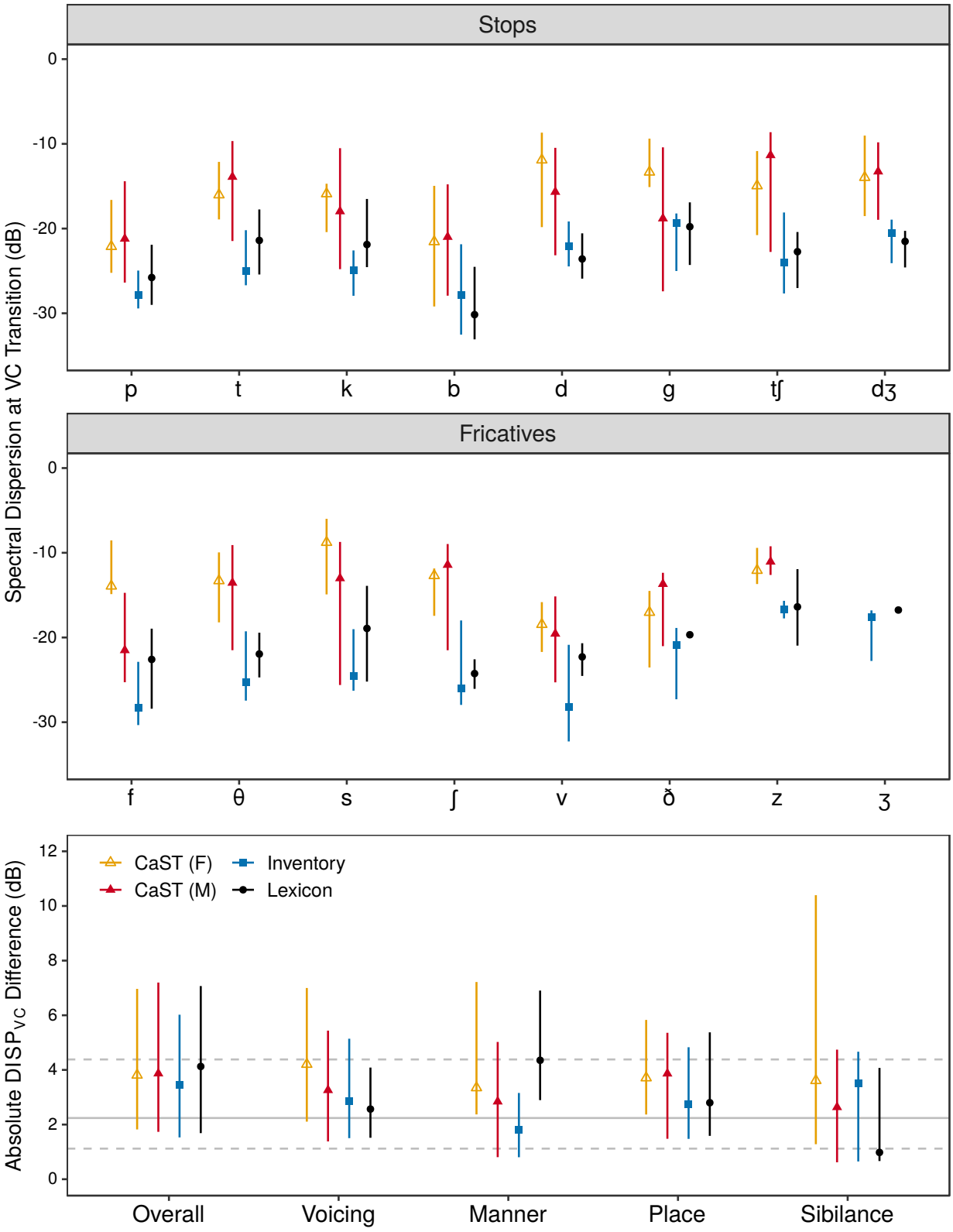

Figure 2.60: Spectral Dispersion at VC Transition ($DISP_{VC}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $DISP_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

from adjacent sibilants. Voicing effects are similar though less consistent across positions, with generally lower dispersion values at voiceless obstruent transitions than at voiced transitions due to the greater damping of upper formant amplitudes in the former.

2.6.7 Low Frequency Energy (LF)

2.6.7.1 Background and physiological basis

The amplitude of low-frequency energy in the spectrum was initially studied in Hughes & Halle (1956), and notably later used in the discrimination of a large set of English fricatives in McMurray & Jongman (2011), and simply reflects the relative strength of low-frequency components of the acoustic signal. This measure has primarily been used to distinguish voiced from voiceless fricatives, as voicing generates a high-amplitude signal in the low-frequency region, but this measure is also theoretically linked to differences in noise source—aspiration yielding higher concentrations of low-frequency energy than supralaryngeal frication. For this reason, LF may be adopted as an index of physiological and acoustic differences in both voicing and manner of articulation.

2.6.7.2 Definition and measurement

Low frequency energy is defined as the mean amplitude in the spectrum below 550 Hz, and is computed by taking the average of the amplitudes of frequency components between 0 and 550 Hz; i.e., $LF = \frac{1}{M}\sum_{n=0}^{M-1} x(n)$, where n is the bin number, M is the number of bins in the spectrum between 0 and 550 Hz, and $x(n)$ is the amplitude of the bin, measured in decibels. That is, complementary to the filter characteristics examined above in measures such as peak frequency, peak amplitude, and spectral tilt, among others, which examined the noise spectrum above 550 Hz, the measurement of mean energy below 550 Hz is intended to capture characteristics of the laryngeal source, where voiced obstruents are expected to exhibit higher low-frequency energy than voiceless obstruents.

2.6.7.3 Category and contrast distributions

Below we review low-frequency energy distributions by category and contrast in the lexicon, inventory, and reference data. Results are presented separately for word-initial, word-medial, and word-final contrasts.

Word-initial position (CV). Figure 2.61 shows distributions of low-frequency energy among word-initial obstruent contrasts. Here we see that voiced and voiceless stops are relatively similar in LF energy, with only [p, b] exhibiting a notable distinction of around 5 dB. This result reflects in part the fact that the word-initial stop voicing distinction in English is not a true voicing distinction, with the majority of phonologically voiced stops occurring without any prevoicing or voicing during the noise interval. Voiced fricatives, on the other hand, are generally fully voiced word-initially (see Figure 2.18 for near-ceiling voicing percentages among voiced fricatives in CV position), and therefore the difference in low-frequency energy between voiced and voiceless fricatives is substantia—on the order of 10–15 dB—with the glottal fricative [h] exhibiting LF values intermediate between the two. These results, though restricted to fricatives, lead to a robust overall voicing effect in all four data sets.

The other major distinction indexed by the mean energy below 550 Hz is for manner of articulation, where the aspiration noise comprising the majority of the noise interval in plosives leads to LF values between 5 and 10 dB higher than in affricates, and up to a 15 dB difference between voiceless plosives and fricatives. These effects are most pronounced in the lexicon, though the inventory and reference data also exhibit manner contrast effects well above chance levels. Finally, obstruent sibilance and place of articulation are minimal in their impact on low-frequency energy, lending further support to the viability of the 550 Hz threshold as a general cutoff between source and filter components in obstruent noise spectra.

Word-medial position (VCV). Intervocalically, the voicing effect remains robust but the patterns in LF energy distinctions between voiced and voiceless obstruents differ notably from those

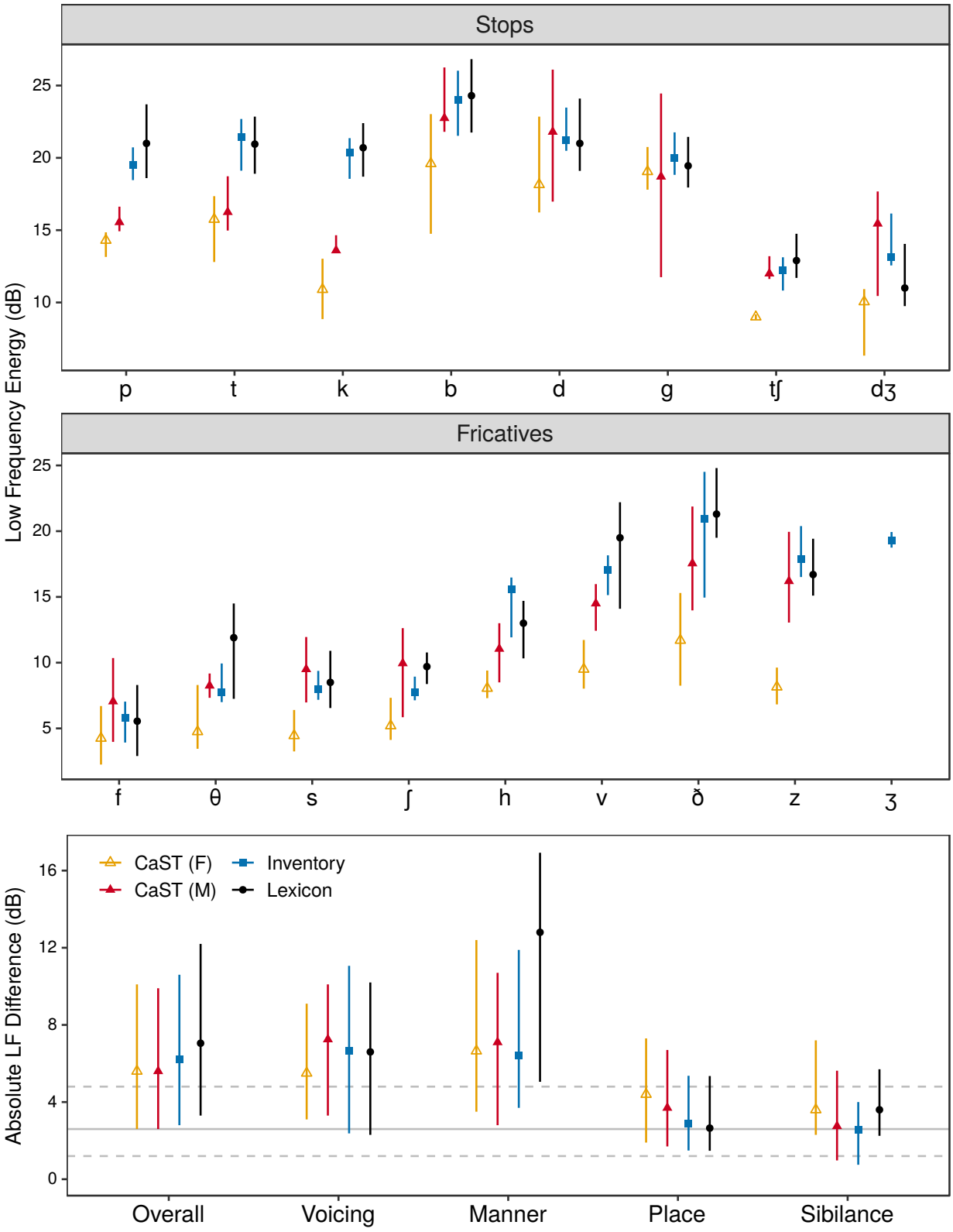

Figure 2.61: Low Frequency Energy (LF) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in LF in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

in CV position. Figure 2.62 shows that while the voicing distinction narrows among fricatives (though there remains a sizeable 5–10 dB difference between voiced and voiceless), in VCV position with [b, d, g, ʤ] now truly voiced there is a consistent difference of around 5 dB between voiced and voiceless stops. These results are most robust in the target data, though the reference data shows consistent, albeit reduced, effects of voicing on LF energy.

The manner effect observed in word-initial position is retained intervocalically, as VCV plosives exhibit greater energy in the low-frequency region than affricates or their nonsibilant fricative counterparts (in the case of labials and [LOW] coronals). Further, the alveolar flap [ɾ] is distinct from all other intervocalic obstruents except [h] due to its approximant-like character and consequently greater LF energy. Because the flap does not occur in 'pure' manner contrasts with other obstruents it does not contribute to the manner effect in the bottom panel of Figure 2.62; nevertheless, we expect [ɾ]'s outlier LF values to contribute positively to the discriminability of multi-feature contrasts in the lexicon. Finally, as in CV position there is little-to-no effect of place or sibilance on LF energy distributions among intervocalic obstruent contrasts.

Word-final position (VC). At word/syllable-offset, low-frequency energy distributions are more variable. As in CV position, there is little difference between voiceless and voiced plosives, and while there is a fairly robust distinction among affricates in the controlled syllable data, in the lexicon this distinction disappears. Among fricatives there is much more overlap in low-frequency energy between the voiced and voiceless series, particularly among sibilants. These results reflect the much greater occurrence of word-final devoicing in the lexicon than in the inventory and reference data, where the greater voicing observed in the latter is likely a consequence of the hyperarticulation of controlled syllables relative to real words. Regarding manner, plosives are higher in low-frequency energy than affricates and fricatives, a result which again reflects differences in the spectral characteristics of aspiration noise relative to frication. Finally, as in CV and VCV positions, sibilance and place effects are minor, though there are consistent differences in the lexicon where among fricatives, sibilants tend to exhibit greater low-frequency energy than nonsibilants,

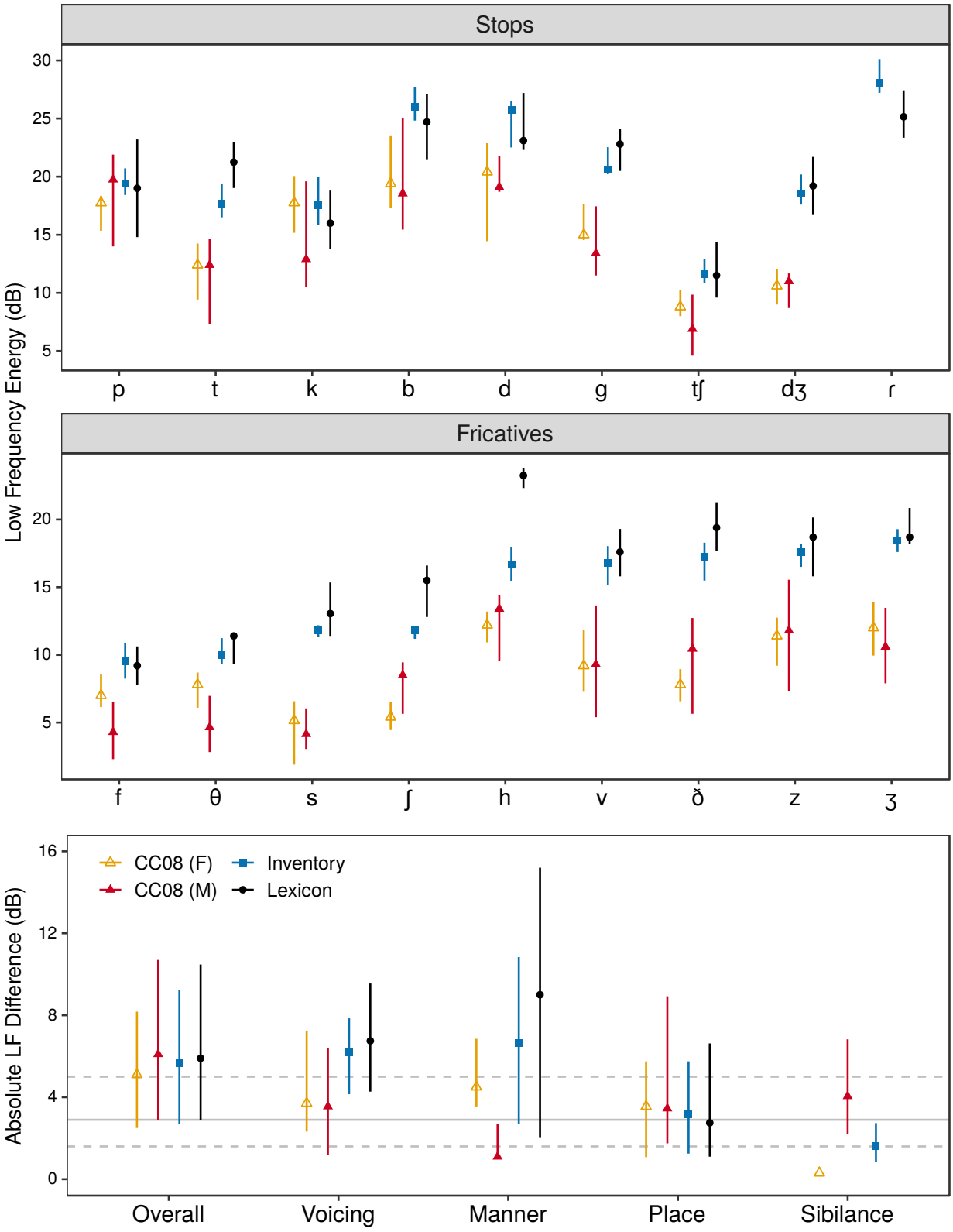

Figure 2.62: Low Frequency Energy (LF) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in LF in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

with the labials [f, v] further lower than their dental counterparts, and among plosives the *alveolar* < *labial/velar* relation holds across both voiced and voiceless sets.

2.6.7.4 Summary

Overall, low-frequency energy serves the anticipated role in distinguishing most voiced obstruents from their voiceless counterparts, where the only notable exceptions—namely in stops in CV and VC positions—occur due to predictable devoicing patterns and thus remain phonetically consistent despite LF providing less information about the phonological voicing contrast. An unanticipated result was the even more robust manner effect wherein voiceless plosives exhibit greater low-frequency energy than their fricative and affricate counterparts. This pattern is no-less physiologically motivated than the voicing effect, however, as aspiration yields a greater amplification of low-frequency components in the spectrum than frication. Finally, contrary to the robust voicing and manner effects reported above, the role of sibilance and place of articulation was relatively minor, suggesting the low-frequency component of the spectrum primarily reflects source characteristics that are largely independent of changes in the configuration of the vocal tract filter.

2.6.8 Relative Amplitude of F3 (AMP_{F3})

2.6.8.1 Background and physiological basis

The amplitude of the noise spectrum relative to that in the vowel in the region of the third formant frequency was initially proposed in Stevens (1985), and later confirmed perceptually in Hedrick & Ohde (1993) as an index of the place distinction between alveolar and postalveolar sibilants, as noise at the postalveolar constriction generally excites F3 and provides greater similarity in F3 amplitudes with the following vowel. Given that postalveolar obstruents are quite distinct in both the location and amplitude of noise excitation, relative F3 amplitude may also be used as a sibilance cue to distinctions between [ʃ, ʧ, ʒ, ʤ] and nonsibilant obstruents, though as such distinctions are ultimately confounded with place of articulation, we will consider AMP_{F3} as primarily a place cue.

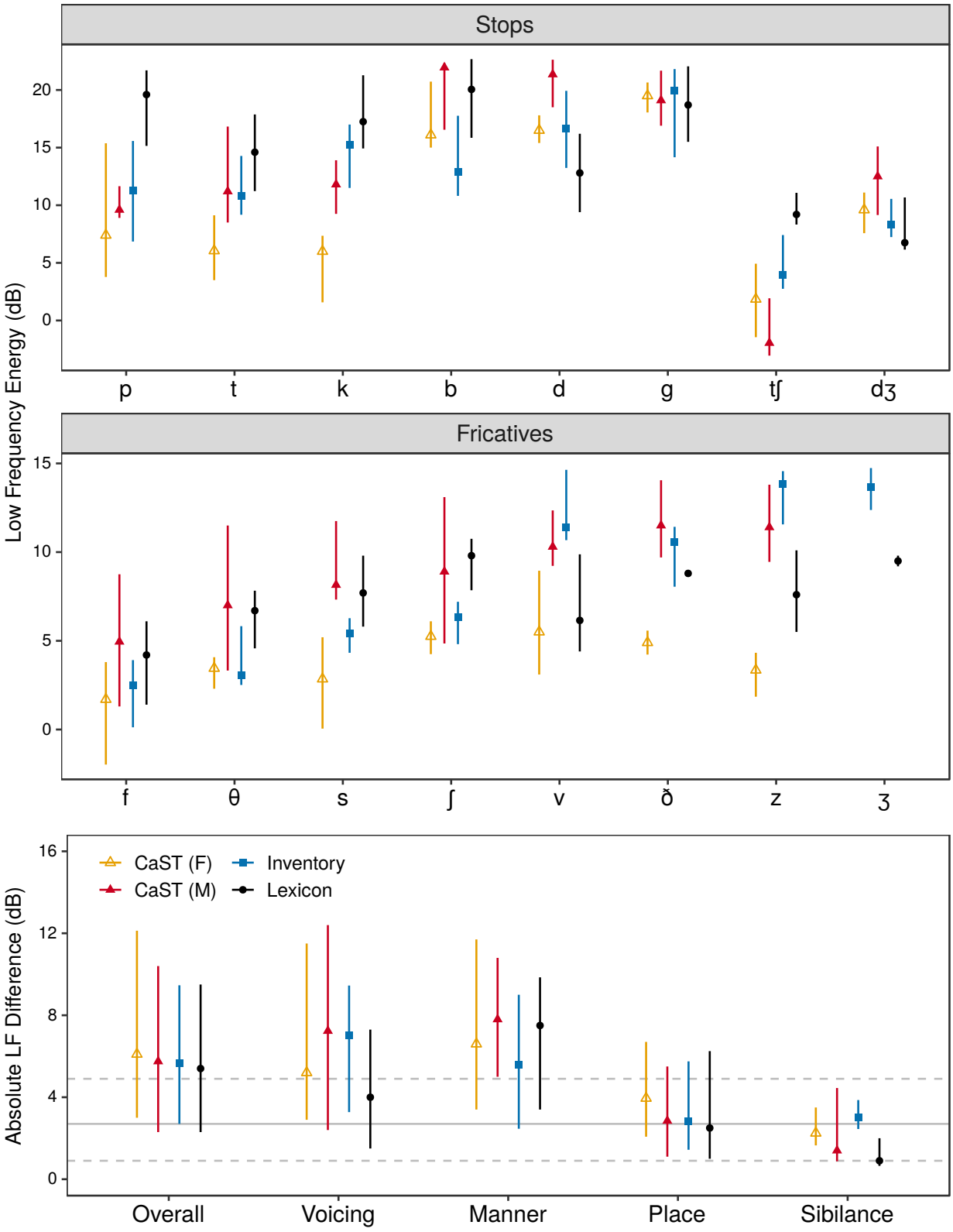

Figure 2.63: Low Frequency Energy (LF) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in LF in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

2.6.8.2 Definition and measurement

Relative F3 amplitude (AMP_{F3}) is defined as the difference between the amplitude of the noise spectrum and the amplitude of the vowel onset/offset spectrum in the frequency region centered on the third formant at the CV/VC boundary. AMP_{F3} is measured by first determining the frequency of the third formant at vowel onset, where vowel onset measurements are taken from the point 10% into the vowel. Formant frequencies were automatically estimated using the Burg algorithm in Praat (Boersma & Weenink, 2016), with estimation errors later hand-corrected via visual inspection of the spectrogram. The mean amplitudes in the F3 region of the consonant noise spectrum ($AMP_{F3[C]}$; taken from the same windows used for the spectral measures above and illustrated in Figure 2.22) and the vowel onset spectrum ($AMP_{F3[V]}$; a 20 ms half-Hamming window aligned at vowel onset and tapering off toward the vowel nucleus) were then measured, where the F3 region is defined as $F3_{CV/VC} \pm 50$ Hz, and $AMP_{F3} = AMP_{F3[C]} - AMP_{F3[V]}$. A fixed bandwidth of 100 Hz was used for simplicity and measurement stability given variation in bandwidth estimation (Fant, 1962; Klatt & Klatt, 1990). In CV and VCV contrasts, AMP_{F3} is measured relative to the following vowel, while in VC position the preceding vowel is used for the reference $AMP_{F3[V]}$ value. See Figure 2.64 for sample spectrograms showing approximate F3 regions and measurement windows over which AMP_{F3} is calculated.

2.6.8.3 Category and contrast distributions

Below we review AMP_{F3} distributions in the lexicon, inventory, and reference data, with results presented separately for word-initial, word-medial, and word-final contrasts.

Word-initial position (CV). Figure 2.65 shows relative F3 amplitude distributions among word-initial obstruent contrasts, and illustrates that AMP_{F3} primarily distinguishes obstruents according to place of articulation. The two most robust place effects which are largely consistent across databases are the [f, θ, s, v, ð, z] < [ʃ, ʒ] relation among fricatives, and the *labial* < *velar* < *alveolar* relation among plosives. These results are broadly consistent with the spectral peak fre-

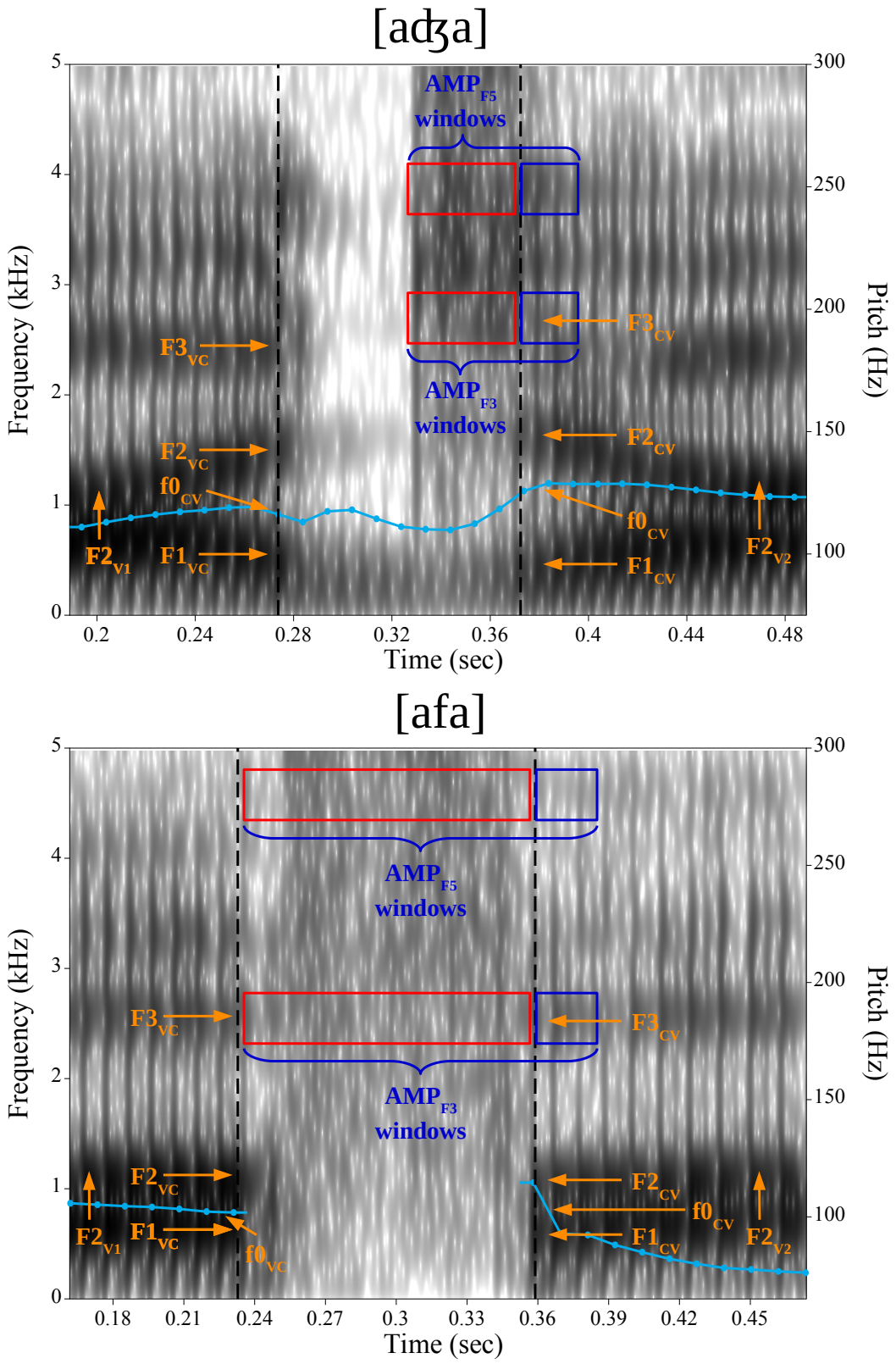

Figure 2.64: Sample measurement of spectral parameters (Set II): Relative F3 Amplitude (AMP_{F3}), Relative F5 Amplitude (AMP_{F5}), Fundamental Frequency ($f0_{VC/CV}$), First Formant Frequency ($F1_{VC/CV}$), Second Formant Frequency ($F2_{VC/CV/V1/V2}$), and Third Formant Frequency ($F3_{VC/CV}$).

quency results in Section 2.6.1 in that AMP_{F3} isolates the obstruents with energy concentrated in the mid-frequency region from those exhibiting greater energy at high (e.g., alveolar sibilants) or low (e.g., labials and dentals) frequencies. Figure 2.65 also shows a modest effect of manner of articulation wherein nonsibilant fricatives, particularly the voiceless nonsibilants [f, θ], tend to be lower in AMP_{F3} values than their plosive counterparts. Voicing effects are less robust, as relative F3 amplitude primarily reflects filter characteristics, as is the sibilance effect which depends on the [s, θ] and [z, ð] contrasts which are all low in AMP_{F3} due do the fact that the frequency regions excited by such constrictions lie well outside of the F3 range.

Word-medial position (VCV). Intervocalically, the place effect remains robust among fricatives, though AMP_{F3} distributions exhibit much greater overlap between plosive places of articulation. However, despite this narrowed scope of AMP_{F3}-based place distinctions in VCV position, the contrast effect for place of articulation is in fact larger than in CV position, at around 10 dB on average. Regarding manner of articulation, Figure 2.65 illustrates that nonsibilant fricatives remain lower in relative F3 amplitude than their plosive counterparts, and this effect is more consistent across voicing classes than in CV position. Conversely, postalveolar fricatives exhibit larger AMP_{F3} values than their affricate counterparts, a result that is consistent with the spectral peak amplitude results in Figure 2.39, wherein [ʃ] and [ʒ] have more defined mid-frequency peaks than [ʧ] and [ʤ], respectively. Finally, as in CV position, sibilance and voicing do not greatly impact the relative amplitude of F3, though regarding the former there is a somewhat greater distinction between [s, z] and [θ, ð] intervocalically.

Word-final position (VC). Figure 2.67 shows AMP_{F3} distributions among word-final obstruent contrasts. The results are broadly consistent with the patterns observed among CV and VCV contrasts, though as in VCV position the place distinction is primarily driven by contrasts between alveolar and postalveolar fricatives, where [s, z] $<$ [ʃ, ʒ]. Place contrasts among plosives are less consistent, a result which could reflect the fact that while fricatives are symmetric in the adjacency between noise and vowel offset/onset intervals, allowing AMP_{F3} to reflect transition characteris-

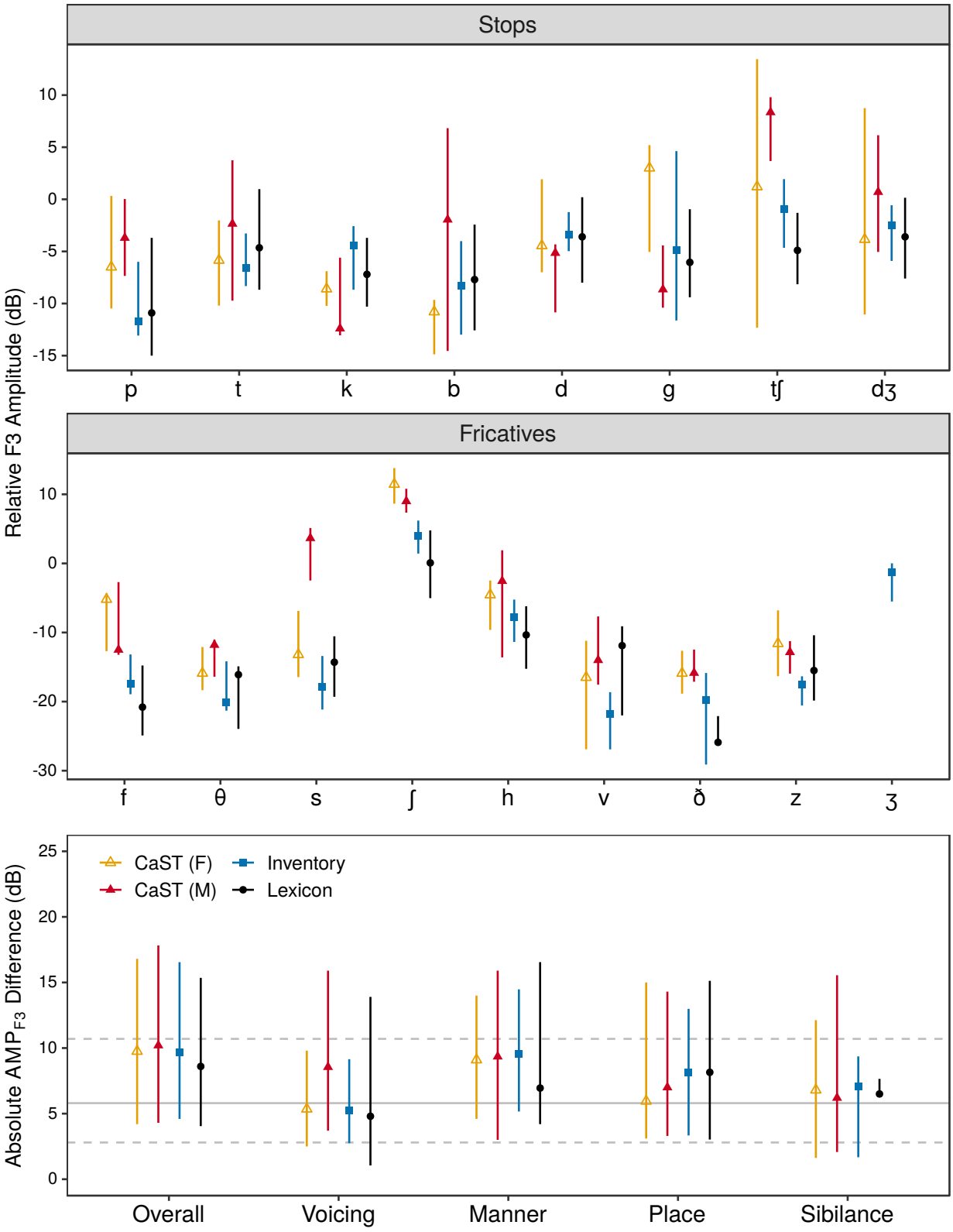

Figure 2.65: Relative F3 Amplitude (ΔAMP_{F3}) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in ΔAMP_{F3} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

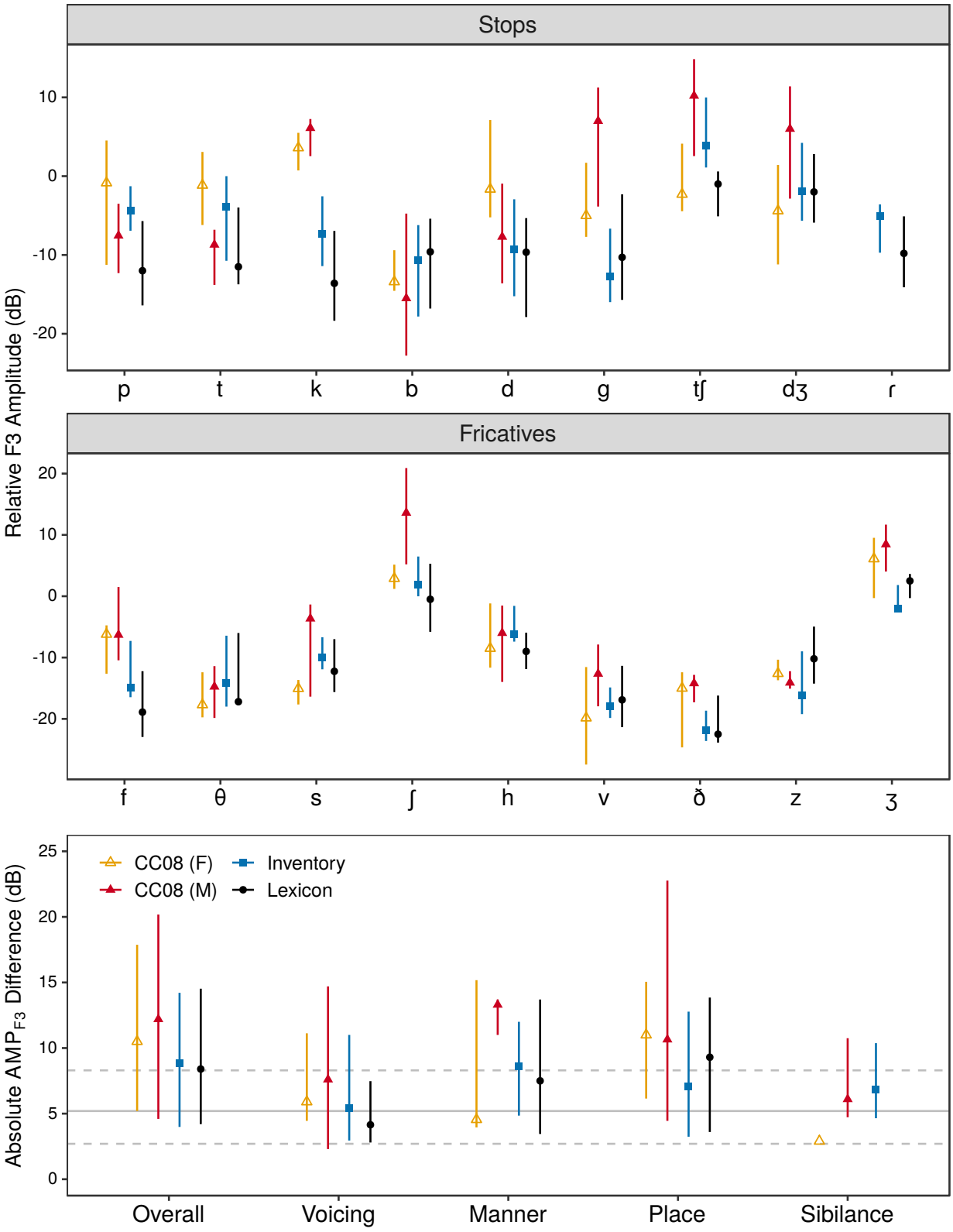

Figure 2.66: Relative F3 Amplitude (ΔAMP_{F3}) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in ΔAMP_{F3} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

tics between the consonant noise and the vowel, word-final stops exhibit a gap between vowel and noise intervals due to the consonant closure which disrupts such transitions. This distinction may also account for the sizeable manner effect word-finally, which as in CV and VCV positions corresponds to the *fricative* > *affricate* distinction among postalveolars, and the *fricative* < *plosive* distinction among nonsibilants. However, this word-final manner effect—on the order of 15–20 dB—is largely restricted to lexical contrasts. In the inventory and reference data, manner effects are almost completely contained within the estimated chance range based on within-item variation in relative F3 amplitude. Finally, AMP_{F3} is largely unaffected by the voicing or sibilance status of obstruents in VC position, a result which is consistent with the CV and VCV patterns above.

2.6.8.4 Summary

Overall, relative F3 amplitude provides a consistent cue to the place and manner of a subset of obstruent contrasts across word-initial, word-medial, and word-final positions. The information provided by AMP_{F3} is not unique, however, and reflects a combination of spectral characteristics such as spectral peak frequency/amplitude and noise amplitude that similarly index obstruent place and manner. Nevertheless, given that it is an open question whether listeners interpret spectral and amplitudinal characteristics of obstruent noise in absolute terms or relative to adjacent phones such as the preceding/following vowel, AMP_{F3} remains useful as a potential cue in the perceptual discrimination of obstruent contrasts.

2.6.9 Relative Amplitude of F5 (AMP_{F5})

2.6.9.1 Background and physiological basis

Stevens (1985) and Hedrick & Ohde (1993) also defined a parallel relative amplitude measure in the F5 region as an index of obstruent sibilance, aimed primarily at distinguishing [f] from [s], the latter being larger in AMP_{F5} than the former. As with AMP_{F3}, this measure reflects differences in noise excitation, but is more dependent on noise source differences—i.e., between obstacle and

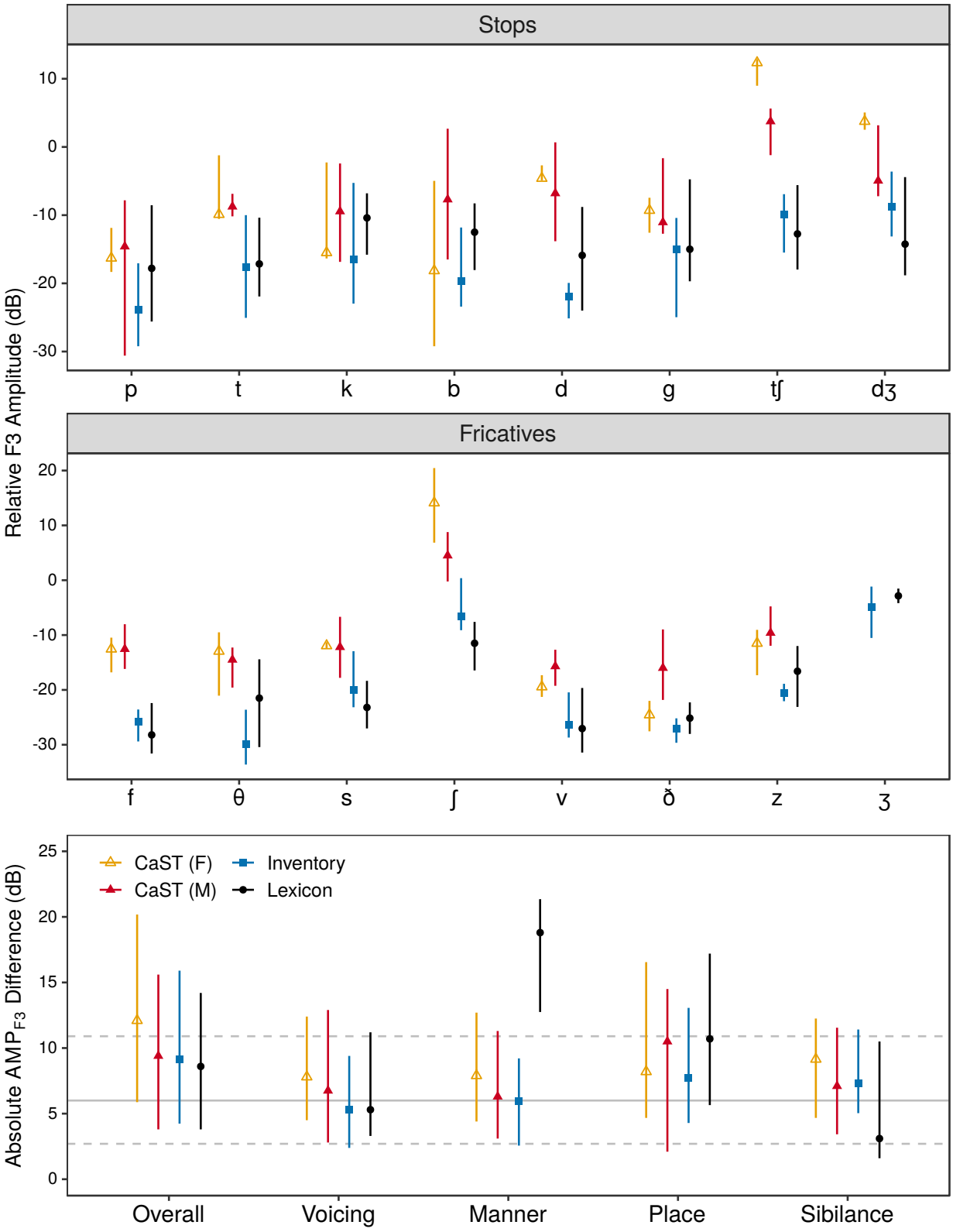

Figure 2.67: Relative F3 Amplitude (ΔAMP_{F3}) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in ΔAMP_{F3} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

non-obstacle turbulence—than place distinctions.

2.6.9.2 Definition and measurement

Relative F5 amplitude (AMP_{F5}) is defined similarly to AMP_{F3}, but with the critical frequency region centered on the fifth formant of the vowel rather than the third. Further, given the much greater measurement uncertainty for F5 at vowel onset (due to the longer spread of high-frequency noise into the vowel), as well as the theoretical expectation that F5 should be relatively constant over the vowel interval, F5 was measured at vowel midpoint rather than vowel onset/offset, though automatic tracking errors were still occasionally present and required hand correction from visual inspection of the spectrogram. Lastly, a 200 Hz bandwidth was used for F5 amplitude measurement based on previous literature and theoretical expectations of increasing bandwidth at higher formant frequencies (Fant, 1962; Klatt & Klatt, 1990). All other measurement procedures matched those of AMP_{F3} described in the previous section. See Figure 2.64 for sample spectrograms showing approximate F5 regions and measurement windows over which AMP_{F5} is calculated.

2.6.9.3 Category and contrast distributions

In the following sections, AMP_{F5} distributions are presented for word-initial, word-medial, and word-final obstruent contrasts in the lexicon, inventory, and reference data. As with AMP_{F3}, CV and VCV results reflect the amplitude of the noise in the F5 region relative to the following vowel, while the VC results reflect F5 amplitude differences between the noise spectrum and the spectrum generated at the offset of the preceding vowel.

Word-initial position (CV). Figure 2.68 shows relative F5 amplitude distributions among word-initial contrasts, and immediately illustrates the broader featural utility of AMP_{F5} relative to AMP_{F3}, where voicing is the only feature with minimal impact. Regarding manner, nonsibilant fricatives tend to exhibit lower relative F5 amplitudes than their plosive counterparts, particularly in the [LOW] coronal contrasts [t, θ] and [d, ð]. Among sibilants the converse relation is obtained;

namely, [ʃ] > [ʧ] and [ʒ] > [ʤ]. The place effect in Figure 2.68 is largely consistent with the spectral peak frequency distributions in Figure 2.35 in that among plosives, alveolars exhibit the highest relative F5 amplitudes, followed by velars, and then labials. This *labial* < *velar* < *alveolar* relation is broadly consistent across databases, though voiceless plosives show more variability in this regard than voiced plosives. Among fricatives, the primary distinction marked by AMP_{F5} is not place but rather sibilance, where sibilants are higher in relative F5 amplitude than their nonsibilant counterparts on the order of 20–25 dB. Overall, AMP_{F5} contrast differences (ΔAMP_{F5}) are robust across all four data sets and marginally more discriminative word-initially than relative F3 amplitude, though the patterns among fricatives are complementary and consistent with Hedrick & Ohde (1993) in showing that AMP_{F3} reliably distinguishes [s] from [ʃ], while AMP_{F5} distinguishes [s] from [θ], with parallel effects further found among voiced fricatives.

Word-medial position (VCV). Relative F5 amplitude is even more discriminative in VCV position than in CV position, as Figure 2.69 shows that in addition to the manner, place, and sibilance effects that are largely consistent with those among word-initial contrasts, there is a further effect of voicing in the target data, wherein across manner classes, voiceless obstruents exhibit higher relative F5 amplitudes than their voiced counterparts. Overall, because of this extension of relative F5 amplitude distinctions to voicing contrasts, ΔAMP_{F5} values are moderately greater in VCV position than in CV position, particularly in the lexicon where they average around 15 dB, well above the mean estimated chance level of 6 dB.

Word-final position (VC). Figure 2.70 shows AMP_{F5} distributions among word-final contrasts. As in CV and VCV positions, obstruent sibilance exerts the greatest effect on relative F5 amplitude given both sibilants' more consistent concentration of energy in higher frequencies, and the greater overall noise amplitude of sibilants relative to nonsibilants (see Figures 2.26–2.28 for details). The place effect is also present word-finally, though here it is largely restricted to voiceless plosives, which are both more variable in relative F5 amplitudes and less consistent in place of articulation patterns across databases. This result is similar to the AMP_{F3} result in Figure 2.67, where all place

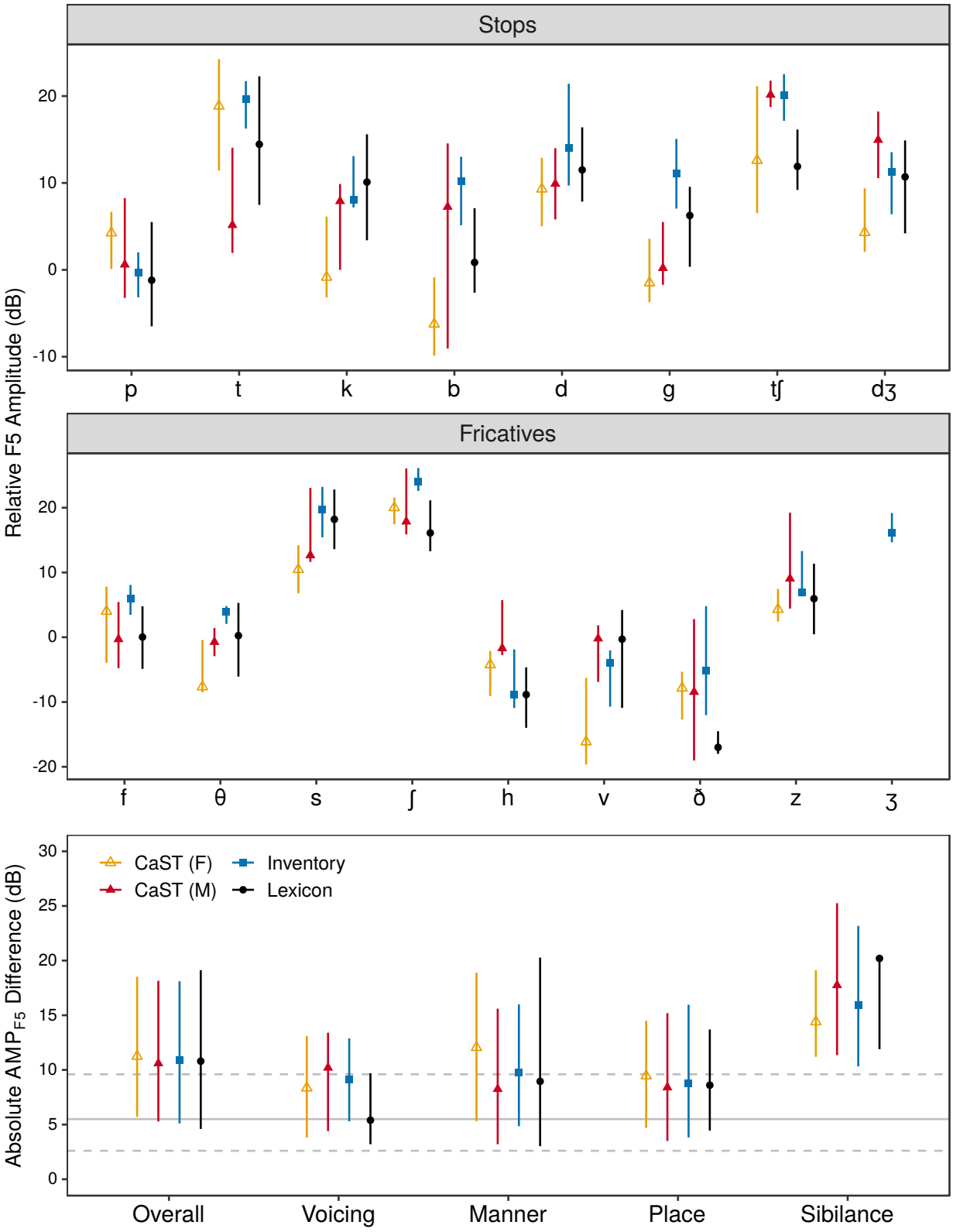

Figure 2.68: Relative F5 Amplitude (ΔAMP_{F5}) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in ΔAMP_{F5} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

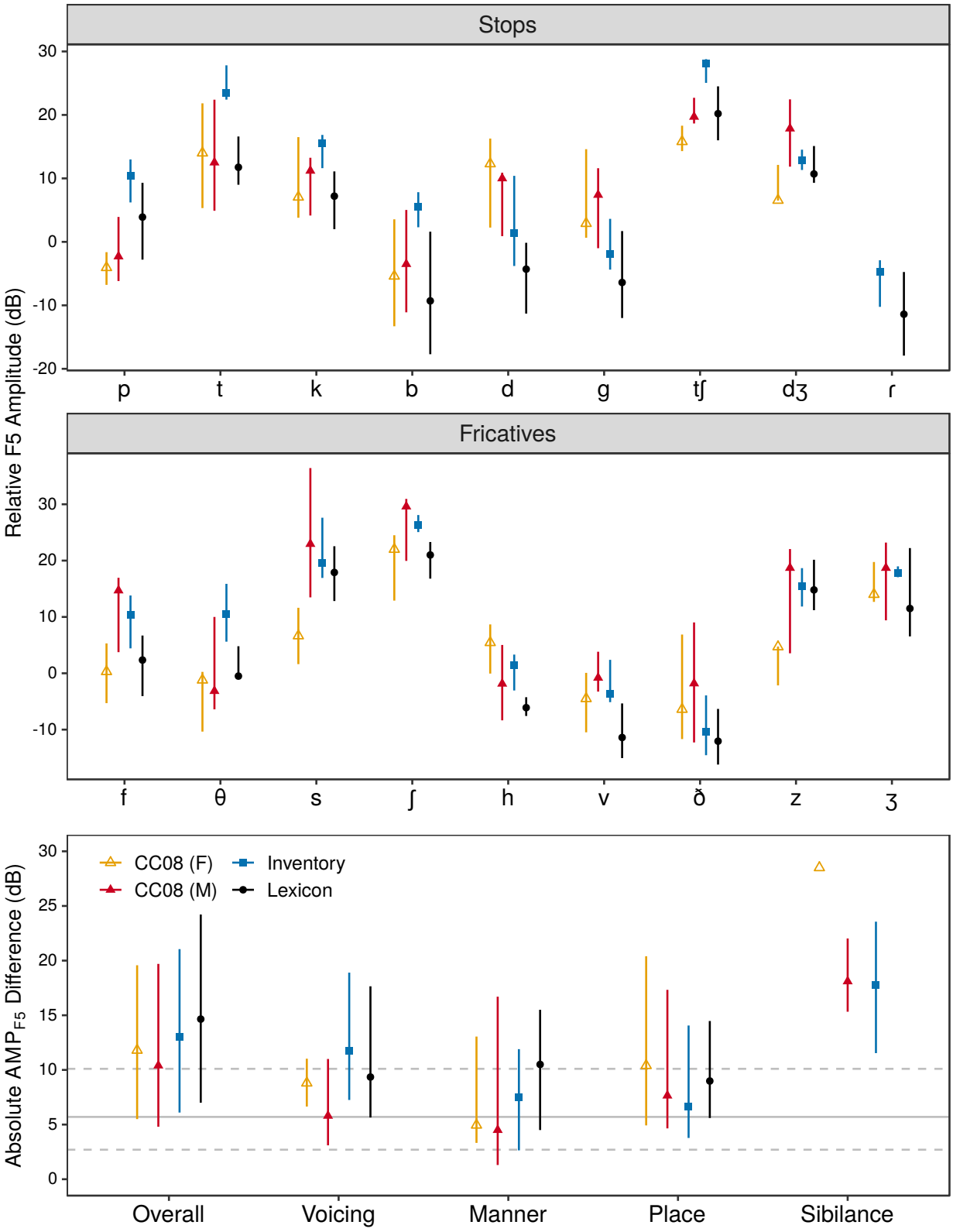

Figure 2.69: Relative F5 Amplitude (ΔAMP_{F5}) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in ΔAMP_{F5} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

distinctions are reduced relative to AMP_{F3} differences word-initially and word-medially. Regarding voicing, as in VCV position, voiceless obstruents exhibit higher relative F5 amplitudes than their voiced counterparts. Finally, no overall manner effect is present among word-final contrasts, though the voiceless alveolar plosive [t] remains notably higher in AMP_{F5} than the voiceless dental fricative [θ]. Overall, the discriminative power of relative F5 amplitude word-finally is intermediate between the ΔAMP_{F5} effect in CV and VCV positions.

2.6.9.4 Summary

In addition to cueing obstruent sibilance among voiceless fricatives, as in Hedrick & Ohde (1993), relative F5 amplitude exhibits structured variation along voicing, manner, and place dimensions in most contrast positions. Thus, in comparison with relative F3 amplitude, AMP_{F5} is more broadly discriminative and therefore may be of greater utility in speech perception. However, as discussed above, some of the contrasts distinguished by AMP_{F3}, such as [s, ʃ], are poorly distinguished by AMP_{F5}, suggesting that the optimal model is one that combines both cues. Further, if listeners are tracking the relative amplitude of the noise in one frequency region, such as around F5, it is reasonable to assume they are tracking other frequency regions as well, even if regions differ in their independent contribution to obstruent discrimination.

2.6.10 Fundamental Frequency ($f0_{VC/CV}$)

2.6.10.1 Background and physiological basis

Fundamental frequency is the acoustic measure of voice pitch, and corresponds to the frequency of vocal fold vibration. As a cue to obstruent consonant distinctions, House & Fairbanks (1953) found initial acoustic evidence of the influence of obstruent voicing on the f0 of the following vowel, voiced obstruents exhibiting a low onset f0, and voiceless obstruents a relatively high onset f0. Perceptual evidence that this perturbation of vowel pitch can be used as a cue to obstruent voicing distinctions was provided in Haggard et al. (1970), and while research has generally focused

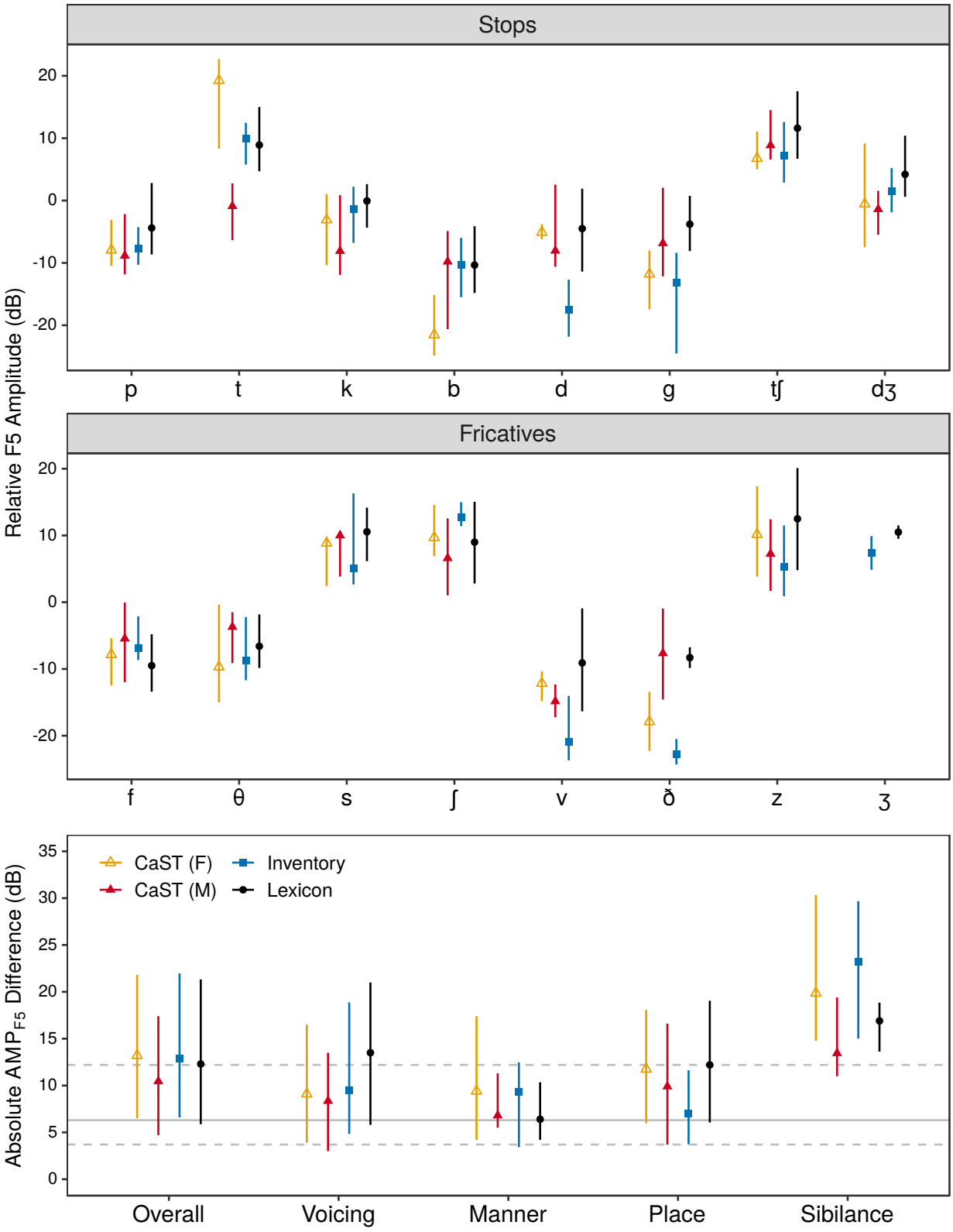

Figure 2.70: Relative F5 Amplitude (ΔAMP_{F5}) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in ΔAMP_{F5} in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

on plosives, the physiological motivation for these effects (voiceless obstruents exhibiting higher airflow and therefore a higher rate of vocal fold vibration at vowel onset) makes the measure relevant for the analysis of fricative and affricate voicing distinctions as well, an extension which will be pursued further in the present study.

2.6.10.2 Definition and measurement

Fundamental frequency ($f0_{VC/CV}$) is defined as the frequency of vocal fold vibration at either the offset of the preceding vowel ($f0_{VC}$) in the case of VCV and VC contrasts, or the onset of the following vowel ($f0_{CV}$) in the case of CV and VCV contrasts. Vowel-onset f0 is measured at the point 10% into the vowel following the target obstruent, and is measured from the pitch-tracking algorithm in Praat (Boersma & Weenink, 2016), with measurement errors hand-corrected based on the duration of the first pitch period of the vowel. Similarly, vowel-offset f0 is measured at the point 10% from the end of the preceding vowel; i.e., at 90% of V1 duration. See Figure 2.64 for sample measurements of $f0_{VC}$ and $f0_{CV}$ in word-medial obstruents.

2.6.10.3 Category and contrast distributions

Below we review vowel offset/onset f0 distributions in obstruent contrasts in the lexicon, inventory, and reference data. Results are presented separately for word-initial ($f0_{CV}$), word-medial ($f0_{VC}$, $f0_{CV}$), and word-final ($f0_{CV}$) contrasts.

Word-initial position (CV). Figure 2.71 shows vowel-onset f0 distributions in word-initial obstruent contrasts, and is consistent with expectations from the literature in showing a raised pitch following voiceless obstruents relative to their voiced counterparts. This effect is broadly present in all four data sets, though the reference data is less consistent in this regard than the target data. There is also a modest manner effect in Figure 2.71 wherein f0 is raised slightly following stop consonants relative to fricatives. Obstruent place and sibilance, on the other hand, show no reliable effect on vowel-onset f0. Ultimately, given the large within-item variance in f0 (shown in the solid

and dashed gray lines in the bottom panel of Figure 2.71), it is difficult from the acoustics alone to determine the reliablility of f0 as a cue to obstruent voicing or manner in the lexicon. This question will be examined more directly in the cue-integration models presented in Chapter 4.

Word-medial position (VCV). For intervocalic contrasts, the consonant may affect the pitch at both the onset and offset of the constriction, and thus both $f0_{VC}$ and $f0_{CV}$ cues are considered for such contrasts. Beginning with vowel-offset f0, Figure 2.72 shows only a modest effect of manner of articulation on $f0_{VC}$, where just as in CV position there is greater pitch lowering at vowel-fricative boundaries than at vowel-stop boundaries. No voicing effects are evident in Figure 2.72, however, while as in CV position, place and sibilance do not impact vowel-offset f0 to a notable degree. Overall, f0 appears to provide less information about the following consonant than it does about the preceding consonant, as evidenced by the comparison of the $f0_{VC}$ results in Figure 2.72 with both the previous $f0_{CV}$ results for word-initial contrasts in Figure 2.71, and the results below for vowel-onset f0 in VCV contrasts.

The fundamental frequency at V2 onset varies largely as a function of obstruent voicing and manner. Both voicing and manner effects are consistent with the CV results; namely, f0 tends to be higher following voiceless obstruents (relative to voiced) and following plosives (relative to fricatives). There is also an apparent place effect in Figure 2.73, but this result is an artifact of the prosodic constraints on the appearance of [t, d, h] word-medially, as this set only occurs intervocalically at the onset of a stressed syllable. Therefore, the f0 rise on vowels following [t, d, h], particularly the voiceless pair [t, h], is due largely to stress, and not to the place of articulation of the consonant. Finally, as in CV position, sibilance has no notable impact on $f0_{CV}$. Overall $\Delta f0_{CV}$ contrast effects are comparable to $\Delta f0_{VC}$ while being reduced slightly in word-medial position relative to word onset. However, as in CV position, $\Delta f0_{CV}$ exhibits a wide range of within-item variation that poses a challenge to the robustness of f0 as a cue to obstruent voicing or manner. Again, this question of cue utility will be revisited in Section 4.4 of Chapter 4, where models of cue integration in the prediction of listener contrast recognition are discussed.

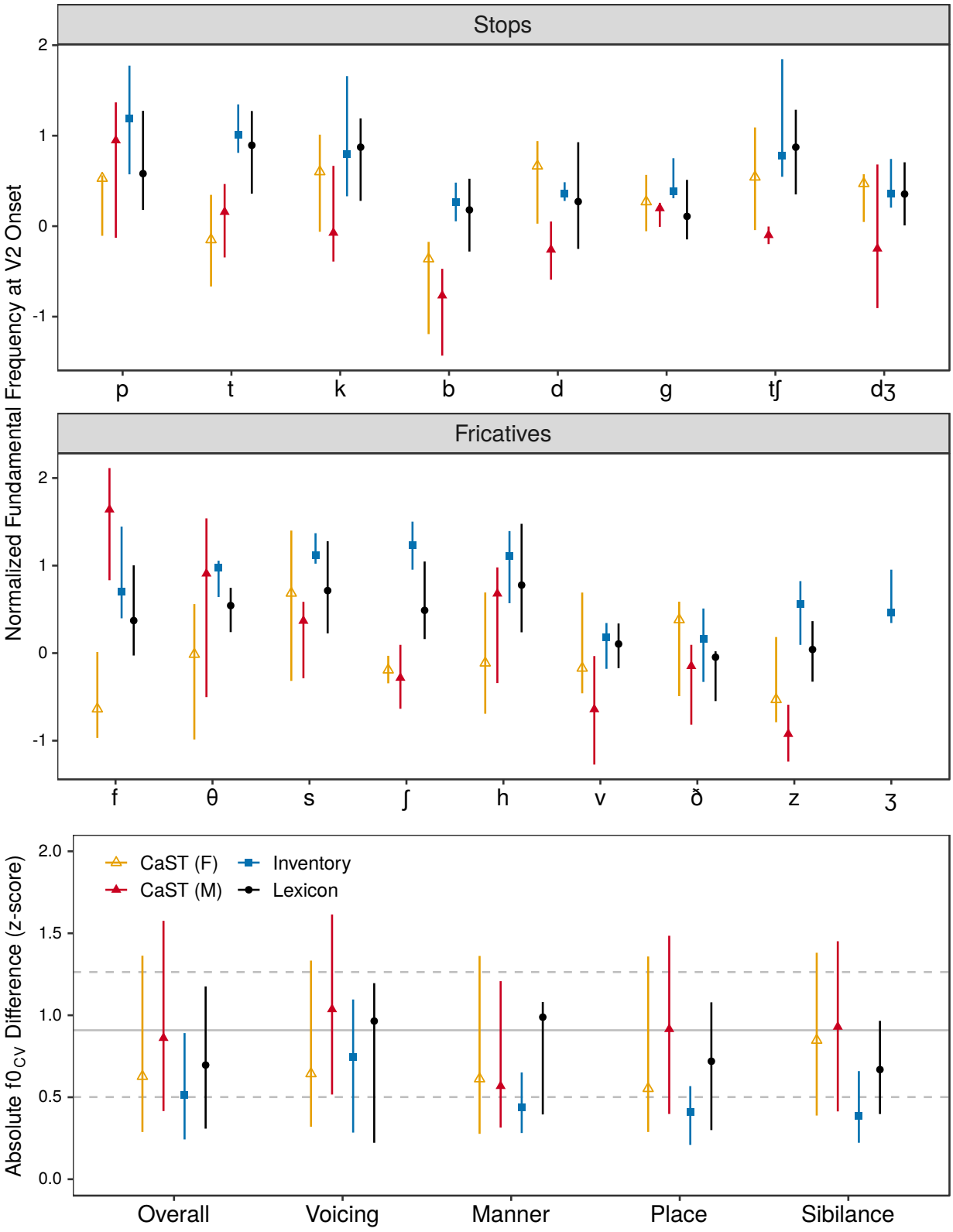

Figure 2.71: Fundamental Frequency at Vowel Onset ($f0_{CV}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $f0_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

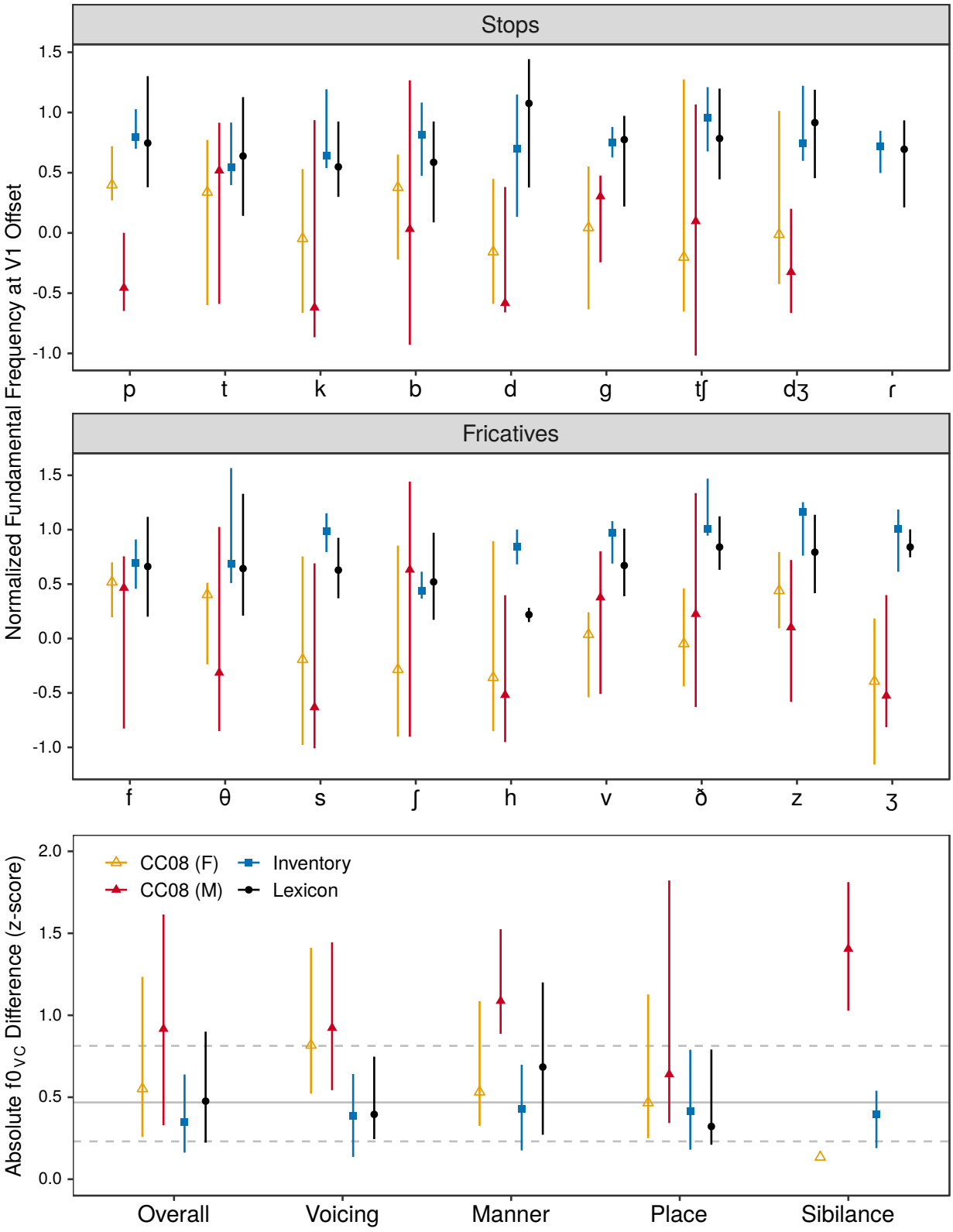

Figure 2.72: Fundamental Frequency at Vowel Offset ($f0_{VC}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $f0_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

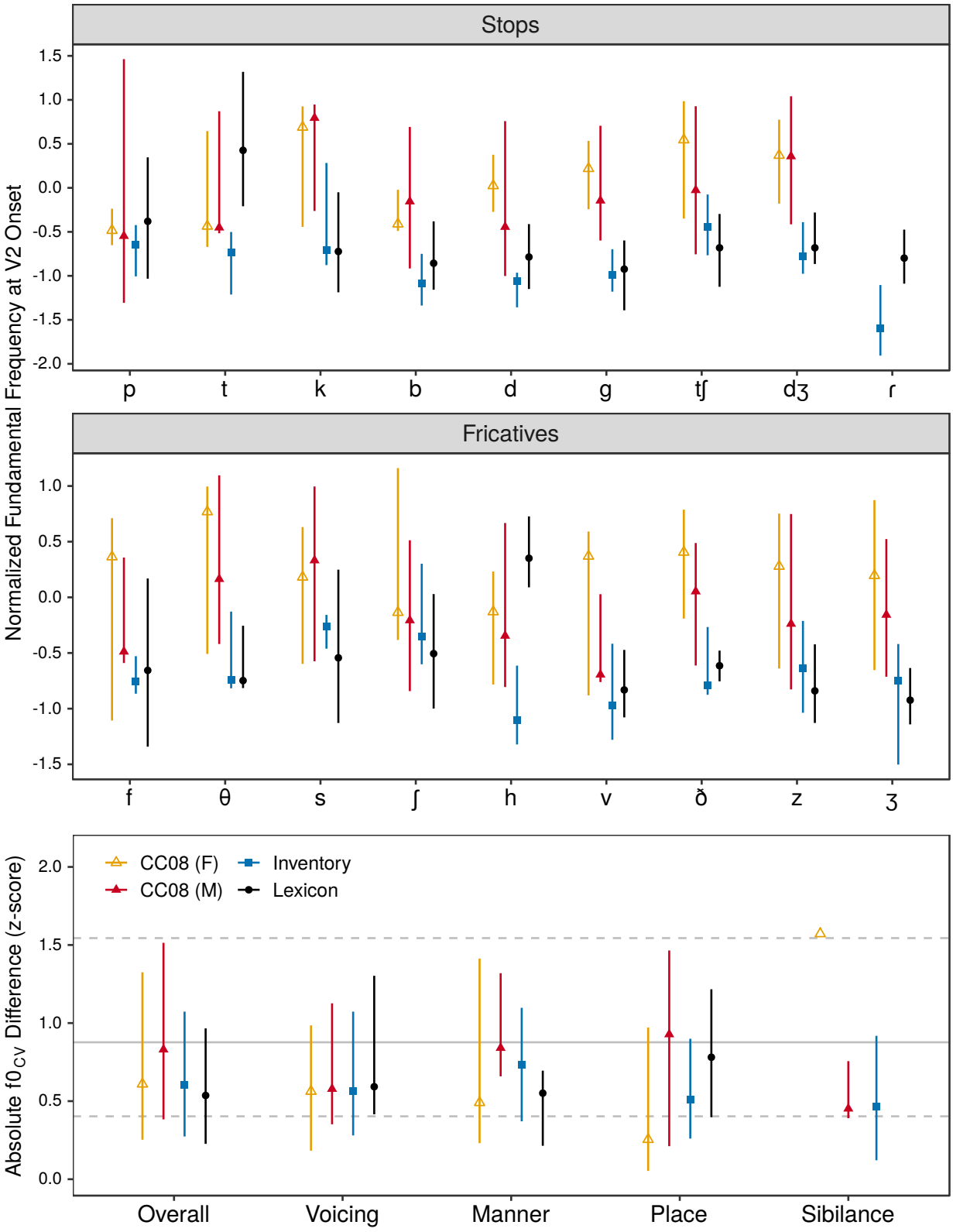

Figure 2.73: Fundamental Frequency at Vowel Onset ($f0_{CV}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $f0_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

Word-final position (VC). Figure 2.74 shows vowel-offset f0 distributions in word-final obstruent contrasts in the lexicon, inventory, and reference data. Results for voicing and manner of articulation are comparable to the $f0_{VC}$ results in word-medial position. Namely, the voicing effect is not consistent across manners, with the only clear distinction occurring between voiced and voiceless fricatives in the lexicon. Regarding manner of articulation, $f0_{VC}$ is generally lowest preceding fricatives, with the effect more robust word-finally than word-medially. Finally, there is a modest place effect among VC contrasts in the lexicon that is not present intervocalically. Among plosives, f0 is the lowest at transitions into alveolars, while the most notable place distinction among fricatives is that between alveolars and postalveolars; i.e., [s] < [ʃ], [z] < [ʒ].

2.6.10.4 Summary

The fundamental frequency of the vowel at the edge of the transition into or out of a consonant constriction, while consistent with the literature on f0 perturbation as a function of voicing, also exhibits considerable variablity, both between and within items. Across CV, VCV, and VC positions, the most reliable dimension distinguished by $f0_{VC/CV}$ was manner of articulation, wherein fricatives induce lower f0 values at vowel offset and onset relative to stops. On the other hand, the voicing effect—i.e., *voiced* < *voiceless*—was primarily restricted to f0 at vowel onset (i.e., $f0_{CV}$), meaning f0 is only viable as a voicing cue in word-initial and word-medial contrasts. Further, this effect is notably more robust in CV position than in VCV position, a result which though consistent with much of the focus in the literature on word/syllable-onset voicing contrasts, reduces the overall discriminative power of f0 in the lexicon.

2.6.11 First Formant Frequency ($F1_{VC/CV}$)

2.6.11.1 Background and physiological basis

The first formant frequency at vowel onset was initially studied as a cue to stop place of articulation in early work at Haskins Laboratories (Liberman et al., 1954; Delattre et al., 1955), but from this

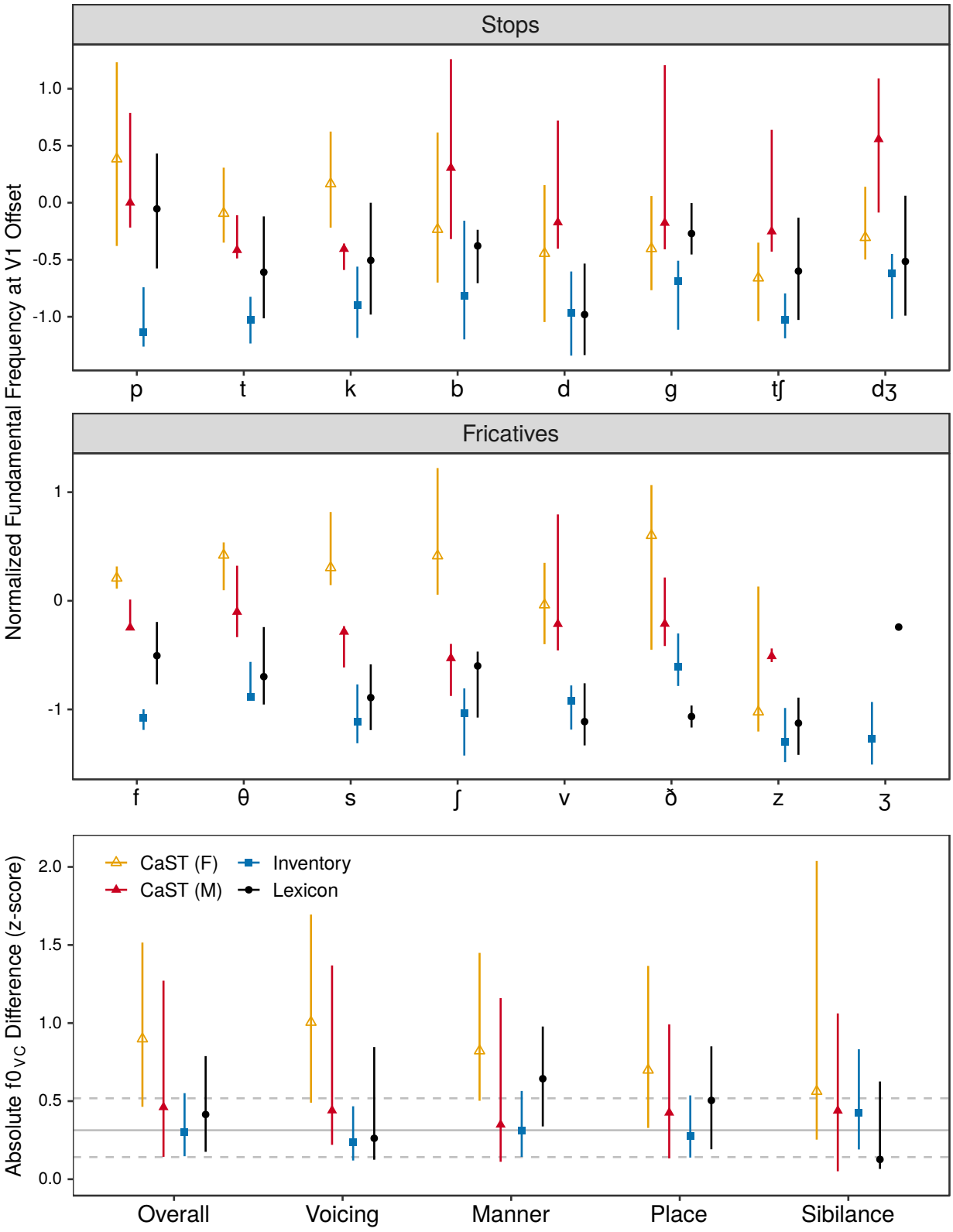

Figure 2.74: Fundamental Frequency at Vowel Offset ($f0_{VC}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $f0_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

work it became apparent that F1 is far more indicative of stop voicing distinctions than place of articulation. The physiological basis of this relation is both the effect of voicing on the size of the posterior cavity in the vocal tract (voiced obstruents typically exhibiting pharyngeal expansion and larynx lowering to preserve the pressure differential across the larynx necessary to maintain voicing), and the relative onset of the appearance of F1 in the spectrogram. This latter measure has been referred to as the F1 'cutback time' (Liberman, 1993), and refers to the fact that the greater aspiration following voiceless plosives tends to obscure the appearance of F1 in the spectrogram. This means that by the time F1 is measurable following voiceless obstruents, it has transitioned further into the steady-state of the vowel, whereas in voiced plosive contexts F1 can be measured almost immediately after the release burst. This relationship between F1 and aspiration duration means that F1 may in fact be able to index place of articulation to some degree, but as a direct function of differences in noise duration rather than reflecting differences in vocal tract resonance for different constriction locations.

2.6.11.2 Definition and measurement

The first formant frequency ($F1_{VC/CV}$) as a cue in obstruent contrast discrimination is defined as the first vocal tract resonance at either the offset of the preceding vowel ($F1_{VC}$) in the case of VCV and VC contrasts, or the onset of the following vowel ($F1_{CV}$) in the case of CV and VCV contrasts. Vowel-onset F1 is measured at the point 10% into the vowel following the target obstruent, and is measured from the Burg formant-tracking algorithm in Praat (Boersma & Weenink, 2016), with measurement errors hand-corrected based on visual identification of F1 in the first pitch period of the vowel. Similarly, vowel-offset F1 is measured at the point 10% from the end of the preceding vowel. See Figure 2.64 for sample measurements of $F1_{VC}$ and $F1_{CV}$ in word-medial obstruents.

2.6.11.3 Category and contrast distributions

As in the previous section on fundamental frequency, the presentation of vowel offset/onset F1 distributions is organized by contrast position, with vowel-onset F1 ($F1_{CV}$) examined word-initially

and word-medially, and vowel-offset F1 ($F1_{VC}$) examined word-medially and word-finally.

Word-initial position (CV). Figure 2.75 shows F1 distributions following word-initial obstruent categories and contrasts. Before assessing the impact of different obstruent features on $F1_{CV}$ distributions, it is worth emphasizing that vowel-onset F1 naturally depends on the category of the vowel. For this reason the category distributions in Figure 2.75 exhibit large variances, and therefore any patterns obtained in aggregate may not generalize across all vowel contexts.

As expected from the literature on F1 transitions in CV position, the primary factor in determining vowel-onset F1 is the voicing status of the preceding consonant, where voiced obstruents exhibit lower $F1_{CV}$ values than their voiceless counterparts. This effect is most pronounced among stops, where distinctions in the target data range between 50 and 100 Hz on average. Among fricatives the effect of voicing is much weaker and more variable between fricative pairs. These effects are consistent between the lexicon and inventory data, and while the reference data patterns are less clear in this regard, this result is expected due to both individual variation in F1 from differences in vocal tract sizes, and differences in vowel contexts: 10 vowels are used in the inventory data, as compared with the 3 corner vowels in Woods et al. (2010).

In addition to the predicted effect of voicing on vowel-onset F1, there is also a minor place effect evident in the ΔF1 distributions in the bottom panel of Figure 2.75, which derives from the raised F1 frequencies following labial plosives relative to alveolars/velars. This result is consistent with the greater consonant-vowel coarticulation in labials, which should raise F1 onset levels given that F1 is generally higher at vowel midpoint than at vowel onset, and greater coarticulation should flatten this transition. Finally, neither manner nor sibilance exhibited notable effects on F1. Overall, because of the large voicing effect and the frequency of word-initial stop contrasts that differ in voicing (among other features), $F1_{CV}$ is quite informative, both in the lexicon and inventory.

Word-medial position (VCV). The voicing effect marked by F1 at preceding vowel offset in VCV contrasts is similarly robust, but reflects different voicing patterns as a function of manner of articulation. Word-initially, the *voiceless* > *voiced* relation for $F1_{CV}$ was primarily restricted

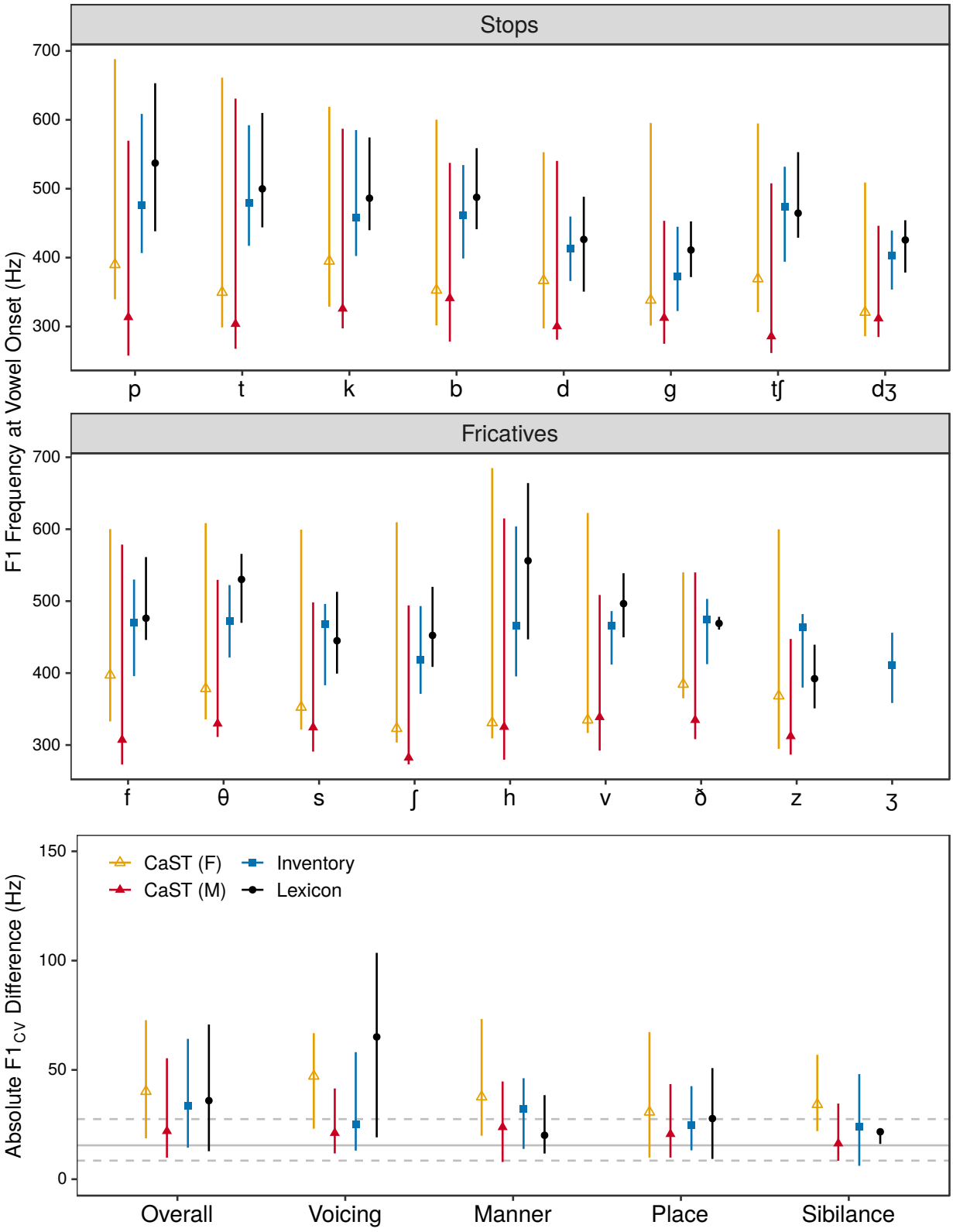

Figure 2.75: F1 Frequency at Vowel Onset ($F1_{CV}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F1_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

to stop consonants; however, word-medially $F1_{VC}$ primarily distinguishes voiced and voiceless fricatives and affricates, with no consistent effect among plosives. There are also minor manner and place effects in the lexicon, where the latter is consistent with the $F1_{CV}$ patterns word-initially, and the former derives from the greater F1 lowering at vowel-stop boundaries than at vowel-fricative boundaries, a result which directly reflects the impact of the degree of obstruent constriction on the first formant frequency. Finally, regarding sibilance, F1 is typically lower preceding the alveolar sibilants [s, z] than in their nonsibilant counterparts [θ, ð], which could reflect differences in tongue position (apical gestures typical of alveolars require greater tongue body lowering than the laminal gestures more common of dental fricatives) rather than true differences in sibilance that derive from the impact of the supralaryngeal sound source on the acoustic signal.

Figure 2.77 shows F1 distributions at the onset of the following vowel in VCV contrasts, and though consistent with the CV results in showing F1 lowering following voiced stops, there is also a moderate voicing-induced F1 lowering among fricatives that is more aligned with the $F1_{VC}$ patterns in the same VCV contrasts. Thus, for fricatives there is a symmetry in F1 perturbation at the initiation and release of the constriction that is not present for stop consonants. This asymmetry, combined with the large F1 differences among stops word-initially, suggests that the lowering of F1 after voiced plosives may simply be an artifact of the impact of aspiration on formant transitions, in that aspiration obscures the formants of the following vowel, and thus longer aspiration intervals will obscure more of the low-frequency portion of the F1 transition than shorter intervals, resulting in a vowel-onset F1 that is relatively higher following voiceless stops.

There are also minor manner, place, and sibilance effects in Figure 2.77, all of which are consistent with the vowel-offset F1 distributions in Figure 2.76: namely, *stops* < *fricatives*, *alveolar/velar* plosives < *labial* plosives, and *sibilant* fricatives < *nonsibilant* fricatives. Overall, $\Delta F1_{CV}$ is similarly discriminative of obstruent contrasts in VCV position as in CV position, though the information captured in vowel-onset F1 is more evenly distributed among different featural dimensions word-medially than word-initially.

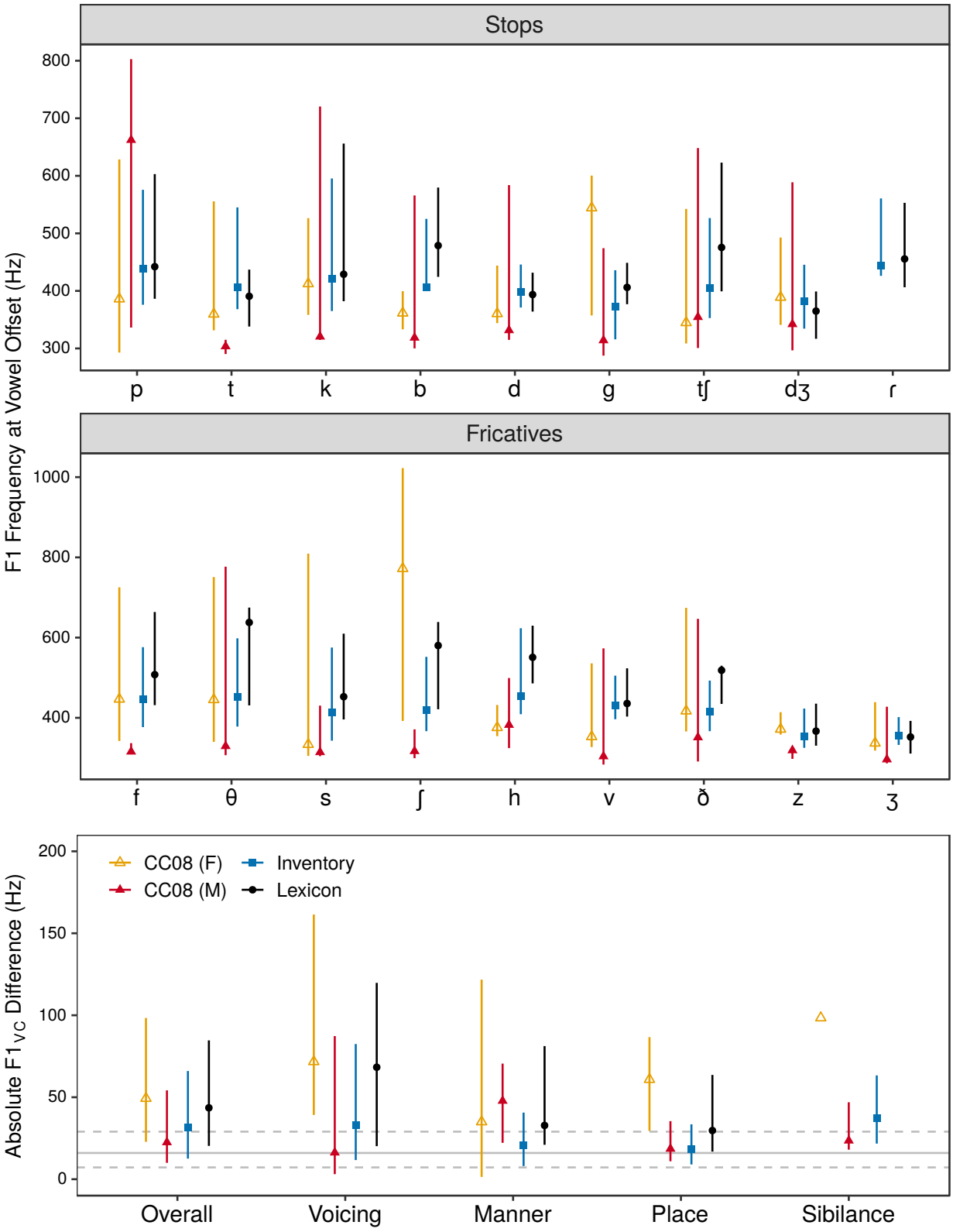

Figure 2.76: F1 Frequency at Vowel Offset ($F1_{VC}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F1_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

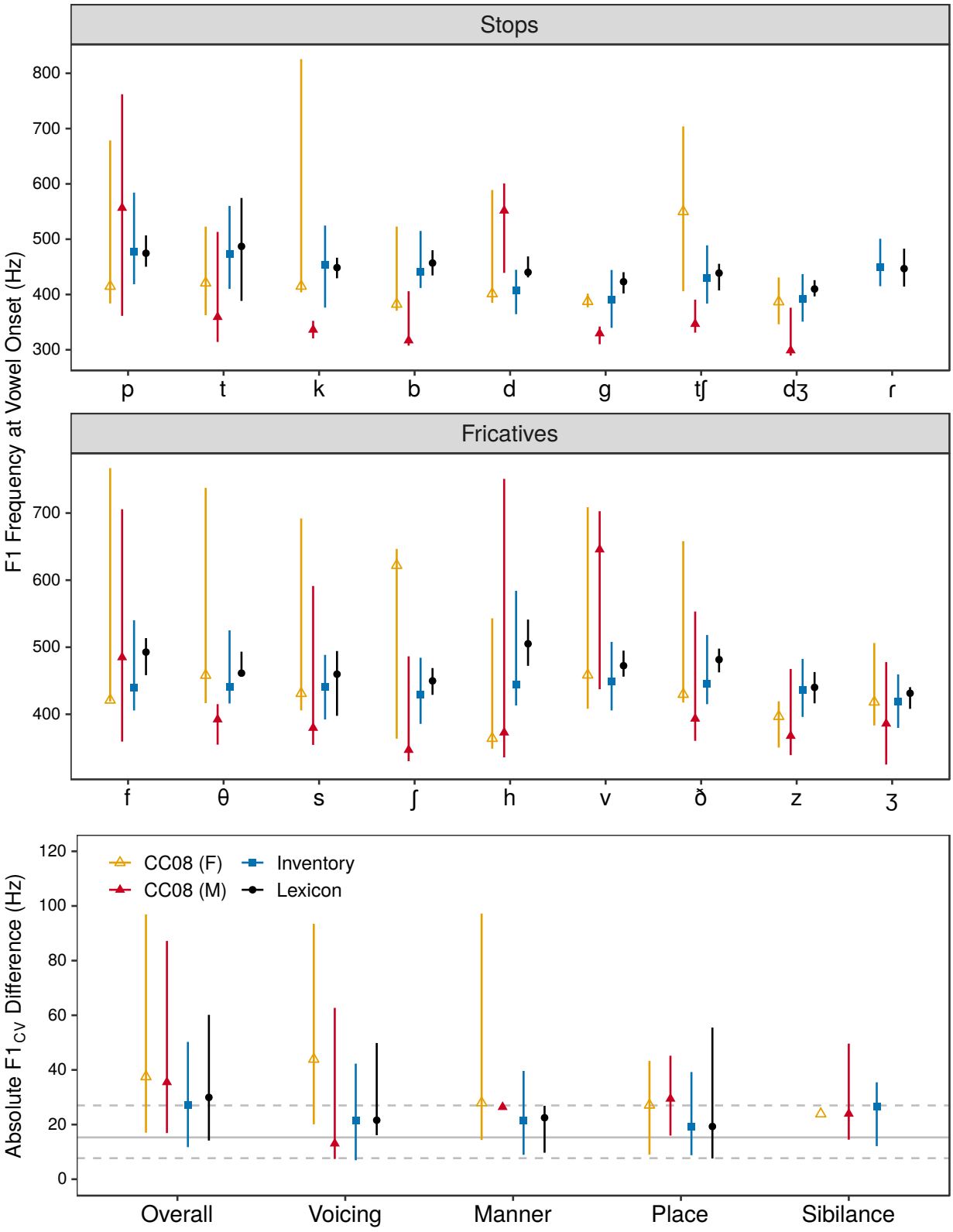

Figure 2.77: F1 Frequency at Vowel Onset ($F1_{CV}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F1_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

Word-final position (VC). $F1_{VC}$ distributions among word-final obstruent contrasts are shown in Figure 2.78, and as in VCV position there is little effect of voicing among plosives in the lexicon, though there is a plosive voicing effect in the inventory, while fricatives and affricates show a consistent *voiced* $<$ *voiceless* relation. Further, the voicing effect among word-final fricatives is more robust than in VCV position, though as in VCV position this pattern is largely restricted to the target data. Other results evident in Figure 2.78 are the lower $F1_{VC}$ values preceding stop closures than preceding fricative constrictions, a reduced sibilance effect relative to VCV position (i.e., [s] $<$ [θ], but [z] $\approx$ [ð]), and a place effect that is due rather to distinctions between fricatives (namely, [θ] $<$ [f], [ð] $<$ [v]) than between plosives. Overall, $F1_{VC}$ is similarly discriminative word-finally as word-medially, though with generally greater consistency across databases.

2.6.11.4 Summary

Unlike $f0_{VC/CV}$, the first formant frequency at vowel onset/offset is much more reliable and consistent across contrast positions as a cue to obstruent voicing. What varies as a function of position, however, is the distribution of voicing effects among different manners of articulation. Vowel-onset F1 ($F1_{CV}$) primarily reflects stop voicing, particularly in word-initial position, whereas $F1_{VC}$ is much more discriminative of fricative voicing contrasts, with $\Delta F1_{VC}$ effects similarly reduced intervocalically relative to word-final position. The other consistent effect is due to manner of articulation, where at both vowel offset and vowel onset, across positions, stops lower F1 to a greater degree than fricatives. Other effects of place and sibilance appear sporadically but are not as consistent or robust as the voicing and manner effects.

2.6.12 Second Formant Frequency ($F2_{VC/CV/V1/V2}$)

2.6.12.1 Background and physiological basis

The second formant frequency has been used in numerous studies of place of articulation distinctions among obstruent consonants, where early work focused directly on the F2 transitions into

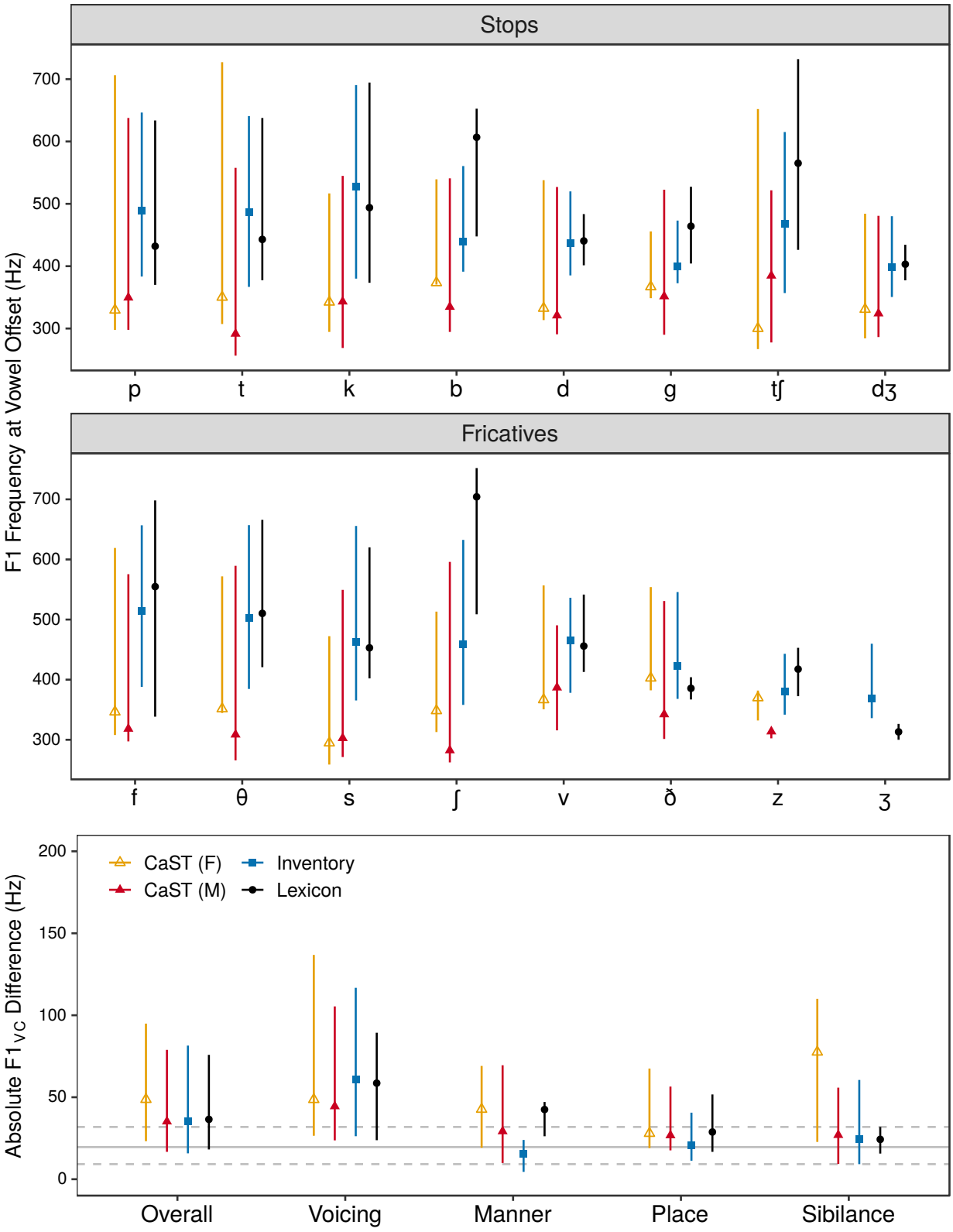

Figure 2.78: F1 Frequency at Vowel Offset ($F1_{VC}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F1_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

and out of the consonant (Liberman et al., 1954; Delattre et al., 1955), while later work in Krull (1989), Nearey & Shammass (1987), Sussman et al. (1991), and Sussman et al. (1993) focused on relationships between F2 at vowel onset and midpoint via the *locus equation*. In both measures, F2 tracks the place of articulation of the consonant constriction because F2 generally corresponds to the cavity resonance anterior to the constriction. Further, as consonants vary in their coarticulatory propensity with adjacent vowels, the degree of influence of the consonant on vowel-midpoint F2 can also be used as a cue to place of articulation. Previous research has primarily focused on the effect of stop consonants on F2, but fricatives should theoretically differ as well in F2 dynamics as a function of place of articulation, though such effects are generally weaker than those found for stops (Sussman & Shore, 1996; Jongman et al., 2000).

2.6.12.2 Definition and measurement

The second formant frequency is measured at both vowel offset/onset ($F2_{VC/CV}$), as with the first and third formant frequencies, and at vowel midpoint ($F2_{V1/V2}$), where the latter set of vowel-nucleus parameters is only defined for F2. This is done because the analysis of F2 as a cue to consonant place of articulation, from early research with the Pattern Playback at Haskins (Liberman et al., 1957) to later work on locus equations (Nearey & Shammass, 1987; Krull, 1989; Sussman et al., 1991, 1993) and target-locus scaling (Broad & Clermont, 1987, 2002, 2010, 2014), typically takes into account either the full transition from vowel onset to midpoint, or endpoints of the transition we have done here with $F2_{VC/CV}$ and $F2_{V1/V2}$. All other procedures for F2 measurement follow those described above for F1.

2.6.12.3 Category and contrast distributions

F2 distributions are presented below for obstruent contrasts in the lexicon, inventory, and reference data. As in the previous two sections, results are separated by contrast position so that in addition to assessing each cue independently, dependencies between onset/offset and midpoint values such as those formalized in the locus equation can also be addressed.

Word-initial position (CV). Figure 2.79 shows $F2_{CV}$ distributions in word-initial obstruent contrasts. Overall, $F2_{CV}$ is highly discriminative, exhibiting systematic variation according to obstruent voicing, manner, and place. Beginning with the place effect, which is the most discussed in the literature and most physiologically informative, among plosives F2 is lowest following labials, and approximately equivalent between alveolars and velars. Note, however, that just as in the previous section, vowel context matters, and thus any patterns that emerge among obstruent categories in the first two panels of Figure 2.79, given that they represent results aggregated over a range of vowels, cannot be assumed to hold in all contexts. This is one of the reasons the within-category variance in Figure 2.79 is so large. Among fricatives, there is a consistent raising of $F2_{CV}$ with more posterior articulations that is present in both the voiceless and voiced series.

Regarding voicing, vowel-onset F2 is notably higher following voiced plosives (primarily [d, g]), an effect which extends to affricate and fricative voicing contrasts but is much reduced relative to the 100–300 Hz distinction among plosives. This result is consistent with the voicing effect for $F1_{CV}$ in Figure 2.75 in that the majority of F2 transitions into the following vowel descend from a higher F2 frequency at consonant offset, and thus greater perturbation of the early portion of the transition from aspiration results in lower F2 values at vowel onset. Voiceless frication can yield similar effects though to a lesser degree than aspiration.

Finally, there is a modest effect of manner of articulation on vowel-onset F2 wherein the second formant is slightly raised after stops relative to their fricative counterparts. As with the manner effects on F1, this relation derives from the greater perturbation of F2 by complete consonant closure, which results in less coarticulation with the following vowel than in fricative-vowel sequences. Overall, $\Delta F2_{CV}$ is robust across obstruent contrasts in the lexicon, inventory, and reference data.

Turning next to F2 at vowel midpoint ($F2_{V2}$), consonantal effects are expected to be notably reduced relative to those at vowel onset ($F2_{CV}$). Nevertheless, some effects persist well past vowel onset, particularly voicing and place of articulation. Figure 2.80 shows that the voicing effect at vowel midpoint is generally consistent with that at vowel onset—i.e., *voiceless* < *voiced*—though it is much reduced and with less clear physiological or acoustic origins. Regarding place, labials

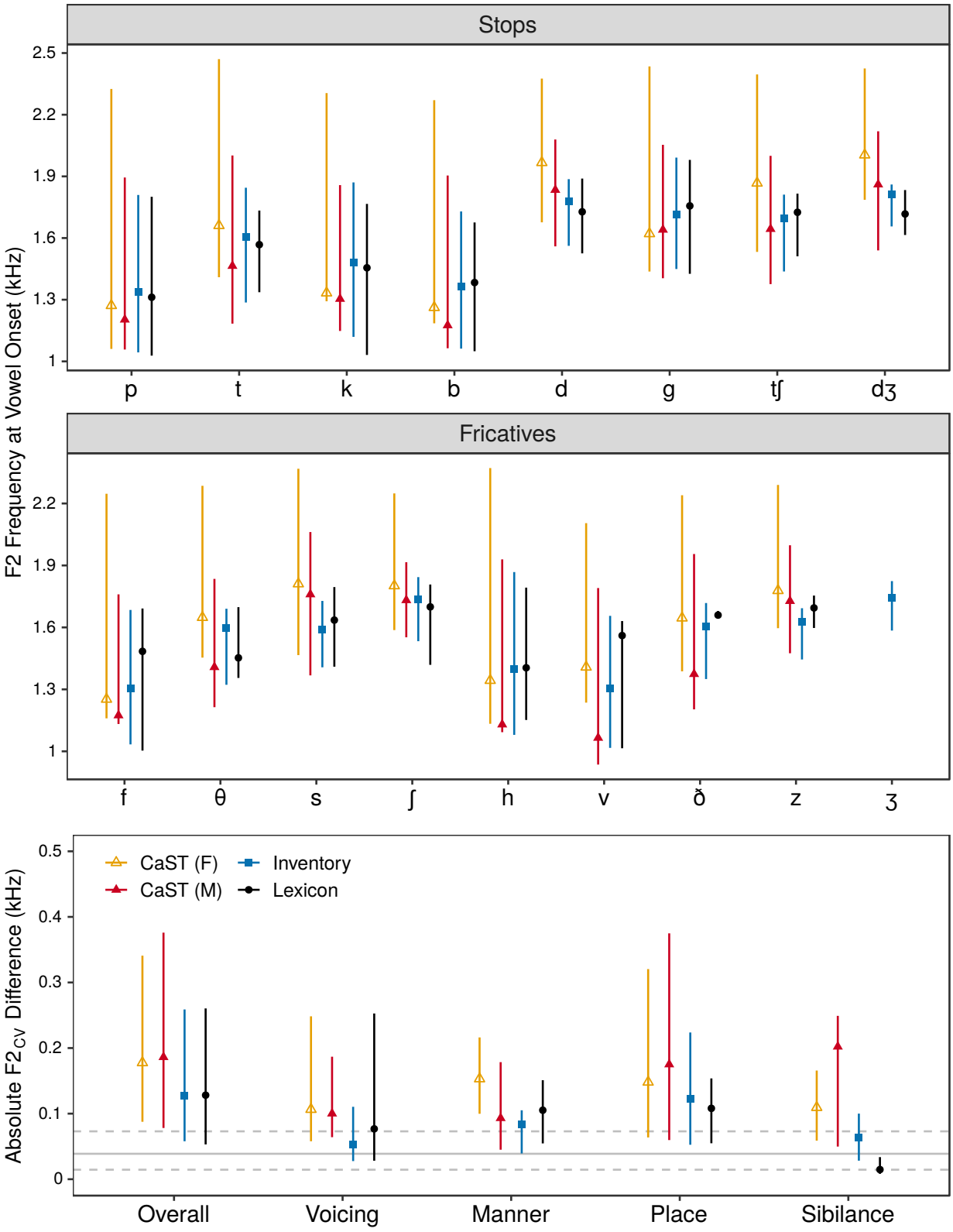

Figure 2.79: F2 Frequency at Vowel Onset ($F2_{CV}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

exhibit the lowest $F2_{V2}$ values, among both plosives and fricatives (excepting [h]), alveolars and postalveolars exhibit the highest $F2_{V2}$ values, and velars and dentals are intermediate between the two. However, these effects are much narrower and more variable than at vowel onset, and thus while vowel-midpoint F2 provides some information about the identity of the preceding consonant, $F2_{V2}$ may be of greater utility in the cue-integration model as a reference against which $F2_{CV}$ may be more reliably evaluated in speech perception.

Word-medial position (VCV). Intervocalically, F2 transitions from both the preceding and following vowel are incorporated in the discrimination of obstruent contrasts, meaning there are four F2 cues in VCV position. Beginning with F2 at the midpoint of the preceding vowel, Figure 2.81 shows relatively constant $F2_{V1}$ distributions across obstruent categories; however, there is still some information about the featural characteristics of the following obstruent that emerges as early as V1 midpoint. Vowels preceding voiced obstruents exhibit slightly higher $F2_{V1}$ values than their voiceless counterparts, and there is also a moderate place effect wherein $F2_{V1}$ is raised preceding alveolar/velar stops. These effects are relatively robust in the lexicon, though in the inventory they are closer to chance levels, suggesting some results may be particular to the kind of vowel contexts comprising VCV minimal pairs in the lexicon.

At vowel offset, F2 distributions are much more distinct (see Figure 2.82, though they remain primarily informative as to the place and voicing of the following consonant. Regarding place, labial obstruents (particularly plosives) exhibit notably lower F2 offsets than alveolars, postalveolars, or velars, again reflecting the greater influence of lingual constrictions on the preceding vowel. There is also a modest voicing effect on $F2_{VC}$ compared to the effect on $F2_{CV}$ in word-initial position, but the *voiceless* $<$ *voiced* relation remains consistent and is present in all manner classes and all places except labials. Thus, while the impact of voicelessness in the following consonant is a loss of formant energy earlier in the VC transition, resulting in higher mean $F2_{VC}$ frequencies in voiced contexts, this effect is much reduced relative to the formant damping that results from the spread of voiceless noise into the following vowel in word-initial position. Next we investigate

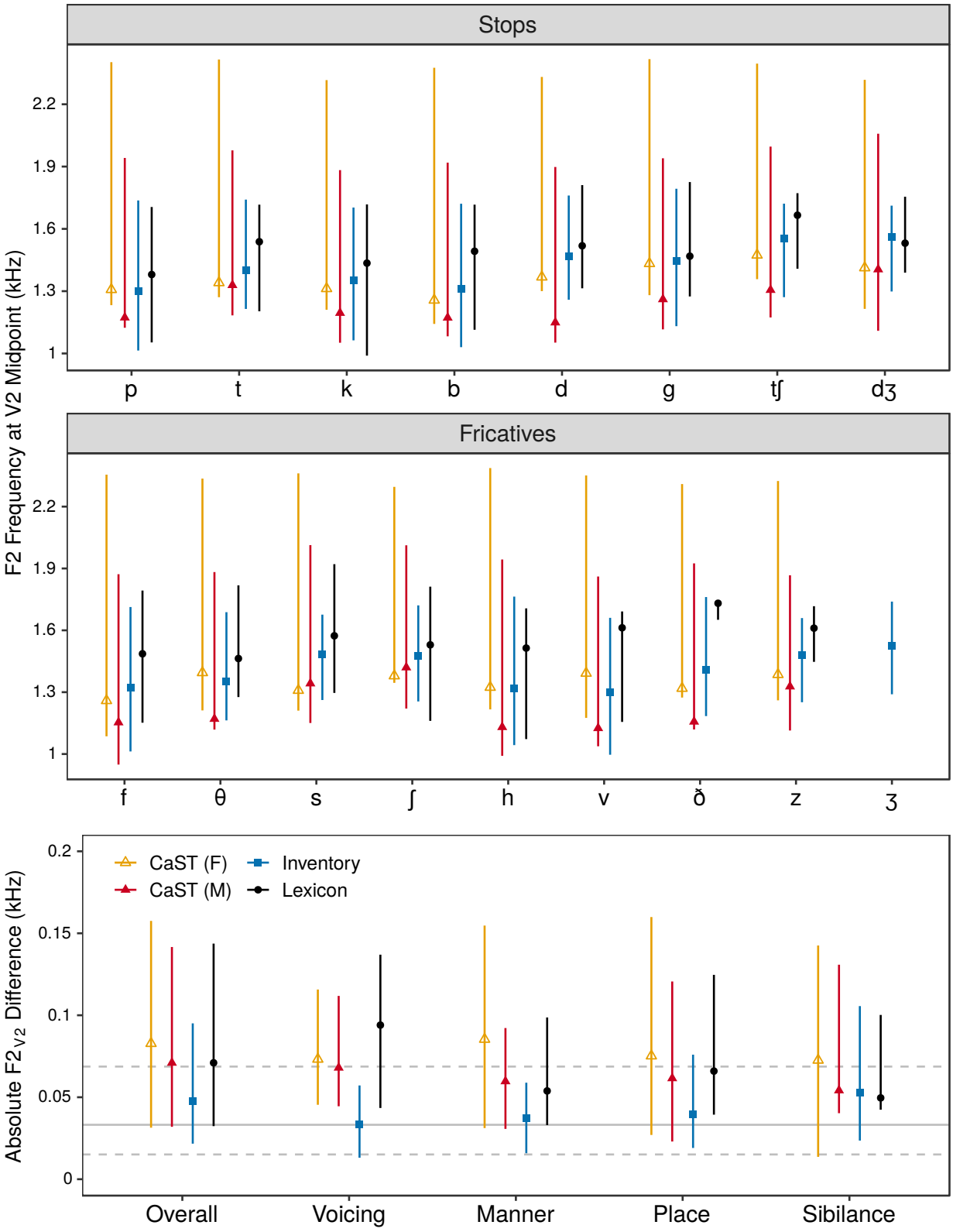

Figure 2.80: F2 Frequency at V2 Midpoint ($F2_{V2}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{V2}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

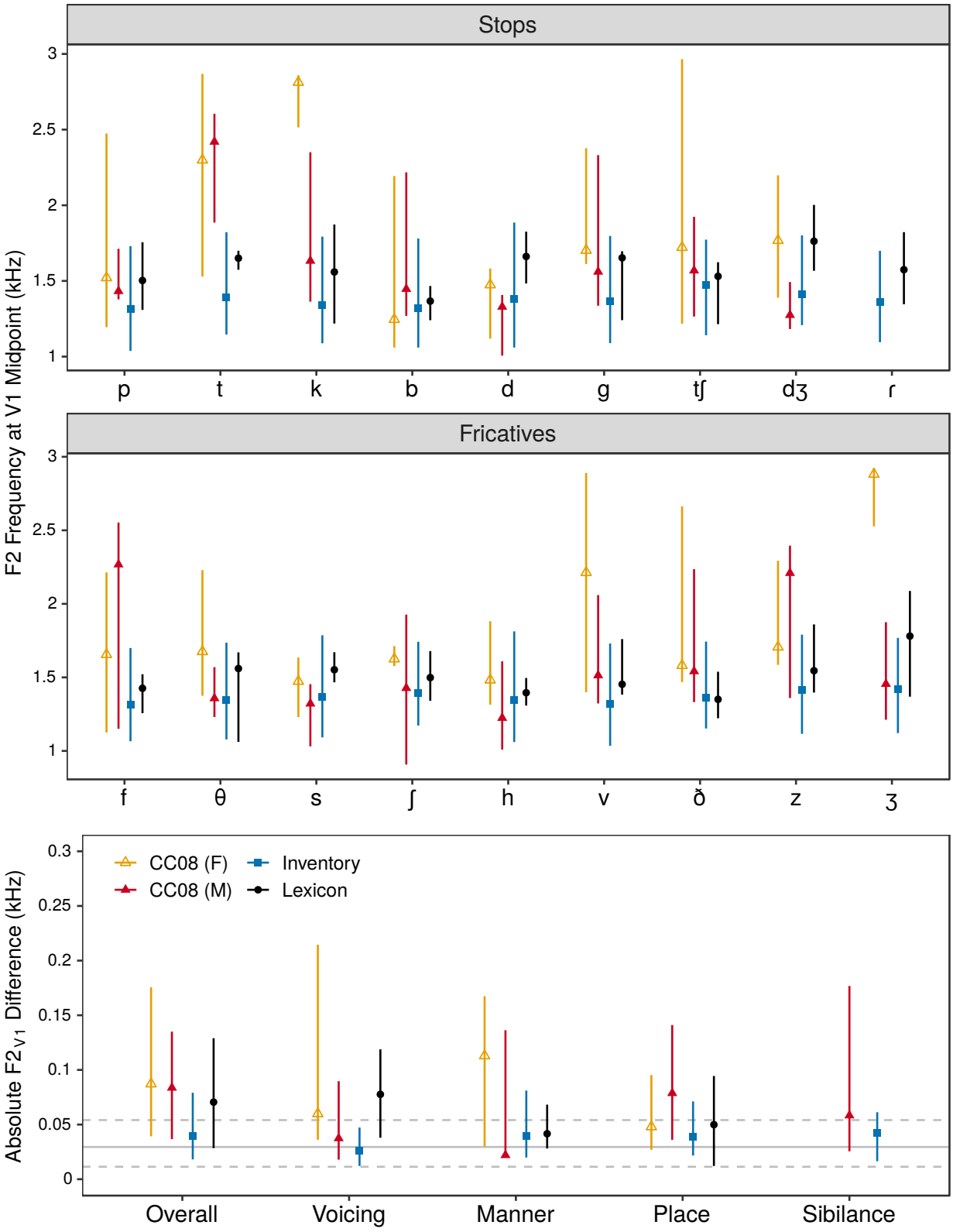

Figure 2.81: F2 Frequency at V1 Midpoint ($F2_{V1}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{V1}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

$F2_{CV}$ distributions in VCV position to determine whether this is a simple serial ordering effect or if it depends more on differences in prosodic structure word-initially and word-medially.

Figure 2.83 confirms the prosodic source of the difference in voicing effect sizes discussed above. There is only a weak voicing effect intervocalically, which suggests that the reduction in the degree of aspiration of voiceless plosives (see Figures 2.11 and 2.12, for instance)—both from their word-medial location and their position in primarily unstressed syllables due to the typical trochaic stress pattern in English—has also reduced the difference in F2 transitions (and thereby their onsets) as a function of voicing. The effect of place of articulation, however, is more robust intervocalically than word-initially. Among all obstruents there is a consistent raising of F2 following more posterior constrictions, with the *labial* < *alveolar* < *velar* relation observed among both voiced and voiceless plosives, and the [f, v] < [θ, ð] < [s, z] < [ʃ, ʒ] relation observed among fricatives. Further, this effect is consistent between the lexicon and inventory, and to a large extent the reference data as well, though the plosives are more variable in this regard. Overall, the discriminative power of $F2_{CV}$ is similar between CV and VCV positions, though the relative utility of $F2_{CV}$ is not similarly distributed according to different features.

Finally, we examine F2 at the midpoint of the following vowel in VCV contrasts. Figure 2.84 shows that the place effects observed at vowel onset, though reduced by vowel midpoint, are still present; namely, $F2_{V2}$ is generally higher following alveolars, postalveolars, and velars than it is following labials and dentals, which given the typically declining F2 transitions from the former suggests that even at its midpoint the vowel shows some influence of the preceding consonant. Other effects such as voicing, manner, and sibilance are not robust.

Word-final position (VC). F2 transitions at word offset are largely consistent with V1 transitions in VCV contrasts; namely, place and voicing effects emerge early in the transition at vowel midpoint, and increase in robustness at vowel offset while also shifting in relative discriminative power ($F2_{V1}$: *voicing* > *place*, $F2_{VC}$: *place* > *voicing*). Figure 2.85 shows $F2_{V1}$ distributions in word-final position. Just as in VCV position, $F2_{V1}$ shows high within-category variance, but still

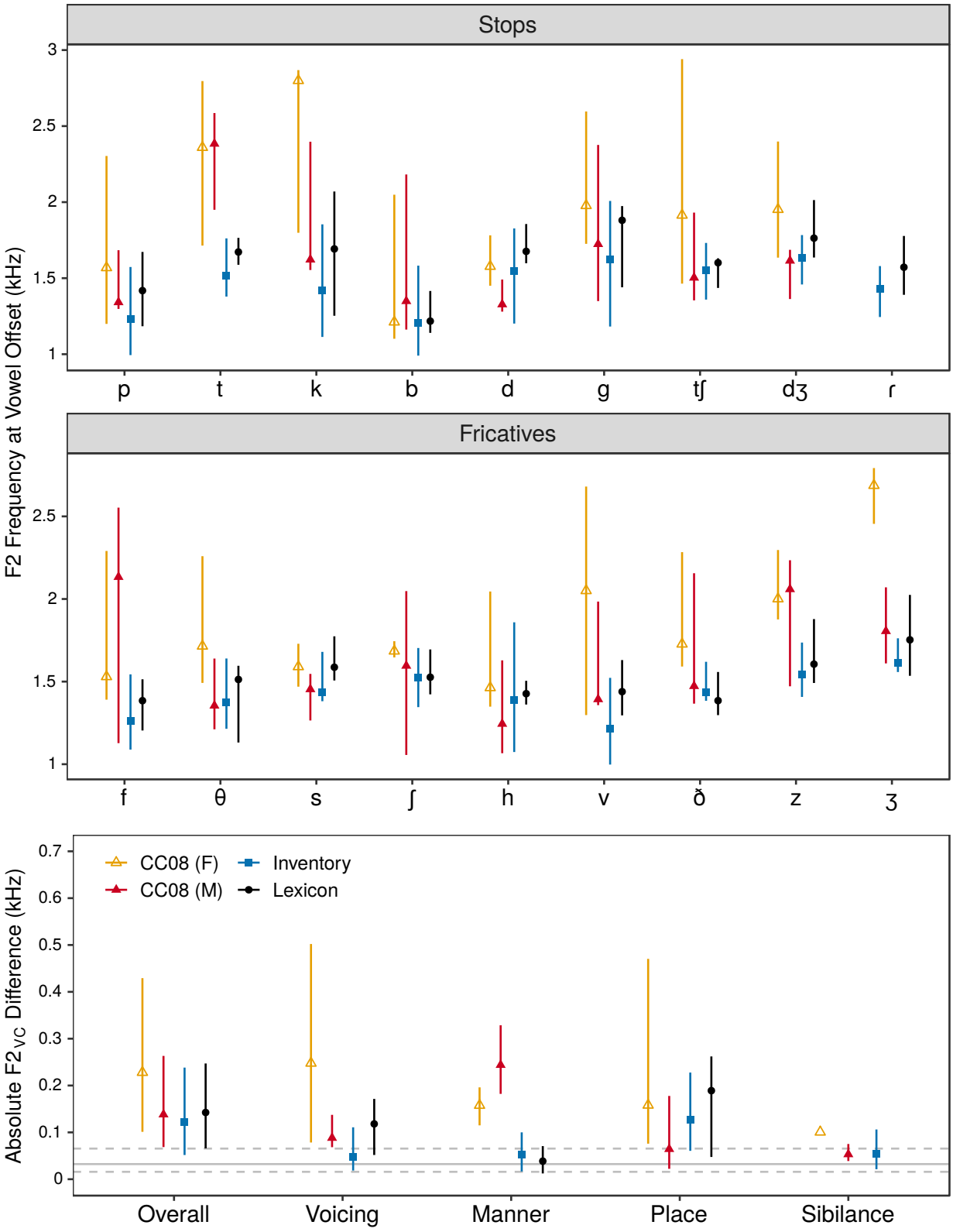

Figure 2.82: F2 Frequency at Vowel Offset ($F2_{VC}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

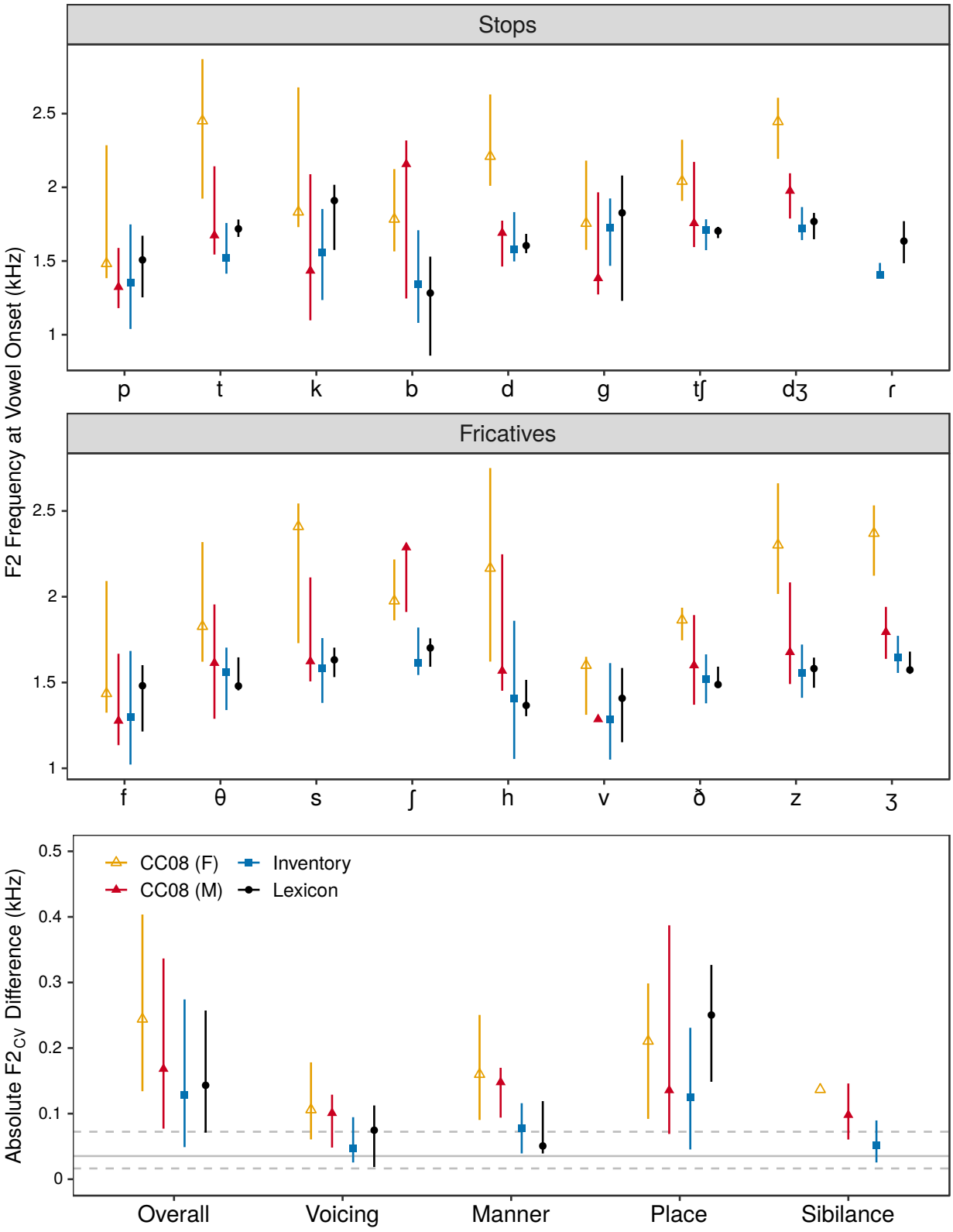

Figure 2.83: F2 Frequency at Vowel Onset ($F2_{CV}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

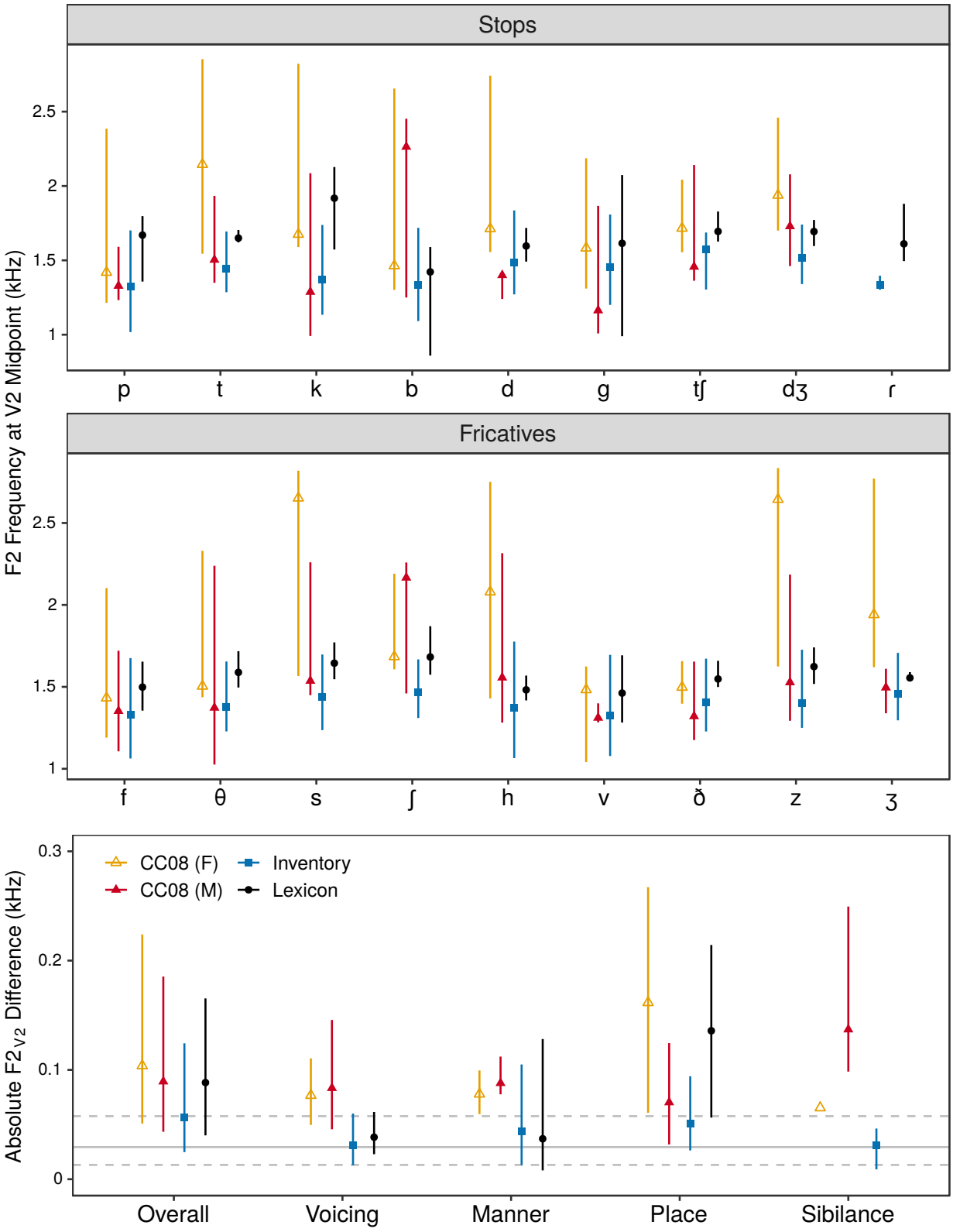

Figure 2.84: F2 Frequency at V2 Midpoint ($F2_{V2}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{V2}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

shows the consistent place and voicing trends for most obstruents. Further, there are effects of sibilance word-finally that were not present word-medially, as both [s] and [z] exert greater influence on the preceding vowel than their nonsibilant counterparts. This effect does not extend to the inventory data, however, and given that no pure sibilance contrasts occur in VCV position in the lexicon, it is difficult to determine whether this result is in fact prosodically driven, or if it is just an artifact of item differences at the two positions.

Figure 2.86 shows vowel-offset F2 distributions among word-final contrasts, and exhibits substantial place effects on $F2_{VC}$, particularly among voiced plosives and fricatives. These effects are consistent with those reported above for CV and VC transitions intervocalically and word-initially; i.e., among plosives, *labials* $<$ *alveolars* $\leq$ *velars*, while fricatives are more variable in this regard but in general $F2_{VC}$ is higher preceding more posterior constrictions. There are also effects of voicing and sibilance that are consistent with previous results, but these effects are less consistent across databases. Overall, $F2_{VC}$ is similarly discriminative word-finally as $F2_{CV}$ is word-initially, though with a greater dependence on place contrasts than in word-initial position.

2.6.12.4 Summary

Characteristics of the second formant transition in vowels adjacent to obstruent contrasts, though highly variable due to their dependence on the category of the vowel, are fairly consistent across CV, VCV, and VC positions in their differentiation of place and voicing contrasts. These effects—generally lower F2 values in anterior versus posterior constriction contexts, and in voiceless versus voiced contexts—are larger at the VC/CV boundary than at vowel midpoint, though both points in the trajectory exhibit overall contrast effects that are above chance. Further, each of the above patterns is physiologically or acoustically explanatory in that the place effect derives from coarticulatory differences that depend on the articulators used in both consonant and vowel gestures, and the voicing effect derives primarily from the degree to which voiceless obstruent noise obscures part of the F2 transition. Therefore, given its discriminative power, its consistency across contrast positions, and its explanatory value, F2 is an important component of the acoustic

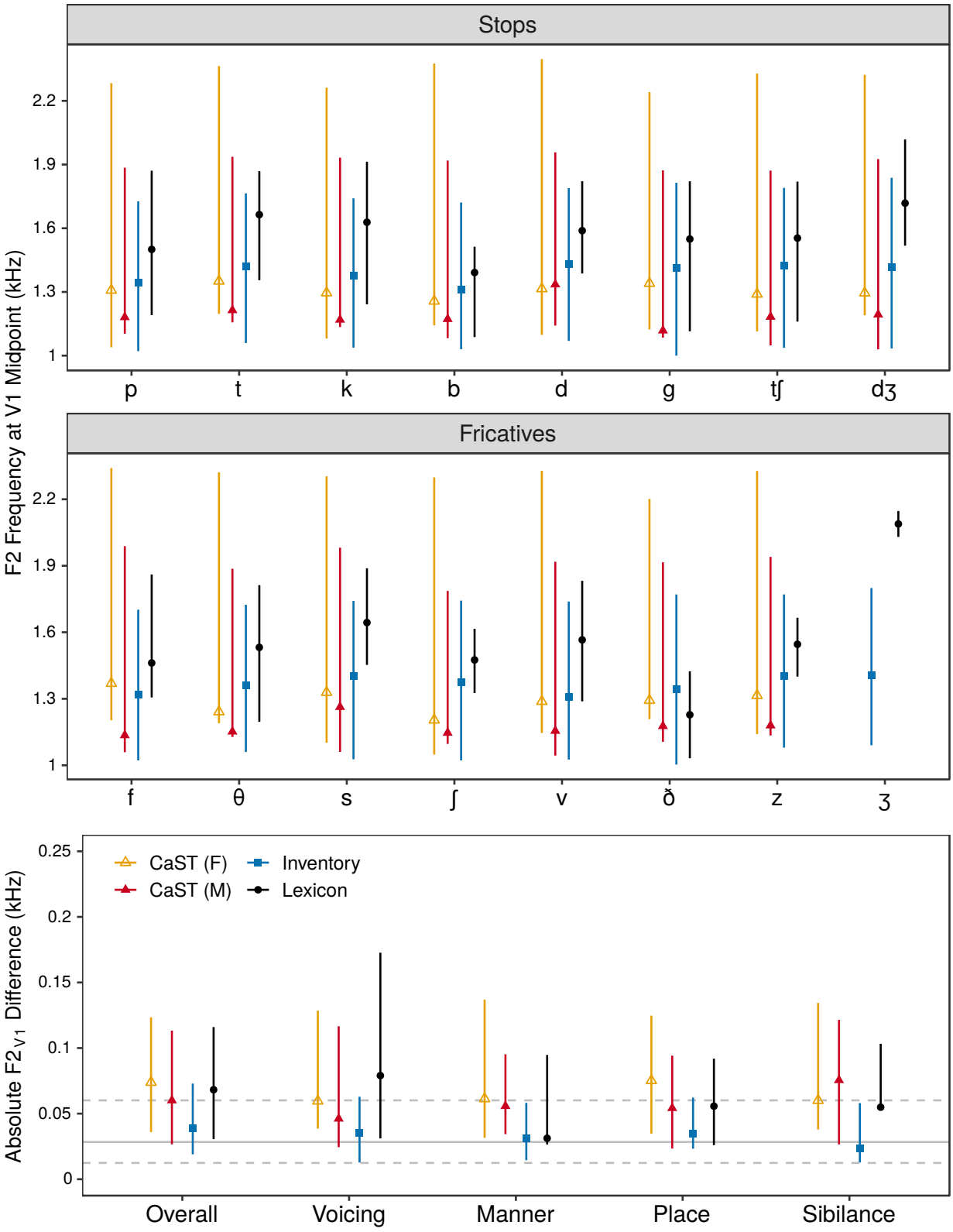

Figure 2.85: F2 Frequency at V1 Midpoint ($F2_{V1}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{V1}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

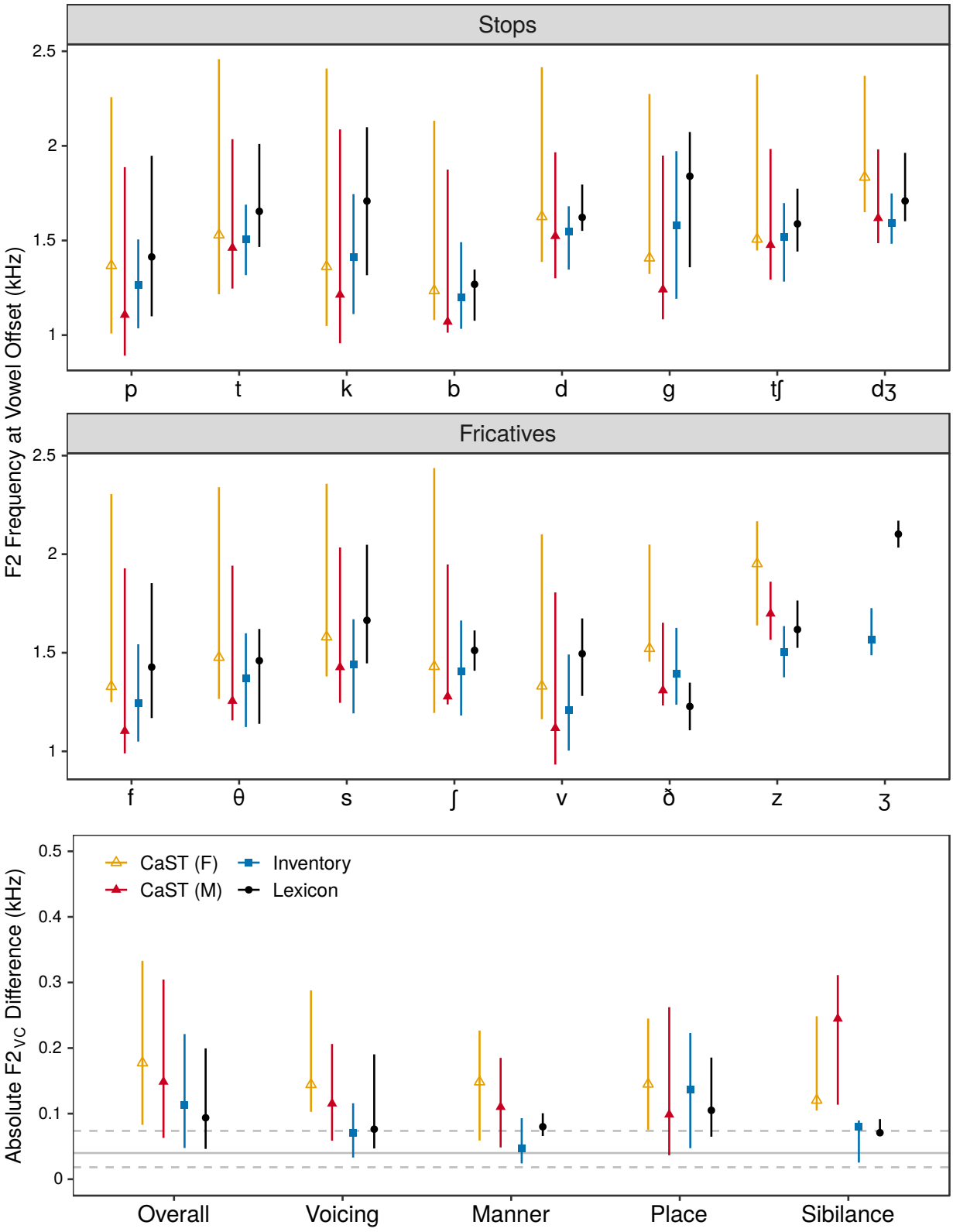

Figure 2.86: F2 Frequency at Vowel Offset ($F2_{VC}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F2_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

description of obstruent contrasts.

2.6.13 Third formant frequency ($F3_{VC/CV}$)

2.6.13.1 Background and physiological basis

The influence of obstruent place of articulation on the third formant frequency at vowel onset/offset is similar to that for F2, but reflects the impact of different vocal tract configurations on the third resonance rather than the second (Öhman, 1966; Fant, 1973; Smits et al., 1996). The most notable place effects on F3 onset are the lowering of F3 due to lip rounding and at velar and postalveolar constrictions (due to the similarity in F2 and F3 frequencies), and the relative raising of F3 at alveolar constrictions, all of which reflect differences vocal tract cavity size and configuration, and cavity coupling as a function of constriction location and degree, consistent with the acoustic theory of speech production in Fant (1960).

2.6.13.2 Definition and measurement

The third formant frequency at vowel offset/onset is defined and measured identically to $F1_{CV/VC}$; namely, the frequency of the third vocal tract resonance is estimated automatically from the Burg formant tracker in Praat (Boersma & Weenink, 2016) and measured at points 10% from vowel offset ($F3_{VC}$) and 10% from vowel onset ($F3_{CV}$), where the percentages refer to time as a percentage of the preceding/following vowel duration. Formant estimation errors were then hand-corrected via visual inspection of the spectrogram, just as in the measurement of F1 and F2 above.

2.6.13.3 Category and contrast distributions

Here we present F3 distributions at vowel offset/onset in the lexicon, inventory, and reference data, with results separated by word-initial, word-medial, and word-final contrast positions.

Word-initial position (CV). Figure 2.87 shows $F3_{CV}$ distributions among word-initial obstruent contrasts, and illustrates that while the discriminative power of F3 is lower overall than F2, there

are some notable distinctions. Foremost among these is the effect of place of articulation on vowel-onset F3. In both voiceless and voiced plosives, velars exhibit the lowest $F3_{CV}$ values, consistent with the 'velar pinch' that is often described in the literature and refers to the simultaneous lowering of F3 and raising of F2 at velar constrictions. This pattern is also consistent with the spectral peak frequency result in Figure 2.35, wherein velars have a large mid-frequency peak due to the merging of F2 and F3. Among fricatives a similar place effect is obtained: F3 is lowered slightly following postalveolars relative to alveolars, with the effect larger in the voiced pair [z, ʒ], just as in the plosive results. This lowering is due in part to the larger sublingual cavity in [ʃ, ʒ] than in [s, z], as the resonance of this cavity corresponds to F3, and thus the larger cavity resonates at a lower frequency, causing a depression of F3 at vowel onset. Other effects present in Figure 2.87 are those due to voicing and manner, but these results are relatively minor and are less consistent across databases.

Word-medial position (VCV). At vowel offset preceding intervocalic obstruents, a similarly robust effect of place of articulation is shown in Figure 2.88, though the pattern among plosives is not consistent with that in word-initial position; namely, [p, b] < [t, d] < [k, g], as compared with the [k, g] < [t, d] pattern word-initially. Among fricatives, however, the same [ʃ, ʒ] < [s, z] pattern is obtained word-medially as word-initially. Voicing is also a factor in $F3_{VC}$ distributions, but is primarily restricted to more posterior articulations, where voiceless postalveolar and velar obstruents exhibit slightly lower F3 frequencies at vowel offset than their voiced counterparts (this effect also extends to alveolar fricatives, but not to alveolar plosives). Because of these place distinctions a modest effect of sibilance (nonsibilants < sibilants) emerges in the inventory, but without pure sibilance contrasts in the lexicon we cannot say whether this result extends to real words, though the category distributions in the lexicon are consistent with such a result.

At V2 onset, the same place relation among plosives is present as at V1 offset: *labials* < *alveolars* < *velars*. Thus, this discrepancy is not due to asymmetries in VC versus CV transitions, but rather is due to differences in plosive production word-initially and intervocalically. Closer

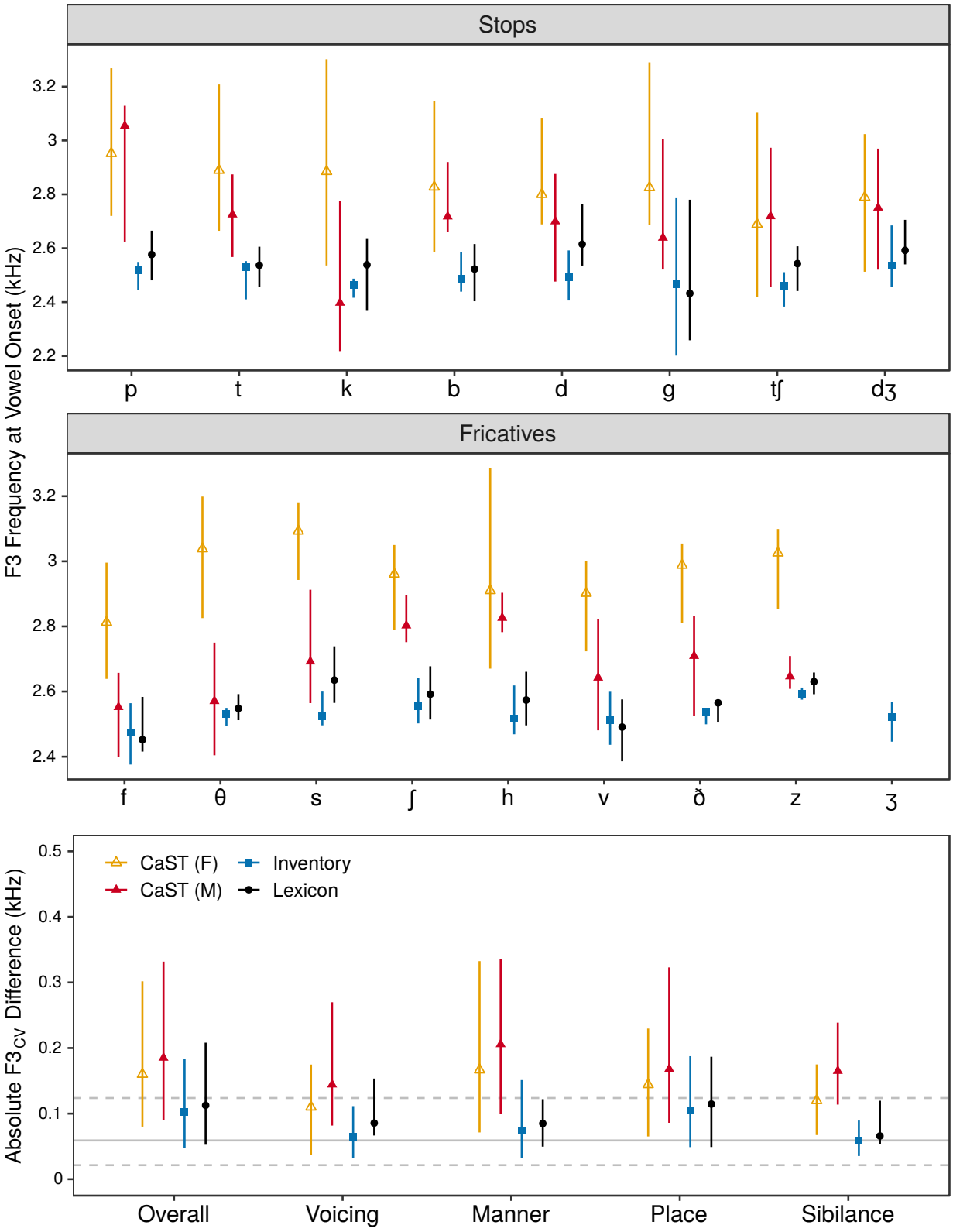

Figure 2.87: F3 Frequency at Vowel Onset ($F3_{CV}$) distributions in CV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F3_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

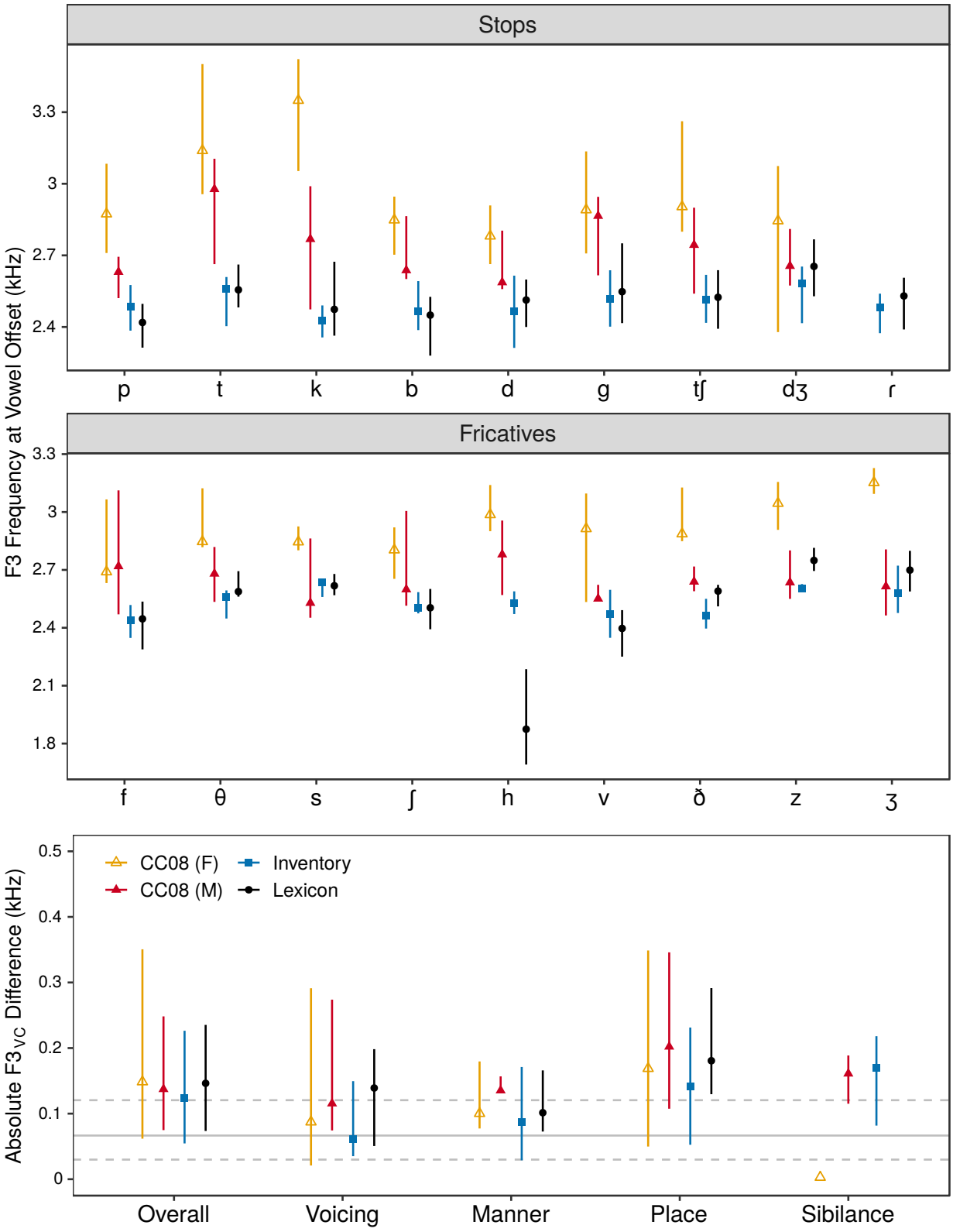

Figure 2.88: F3 Frequency at Vowel Offset ($F3_{VC}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F3_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

examination of the data reveals that the greater propensity to reduce labials and velars to fricatives/approximants intervocalically—a result which is partly reflected in the burst presence rates in Figure 2.24—yields distinct $F3_{CV}$ distributions that cannot extend to the alveolars [t, d], which reduce to [ɾ] categorically via phonological rule. The only other notable featural effect on $F3_{CV}$ is that of voicing, but this effect is in the opposite direction of that observed for VC contrasts; namely, F3 tends to be higher after voiceless obstruents than after voiced obstruents. These differences are quite small, however, and do not extend from the lexicon to the inventory, suggesting this pattern may depend on specific characteristics of vowel contexts such as V_1C[ɚ] and V_1C[ɪ] that are prevalent in VCV contrasts in the lexicon. Overall, both vowel-offset and vowel-onset F3 transitions are informative for intervocalic obstruent contrasts, and though the voicing effect is not consistent between $F3_{VC}$ and $F3_{CV}$, the place effect is.

Word-final position (VC). Figure 2.90 shows vowel-offset F3 distributions in word-final position. Overall, $F3_{VC}$ is less informative word-finally than word-medially. Further, while there is a modest effect of place of articulation in the lexicon, distinctions among obstruent places are quite narrow and inconsistent between the lexicon and inventory, with the inventory data largely reflecting the place relations word-initially, while the lexicon shows distinct patterns among voiced plosives and sibilant fricatives that leads to an overall less clear picture for F3 and place of articulation word-finally. Effects of sibilance, manner, and voicing are all close to chance levels, with voicing the largest of the three but also inconsistent between the lexicon and inventory data.

2.6.13.4 Summary

The third formant frequency at vowel offset/onset, though less informative and consistent in its patterning than F2, does provide cues to obstruent place of articulation that are complementary to the F2 results in the previous section. In particular, F3 provides further information about the plosive contrasts [t, k] and [d, g], as well as the sibilant fricative contrasts [s, ʃ] and [z, ʒ], that are more similar in their formant transitions than they are to their labial counterparts. Other effects

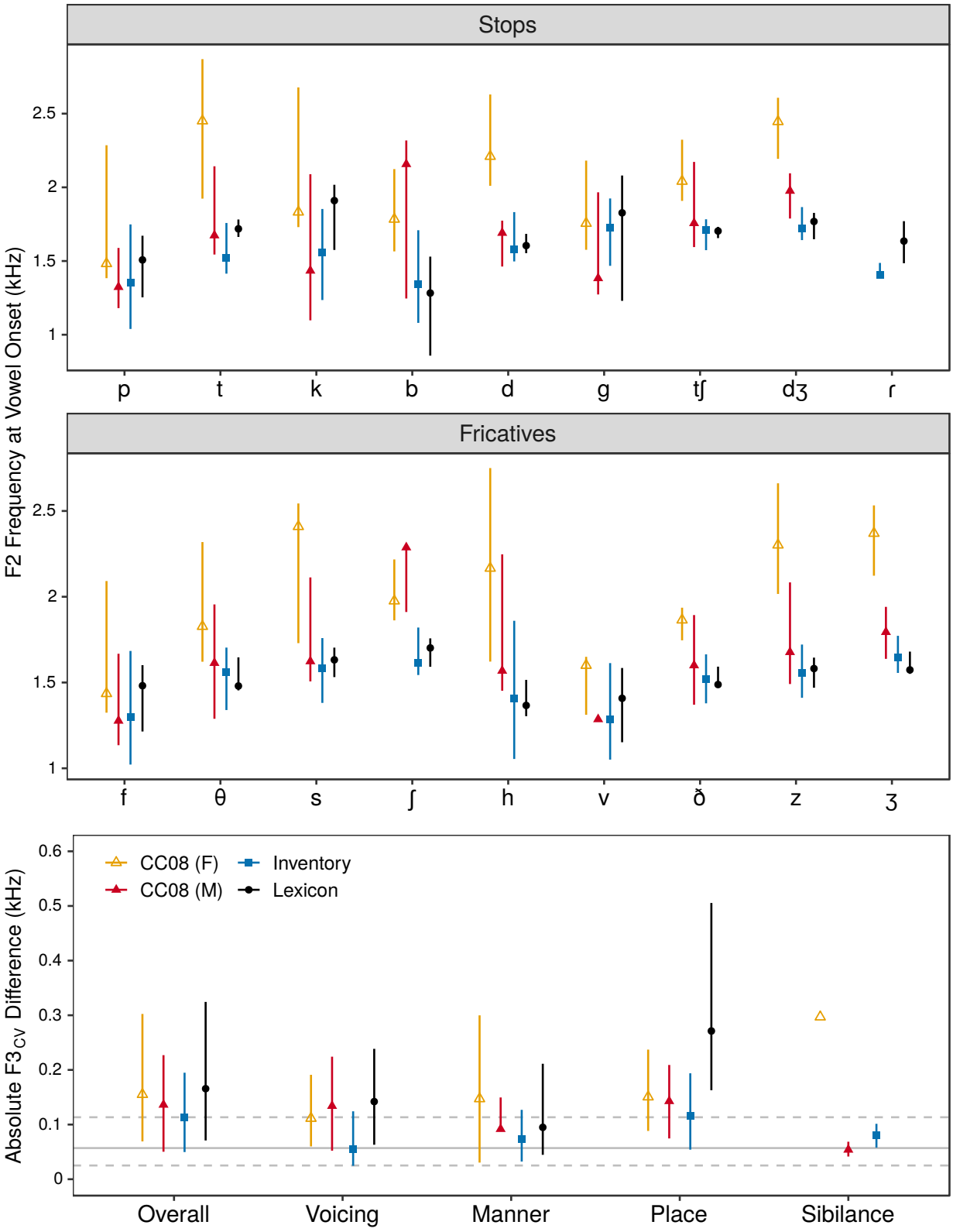

Figure 2.89: F3 Frequency at Vowel Onset ($F3_{CV}$) distributions in VCV position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F3_{CV}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

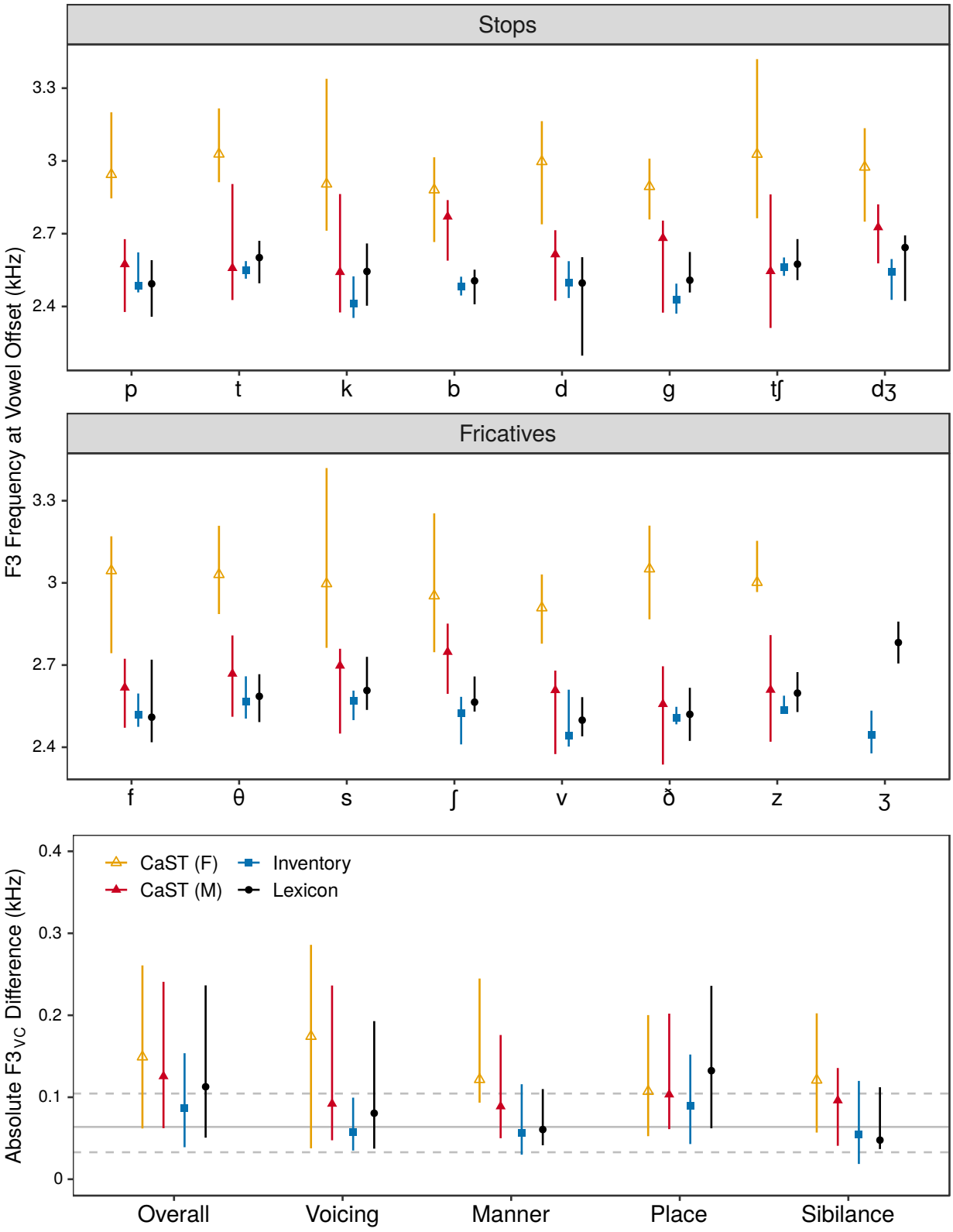

Figure 2.90: F3 Frequency at Vowel Offset ($F3_{VC}$) distributions in VC position. The top two panels show category medians and IQRs. The dashed gray lines indicate medians and IQRs of within-item differences in $F3_{VC}$ in the inventory data (between reps 1 and 2) and serve as a reference for potential chance effects.

of voicing, however, were less reliable, both between VC and CV transitions, and as a function of contrast position. And thus, F3 is considered primarily a cue to obstruent place of articulation, though other features may impact the F3 transition through their influence on the point of vowel onset/offset given that variation in both laryngeal and supralaryngeal source characteristics affects properties of the spectrum—particularly upper formant amplitudes—that define consonant-vowel boundaries. In future work we aim to define parameters that are less dependent on discrete segmental boundaries, and thus may more reliably tied to the articulatory characteristics that define much of the feature system.

2.6.14 Comparative discriminative power of spectral parameters

As with the temporal and amplitudinal parameters, we conclude this section with a comparison of the overall discriminative power of each spectral parameter as a function of contrast position and data source. Figure 2.91 shows normalized contrast effects—the mean of the absolute value of itemwise contrast differences between each scaled (transformed to range between 0 and 1) parameter—of each spectral parameter in the first set (spectral peak frequency – low frequency energy) in both target and reference data sets in CV, VCV, and VC positions. Figure 2.91 shows a remarkable similarity in parameter rankings between the four databases, with the target data particularly closely matched. Spectral peak frequency and consonantal spectral tilt are generally the most robust cues to obstruent contrasts in all three positions. Other parameters of relatively high discriminative power are the low-frequency energy in the consonant noise interval (LF), which is highly ranked across positions, and the spectral dispersion of the consonant noise spectrum ($DISP_C$), which is most discriminative intervocalically.

However, unlike the amplitudinal and temporal parameters, there is much greater parity between spectral parameters in terms of their overall utility in distinguishing obstruent contrasts. With the exception of the vowel spectral tilt parameters, the majority of parameters exhibit normalized contrast effects between 0.2 and 0.25. Finally, regarding discrepancies between the inventory and lexicon data, the largest differences in discriminative power occur word-finally, where

in the lexicon, for example, AMP_{DYN} and $TILT_C$ are equal to or better than $FREQ_{PK}$, while the relative power of these parameters in the inventory is notably reduced. Conversely, SHAPE and $DISP_C$ have relatively higher weight in the inventory than in the lexicon. In the next section, where parameter weights are studied as a function of their utility in a statistical model of acoustic cue integration, we will discuss the extent to which such discrepancies my be due to distributional differences in the occurrence of different obstruent contrasts in the lexicon.

Figure 2.92 shows the overall discriminative power of the second set of spectral parameters (relative F3 amplitude – third formant frequency). Here we see that in general the pitch and formant cues from adjacent vowels are less informative than most characteristics of the noise spectrum, as the former range for the most part between 0.05 and 0.15 in normalized contrast effects, while —AMP_{F3} and AMP_{F5}—generally range between 0.2 and 0.25, similar to the majority of parameters in Figure 2.92. Among the pitch and formant cues, f0 is highly informative in the lexicon, though intervocalically vowel-offset pitch ($f0_{VC}$) is much more discriminative than vowel-onset pitch. These effects are reduced somewhat in the inventory but nevertheless pitch remains relatively more informative than the vowel formant frequencies. Here it is worth noting that these results appear to run counter to those in the f0 and F1–F3 sections above, which reflects in part the fact that we have no measure of estimated chance distinctions in Figure 2.92. For this reason the summary results in this section should be considered alongside the results in the previous sections and the parameter weights derived from the cue-integration model in the next section.

F2 and F3 are similarly informative, followed closely by F1 in CV and VCV positions, though word-finally F1 plays a greater role. This result is predictable given that F2 and F3 primarily cue place distinctions, while F1 primarily cues voicing, and the former plays a greater role than the latter, particularly word-initially and word-medially, due to the more frequent occurrence of within-voicing contrasts both in the lexicon and inventory. Finally, there are few notable discrepancies in discriminative power between the lexicon and inventory data. The one exception is $f0_{CV}$ in word-initial position and $f0_{VC}$ in word-medial position, which are much less discriminative of obstruent contrasts in the inventory than in the lexicon. However, this result is not due to differences in

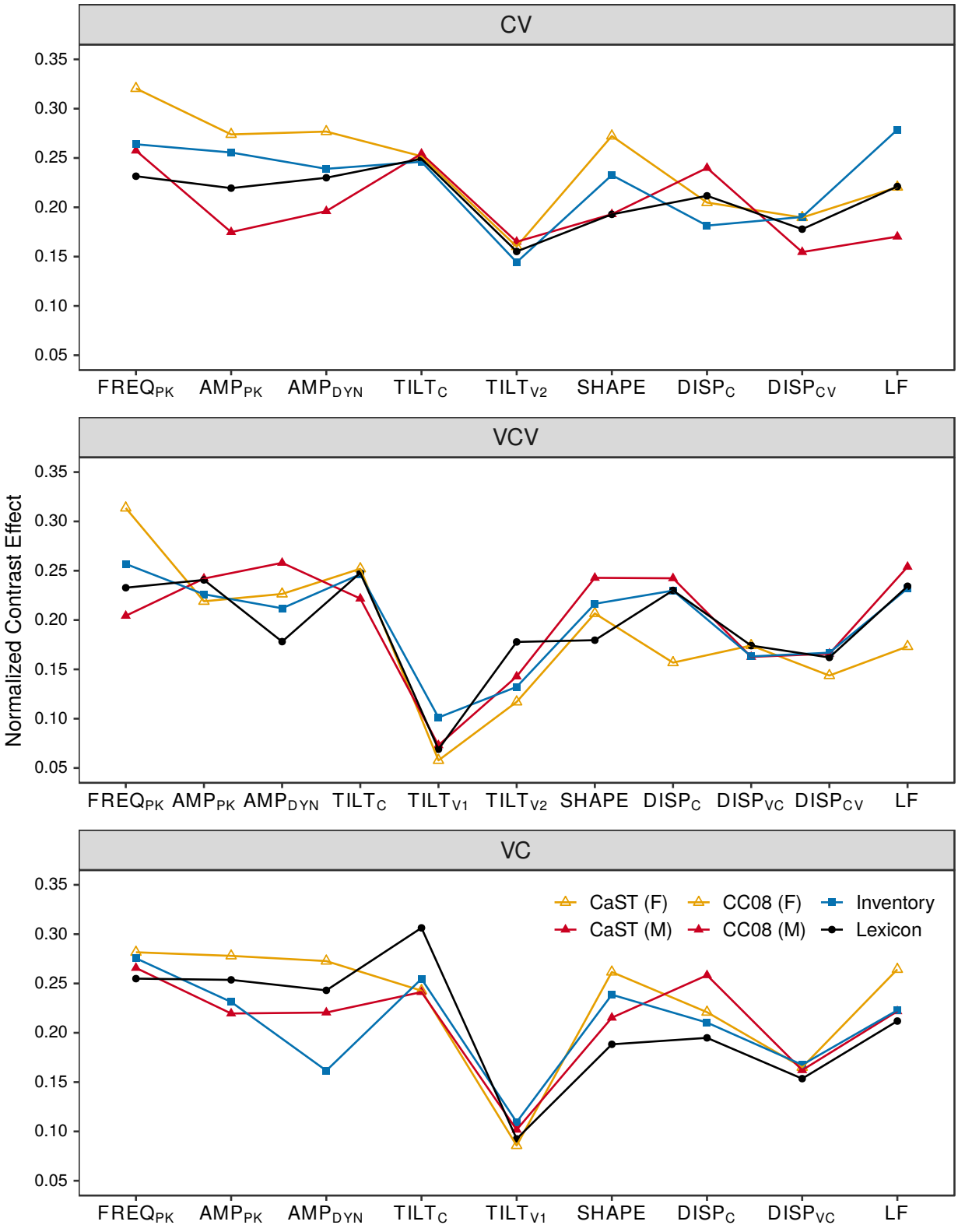

Figure 2.91: Comparative discriminative power of spectral parameters (Set I) in CV, VCV, and VC positions, as measured by the *normalized contrast effect*, the mean of the absolute differences by contrast between parameters that have been scaled to range between 0 and 1. The CaST and CC08 reference data sets share the same color palette because their contexts (CV/VC and VCV, respectively) are mutually exclusive.

voicing contrasts in the two databases, but rather reflects the fact that f0 also varies in the lexicon as a function of manner of articulation, broadening the scope of f0 beyond voicing contrasts alone.

2.7 Discussion

In addition to providing a comprehensive picture of the acoustic characteristics of obstruent contrasts as a function of position (CV, VCV, VC), stimulus type (syllable, word), and system structure (homogeneous inventory, heterogeneous lexicon), this chapter reveals several important aspects of the acoustic scaling problem between the inventory of phones and the system of lexical distinctions those phones serve to encode. First, despite some discrepancies in the two target databases, the category distributions for most acoustic parameters are closely matched between those derived from controlled syllables and those derived from real words. This result is in one respect trivially expected given that both syllable and word transcriptions are based on the same auditory assessment procedure, and thus a [b] in the lexicon cannot deviate too much from the inventory data or it would not have been labeled a [b] in the first place. However, the analysis of contrast distributions does not exhibit the same vulnerability, as it operates over larger classes of items, either featural distinctions, or undifferentiated contrasts where the only information required is that a given interval is an obstruent consonant and that it differs perceptually and articulatorily from a matched interval in another item. Parameters of note in this set are consonant/noise duration (DUR_C, DUR_N), consonantal spectral tilt ($TILT_C$), relative F3/F5 amplitude (AMP_{F3}, AMP_{F5}), and F1 at vowel onset/offset ($F1_{VC}$, $F1_{CV}$).

Second, where discrepancies arose in the behavior of a given cue in the inventory and lexicon, in many cases the cue was more discriminative in the real-word contrasts than in contrasts between controlled syllables, a result which runs counter to expectations of differences in hyperarticulation between the two data sets. However, hyperarticulation not only provides an enhancement of acoustic contrast in many cases, it can also flatten out distinctions that might otherwise be informative, such as subtle amplitudinal or temporal distinctions, or information from coarticulation with the preceding/following vowel that is reduced when hyperarticulated. Finally, the effect of position is

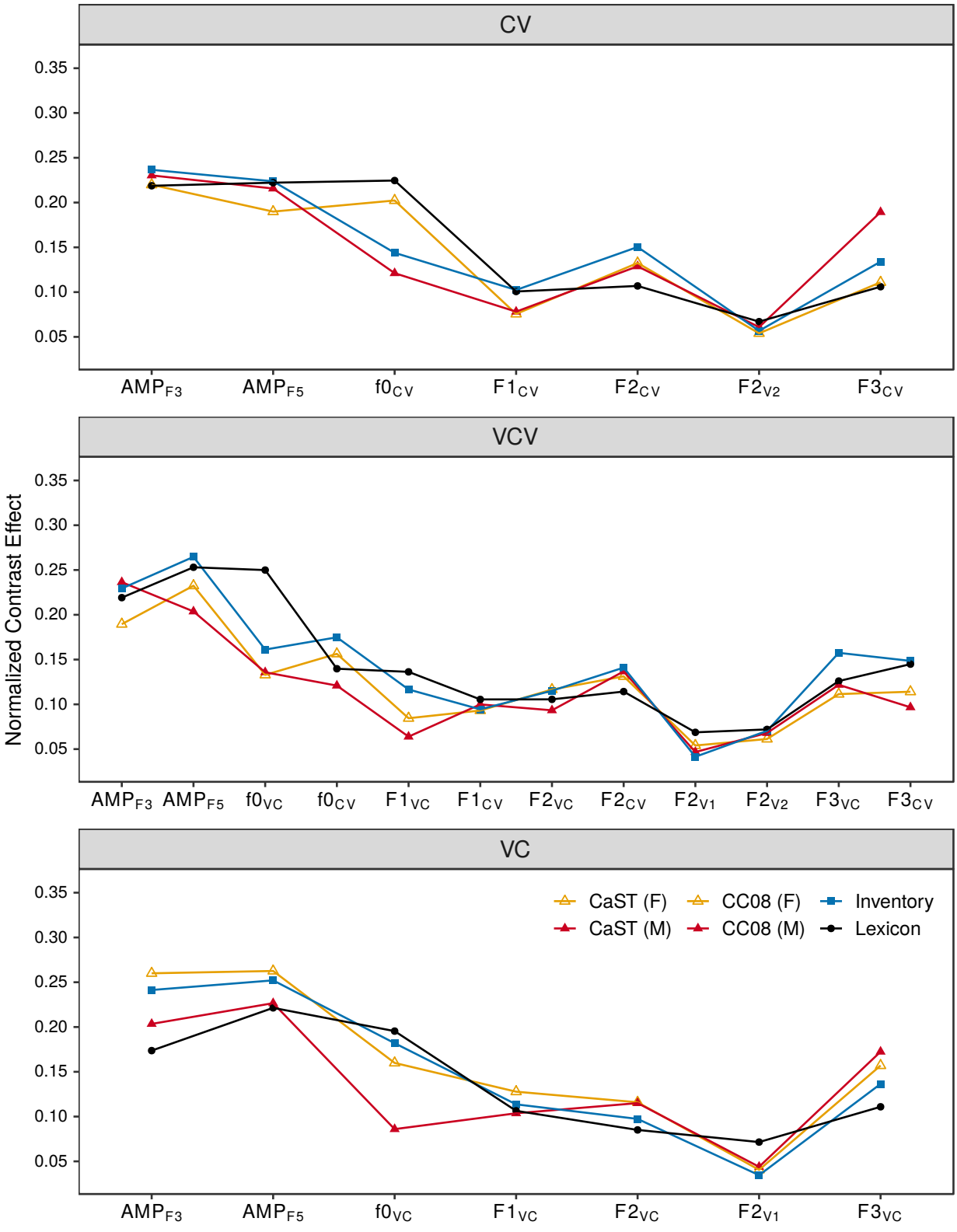

Figure 2.92: Comparative discriminative power of spectral parameters (Set II) in CV, VCV, and VC positions, as measured by the *normalized contrast effect*, the mean of the absolute differences by contrast between parameters that have been scaled to range between 0 and 1. The CaST and CC08 reference data sets share the same color palette because their contexts (CV/VC and VCV, respectively) are mutually exclusive.

an important part of the scaling problem, as word-final contrasts are notably more constrained in the lexicon than in the inventory. This result emerges in relative differences in acoustic agreement between VC and CV positions, and will likely only further diverge in perception.

Chapter 3

Lexical contrast perception

Outline

This chapter describes results of a perception experiment measuring listener word recognition patterns in closed-class (2AFC) identification. In analyzing the results of Experiment 1 we focus primarily on listener accuracy and error distributions as a function of obstruent categories and contrasts in the lexicon, while also accounting for more global factors such as noise level, word length, and word frequency in modulating contrast recognition. This information is critical to the later interpretation of both statistical models of cue integration in speech perception (Chapter 4), and the overall structure of the system of contrast in the lexicon and its response to perturbations of the acoustic signal by noise or cue loss (Chapter 5). Thus as in Chapter 2, this chapter is primarily descriptive, but comprehensive descriptions of obstruent recognition patterns throughout the lexicon are necessary for the development of our approach because such data are sparsely represented in the literature, and where present the analysis is often restricted to a narrow set of contrasts and items that may not be scalable to the lexicon as a whole.

3.1 Introduction

3.2 Pilot experiment: Word recognition by Canadian listeners
A potential confound is addressed regarding the pairing of an American English listener population and stimuli produced by a Canadian English speaker. The behavior of both Canadian and American listener groups was highly correlated.

3.3 Experiment 1: Closed-class recognition
Listener responses to a two-alternative forced choice (2AFC) task between obstruent minimal pairs are analyzed to provide upper-bound estimates of the discriminability of lexical contrasts and to serve as a behavioral benchmark for models of contrast discrimination and system structure in Chapters 4–5.

3.4 General discussion

3.1 Introduction

Chapter 2 provided information on the acoustic properties underlying the obstruent contrasts critical for the maintenance of lexical distinctions in English. This ensemble of parameters, when structured to optimize discrimination of lexical contrasts, may be viewed as comprising an upper limit on the information available to listeners in the acoustic signal. Yet, such an approach is inadequate for any complete account of the structure of the speech system, because it assumes the encoding of linguistic information in the acoustic signal is independent of any constraints that may be imposed by listeners, such as their cumulative language experience (e.g., they may have learned to attend to some cues more than others, based on their utility in communication, or their robustness to noise or inter-speaker variability) or general auditory/neural physiology (some acoustic properties exceed the spectral, temporal, or amplitudinal processing capabilities of the ear/brain). The present chapter addresses this limitation by using data on listener perception of lexical contrasts to build more ecological models of acoustic cue integration, models which will ultimately inform our understanding of the structure of the phonetic system and the encoding of higher-order units in the acoustic signal.

In predicting listener behavior we aim to approximate the relative perceptual weight listeners place on a given property of the acoustic signal. And through the aggregation of these weighted cues, we aim to identify and explain general patterns in the resilience (or vulnerability) of different obstruent contrasts to perturbation by background noise, both acoustically, and as a function of the real words they serve to distinguish. Three experiments are reported in this chapter, each aimed at providing a distinct window on listener perception and word recognition.

Experiment 1 employs a two-alternative forced-choice (2AFC) task, where the two alternatives are members of a minimal pair where the critical contrast distinguishing the two items is between obstruent consonants. This experiment was conducted twice, with the two sub-experiments (a/b) being identical in all respects except in the item and participant sets. This sub-experiment structure serves primarily as an internal replication to test the degree to which estimates on one subspace of the lexicon can scale to another, and ultimately to the lexicon as a whole. Because of the narrow

choice set presented to listeners (the theoretical minimum), estimates of contrast discriminability, and the acoustic properties that underly such decisions, will serve as a lower bound on the ultimate confusability of phonetic contrasts and the utility of acoustic cues in speech communication, while also providing an upper bound on the confusability of particular lexical contrasts (specific pairs of words), as by restricting the set we are shifting some of the error probability to a single competitor. That is, while some errors in a more naturalistic open-class recognition task will shift to the target—e.g., a [ʃ]-like token of a [s]-onset word might result in a misperception if there is a [ʃ]-onset neighbor of the target word (e.g., *sell* vs. *shell*), but not when the alternative provided in the task is more distinct from [s] than [ʃ] is (e.g., when the competitor to *sell* is *fell*)—others will shift to the competitor, especially when lexical factors such as word frequency are taken into account. Therefore, in general we expect the number of errors on a given minimal pair to increase in this closed-class task relative to a more naturalistic open-class task, whereas for a given phonetic contrast such as [s, ʃ] we expect an open-class task to elicit a greater number of errors (i.e., the confusability in the 2AFC is a lower bound) because the lack of a constrained choice set presents more opportunities for such confusions to arise.

3.2 Pilot experiment: Word recognition by Canadian listeners

Given that the speaker-internal system under study in this thesis is based on a large database of recordings from a native speaker of Western Canadian English, and the perception experiment introduced above is of American English listeners responding to stimuli from this speaker, we first obtained pilot open-class recognition data from Canadian listeners to determine the validity of employing this cross-dialectal design. A comparable open-class perception study was then run with American listeners for comparison, though analysis of data from this experiment is restricted to this section as the open-class data does not provide the experimental control on contrast positions, word length, and choice set that is required for the detailed modeling of cue integration in Chapter 4.[1]

[1] It should be noted that both choices—Western Canadian English for production and American English for perception—were made based on convenience, as the single-speaker database developed for the MALD project is the only one of its kind, and the large volume of perception data required for this study could more feasibly be collected at our

	monosyllabic	disyllabic	trisyllabic	Total
CV	40	30	10	80
VCV	NA	65	15	80
VC	50	20	10	80
Total	90	115	35	240

Table 3.1: Distribution of items in the Pilot Experiment by Contrast Position (CV, VCV, VC) and Word Length (mono-, di-, and tri-syllabic).

3.2.1 Methods

3.2.1.1 Participants

Forty native speakers of Canadian English were recruited from the University of Alberta student population for participation in the experiment. Participants were all volunteers, and received either $10 CAD or course credit from the Department of Linguistics as compensation. All participants were administered a language background questionnaire prior to the experiment, and those reporting speech or hearing impairments, non-native speakers of English, and simultaneous bilinguals were excluded from the study.

3.2.1.2 Materials

A list of 240 words was compiled from the MALD lexical database such that all words participated in at least one minimal pair obstruent contrast in word-initial (CV), word-medial (VCV), or word-final (VC) position, and items were between one and three syllables in length. Items were evenly divided between the above three positions—i.e., 80 CV, 80 VCV, 80 VC—and otherwise approximately representative of the word length distribution in the lexicon. See Table 3.1 for a full breakdown of items according to contrast position and word length.

Audio stimuli were drawn from the MALD database (see the description of recording conditions in Section 2.3) and normalized to 70 dB mean amplitude. The background noise, six-talker babble, was then created by randomly selecting three male speakers and three female speakers

home institution in Lawrence, Kansas.

from an eight-speaker (4F) corpus of Canadian Broadcasting Corporation (CBC) radio broadcasts compiled specifically for this study. From each speaker, a random one second interval was selected and normalized to 70 dB. All six samples were then combined and re-normalized to 70 dB. The target word was then padded on either side with silence to match the one-second duration of the noise, following which the two were combined in ratios creating final stimuli at 0 dB and +5 dB signal-to-noise ratios (SNRs). These SNRs were chosen based on a pilot experiment targeting an overall accuracy of approximately 50%. Finally, stimulus amplitudes were manipulated to ramp up linearly from zero over the initial 200 ms (a noise-only interval preceding the onset of the target word) and ramp down to zero over the final 200 ms.

3.2.1.3 Procedure

The experiment was prepared in E-Prime (Psychology Software Tools), and consisted of a five-word practice block, followed by four blocks of 60 trials each. In between blocks listeners were given a one-minute break, and in total the experiment took around 20 minutes. Trial order was randomized by participant, as was the pairing of words and noise levels; i.e., each participant received a different set of 120 words at 0 dB SNR and 120 at +5 dB SNR. On a given trial, listeners heard a stimulus word over headphones, after which they were shown a prompt into which they were to type whichever word they thought they heard in the stimulus. Listeners were instructed that the target was a single English word, but otherwise there were no constraints on how they could respond. No time pressure was applied to the task, and trials proceeded one second after a response was completed on the previous trial (i.e., ITI = 1 sec). Listeners completed the experiment while seated in a sound-attenuated booth in the Alberta Phonetics Lab, and viewed the experimental presentation through a window that allowed them to see the display of a computer monitor located outside of the booth.

3.2.2 Results

Canadian listener perception was used to validate the perception of Canadian speech by American listeners by comparing the pilot response set to responses in a comparable experiment conducted with American listeners at the University of Kansas. These experiments will be referred to hereafter as Experiments 0a and 0b, respectively. The two experiments are not identical, however, primarily because Experiment 0b is part of a future study on the relationship between open- and closed-class recognition in models of cue integration and obstruent system structure. The number of stimuli in Experiment 0b was expanded from 240 to 300 because the average duration of Experiment 0a was under 30 minutes, meaning that more stimuli could be accommodated while adhering to the constraint that the experiment be under 45 minutes in length. In order to facilitate comparison between both experiments, the same set of 240 stimuli from Experiment 0a was included in Experiment 0b and presented over the first four blocks.

Further, different SNRs were used in the two experiments, because the SNRs in Experiment 0b were set to partially match the SNRs in Experiment 1, which were set at $+2$ and -2 dB in order to yield an average Experiment 1 accuracy between 70 and 80%, with a 10% difference between the lower and higher SNR. Therefore, Experiment 2 used SNRs of $+2$ and $+6$ dB, where the lower SNR matches the higher SNR of Experiment 1 (this matching of lower with higher is necessary because the task in Experiment 0b is harder than that in Experiment 1), and the difference in SNRs is the same in both experiments: 4 decibels. Thus Experiments 0a and 0b have slightly different SNRs, so we expect a greater difference in responses at the lower SNR (0 dB vs. +2 dB) than at the higher (5 dB vs. 6 dB), and an overall lower accuracy on Experiment 0a relative to 0b. Nevertheless, our primary interest in comparing the two experiments is that the general accuracy distribution and error patterns are similar, suggesting the dialect mismatch between speaker and listener in Experiments 1–2 does not present a major confound in the analysis of the data therein.

Beginning with the analysis of listener accuracies on shared stimuli between the Experiments 0a and 0b, the overall accuracy of Canadian listeners (43%; 29% at 0 dB, 57% at 5 dB) was less than that of American listeners (62%; 53% at 2 dB, 71% at 6 dB), though accuracies on the higher

SNRs (5/6 dB; accuracy difference of 14%) were more similar than were the lower SNRs (0/2 dB; accuracy difference of 24%), both consistent with expectations. Itemwise accuracies between the two experiments, however, were significantly correlated (Kendall's $\tau = 0.49$, $z = 10.82$, $p < 0.001$), and were consistent at both low ($\tau = 0.44$, $z = 9.223$, $p < 0.001$) and high ($\tau = 0.44$, $z = 9.472$, $p < 0.001$) SNRs. That is, items that the Canadian listeners in the were generally better at recognizing in Experiment 0a were also the items that American listeners recognized well in Experiment 0b; similarly, items that were poorly recognized by Canadian listeners generally showed lower accuracies when presented to American listeners.

Regarding error patterns, 43% of the most common errors made by Canadian listeners (51% at the higher SNR, 37% at the lower SNR)—for each stimulus, the non-target word that listeners most often gave as a response—were shared as the most common errors among American listeners. Considering a wider response set of the top two errors per stimulus, there was a 60% overlap between the two experiments (i.e., for a given stimulus, one or both of the top two errors from Exp. 0a were present in the top two errors on that stimulus in Exp. 0b), 70% at the higher SNR and 50% at the lower SNR. See Table A.1 in the appendix for a full listing of accuracies and common errors by stimulus in Experiments 0a and 0b.

3.2.3 Discussion

The pilot experiment, which examines word recognition from listeners at the University of Alberta who were familiar with the Western Canadian dialect of the target speaker, in showing comparable response patterns to those recorded from American listeners at KU, confirms the validity of the cross-dialectal structure of Experiments 1–2. The distribution of accuracies by item correlates significantly between the two experiments, and many of the most common errors are shared between the two listener groups. This result, that the two populations show similar behavioral patterns indicative of a shared perceptual system, is necessary for the analyses below, and was entirely in line with expectations based on several facts about the speaker and listener speech communities.

First, the speaker's dialect, Western Canadian English, differs very little from the Midwestern

American English dialect of our listeners at KU. The primary phonetic differences are restricted to the low/mid-back vowel space, and further, such differences are variable across speakers and items (Boberg, 2008a,b). And while differences are expected in some aspects of the vocabularies of the two groups, including differences in the relative familiarity and frequency of usage of different words, such differences are not expected to be substantially different from those that might be observed between different regions of the United States. Finally, when participants were asked following the experiment whether they noticed anything about the speaker's accent, no participant perceived that the accent was any different from their own, and in particular no one noticed that the speaker was Canadian. The dialect difference even evaded for several weeks the notice of two undergraduate research assistants tasked with annotating a subset of the stimuli in Praat, a task which involved listening carefully to hundreds of words from the same speaker, all presented in clean listening conditions; i.e., without the background noise obscuring stimuli presented to participants in the experiment. Therefore, we are confident that any perceptual patterns observed below represent significant features of the acoustic-perceptual structure that defines the system of contrastive information in the English lexicon, and are not an artifact of a dialectal mismatch between speaker and hearer.

3.3 Experiment 1: Closed-class recognition

As noted above, the goal of the first experiment, in using a closed-class recognition task, is to obtain a lower-bound estimate on the confusability of different phonetic contrasts in the lexicon, and an upper-bound estimate on the confusability of particular minimal pairs. By constraining listener choices to just two options—a minimal pair contrasting in obstruent phones at a specific position (CV, VCV, VC), and for words of a particular length—we are able to determine the minimal acoustic information separating that pair of words that listeners are able to pick up on, all while accounting for the inherent role of lexical biases (e.g., the relative frequencies of the two items) in modulating listeners' choices in word recognition.

3.3.1 Methods

3.3.1.1 Participants

Eighty native speakers of American English were recruited from the University of Kansas student population for participation in the experiment. Forty participants were assigned to Experiment 1a, and 40 to 1b. Participants received either $10 USD or course credit from the Department of Psychology as compensation for their time. All participants were administered a language background questionnaire prior to the experiment, and those reporting speech or hearing impairments, non-native speakers of English, and simultaneous bilinguals were excluded from the study.

3.3.1.2 Materials

A total of 960 minimal pairs were identified from the model lexicon whose single point of contrast was in a difference between obstruent consonants. Items were evenly divided in the position of this contrast between word-initial (CV), word-medial (VCV), and word-final (VC) contexts, and were otherwise distributed between mono-, di-, and tri-syllabic lengths approximating that observed in the lexicon.[2] Table 3.2 displays the number of minimal pairs in Experiments 1a and 1b by Position and Word Length. These items were chosen from a supervised random selection procedure where items meeting a given position/length requirement were first randomly drawn from the database, and then items were removed and replaced based on the following factors: (1) speech errors or productions inconsistent with the transcription defining the minimal pair; (2) status as a minimal-pair neighbor of another item in the same experiment; and (3) existing presence in the item set, where repetition was not permitted within an experiment (1a or 1b), and was only allowed across experiments when the configuration of minimal pairs made it impossible to avoid, usually due to violation of point (2). In total, 29 items needed to be repeated, reducing the expected count of 1,920 items comprising 960 minimal pairs to 1,891. However, all 960 minimal pairs were distinct, and no item or minimal-pair neighbor was repeated within a given experiment.

[2]Due to a technical error in the stimulus selection, Exp. 1b had 160 CV, 161 VCV, and 159 VC items, as compared with 160 each for Exp. 1a. This discrepancy is reflected in Table 3.2, but should not impact our analysis.

	CV			VCV			VC			
	mono	di	tri	mono	di	tri	mono	di	tri	Total
Exp. 1a	87	59	14	–	137	23	114	36	10	480
Exp. 1b	89	56	15	–	135	26	111	38	10	480
Total	176	115	29	–	272	49	225	74	20	960

Table 3.2: Distribution of minimal pairs by Position (CV, VCV, VC), Length (mono-, di-, and tri-syllabic), and Experiment (1a, 1b).

Finally, regarding the selection of items, it is worth noting that the position of contrasts in the lexicon is not evenly distributed between CV, VCV, and VC contexts (in the database, items in CV, VC, and VCV contrasts occur in a ratio of approximately 6:3:2, respectively). However, each position is frequent enough that it was determined balance was required to obtain sufficient data for an adequate model of acoustic cue integration, as position is a major factor in constraining the set of cues available to the listener.

Audio stimuli were created by first selecting each target word from the database and normalizing its mean amplitude to 70 dB. The background noise, six-talker babble, was then created by randomly selecting three male speakers and three female speakers from an eight-speaker (4F) corpus of CBC radio broadcasts compiled specifically for this study. From each speaker, a random one second interval was selected and normalized to 70 dB. All six samples were then combined and re-normalized to 70 dB. The target word was then padded on either side with silence to match the one-second duration of the noise, following which the two were combined in ratios creating final stimuli at +2 dB and -2 dB SNR. These SNRs were chosen based on a pilot experiment targeting between 70 and 80% accuracy overall, and an approximate 10% difference in accuracies between the two SNRs.

For each listener and each item a unique set of background noise was created (i.e., the multi-talker babble varied randomly by listener and item in both the talkers that it contained, and the stretches of speech that were selected from each talker). Further, a unique set of items were played at a given SNR, such that no two speakers heard the same items in the same noise or at the same SNRs. All listeners heard 240 items at -2 dB and 240 items at +2 dB, and their occurrence

throughout the experiment was random. Finally, each word at a given SNR was responded to by 10 listeners, meaning that each minimal pair at a given SNR was responded to by 20 listeners, with 40 responses recorded in total for each minimal pair.

3.3.1.3 Procedure

Each experiment (1a, 1b) adopted the same procedure. Participants were seated in front of a computer monitor in separate cubicles in the subject testing room of the KU Phonetics & Psycholinguistics Laboratory. The experiment was run in Paradigm (Perception Research Systems) and proceeded as follows. On a given trial, a noise-masked word was presented binaurally over headphones, after which two words appeared on the screen, one being the target word and the other its minimal pair competitor. Each word was associated with left and right buttons on a button box, corresponding to the placement of the words on the screen, and participants were instructed to push the button corresponding to the word they heard (screen position / button order was counterbalanced across participants). No time pressure was applied to this choice, and after selecting an option the next trial began. Listeners were instructed to guess in cases where they were unsure or did not perceive a word in the stimulus.

Ten practice trials were given before the main experiment, all using minimal pairs distinct from the experimental items and not exhibiting obstruent consonant contrasts, though they were otherwise similar in exposing listeners to words of mono-, di-, and tri-syllabic lengths. The 480 experimental trials were divided into five blocks of 96 items each. Between blocks, participants were given up to a 1 minute break, though they were able to start the next block whenever they were ready. In total, the experiment took between 35 and 45 minutes.

3.3.2 General word recognition factors

The goal of this section is to describe more general factors responsible for word recognition performance in Experiment 1, outside of the specific phonetic category/contrast results which are the focus of the remaining sections; namely, how is listener word recognition affected by different

characteristics of the stimuli such as background noise level, word length, and word frequency?

3.3.2.1 Noise Level

The 4 dB difference in SNRs employed in Experiment 1 (+2 dB vs. −2 dB) resulted in an average accuracy difference of 11%, with stimuli at +2 dB SNR recognized correctly approximately 86% of the time (86.5% in Exp. 1a, 86.2% in Exp. 1b), as compared with 75% at −2 dB (75.4% in Exp. 1a, 73.8% in Exp. 1b). This difference was by design, having been determined in initial pilot testing, so the consistency reported above merely confirms the manipulation retained the planned effect throughout both experiments. Further, the fact that the inter-quartile ranges (IQRs) are between 6 and 10% for each noise level, indicating little distributional overlap between the two groups, confirms this difference is significant and would have had the desired effect of disrupting to some extent listeners' ability to accommodate to the background noise as the experiment progressed.[3]

3.3.2.2 Word Length

The overall effect of word length on listener word recognition in Experiment 1 was minor but significant, showing a moderate decline in accuracy with increasing word length, with listeners at 82% accuracy on monosyllables (Exp. 1a = 82.3%, Exp. 1b = 81.5%), followed by 80% on disyllables (Exp. 1a = 80.1%, Exp 1b = 79.3%), and 78% on trisyllables (Exp. 1a = 79.3%, Exp. 1b = 76.9%). In a logistic mixed effects regression with Length (mono [ref], di, tri) as a fixed effect and Subject as a random effect, monosyllables were significantly more accurately perceived than disyllables ($\beta_{m-d} = 0.140$, $z = 4.907$, $p < 0.001$),[4] with disyllables further differentiated from trisyllables ($\beta_{d-t} = 0.107$, $z = 2.305$, $p = 0.021$), though given the much smaller number of trisyllables the latter effect was variable and only emerged in Experiment 1b ($\beta_{d-t} = 0.145$,

[3]From a model regressing accuracy onto block number (1-5), with listener random slopes for Block, there was an average increase in predicted accuracy of 3% from the first block to the last block (within-listener range: [−3%, +8%]), so some accommodation is apparent in the data, though the change is small enough to remain well inside of ceiling and floor effects.

[4]The notation used here, β_{i-j}, refers to the coefficient capturing the difference in mean outcomes $\hat{y}$ between levels i and j of a ternary or greater categorical variable (i.e., in cases where two or more coefficients are required to capture the effect of a given variable).

$z = 2.449$, $p = 0.014$; Exp. 1a: $p > 0.1$). The same result is obtained when word length is dichotomized (as it is in the composite model exploring interactions below), with polysyllables at 79% overall accuracy (80% in Exp. 1a, 79% in Exp. 1b), and monosyllables significantly greater than polysyllables ($\beta = -0.162$, $z = -6.089$, $p < 0.001$), both in Experiment 1a ($\beta = -0.158$, $z = -4.156$, $p < 0.001$) and 1b ($\beta = -0.166$, $z = -4.450$, $p < 0.001$). These results were not anticipated, given that the presentation of minimal pairs in the 2AFC task ensures the length of the stimulus word is always matched by the competitor word; however, longer words might pose a greater phonological memory burden on the listener, or simply exhibit phonetic reduction effects with increasing word length, weakening the salience of the critical contrast differentiating the target and competitor.

3.3.2.3 Word Frequency

Two measures of word frequency are relevant for the prediction of listener performance on the 2AFC task: the absolute frequency of the target word (AF), and the frequency of the target relative to the competitor (RF = F_T – F_C).[5] The reason for this dual representation of frequency is that while the relative frequency of the two items on the screen in the 2AFC task is directly related to listeners' decisions via the Luce choice rule (Luce, 1959), we might also expect differences between cases where target and competitor are relatively high in absolute frequency and cases where they are relatively low in absolute frequency. This latter effect could be due to multiple factors, including the relative salience of frequencies in different ranges (akin to the dependence of JNDs on scale), and the relative reliability of word frequency as an estimate of listener expectations (e.g., words of lower frequency in a corpus may be poorer estimates of a listener's knowledge than are high-frequency words).

Both absolute and relative target frequency correlate significantly with word recognition accuracy (logit-tranformed; r = 0.129 and 0.149, respectively; ps < 0.001), and these effects are con-

[5]Frequency estimates are derived by first normalizing the frequency measurements from four corpora—the SUBTLEX-US corpus (Brysbaert et al., 2012), the written and spoken Corpora of Contemporary American English (COCA, COCAspok; Davies, 2009), and the Google Unigram corpus (Michel et al., 2011)—to counts per million, then taking each item's mean normalized frequency across the four sources, and finally log-transforming the result.

sistent across sub-experiments (Exp. 1a: $r_{AF} = 0.178$, $r_{RF} = 0.184$; Exp. 1b: $r_{AF} = 0.080$, $r_{RF} = 0.112$; ps < 0.05). Further, the two frequency measures interact negatively (overall: $\beta = -0.014$, $z = -7.560$, $p < 0.001$; Exp. 1a: $\beta = -0.016$, $z = -5.930$ $p < 0.001$; Exp. 1b: $\beta = -0.013$, $z = -4.606$, $p < 0.001$) such that at lower target frequencies the difference between target and competitor has a greater effect, and conversely with greater relative frequencies the absolute frequency of the target has less of an effect on listener performance.[6] In the next section, we examine the extent to which these effects interact with background noise level and word length.

3.3.2.4 Interactions between Noise Level, Word Length, and Word Frequency

Among the general stimulus factors reviewed above, several effects may be expected to interact in their influence on listener word recognition, such as the relation between word frequency and noise level (at lower SNRs, top-down information such as word frequency might play a greater role in listener responses than at higher SNRs when more acoustic information is available), and the relation between noise level and word length (longer words might be more robust to noise than shorter words). To test for these potential interactions, a logistic mixed effects model was run with Accuracy (correct = 1, incorrect = 0) as the outcome, Noise Level (+2 dB [ref], -2 dB), Word Length (monosyllabic [ref], polysyllabic), Absolute Target Frequency (AF; continuous), and Relative Target Frequency (RF; continuous) as fixed effects, and Listener as a random intercept.

Only one significant interaction emerged in the model: AF $\times$ RF ($\chi^2(4) = 40.4$, $p < 0.001$). All other interactions were not significant (ps > 0.1), indicating generally additive effects between Noise Level and Word Length, and relatively independent effects of Word Frequency. That is, the effect of Noise Level remains relatively constant across word lengths ($\beta_{mono} = -0.761$, $z = -14.05$, $p < 0.001$; $\beta_{poly} = -0.828$, $z = -19.44$, $p < 0.001$),[7] and conversely, the effect

[6] These estimates were derived from a logistic mixed effects model regressing Accuracy (1 = correct, 0 = incorrect) onto Absolute Target Frequency and Relative Target Frequency, with random intercepts for Listener.

[7] The notation used here refers to the coefficient mentioned prior to the parentheses (here, Noise Level) at a particular level of an interacting categorical variable, indicated in the subscript (here, Word Length). Thus, for example, β_{mono} in this discussion of the effect of Noise Level refers to the β estimate for Noise Level at the reference level of monosyllabic Word Length (because of the model structure we further assume these estimates are for average absolute and relative target frequencies).

of Word Length varies little with SNR ($\beta_{+2} = -0.139$, $z = -2.564$, $p = 0.010$; $\beta_{-2} = -0.206$, $z = -4.842$, $p < 0.001$). Given that both frequency variables interact significantly, all further analysis of interactions between either AF or RF and the other two variables, such as the interaction between AF and Word Length, will be presented in terms of the implications of such effects for the AF $\times$ RF interaction.

Figure 3.1 shows the predicted probabilities of accurate word recognition in Experiment 1 as a function of RF (shown along the *x*-axis), AF (shown via different colored lines, from lowest in red to highest in blue), SNR (column panels), and Word Length (row panels). For both monosyllabic and polysyllabic target words, Absolute and Relative Frequency interact significantly at both SNRs (Mono: $\beta_{-2} = -0.016$, $z = -4.410$, $p < 0.001$; $\beta_{+2} = -0.014$, $z = -3.192$, $p = 0.001$; Poly: $\beta_{-2} = -0.010$, $z = -2.638$, $p = 0.008$; $\beta_{+2} = -0.011$, $z = -2.298$, $p = 0.022$), though effects are moderately larger in monosyllables. This distinction between monosyllabic and polysyllabic items primarily comes from the more consistent positive relationship between AF and accuracy on polysyllabic targets. That is, while the effect of RF is always greater for low frequency targets than for high frequency (shown in the comparatively steeper red lines), listeners remain more accurate in general on the high frequency targets, regardless of their frequency relative to the competitor. For monosyllabic items, on the other hand, the target's relative frequency appears to be the primary determinant of listener performance. At -2 dB, listeners are more accurate even on low frequency items so long as the frequency of the target is greater than that of the competitor, and at $+2$ dB listeners always show an effect of relative frequency, even for targets of low absolute frequency.

This result may stem from many sources, most notably the fact that monosyllabic words tend to exhibit higher frequencies of usage in English than do polysyllabic words (median log-transformed frequencies in Exp. 1 were 3.55 and 3.08, respectively), and so a far greater number of monosyllabic words lie above an assumed threshold, after which the differences in absolute frequency are less apparent than are relative differences between target and competitor frequency when the choice is constrained to two alternatives. Nevertheless, the present study is not designed to tease apart such effects; rather, the results of the above analyses that are most critical to the analyses

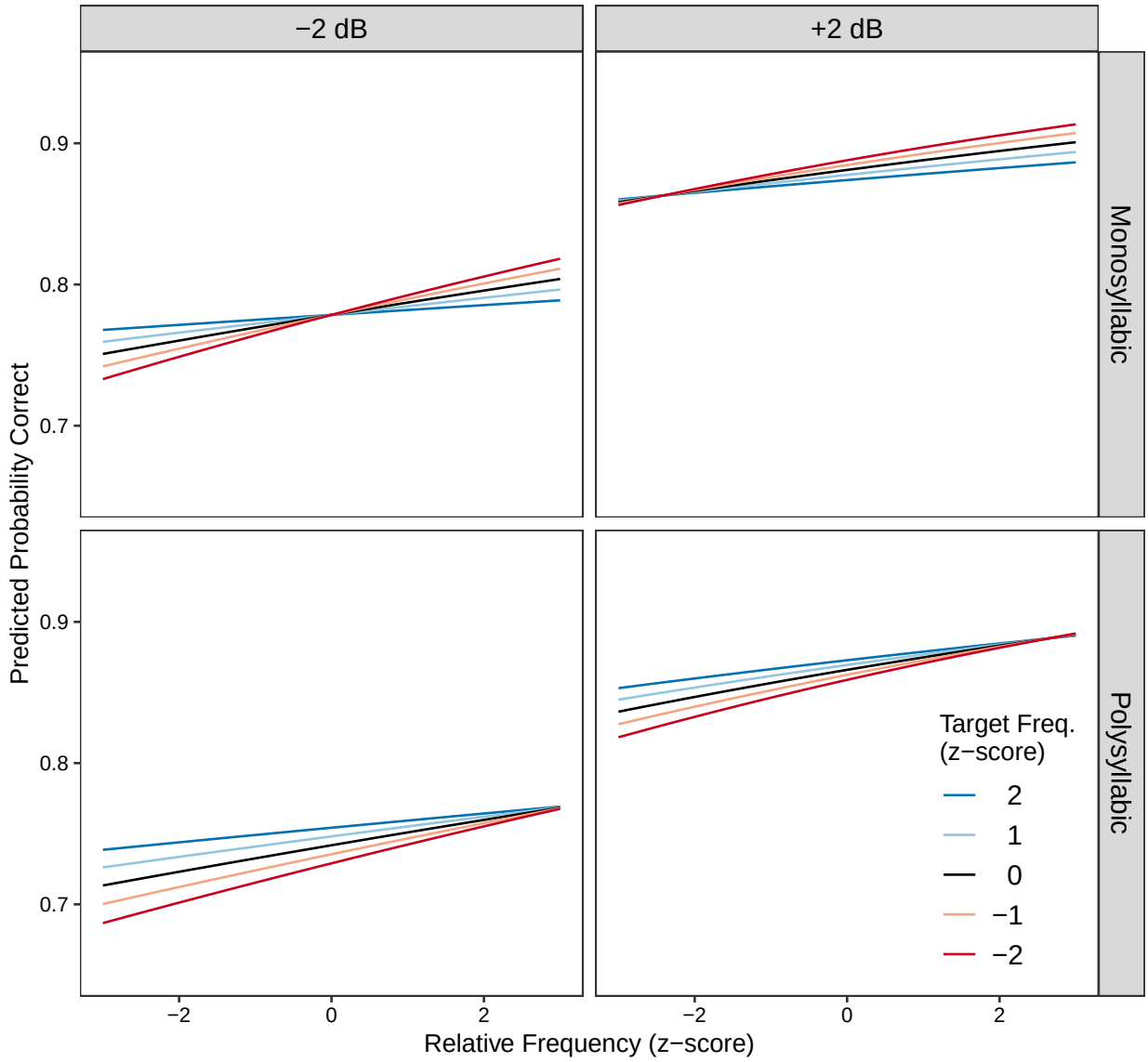

Figure 3.1: Predicted probabilities from model interactions between Absolute and Relative Frequency as a function of SNR (−2, +2 dB) and Word Length (monosyllabic, polysyllabic).

that follow, particularly for the cue-weighting models in Section 4.4 of Chapter 4, are the general findings that both absolute and relative target frequencies significantly impact listener performance across word lengths and noise levels, and that the two variables interact negatively such that RF has the greatest impact at low absolute frequencies, with diminishing effects at higher frequencies.

This result not only confirms the necessity of including frequency effects in any cue-weighting model, but suggests that the relative impact of different cues more globally throughout the lexicon will depend on the degree to which the contrasts where such cues are more salient (e.g., VOT

distinctions among word-initial plosive voicing contrasts) occur in words where such frequency effects are minimized. That is, if a given acoustic cue tends to be most distinct in lexical contrasts where the absolute and relative frequency effects are at a minimum—i.e., at intermediate AF and RF values—then that cue is predicted to have a greater impact on the perceptual maintenance of lexical contrast (due to the general lack of top-down, non-acoustic information) than if it were to occur in words where AF and RF play a greater role in listener performance.

But before analyzing specific acoustic cues and cue-weighting models, we review the general patterns in listener performance as a function of the phonetic categories and contrasts comprising the target and competitor stimuli in the 2AFC task. These patterns will not only be informative for the general perceptual distribution of minimal-pair obstruent contrasts throughout the English lexicon, but will provide us with some expectations as to which acoustic properties are expected to be most critical to determining listener word recognition patterns in Experiment 1.

3.3.2.5 Assessing positional differences in general recognition factors

In the analyses that follow, the three contrast positions—CV, VCV, and VC—are largely treated separately given differences in phonotactic constraints, allophony, and acoustic cue availability by position. The goal of this section is to determine the extent to which the above factors of Noise Level, Word Length, and Word Frequency are consistent or vary in their effects across different contrast positions.

Overall, listeners were 80% accurate on CV contrasts (Exp. 1a = 80.9%, Exp. 1b = 79.5%), 82% accurate in VCV (Exp. 1a = 81.7%, Exp. 1b = 81.3%), and 80% accurate in VC (Exp. 1a = 80.2%, Exp. 1b = 79.0%). This moderate but significant increase in overall accuracy on VCV contrasts (overall: βs > 0.09, zs > 2.58, ps < 0.01; Exp. 1a: $\beta_{vcv-cv} = 0.053$, $z = 1.148$, $p > 0.1$, $\beta_{vcv-vc} = 0.097$, $z = 2.121$, $p = 0.034$; Exp. 1b: ps < 0.05) is consistent with expectations based on the availability of both pre-consonantal and post-consonantal coarticulatory information in VCV position, as opposed to the constrained information at word onset/offset.[8] However, these

[8]The statistical model used to derive the estimates above is a mixed effects logistic regression predicting Accuracy (correct = 1, incorrect = 0) from Position (CV [ref], VCV, VC), with random slopes and intercepts for Listener.

are overall patterns, and there remains considerable variability within each position (IQRs between 7 and 11%) which may derive in part from interactions with other stimulus characteristics.

Beginning with the impact of noise level at different positions, listeners were 86% accurate at +2 dB on CV contrasts (Exp. 1a = 86.1%, Exp. 1b = 86.4%), 88% accurate on VCV (Exp. 1a = 87.7%, Exp. 1b = 88.7%), and 84% accurate on VC (Exp. 1a = 85.5%, Exp. 1b = 83.4%). At -2 dB listeners achieved accuracies of 74% (Exp. 1a = 75.7%, Exp. 1b = 72.6%), 75% (Exp. 1a = 75.7%, Exp. 1b = 74.0%), and 75% (Exp. 1a = 75.0%, Exp. 1b = 74.6%), respectively, on CV, VCV, and VC contrasts. This interaction was significant in a mixed effects logistic regression with Position and Noise Level as fixed effects and Listener as a random intercept ($\chi^2(2) = 23.45$, $p < 0.001$). That is, there was no difference in accuracy by position at -2 dB (ps > 0.1), while at +2 dB the following relation was obtained: VC $<$ CV $<$ VCV (ps < 0.01).

Effects of word length also varied by position, though not substantially. The experiment-wide pattern of decreasing accuracies with increasing word length was present in CV (mono→di→tri = 83→78→75%; Exp. 1a = 82.8→79.1→77.0%, Exp. 1b = 82.5→76.6→73.5%); however, in VCV position this pattern was not consistent across sub-experiments (di→tri = 82→81%, Exp. 1a = 81.5→82.8%, Exp. 1b = 81.7→79.4%), and in VC position only the monosyllabic $>$ polysyllabic pattern was robust (81→76→75%, Exp. 1a = 82.0→76.3→74.5%, Exp. 1b = 80.7→75.1→75.5%). A mixed-effects logistic regression with Length and Position as fixed effects and Subject as a random effect confirmed the above patterns, as all three relations in CV position were significant ($\beta_{m-d} = 0.308$, $z = 6.428$, $p < 0.001$; $\beta_{d-t} = 0.152$, $z = 1.963$, $p = 0.0496$), and only the mono $>$ di/tri relation was significant in VC ($\beta_{m-d} = 0.344$, $z = 6.708$, $p < 0.001$; $\beta_{d-t} = 0.038$, $z = 0.408$, $p > 0.1$), with no significant effect of position in VCV ($p > 0.1$). That is, the most robust effect of word length across positions is the monosyllabic vs. polysyllabic difference, which is confirmed by the consistency of the mono $>$ poly effect in CV and VC positions when word length is dichotomized (ps < 0.001). This leaves the effect of word length absent from VCV position merely by virtue of the intervocalic context being incompatible with monosyllables.

Finally, addressing word frequency effects as a function of contrast position, the significant cor-

relations between absolute and relative target frequency are present in all three positions ($0.107 < r_{AF} < 0.200$, ps < 0.01; $0.125 < r_{RF} < 0.165$, ps < 0.01). However, in CV position these effects are not replicated across sub-experiments (Exp. 1b: rs < 0.05, ps > 0.1), and in VCV position the correlation between AF and accuracy is only present in Experiment 1a ($r_{AF} = 0.136$, $p = 0.014$); both correlations are robust in VC position. Considering next the significant negative interaction between AF and RF, this effect is present overall at each position ($-0.013 < \beta < -0.016$, $-4.582 < z < -3.817$, ps < 0.001), but again with inconsistencies across sub-experiments. All effects are significant in Experiment 1a, but in 1b the interaction in CV position remains significant but reduces substantially in size ($\beta = -0.009$, $z = -2.103$, $p = 0.036$), and in VCV position the effect disappears entirely ($p > 0.1$). This variability suggests the effect of word frequency may be less stable with earlier positions of the target phone in the stimulus, a result which is consistent with the expectation that word frequency should exert greater effects on contrasts later in the word, as further information from the stimulus generates stronger hypotheses about lexical identity, which then influence listeners' online perception, particularly in the face of background noise.

3.3.3 Phonetic category recognition

Here we analyze patterns in the recognition of different phonetic categories and the feature classes they comprise along two lines. First, we describe the distribution of obstruent phones and features in minimal-pair contrasts in the lexicon. Second, we measure the overall accuracy of listeners on those stimulus categories/classes, as well as considering potential modulating effects of noise level, word length, and word frequency on listener performance.

3.3.3.1 Target phone distributions

We begin the analysis of obstruent category distributions with an examination of the frequency of occurrence of *target phones* in Experiment 1; that is, phones that occurred in the stimulus as the target member of the obstruent contrast distinguishing the minimal pair in the 2AFC task. Since the presentation of minimal pairs is symmetric, these distributions merely represent total counts of

	p	t	k	b	d	g	ʧ	ʤ	f	θ	s	ʃ	h	v	ð	z	Total
Exp. 1a	27	30	39	34	22	18	9	17	20	8	37	20	24	10	3	2	320
Exp. 1b	30	32	30	26	17	10	18	14	32	8	38	16	30	15	2	2	320
Total	57	62	69	60	39	28	27	31	52	16	75	36	54	25	5	4	640

Table 3.3: Distribution of CV target phones in minimal pair stimuli in Experiment 1.

obstruents comprising the contrasts defining the minimal pair items in the experiment; however, we retain the *target phone* terminology here for consistency with the analysis of category accuracies in Section 3.3.3.3, which are not symmetric between a phone's occurrence in the target auditory stimulus, and its occurrence in the mental representation of a visually presented competitor.

Word-initial position (CV). Word-initially, all 16 permissible obstruents, [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, h, v, ð, z], were presented in minimal pairs in Experiment 1. The relative frequency of occurrence of each phone, because of the semi-random sampling design that generated the stimuli, approximates the occurrence of phones in word-initial obstruent contrasts in the lexicon. Broadly, plosives and the alveolar sibilant [s] are the most frequent, followed by the other sibilant fricatives and affricates and the voiceless fricatives [f] and [h], with the dental fricatives [θ, ð] and voiced fricatives [v, z] relatively rare word-initially. See Table 3.3 for the complete distribution, both overall and by sub-experiment (note that Exp. 1a and 1b are closely matched in this respect).

Word-medial position (VCV). All 18 English obstruents (including the alveolar flap allophone of /t, d/) were presented word-medially in Experiment 1; namely, [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, h, v, ð, z, ʒ, ɾ].[9] See Table 3.4 for the full item distribution. Table 3.4 illustrates that the distribution of phones in VCV position is moderately less balanced than in CV (normalized entropy $\hat{H}_{VCV} = 0.91$, as compared with 0.94 in CV position), with the alveolar flap [ɾ] by far the most frequent at 94 items overall. The voiceless plosives [p, k] (57 and 75 items, respectively), and the fricatives [f, v, s, z] (44, 59, 58, and 43 items, respectively) are the next most common phones in VCV contrasts.

[9]Flaps were identified in the stimuli based on auditory and visual inspection, and generally conformed with descriptions of the primary context for English flapping (intervocalically, preceding unstressed syllables). Cases of intervocalic [t, d] not occurring in licit environments (i.e., not preceding stressed syllables), appear to be artifacts of hyperarticulation due to the laboratory environment in which the database was recorded.

	p	t	k	b	d	g	ʧ	ʤ	f	θ	s	ʃ	h	v	ð	z	ʒ	ɾ	Total
Exp. 1a	27	11	37	11	5	13	17	16	22	3	25	16	1	35	6	22	3	50	320
Exp. 1b	30	15	38	26	7	16	12	13	22	2	33	9	3	24	2	21	5	44	322
Total	57	26	75	37	12	29	29	29	44	5	58	25	4	59	8	43	8	94	642

Table 3.4: Distribution of VCV target phones in minimal pair stimuli in Experiment 1.

	p	t	k	b	d	g	ʧ	ʤ	f	θ	s	ʃ	v	ð	z	ʒ	Total
Exp. 1a	25	42	22	10	55	14	9	9	11	6	25	12	16	2	60	2	320
Exp. 1b	18	45	28	4	65	9	13	7	14	8	21	11	18	0	57	0	318
Total	43	87	50	14	120	23	22	16	25	14	46	23	34	2	117	2	638

Table 3.5: Distribution of VC target phones in minimal pair stimuli in Experiment 1.

The fricatives [θ, ð, h, ʒ], and the voiced plosive variant of the alveolar flap, [d], on the other hand, are quite infrequent, appearing in contrasts in under 20 items across Experiment 1.

Word-final position (VC). Word-finally, 16 obstruents were presented in Experiment 1, comprising the set [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, v, ð, z, ʒ], though the occurrence of the fricatives [ð, ʒ] in word-final contrasts is sparse, with the two being entirely absent from the items in Exp. 1b, and only present twice each in Exp. 1a (see Table 3.5 for the complete distribution). As Table 3.5 shows, the distribution of phones in VC position is even less balanced than in VCV ($\hat{H}_{VC}$ = 0.85, as compared with 0.91 in VCV, and 0.94 in CV). This result appears to be due primarily to the influence of morphology on the distribution of obstruent contrasts in English, as the two most frequent phones in VC position, [d, z], represent multiple inflectional suffixes—past tense in the case of [d], present tense and plural in the case of [z]—and as such appear widely in English.[10] Other effects such as articulatory and perceptual constraints in word-final position have been discussed elsewhere and are likely also at play in the lexical distribution of VC contrasts.

[10]These results would not be obtained under models of the lexicon where items are decomposed morphologically (e.g., the lexicon contains only irreducible morphemes, not fully inflected or derived forms); however, exclusion of inflected items would have radically reduced the number of word-final obstruent minimal pairs in English. Further, doing so would have prevented our models from needing to address the problem of suffix discrimination, which we consider an important part of speech perception.

3.3.3.2 Target feature distributions

Next we consider the distribution of target phones by feature class. Here we examine four features fundamental to the articulatory, acoustic, and perceptual characteristics of obstruent consonants: *voicing* (voiced [b, d, g, ʤ, v, ð, z, ʒ, ɾ], voiceless [p, t, k, ʧ, f, θ, s, ʃ, h]), *manner of articulation* (plosive [p, t, k, b, d, g], affricate [ʧ, ʤ], fricative [f, θ, s, ʃ, h, v, ð, z, ʒ], flap [ɾ]), *place of articulation* (labial [p, b, f, v], coronal (low) [t, d, θ, ð, s, z, ɾ], coronal (high) [ʧ, ʤ, ʃ, ʒ], dorsal [k, g], glottal [h]), and *sibilance* (sibilant [s, z, ʃ, ʒ, ʧ, ʤ], nonsibilant [p, b, t, d, k, g, f, v, θ, ð, h, ɾ]).[11] Note that here we refer to feature *classes* as a way of evaluating how groups of phones are perceived, where accuracy is determined by averaging over the phones in a given class. This is different from the typical analysis of distinctive feature perception, which considers the transmission of information through a binary feature channel (see Miller & Nicely, 1955) such that, for example, voicing perception is measured through the distribution of correct and incorrect responses to contrasts differing (at least) in voicing (e.g., [s] vs. [z], but also [f] vs. [b]). The analysis of information transmitted by feature will be discussed later in this section when listener response distributions by *contrast* are considered.

Word-initial position (CV). Table 3.6 displays the distribution of target phones in CV contrasts by feature class. Voiceless obstruents, while comprising approximately half of the inventory (due to symmetries involving all but the glottal fricative [h]), comprise 70% of the obstruents in minimal pair contrasts in Experiment 1. Other distributions, such as the relative rarity of affricates and dorsal/glottal places of articulation, are more consistent with the number of phones comprising a given class. The same is also true for the nonsibilant class, which, as noted above, is a set complementary to the more well-defined sibilant set [s, ʃ, ʧ, z, ʒ, ʤ], and as such includes many sounds not traditionally referred to as nonsibilants, such as the plosives. Finally, it is worth noting that plosives carry a substantial weight in the lexicon, as despite comprising 35% of the obstruent

[11] Here we incorporate the *height* feature of Lahiri & Reetz (2002) to subdivide the large coronal set into [LOW] coronals—the dentals and alveolars, which are produced with a relatively lower tongue body, and [HIGH] coronals, the postalveolars, which are produced with a relatively high tongue body.

	Voicing		Manner			Place					Sibilance	
	vl.	vd.	plos.	affr.	fric.	lab.	cor_L	cor_H	dor.	glot.	sib.	nsib.
Exp. 1a	214	106	170	27	123	91	102	46	57	24	85	235
Exp. 1b	234	86	145	31	144	103	99	48	40	30	88	232
Total	448	192	315	58	267	194	201	94	97	54	173	467

Table 3.6: Distribution of CV target phones by feature class (Voicing: voiceless [vl.], voiced [vd.]; Manner: plosive [plos.], affricate [affr.], fricative [fric.]; Place: labial [lab.], low coronal [cor_L], high coronal [cor_H], dorsal [dor.], glottal [glot.]; Sibilance: sibilant [sib.], nonsibilant [nsib.]) in Experiment 1.

inventory, they feature in half of all minimal-pair obstruent contrasts word-initially. In light of this fact, and the occurrence frequencies of [f, θ, v, ð] in Table 3.3, we can see that the canonical nonsibilant fricative set is actually quite rare among minimal pairs in the lexicon, accounting for only 15% of such phones despite representing 24% of the obstruent inventory.

Word-medial position (VCV). Considering next the featural distribution in VCV position in Table 3.7, we find that voicing classes are substantially more balanced than in CV position (323/319 for voiceless/voiced, as compared with 448/192 in CV). This result appears to be due largely to the frequent occurrence of the voiced alveolar flap intervocalically. Regarding manner of articulation, plosives remain quite frequent, though slightly less so than fricatives due to the flapping of the majority of intervocalic alveolar plosives. As noted above, the ubiquity of the flap in VCV position (15% of all intervocalic obstruent contrasts) is a critical fact for any model of the acoustic/perceptual structure of the English obstruent system. We will return to this point in Section 4.4 in the context of the structure of cue-weighting models. Regarding place of articulation, the coronals, particularly the [LOW] coronals (246 items, as compared with 91 for [HIGH] coronals), remain dominant at over 50% of items, while glottals nearly disappear from VCV contrasts, participating in only 4 minimal pairs, or less than 1% of all target phones. Finally, the sibilant distribution remains similar between CV and VCV positions, though the number of sibilants increases somewhat (192 as compared with 173 in CV). Decomposing the [−sibilant] class into the canonical nonsibilant fricative group [f, v, θ, ð], we find a slightly greater occurrence of this set at 18%, as compared with 15% in CV position, though this figure remains below the inventory expectation of 24%.

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	affr.	fric.	flap	lab.	cor_L	cor_H	dor.	glot.	sib.	nsib.
Exp. 1a	159	161	104	33	133	50	95	122	52	50	1	99	221
Exp. 1b	164	158	132	25	121	44	102	124	39	54	3	93	229
Total	323	319	236	58	254	94	197	246	91	104	4	192	450

Table 3.7: Distribution of VCV target phones by feature class (Voicing: voiceless [vl.], voiced [vd.]; Manner: plosive [plos.], affricate [affr.], fricative [fric.], flap [flap.]; Place: labial [lab.], low coronal [cor_L], high coronal [cor_H], dorsal [dor.], glottal [glot.]; Sibilance: sibilant [sib.], nonsibilant [nsib.]) in Experiment 1.

	Voicing		Manner			Place				Sibilance	
	vl.	vd.	plos.	affr.	fric.	lab.	cor_L	cor_H	dor.	sib.	nsib.
Exp. 1a	152	168	168	18	134	62	190	32	36	117	203
Exp. 1b	158	160	169	20	129	54	196	31	37	109	209
Total	310	328	337	38	263	116	386	63	73	226	412

Table 3.8: Distribution of VC target phones by feature class (Voicing: voiceless [vl.], voiced [vd.]; Manner: plosive [plos.], affricate [affr.], fricative [fric.]; Place: labial [lab.], low coronal [cor_L], high coronal [cor_H], dorsal [dor.]; Sibilance: sibilant [sib.], nonsibilant [nsib.]) in Experiment 1.

Word-final position (VC). Table 3.8 shows the distribution of target phones by feature class. As in VCV position, voicing is evenly balanced word-finally between voiceless (49%) and voiced (51%), though the aforementioned prevalence of voiced [d, z] indicates that the distribution of voiced obstruents is heavily skewed (as Table 3.5 illustrates, the voicing distribution is primarily balanced out by the greater overall frequency of voiceless plosives over voiced). Manner contrasts are also relatively consistent with CV and VCV positions, with plosives representing a large portion of target phones (53%), followed closely by fricatives (41%). The place distribution word-finally is also heavily skewed toward coronal obstruents, particularly [LOW] coronals, which comprise 61% of all target phones in VC position. Finally, there are moderately more sibilants in VC than in VCV (35% as compared with 30%), with the nonsibilant fricatives [f, v, θ, ð] least represented among the three positions at 12%, as compared with 15% in CV and 18% in VCV.

3.3.3.3 Target phone accuracy

Before examining more directly the impact of these distributions on the overall contribution of each phone/feature to listener word recognition, we present data on mean listener accuracies by category/class. Further, we explore the effects of global factors such as noise level, word length and word frequency on featural accuracy. However, the distributional data above remains relevant for the accuracy analysis in so far as it reveals that estimates for certain phones and feature classes should be less reliable due to their relatively rare occurrence in the data.[12]

Word-initial position (CV). Figure 3.2 shows listener accuracies on target phones in CV position. In Figure 3.2, overall accuracies are plotted in rank order with accuracy on the *y*-axis, meaning the phone furthest to the right, [s], is the most accurately perceived at 91%, and the left-most one, [v], is least accurate at 70%. Shown alongside this aggregate line are separate lines for $+2$ and -2 dB SNRs, which indicate the relative impact of noise level on target phone accuracy. The post-alveolar sibilant [ʃ], for example, is similarly accurate at both SNRs (72% and 80% at -2 and $+2$ dB, respectively), as are the other voiceless fricatives/affricates [f, s, ʧ, h], while [ʤ] varies considerably (70% at -2 dB, 85% at $+2$ dB), as do several other voiced obstruents, including [b, ð, d, z]. Examining the accuracy ranks of sounds in different feature classes, while there do not appear to be any clear alignments in Figure 3.2 according to manner or voicing, place of articulation is apparent as [LOW] coronal and dorsal obstruents are more accurately perceived than other places, which could be due to their more distinct spectral characteristics (relative to labials) and more informative formant transitions (relative to labials and dental fricatives). The lower accuracies observed on [HIGH] coronals are not consistent with the above predictions, however, and will be addressed later in this section when listener performance on specific contrasts is examined.

The two sub-experiments, 1a and 1b, largely replicate the above patterns. The voiceless alveolar sibilant [s] is the most accurate in both, and is consistently robust to noise, while [v] is least

[12] Since these distributions reflect the type distribution in the lexicon, an examination of a larger sample of such tokens in English would be required to obtain more reliable estimates of listener accuracy on more marginal phones and feature classes.

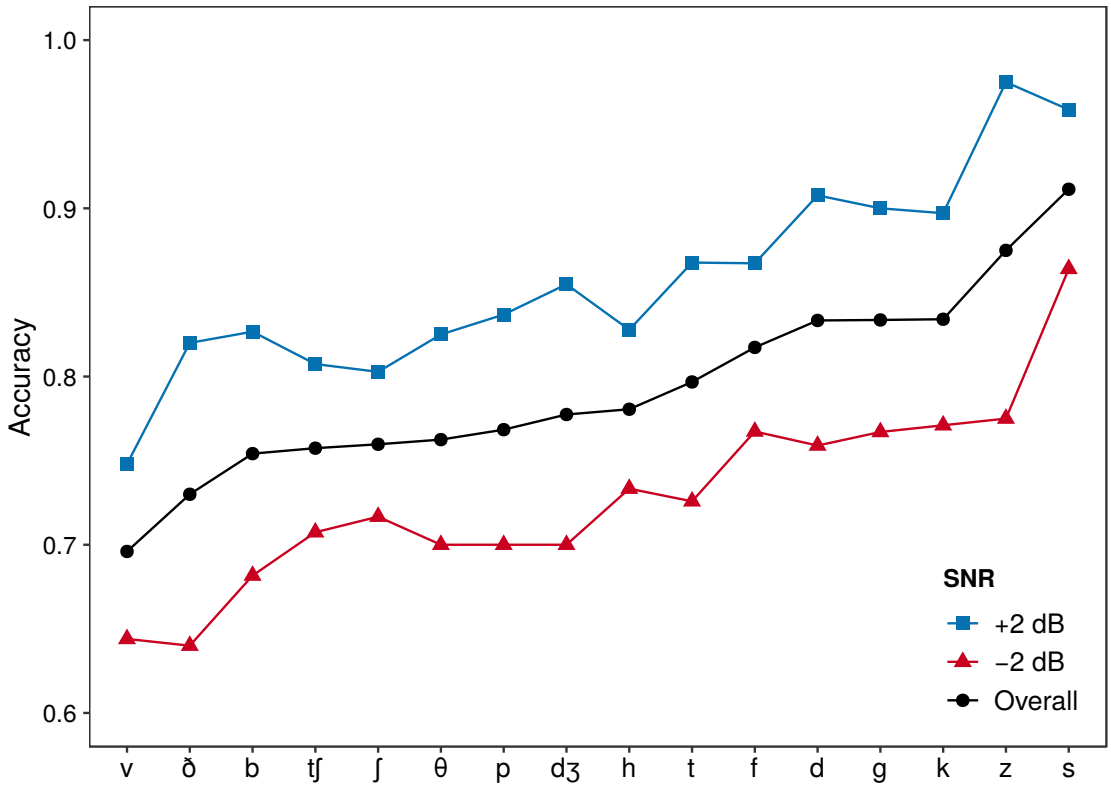

Figure 3.2: Target phone accuracies in CV position in Experiment 1, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

accurate in Experiment 1a, and second-to-last in Experiment 1b, where the least accurate phone in 1b, [ð], is not present in 1a. Among the phones which are less robust to noise, there is greater variability between the two experiments, though [z] and the voiced plosives [b, d, g] all show substantial drops in accuracy from +2 to −2 dB SNR in Experiments 1a and 1b. Finally, the place effects observed above—dorsals and most [LOW] coronals (excluding [θ, ð]) being generally more accurate than labials—are consistent across sub-experiments, and thus likely reflect robust perceptual distinctions among word-initial obstruents. See Figures A.1 and A.2 in the appendix for complete target phone accuracy results in Experiments 1a and 1b.

Word-medial position (VCV). Listener accuracies on target phones in VCV contrasts are shown in Figure 3.3, where results are given both overall and separately by SNR. As in CV position, the alveolar sibilants [s, z] are most accurately perceived, while the voiced nonsibilant fricatives [v, ð]

are least accurate. Unlike in CV position, however, there is a clearer grouping of phones according to voicing. Voiceless obstruents are generally perceived more accurately in intervocalic position (only [p] is below the median ranking, and only [z] is above), with the difference in performance particularly stark at the lower SNR. Effects of place of articulation are less clear, while manner of articulation appears to interact with voicing in that all voiceless fricatives/affricates are above the median accuracy rate. This result may be due to the greater salience of obstruent consonants when distinguished from adjacent vowels in voicing, or it could be a consequence of the process of lenition, whereby voiced consonants are more likely to be weakened (to fricatives or approximants) in intervocalic position than are voiceless consonants (Kirchner, 1998).

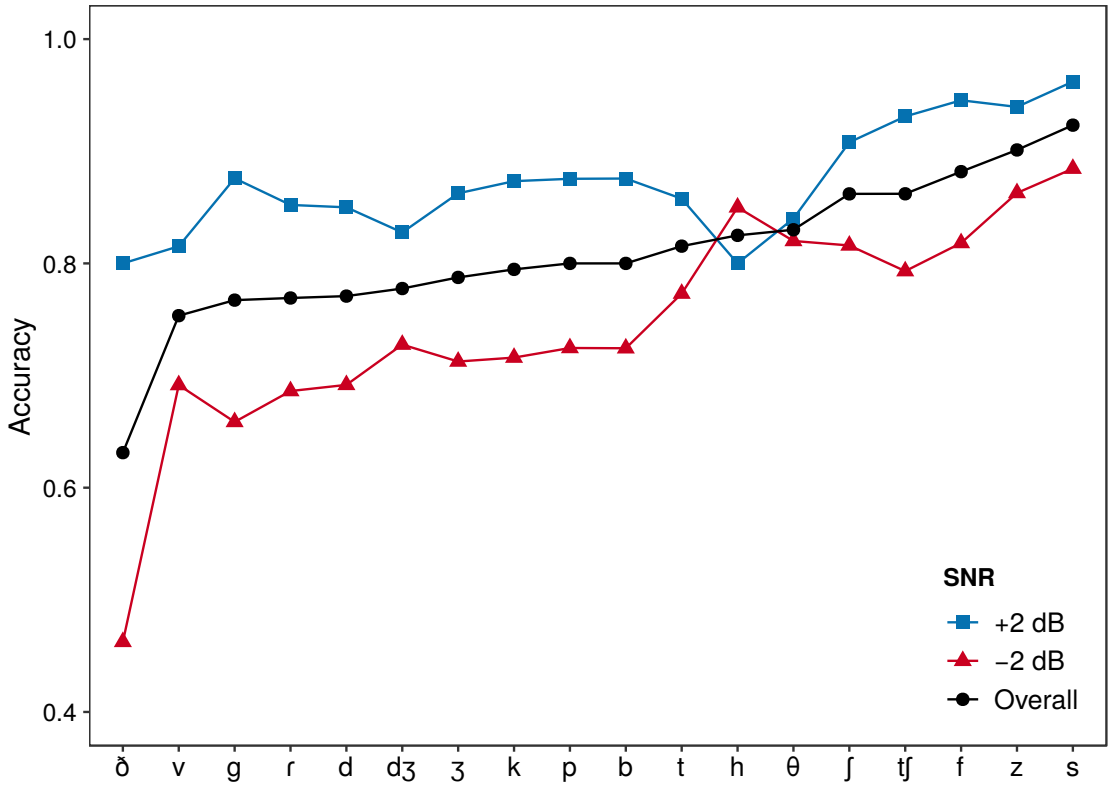

Figure 3.3: Target phone accuracies in VCV position in Experiment 1, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

As for the relative vulnerability of different phones to noise masking, we must first address the seemingly anomalous cases of [h] and [θ], the former showing greater accuracy at −2 dB than at +2 dB, and the latter showing a much narrower difference between the two SNRs than

any other phone. These are the two least frequent sounds in our sample, at 4 and 5 occurrences, respectively, making them more susceptible to acoustic or lexical peculiarities of the items they occur in; however, on the lexical end no notable frequency biases were evident that might have made them exceptionally robust to noise. On the acoustic end, we will return to this issue in Section 4.4, where cue weighting models of listener perception in Experiment 1 are presented. Among the more well-represented sounds in Experiment 1 that show relatively greater robustness to noise are the sibilants [s, z, ʃ, ʤ], the voiced labiodental fricative [v], and the voiceless alveolar plosive [t]. However, with the exception of the above set, and two sounds that exhibit relatively greater vulnerability to noise masking, [g, ð], the effect of the 4 dB difference in SNR remains relatively constant across phones at between 13 and 16%.

The results above largely replicate across sub-experiments. Both Experiments 1a and 1b show a robust advantage for voiceless obstruents over voiced obstruents, as well as an overall accuracy advantage and robustness to noise of the alveolar sibilants [s, z]. The interaction between voicing and manner is also present in both experiments. Finally, Experiments 1a and 1b each show the sensitivity of [g] and [ð] to the amplification of background noise, and surprisingly even replicate the [h] and [θ] anomalies reported above. See Figures A.3 and A.4 in the appendix for the full target phone accuracy results in Experiments 1a and 1b.

Word-final position (VC). Listener mean accuracies on target phones in VC contrasts are shown in Figure 3.4. As in CV and VCV position, the alveolar sibilant [s] is the most accurately perceived, both overall and at each SNR. At the other end of the spectrum is the voiced labial plosive [b], which is more consistent with the CV pattern than VCV. As in word-medial position, voiceless obstruents tend to occupy the higher accuracy ranks, particularly voiceless fricatives and affricates. Plosives are among the least accurately perceived obstruents word-finally, which could be due in part to their tendency to be unreleased, leaving listeners only characteristics of the vowel to use in consonant identification. No clear effects of place of articulation or sibilance are present in the individual phone rankings in Figure 3.4 (though voiceless sibilants are more accurately perceived

than voiced sibilants), but general feature effects will be clarified in the next section when feature classes are analyzed directly.

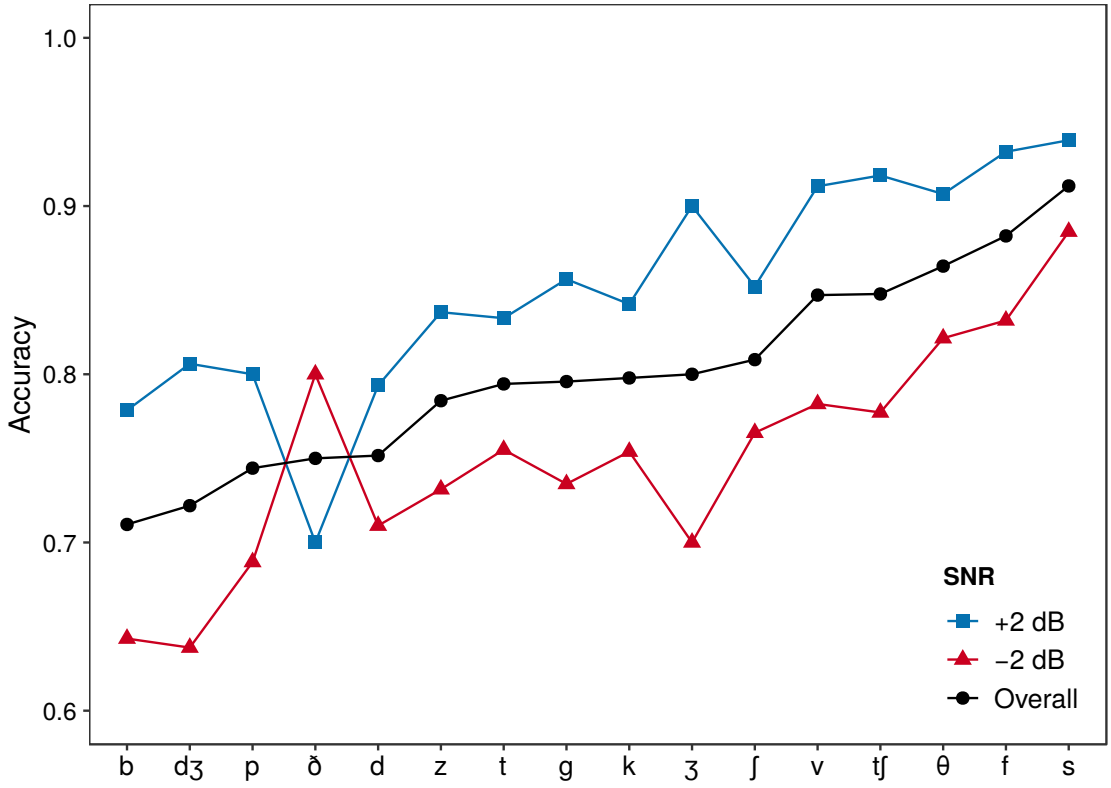

Figure 3.4: Target phone accuracies in VC position in Experiment 1, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

Regarding the relative resilience or vulnerability of different word-final obstruents to noise, Figure 3.4 indicates several phones, particularly [s], that show little decline in accuracy ($< 10\%$) with a 4 dB increase in the relative level of background noise. That set, [s, f, θ, ʃ, k, t, d] largely comprises voiceless fricatives and plosives, [d] being the one exception, and [p] being excluded as it exhibits a relatively greater accuracy difference between +2 and −2 dB, though the greatest differences are observed in [ʒ] and [ʤ]. This latter result could again be due to uncertainty in the data given that [ʤ], and particularly [ʒ], are both relatively uncommon word-finally, at 16 and 2 occurrences, respectively. The voiced dental fricative [ð] also shows a seemingly anomalous SNR pattern, but only occurs twice in Experiment 1, so these estimates are unreliable on their own, and could reflect idiosyncrasies in the items or in their lexical competitor relations.

Finally, considering the replicability of the above results across different item and participant sets, the two sub-experiments agree in the salience of [s], the generally poorer perception of plosives than fricatives or affricates. In terms of resilience to noise, there is less agreement between Experiments 1a and 1b, as the stability of the set [s, f, ʃ, k] is driven by Experiment 1a, while [θ, t, d] are mostly found to be robust in Experiment 1b. The voicing effect is partially preserved across sub-experiments, as the [s, f, θ, ʧ] set ranks in the top 5 most accurate phones in both experiments. Experiment 1a, however, shows a greater generalization of this ranking priority to other voiceless obstruents than Experiment 1b. See Figures A.5 and A.6 in the appendix for the full target phone accuracy results in Experiments 1a and 1b.

3.3.3.4 Target feature accuracy

We turn now to accuracy patterns by the feature class of the target phone, and as a function of the general stimulus recognition factors: noise level, word length, and word frequency. As before, results for CV, VCV, and VC positions are considered separately within each analysis.

Accuracy by Noise Level. The analysis of listener recognition of target obstruent features as a function of noise level is focused primarily on two questions: (1) within a given feature what is the relation between its constituent classes at both low and high noise levels; (2) how does the manipulation of background noise affect different feature classes, specifically with regard to their relative robustness or vulnerability to signal masking.

Word-initial position (CV). Listener accuracies by stimulus feature class in word-initial contrasts are shown in Figure 3.5. Voiced and voiceless obstruents are similar in accuracy +2 dB SNR; However, at −2 dB, voiced obstruents drop in accuracy relative to voiceless (voiced = 71%, as compared with 76% for voiceless), indicating voiced obstruents are more vulnerable to perturbation by noise than voiceless, though this interaction was not significant ($\beta = 0.102$, $z = -1.024$,

$p > 0.1$).[13] The different manner classes also show differential susceptibility to noise, with fricatives substantially more accurately perceived at -2 dB (76%, as compared with 73% for plosives and 70% for affricates), but this distinction is reduced at the higher SNR, to where fricatives, at 86% accuracy, are relatively evenly perceived with plosives (87%) and much more narrowly distinguished from affricates (83%). Yet, as with voicing, the apparent effect of SNR on manner distinctions fails to reach significance in aggregate ($\chi^2(2) = 4.93,\ p = 0.085$), though the binary distinction between fricatives and plosives does significantly vary by noise level ($\beta = 0.220$, $z = 2.218$, $p = 0.027$).

Regarding place of articulation, Figure 3.5 shows a consistent pattern at both SNRs, where labials, [HIGH] coronals, and glottals are less accurately perceived than [LOW] coronals and dorsals. There is in fact a significant interaction between place and noise level ($\chi^2(4) = 11.3,\ p = 0.023$), but this interaction reflects differences of degree (coronal [LOW] $\times$ labial: $\beta = -0.269$, $z = -2.175$, $p = 0.030$; coronal [LOW] $\times$ coronal [HIGH]: $\beta = 0.356$, $z = 2.416$, $p = 0.016$; coronal [LOW] $\times$ glottal: $\beta = 0.468$, $z = 2.617$, $p = 0.009$; dorsal $\times$ glottal: $\beta = 0.409$, $z = 2.031$, $p = 0.042$), rather than differences in directionality, as in the manner effects reported above. Finally, sibilants are significantly better perceived than nonsibilants (overall: 84%, as compared with 80% for nonsibilants; -2 dB: 80% vs. 74%; $+2$ dB: 90% vs. 88%), and this distinction, though greater at the lower SNR, is not significantly lesser at $+2$ dB than at -2 dB ($\beta = -0.006$, $z = -0.058$, $p > 0.1$).

In general, the majority of the patterns above replicate across sub-experiments. Dorsals and [LOW] coronals show advantages over the remaining places of articulation in both Experiment 1a and Experiment 1b, and the sibilant advantage holds for both SNRs in each sub-experiment. Voicing is also generally robust, though listeners in Experiment 1b are nearly even in their recognition of voiceless and voiced obstruents at $+2$ dB. Finally, the overall disadvantage for affricates word-initially is notably reduced in Experiment 1b, which similar to the voicing effect shows no

[13] This analysis, and each of the remaining featural analyses in this and the following paragraph, comes from a logistic mixed effects model with Noise Level ($+2$ dB [ref], -2 dB), Feature (e.g., Voicing, Place; the leftmost category in Figure 3.5 is the reference level), and their interaction as fixed effects, and Listener as a random intercept.

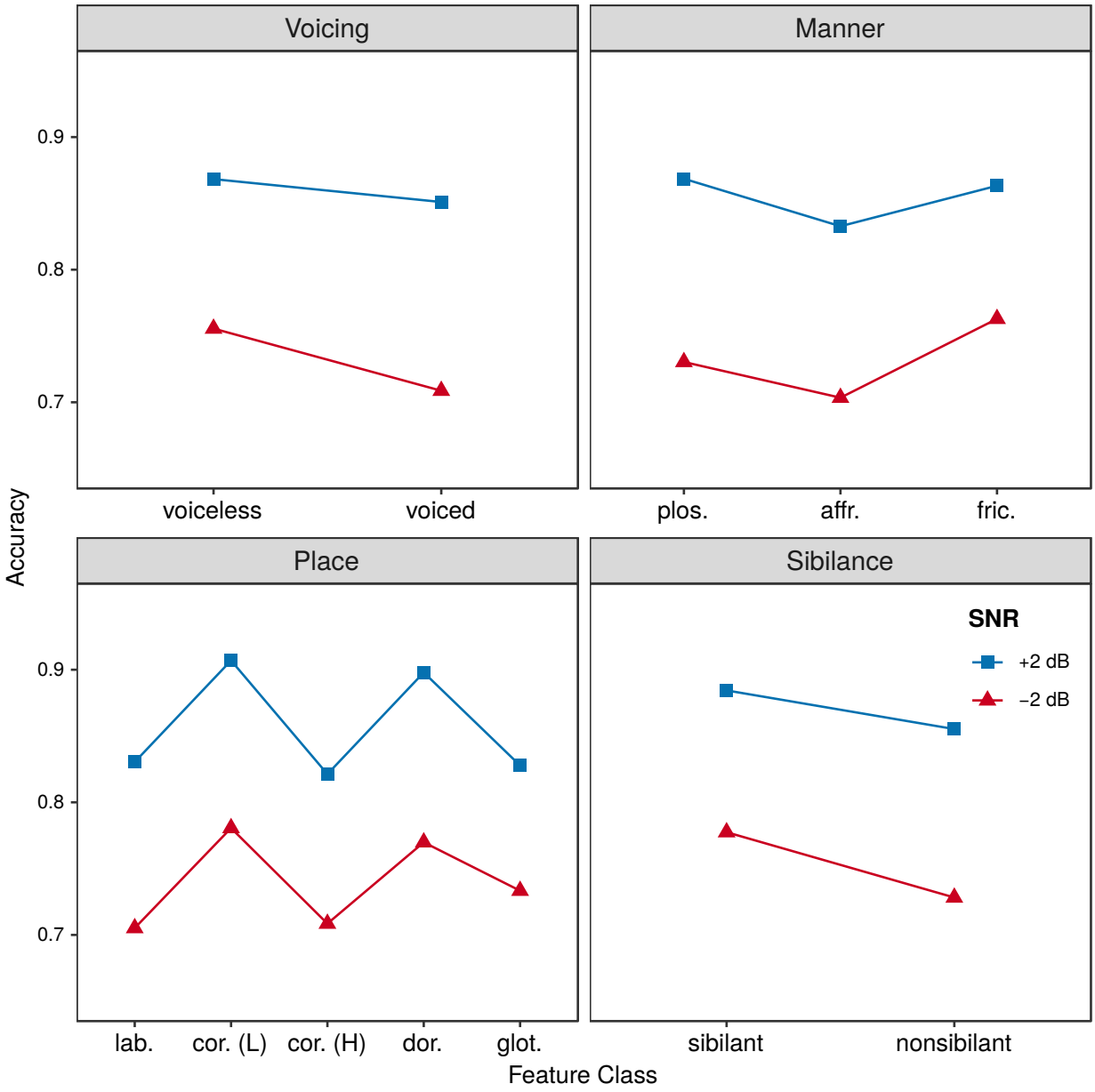

Figure 3.5: Target feature accuracies by SNR in CV position in Experiment 1.

clear manner effect at the higher SNR. See Figures A.7 and A.8 in the appendix for the full results of target feature accuracy by noise level in Experiments 1a and 1b.

Word-medial position (VCV). Figure 3.6 shows listener accuracies on VCV contrasts by voicing, manner, place, and sibilance. Beginning with voicing, compared with word-initial position, voiceless obstruents show a considerably greater advantage over voiced obstruents in VCV posi-

tion, both at +2 dB ($\beta = -0.449$, $z = -5.674$, $p < 0.001$) and −2 dB ($\beta = -0.399$, $z = -6.773$, $p < 0.001$), though as in CV position, there is a moderate though non-significant ($p > 0.1$) increase in the voiceless advantage at the lower SNR. A similar exaggeration of CV effects in VCV position is evident in the sibilance feature, where sibilants remain both more accurately perceived than nonsibilants ($\beta_{+2} = -0.607$, $z = -6.349$, $p < 0.001$; $\beta_{-2} = -0.641$, $z = -9.224$, $p < 0.001$), and more robust to noise ($\beta_{sib} = 0.927$, $z = 8.868$, $p < 0.001$; $\beta_{nsib} = 0.961$, $z = 17.32$, $p < 0.001$), though again, the interaction between sibilance and noise level was not significant ($p > 0.1$).

Manner and place of articulation in VCV contrasts, however, diverge from some of the key patterns observed in CV position. Plosives, for instance, are as accurate or worse than affricates on average, though fricatives remain relatively well perceived. The primary distinction between the four manner classes is between the more accurate {affricate, fricative} set and the less accurate {plosive, flap} set; i.e., at +2 dB, fricatives are more accurate than plosives ($\beta_{fr-pl} = 0.326$, $z = 3.553$, $p < 0.001$) and flaps ($\beta_{fr-fl} = 0.482$, $z = 4.203$, $p < 0.001$), and at −2 dB, fricatives are more accurate than plosives ($\beta_{fr-pl} = 0.452$, $z = 6.606$, $p < 0.001$), and both fricatives and affricates are more accurate than flaps ($\beta_{fr-fl} = 0.588$, $z = 6.730$, $p < 0.001$; $\beta_{af-fl} = 0.347$, $z = 2.842$, $p = 0.005$). As a result of this relative consistency at each SNR, there was no significant interaction between manner and noise level ($p > 0.1$), though it is quite plain from Figure 3.6 that all distinctions are enhanced at the lower SNR.

The pattern of place distinctions is notably different intervocalically, the only constant with CV position being listeners' relatively low accuracy on labials. At +2 dB there are no significant differences between places, while at −2 dB the [LOW] coronals are significantly more accurately perceived than labials ($\beta_{lc-l} = 0.195$, $z = 2.738$, $p = 0.006$) and both coronal classes are more accurate than dorsals ($\beta_{lc-d} = 0.356$, $z = 4.225$, $p < 0.001$; $\beta_{hc-d} = 0.343$, $z = 3.250$, $p = 0.001$), but no other significant distinctions emerge. Further, no significant omnibus interaction between place and noise level emerged ($p > 0.1$). Incorporating these results with the manner and sibilance effects reported above, we can interpret the superiority of coronal obstruents in VCV position as being primarily driven by the robustness of sibilant fricatives and affricates, where by comparison,

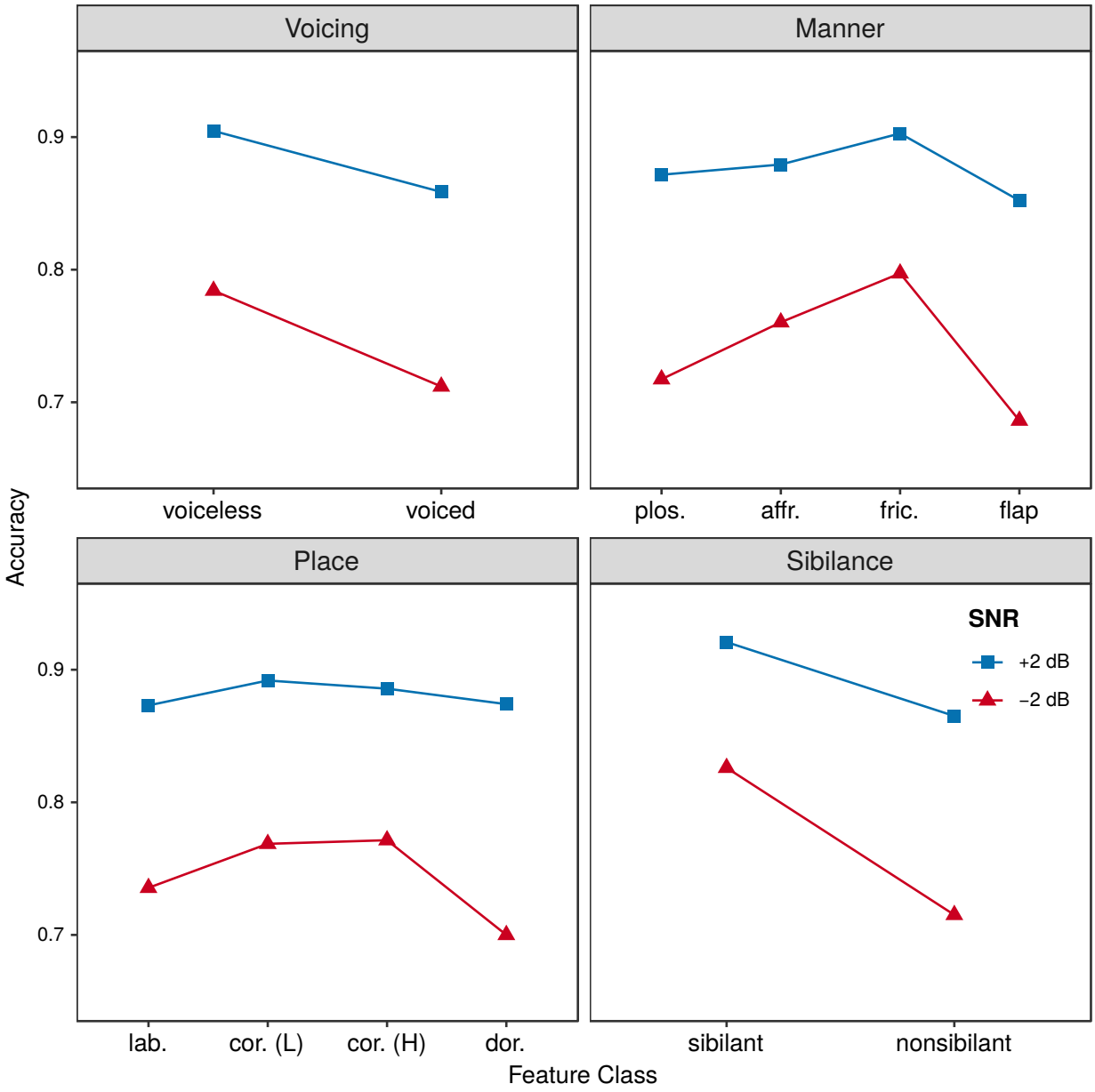

Figure 3.6: Target feature accuracies by SNR in VCV position in Experiment 1. The glottal fricative has been omitted from the place of articulation results due to sparsity of data.

the labial POA is composed of plosives and nonsibilants, both poorly perceived classes, and the dorsal POA similarly is only represented in plosive obstruents. Thus, it is not clear that this result is a place effect *per se*, and not rather an indirect effect of the other features that comprise coronal articulations; nevertheless, it is no accident that sibilants occur exclusively in coronals, nor is it a coincidence that the majority of English fricatives are coronals (this is the cross-linguistic norm;

Maddieson, 1992).

Finally, the effects of voicing, manner, and sibilance are consistent across sub-experiments, while for place of articulation, the advantage for coronals and disadvantage for labials and dorsals is generally robust, though listeners in Experiment 1b show no clear distinction between the four places at +2 dB. See Figures A.9 and A.10 in the appendix for the full results of target feature accuracy by noise level in Experiments 1a and 1b.

Word-final position (VC). Figure 3.7 shows listener accuracies by feature class in word-final position, and replicates several key patterns observed in CV and VCV contexts; namely, the voiceless > voiced ($\beta_{+2} = 0.310$, $z = 4.406$, $p < 0.001$; $\beta_{-2} = 0.302$, $z = 5.123$, $p < 0.001$) and sibilant > nonsibilant ($\beta_{+2} = 0.260$, $z = 3.469$, $p < 0.001$; $\beta_{-2} = 0.145$, $z = 2.352$, $p = 0.019$) advantages, and the general robustness of fricatives over plosives ($\beta_{+2} = 0.503$, $z = 6.710$ $p < 0.001$; $\beta_{-2} = 0.329$, $z = 5.319$ $p < 0.001$). Affricates have generally been more poorly perceived than fricatives, but at +2 dB the two are equivalent ($\beta_{a-f} = -0.086$, $z = -0.519$, $p > 0.1$). Regarding place of articulation, no clear patterns emerged (ps > 0.05), similar to the VCV results at +2 dB but with even greater uniformity across noise levels.

Finally, these patterns are largely consistent across sub-experiments, with the one exception being the absence of a sibilance effect at −2 dB in Experiment 1b. See Figures A.11 and A.12 in the appendix for the full results of target feature accuracy by noise level in Experiments 1a and 1b.

Accuracy by Word Length and Frequency. Next we examine what effects, if any, additional factors such as word length and word frequency have on the recognition of different feature classes. Here our concerns are both perceptual and structural, where the former is consistent with the analysis in the previous section in examining both bottom-up and top-down influences on signal parsing, while the latter has implications for the ultimate role of different components of the obstruent system in the maintenance of contrast in the lexicon, as both word length and frequency affect the size and rate of competition in lexical access.

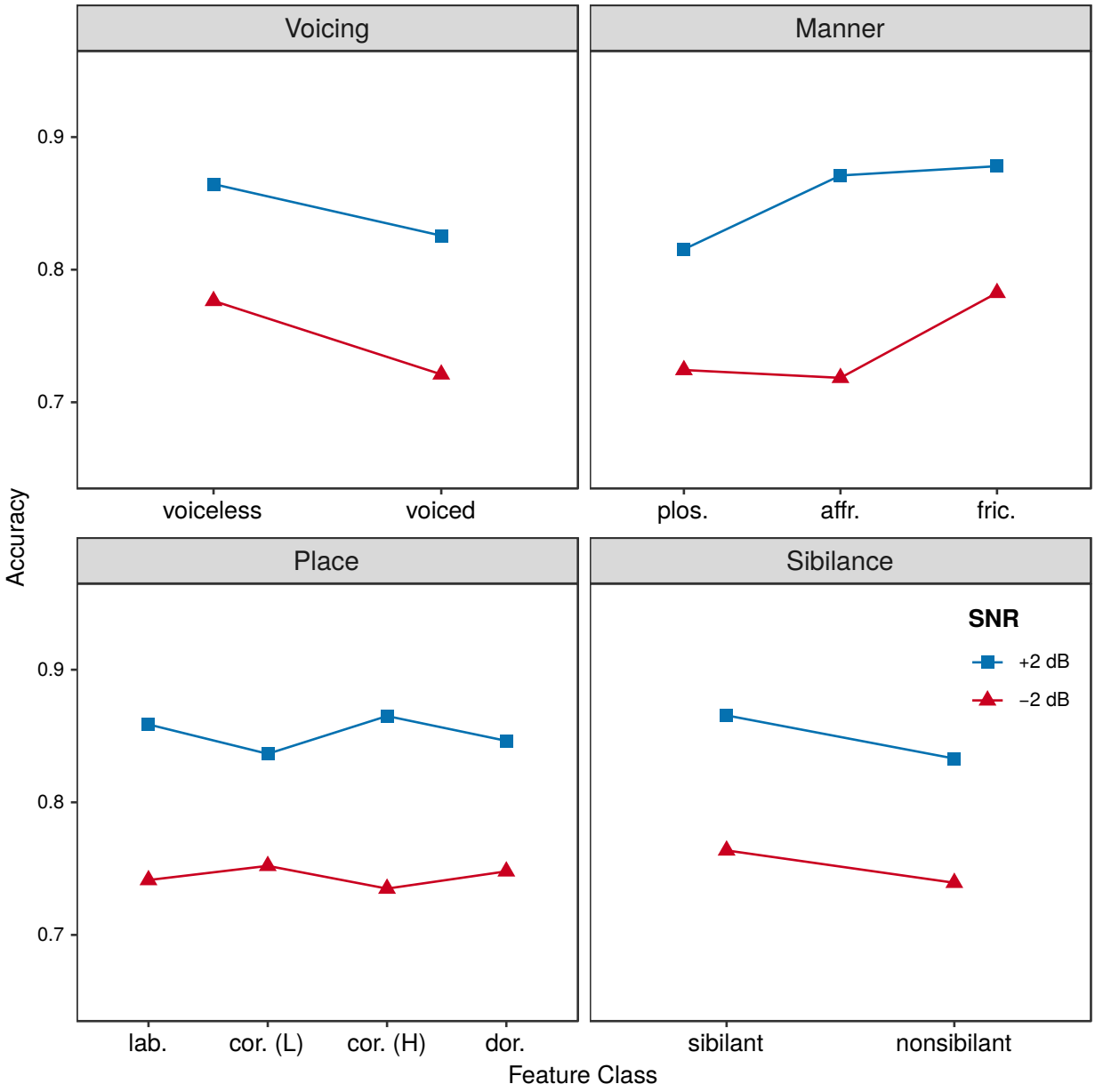

Figure 3.7: Target feature accuracies by SNR in VC position in Experiment 1.

Word-initial position (CV). Figure 3.8 shows listener accuracies by Feature—voicing, manner, place, and sibilance (shown along the columns)—and Word Length (monosyllabic, polysyllabic), Absolute Target Frequency (AF), Relative Target Frequency (RF), and the AF×RF interaction (shown along the rows), where the latter three word frequency variables are each discretized into terciles (< 0.33, $0.33 - 0.67$, > 0.67) for visualization purposes.

Beginning with obstruent voicing, Figure 3.8 illustrates the advantage for voiceless obstruents

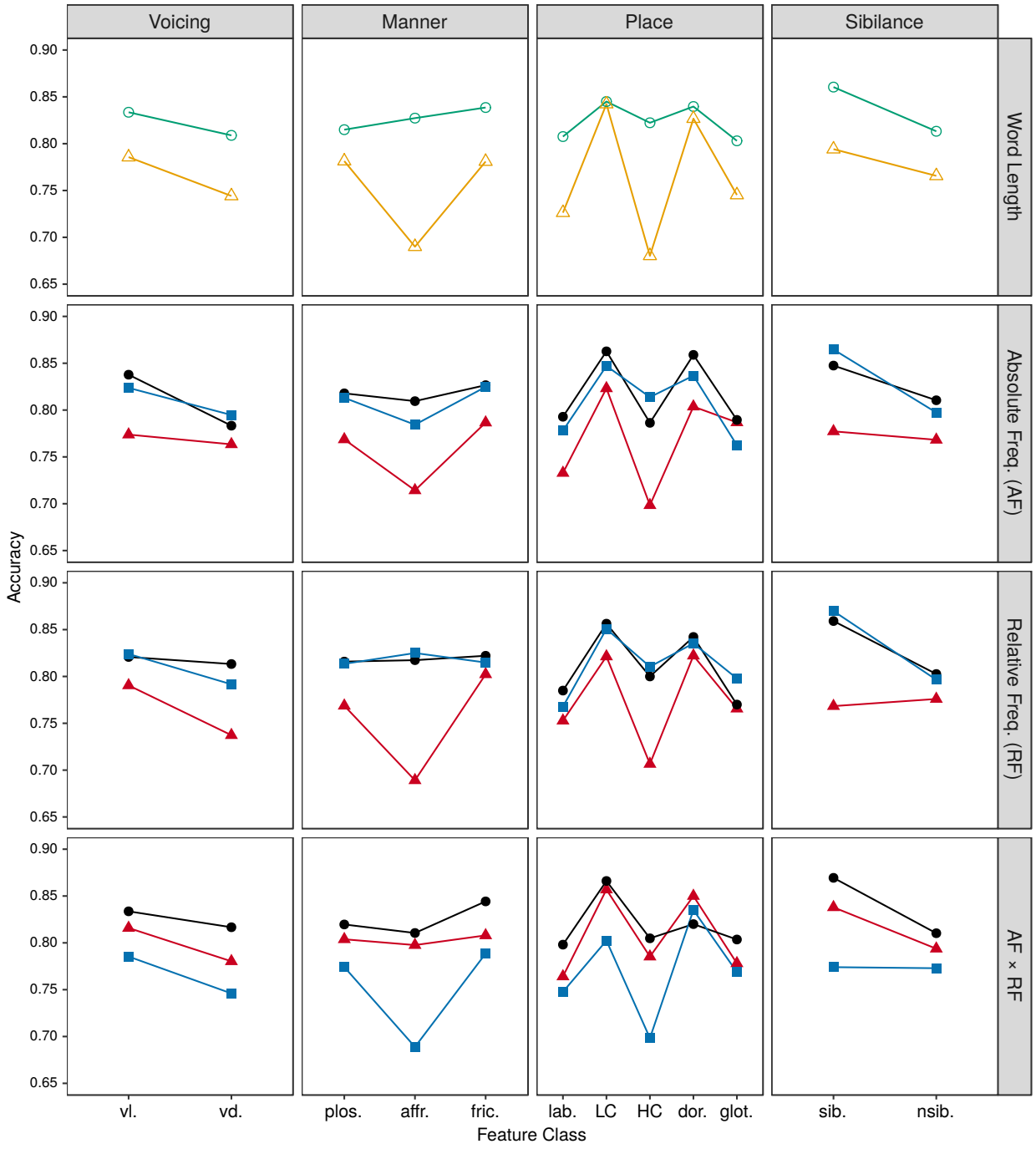

Figure 3.8: Target feature accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in CV position in Experiment 1. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

is quite robust, with no significant interactions with Word Length, Absolute Frequency, Relative Frequency, or AF×RF (ps > 0.1) in a mixed-effects logistic regression with Listener random intercepts. That is, listener accuracy on voiceless obstruents is between 2 and 8% better than on voiced obstruents over a range of word lengths and frequencies. The pattern in word recognition as a function of the sibilance of the target phone is similarly robust to differences in word length ($\chi^2(1) = 3.007$, $p = 0.083$), though as Figure 3.8 indicates, the sibilant advantage increases significantly as both absolute ($\beta = 0.195$, $z = 3.587$, $p < 0.001$) and relative ($\beta = 0.207$, $z = 3.701$, $p < 0.001$) frequencies increase. Further, there was a marginal but significant three-way interaction between AF, RF, and Sibilance ($\beta = -0.093$, $z = -1.981$, $p = 0.048$), consistent with the pattern in Figure 3.8, which shows a notable change when the AF×RF interaction term is high (i.e., when both absolute and relative frequencies are high, or both are low): sibilants and nonsibilants become equivalent in accuracy. This is because in such cases frequency biases are the greatest and likely to even out any acoustic differences captured in phonological features. What, then, do the interactions between sibilance of the target phone and the absolute or relative frequency of the target word mean? One possibility, if we examine the fact that the nonsibilant accuracies all hover around the CV average of 80% accuracy and only increase slightly with greater AF/RF, is that this group is simply too diverse to show major benefits from word frequency. Sibilants, on the other hand, show accuracies that are generally above the mean and only drop down around or below the mean when they are disadvantaged by low absolute frequencies or low frequencies relative to the competitor.

Place and manner of articulation exhibit patterns which are more complex to discern as a function of interacting variables in a display such as that in Figure 3.8, however, in combination with numerical analyses a number of key results emerge. The first concerns the impact of word length on both manner and place distinctions. As Figure 3.8 illustrates, the overall patterns shown earlier in Figure 3.5 are primarily driven by the polysyllabic items. For example, the decline in accuracy from dorsals and [LOW] coronals to labials, [HIGH] coronals, and glottals averages 12% in polysyllables, compared with 4% in monosyllables. Similarly, the lower accuracy on affricates shown in Figure 3.5 is only present in polysyllabic items. These effects are confirmed in signif-

icant Length×Place ($\chi^2(4) = 36.5,\ p < 0.001$) and Length×Manner ($\chi^2(2) = 13.4,\ p = 0.001$) interactions. These patterns, including the trend that fricatives are more accurately perceived than plosives in monosyllables (by approximately 3%), but not in polysyllables (accuracy difference $<$ 0.05%), are all consistent with the general phenomenon of faster rates of production and greater reduction in articulatory gestures with increasing linguistic unit size. For example, labials in general are more difficult to perceive than coronals or velars because both the release burst in the case of plosives and the friction in the case of fricatives are weak in amplitude and lack a front resonating cavity to provide clear place information. Further, their relatively greater coarticulation with the following vowel means there is potentially less information in the formant transition than for coronals or velars. Increasing the speed and reduction of articulation should further constrain the already sparse set of cues to labial obstruents, while coronals and dorsals, though also impacted by these effects, are likely to require greater perturbations to achieve the same losses in perceptibility, simply because they have a wider base of acoustic information to begin with. Again, it is unclear why affricates, and postalveolars in general, diverge from this pattern; these results should be clarified in the analysis of specific contrasts and error patterns in the following section.

Regarding word frequency, both AF and RF show a similar effect of yielding the greatest place and manner distinctions at the lowest word frequencies. This effect is significant for both interactions with Relative Frequency (Place×RF: $\chi^2(4) = 13.44,\ p = 0.009$; Manner×RF: $\chi^2(2) = 15.1,\ p < 0.001$); however, neither Place nor Manner interact significantly with Absolute Frequency (ps > 0.1). Examining first place of articulation, labials, glottals, and [HIGH] coronals are consistently more poorly perceived than [LOW] coronals and dorsals, with the latter distinction (between [HIGH] coronals and dorsals / [LOW] coronals) exhibiting a notable reduction at higher absolute/relative frequencies. Ultimately, however, these patterns point to a fairly consistent effect of place of articulation that merely increases at lower frequencies. Manner of articulation shows a more dramatic leveling off of differences between plosives, affricates, and fricatives at higher absolute and relative target frequencies; however, the general result remains that under the extreme loss of facilitating top-down information (i.e., at the lowest absolute and relative frequencies), fun-

damental differences in obstruent features, and thereby the acoustic characteristics that define such features, emerge as the primary determinants of listener behavior.

Finally, there is a significant interaction between Place, AF, and RF ($\chi^2(4) = 9.867$, $p = 0.043$) which goes in the opposite direction of the sibilance effect; namely, at extremes of target/relative frequency (high AF and RF, and low AF and RF), the manner and place distinctions described above are at their greatest. This apparent inconsistency, however, can be accounted for along with the sibilance results by pointing to the behavior of each feature at low relative and absolute frequencies. In the case of the sibilance feature, at low frequencies the difference between sibilants and nonsibilants disappears, whereas for place and manner it is at the lowest frequencies that distinctions are the greatest. Therefore, these interactions appear to be driven primarily by listener behavior at the lowest absolute and relative target frequencies. As noted earlier, for sibilants, the effect of lowering the frequency of the target word is to reduce the sibilant advantage while maintaining a relatively constant recognition accuracy for nonsibilants, which again represent the larger phonetic class. For manner and place we see notable declines in accuracy on affricates and coronals, respectively, which serve to enhance rather than reduce featural contrast. But these effects are equivalent in that the combination of absolute and relative frequency effects serves to enhance any distinctions that were brought about by the two independently. It just happens to be the case that for sibilance, since the class affected started out at such an advantage the end result is a collapse in the distinction rather than an enhancement.

With the exception of manner distinctions, which though present are substantially reduced in Experiment 1b, the patterns above are robust across sub-experiments. See Figures A.13 and A.14 in the appendix for the complete featural accuracy results by word length and frequency in Experiments 1a and 1b.

Word-medial position (VCV). Figure 3.9 displays featural accuracy patterns in VCV position as a function of word length and frequency (AF, RF, AF×RF). The results for voicing and sibilance in Figure 3.9 extend the overall patterns reported in Figure 3.6, in that the perceptual advantage for

VCV contrasts with voiceless or sibilant target phones is not only more robust than in CV position, but the advantage shows less variability across word lengths and frequencies.[14] That is, words in VCV minimal pairs with voiceless obstruent target phones are consistently more accurately perceived than voiced obstruents by between 4 and 8%, while those with sibilant targets exhibit a 4-11% advantage over nonsibilants. Statistical analyses generally confirmed this visual constancy, and with the exception of a significant interaction between sibilance and word length, where the sibilant advantage was significantly reduced in trisyllabic items ($\beta = -0.494$, $z = -2.246$, $p = 0.025$), no significant interactions were found between voicing/sibilance and word length, absolute frequency, relative frequency, or AF×RF (ps > 0.1).

Regarding manner of articulation, there is a relatively consistent advantage of fricatives over the remaining classes, and a consistent disadvantage for plosives and flaps. Two notable exceptions are the flap > fricative relation in trisyllables, which though not significant does represent a significant change from the disyllabic relation between the two classes (di: $\beta_{fl-fr} = -0.617$, $z = -8.478$, $p < 0.001$; tri: $\beta_{fl-fr} = 0.128$, $z = 0.482$, $p > 0.1$; $\beta_x = 0.745$, $z = 2.710$, $p = 0.007$), and the ultimate advantage of affricates over fricatives and plosives at high absolute and relative frequencies. The latter effect derives from the significantly greater influence of AF ($\beta_x = 0.259$, $z = 2.563$, $p = 0.010$), and the consistent though non-significant effect of RF ($\beta_x = 0.143$, $z = 1.689$, $p = 0.091$) on affricate target phones than on the other manners of articulation.

Finally, place of articulation also does not show an overall interaction with word length ($\chi^2(3) = 5.94$, $p > 0.1$), though there is a trending reversal in accuracies between [HIGH] coronals and dorsals as a function of word length ($\beta_x = 0.500$, $z = 1.875$, $p = 0.061$), the former being more accurate than the latter in disyllables ($\beta_{hc-d} = 0.328$, $z = 3.669$, $p < 0.001$), while in trisyllables the two are approximately equal ($\beta_{hc-d} = -0.172$, $z = -0.684$, $p = 0.494$).[15] Place does, however, interact significantly with both AF ($\chi^2(3) = 31.6$, $p < 0.001$) and RF ($\chi^2(3) = 9.1$, $p = 0.028$).

[14] To be clear, the comparison of word length effects in CV and VCV is not one-to-one because the CV distinction is between mono- and polysyllables, whereas the VCV distinction is between di- and tri-syllables, given that monosyllables cannot contain VCV contrasts.

[15] As in Figure 3.9, glottals were removed from the place of articulation analysis due to their sparsity in the data (glottals occur as target phones in only 4 items, as compared with greater than 90 items for the remaining places).

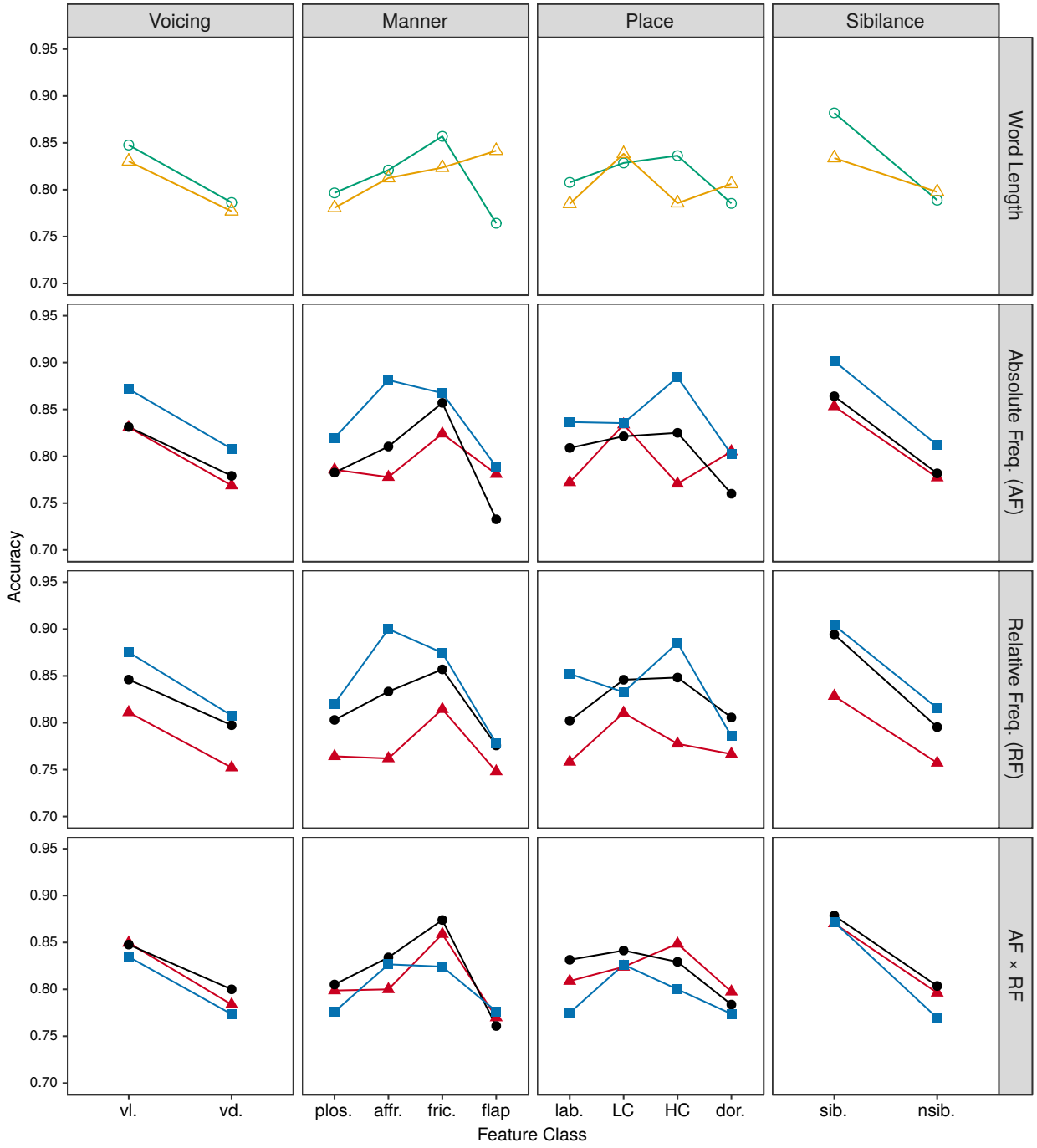

Figure 3.9: Target feature accuracies by Word Length and Word Frequency (AF, RF, RF×AF) in VCV position in Experiment 1. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively. Glottals have been omitted from the place results due to their sparsity in the data.

The three-way interaction between Place, AF, and RF was not significant ($\chi^2(3) = 3.1$, $p > 0.1$). Examining Figure 3.9, we see that the effect of word frequency on place perception is largely restricted to labials and [HIGH] coronals, both of which show substantial increases in accuracy at higher absolute and relative target frequencies. While the two classes are among the least accurate places at low frequencies, they are the most accurate at high frequencies. This general susceptibility of [HIGH] coronals to perturbation from top-down information is consistent with the results word-initial results discussed earlier, while the effect of word frequency on labial recognition accuracy is novel word-medial contrasts.

Among the featural accuracy patterns described above, the robust voicing and sibilance effects are consistent across sub-experiments, as are the general modulating effects of word length and frequency on place and manner of articulation. The one notable discrepancy is in the effect of word length on dorsal obstruent perception, where the significant advantage for dorsals in trisyllables is restricted to Experiment 1a. Given the relative rarity of trisyllabic contrasts in Experiment 1, this pattern is likely idiosyncratic to the items bearing the contrast, rather than a more general characteristic of intervocalic dorsal obstruent perception. See Figures A.15 and A.16 in the appendix for the full target feature by length/frequency results in Experiments 1a and 1b.

Word-final position (VC). Patterns in listener recognition of VC contrasts by feature class, word length, and several measures of word frequency (AF, RF, AF×RF) are shown in Figure 3.10. Right away, an examination of word length effects across the four features reveals that the behavior of word-final contrasts is notably distinct from CV and VCV contrasts. The robust voicing and sibilance effects observed in CV and VCV positions, for instance, are primarily restricted to monosyllables in VC position (mono: voiceless − voiced = 4%, poly: voiceless − voiced = 2%; mono: sibilant − nonsibilant = 6%, poly: sibilant − nonsibilant = −1%), though the interaction is only significant for the sibilance distinction ($\beta_x = 0.504$, $z = 5.071$, $p < 0.001$). Regarding manner of articulation, listeners are significantly more accurate on fricative targets than on plosives in both monosyllables ($\beta_{f-p} = 0.619$, $z = 9.902$, $p < 0.001$) and polysyllables ($\beta_{f-p} = 0.208$, $z = 2.665$,

$p = 0.008$). Further, in monosyllabic targets, affricates are intermediate between the two, being more accurate than plosives ($\beta_{a-p} = 0.204$, $z = 2.001$, $p = 0.045$) and less accurate than fricatives ($\beta_{f-a} = 0.415$, $z = 3.790$, $p < 0.001$), consistent with the general expectation that obstruents with greater noise durations are more perceptible word-finally because they reduce listeners' dependence on VC formant transitions.[16] Finally, patterns in word recognition according to target place of articulation are completely different in mono- and poly-syllabic items ($\chi^2(3) = 47.1$, $p < 0.001$), with the only significant distinction in monosyllables being the reduced accuracy on labials relative to [LOW] coronals ($\beta_{lc-l} = 0.241$, $z = 3.504$, $p < 0.001$), while polysyllables show a substantial labial $>$ [LOW] coronal $\geq$ [HIGH] coronal $>$ dorsal relation (ps < 0.05). As for why these patterns emerge in VC position, the manner result shows a significant reduction in the fricative $>$ plosive advantage from mono- to poly-syllables ($\beta_x = -0.412$, $z = -4.121$, $p < 0.001$) that is consistent with the CV result, and could be due to a weakening of fricative and plosive articulations in longer words, which has the consequence of making fricatives more plosive-like by reducing their duration and noise amplitude, and vice versa as plosives become less likely to exhibit release bursts, and more likely to show flatter formant transitions (consistent with fricative VC transitions) with greater vowel and consonant reduction. However, the place relation in polysyllables does not have any obvious articulatory explanation, and may reflect other characteristics of the particular phonetic contrasts and lexical items involving each place of articulation in VC position. We will return to this point in the next section when listener accuracies by phonetic *contrast*, rather than *category*, are presented.

Regarding frequency effects, Figure 3.10 illustrates both consistencies and discrepancies with the CV/VCV results. First, word-final voicing contrasts are consistent with CV and VCV positions in showing robustness to changes in absolute frequency ($\beta_x = 0.030$, $z = 0.655$, $p > 0.1$), but as for relative frequency effects, VC position is distinct in only preserving the voiceless $>$ voiced pattern at intermediate values of RF; i.e., when top-down disambiguating information is at a minimum. More precisely, there is a significant interaction between target obstruent voicing and relative fre-

[16] Affricates in polysyllables were not analyzed due to their sparsity in the data.

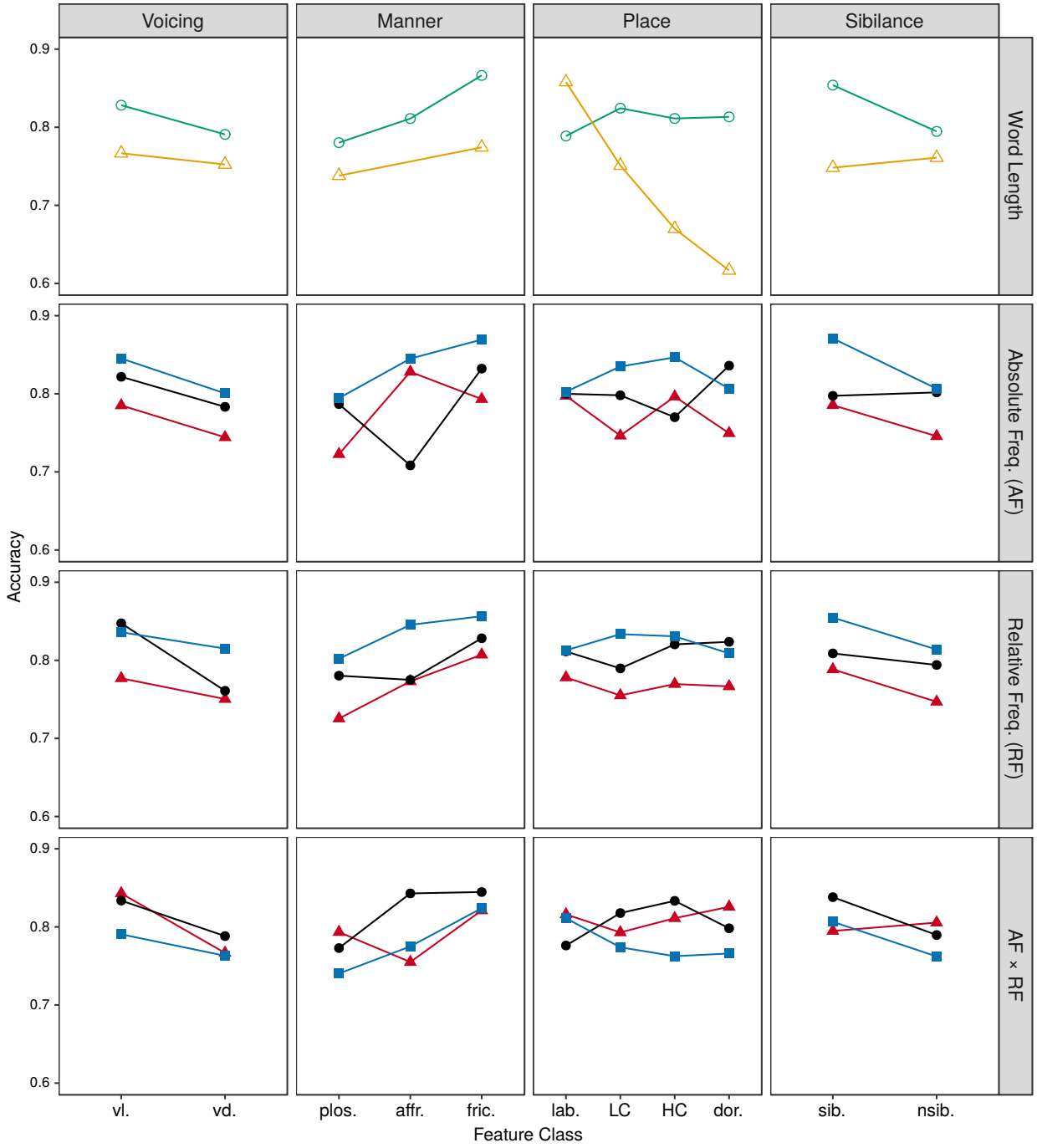

Figure 3.10: Target feature accuracies by Word Length and Word Frequency (AF, RF, RF×AF) in VC position in Experiment 1. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively. Affricates in polysyllabic items have been omitted from the manner results due to their sparsity in the data.

quency ($\beta_x = 0.120$, $z = 2.646$, $p = 0.008$) such that RF has a greater impact on voiced items ($\beta = 0.244$, $z = 7.322$, $p < 0.001$) than on voiceless items ($\beta = 0.124$, $z = 4.036$, $p < 0.001$). It is not clear whether this result is acoustic (word-final voiced obstruents tend to lengthen the vowel and the word in general, allowing listeners greater time to access the lexicon and apply frequency biases than in shorter, voiceless-final words) or lexical (voiced-final words are dominated by morphological distinctions between [d]- and [z]-suffixed words, which could behave differently with respect to the use of top-down information than in monomorphemic contrasts). Nevertheless, the general result that the voiceless advantage is at its maximum when lexical frequency biases are at a minimum points to a robust difference between word-final voiced and voiceless obstruents in terms of their acoustic perceptibility.

Sibilance, on the other hand, shows an opposing pattern to the relative frequency effects on voicing, and the general sibilance trends in CV and VCV positions; namely, the sibilant $>$ nonsibilant distinction is at a minimum at intermediate absolute and relative frequencies. This result is captured in a significant three-way interaction between Sibilance, AF, and RF ($\beta_x = -0.144$, $z = -3.216$, $p = 0.001$), where there is a significant negative relation between AF and RF in nonsibilant-final words ($\beta = -0.095$, $z = -5.057$, $p < 0.001$) that is not present in sibilant-final words ($\beta = -0.083$, $z = -1.327$, $p > 0.1$). Thus, it appears that in word-final contrasts where the target is a sibilant obstruent, frequency effects are significantly reduced, a result which is consistent with the greater robustness of sibilants acoustically, making bottom-up information less susceptible to top-down modulation.

Manner effects are relatively consistent across absolute and relative target frequency ranges, and the more robust fricative $>$ plosive relation is preserved regardless of AF, RF, or the interaction between the two (ps > 0.1). The only significant interaction between word frequency and manner of articulation occurs with affricates, which show a significantly greater effect of AF$\times$RF than do plosives ($\beta_x = 0.269$, $z = 2.315$, $p = 0.021$), where affricates are much better recognized when frequency biases are reduced than when they are enhanced. Nevertheless, given the relatively minor role of affricates in word-final contrasts relative to fricatives and plosives, the more critical

result of the manner analysis is the robustness of fricative perception over that of plosives in the face of variation in word length and word frequency.

Finally, place distinctions in word-final contrasts are generally minor compared to those in CV and VCV positions, and therefore the role of word frequency in modulating such distinctions is constrained as well. Only absolute target frequency interacts significantly with place of articulation ($\chi^2(3) = 14.0$, $p = 0.003$), an effect which derives from the significant positive impact of AF on recognition of [LOW] coronal-final ($\beta = 0.256$, $z = 9.034$, $p < 0.001$) and dorsal-final ($\beta = 0.174$, $z = 2.739$, $p = 0.006$) words, but no such effect for labials or [HIGH] coronals (ps > 0.05).

Among the aggregate effects of word length/frequency on target feature recognition in Experiment 1, the patterns for voicing, manner, and sibilance show the greatest replicability across sub-experiments (Experiment 1a does show a more robust effect of voicing than 1b, but the general patterns remain). Place effects are less consistent, particularly the interaction between place and word length, where the only constant patterns between Experiments 1a and 1b are the relatively narrow differences among places in monosyllables, and the significant enhancement of labials over the remaining classes in polysyllabic items. Results for [HIGH] coronals and dorsals in polysyllabic items are inconsistent across sub-experiments, likely due to their much greater sparsity in word-final contrasts in the lexicon relative to labials or [LOW] coronals. See Figures A.17 and A.18 in the appendix for the full VC results in Experiments 1a and 1b.

3.3.3.5 Summary of category recognition results

The sections above provide a complex array of patterns in listener word recognition as a function of the category (phone/feature) of the target member of the obstruent contrast distinguishing each minimal pair in the 2AFC task of Experiment 1. However, these results can be summarized into several key outcomes that will ultimately impact both the relative importance of different acoustic cues in distinguishing minimal obstruent contrasts in the lexicon, and the structural dependence of the system of lexical contrasts on particular phones and feature classes.

First, sibilants, particularly sibilant fricatives, stand out as more robust to background noise

and more easily perceived than any other subset of the obstruents. This relation holds across CV, VCV, and VC positions, and is evident in both the accuracy ranks of individual phones, and the aggregate accuracy of the sibilant feature class. Sibilants are further relatively insensitive to effects of word length and frequency, though as with most classes they do tend to generally decline in accuracy with increasing word length (this holds across positions and increases in magnitude from CV to VCV to VC), and the impact of word frequency is generally to bring them toward the mean. Further, since sibilants start off at a much higher accuracy than most other obstruents, these declines in accuracy with more negatively biasing word frequency information can in many cases be more sizable than on other classes. This general robustness of sibilants suggests that certain cues that we expect to be prominent in sibilants, such as the amplitude and frequency of the spectral peak, must be well-perceived by listeners and therefore highly weighted in models of obstruent cue integration.

At the other end of the spectrum are the labial obstruents, which, with the exception of the fricatives [f, v] word-finally, consistently rank in the lower half of the obstruent phones in terms of listener accuracy. These effects are particularly pronounced word-initially, where listeners have weak cues from the consonant (the burst/noise spectrum), and where the formant transitions in the vowel are less distinct than they are for labials in post-vocalic position. This distinction is further widened in CV position as word length increases, and is not notably mitigated by any place-specific frequency biases. Therefore, we expect labials as a class to be less stable in perception and pose a greater risk to the system of lexical contrasts as a whole. Further, cues to labial obstruents might be down-weighted in perception due to their greater vulnerability to changes in stimulus conditions, particularly the background noise level, word length, and position of the contrast.

Returning to points of robustness, voiceless obstruents are consistently well perceived in Experiment 1, occupying most of the higher ranks of phone accuracies, and outperforming voiced obstruents in nearly all stimulus conditions, with the distinction particularly notable in VCV contrasts. This result may reflect both *paradigmatic* effects—voiceless obstruents are more energetic and therefore more likely to transmit identifying information in a noisy signal—and *syntagmatic*

effects—voiceless obstruents, relative to voiced, are more easily separable from adjacent voiced vowels due to the availability of segmentation cues from the voicing feature change.

Finally, regarding manner of articulation, among the two major obstruent manner classes, fricatives are generally better perceived than plosives, particularly in VCV and VC positions, and in monosyllabic words. This result is likely due to both the greater length of fricatives, allowing more time for listeners to pick up on spectral information from the consonant, and the greater average amplitude of fricatives, making them more robust to interference from background noise. This result has implications for both the cues listeners use in word recognition and the ultimate contribution of both fricative and plosive contrasts to the stability of the contrast system as a whole.

3.3.4 Phonetic contrast recognition

The above results give a general picture of how the inventory of obstruent phones maps onto lexical contrast perception. However, lexical contrast is ultimately dependent on the perception of acoustic distinctions between pairs of phones, with some pairs (e.g., [b, v]) substantially more similar, and thus more confusable, than others (e.g., [b, ʃ]). Therefore, in order to more precisely understand the phonetic system as distributed over a system of distinctions in the lexicon, we next consider listener word recognition by contrast. As in the category analysis, we first outline the distribution of CV, VCV, and VC contrasts presented in Experiment 1, following which listener performance is analyzed by segmental and featural contrast.

3.3.4.1 Phone distributions

We begin with a summary of contrast distributions by obstruent phone, with the primary goals being (1) to identify which contrasts are most prevalent in a given position (and conversely which are rare or absent entirely), and (2) to provide further detail on how the target phone distributions presented earlier are distributed among particular contrasts.

Word-initial position (CV). Table 3.9 displays the distribution of word-initial contrasts presented in the minimal pair stimuli of Experiment 1, shown in matrix format where the total in each cell represents the number of items exhibiting a contrast between the phones in the corresponding row and column. The most frequent contrasts in word-initial position are [p, f], [b, k], [k, h], and [s, ʃ], while there are many contrasts absent entirely or only occurring once. Given the large number of stimuli in Exp. 1, and the sampling design employed in choosing items, these uncommon contrasts are understood to represent real gaps in the lexicon.

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**h**	**v**	**ð**	**z**
p	4	7	7	2	5	5	3	11	1	5	3	3	0	0	1
	t	8	9	5	2	0	6	4	4	7	4	5	3	0	1
		k	10	8	4	3	1	3	0	9	2	10	3	1	0
			b	4	2	1	2	6	5	6	1	5	2	0	0
				d	0	3	3	5	0	3	1	4	0	1	0
					g	2	4	1	1	2	3	1	1	0	0
						ʧ	0	2	0	3	2	4	2	0	0
							ʤ	3	0	6	0	1	1	1	0
								f	1	7	4	2	3	0	0
									θ	3	0	1	0	0	0
										s	10	9	5	0	0
											ʃ	4	1	1	0
												h	3	0	2
													v	1	0
														ð	0

Table 3.9: Distribution of CV contrasts in minimal pair stimuli in Experiment 1.

Recalling the target phone distribution in Table 3.3, the most frequent phones in CV position are [s, k, t, b, p, h], in that order, while [θ, ð, z] are comparatively rare. Looking at the contrast distribution in Table 3.9, we see that the occurrence of [s] in contrasts with other phones is relatively balanced ($\hat{H} = 0.871$), with [ʃ, h] the most frequent competitors, and no contrasts between [s] and [ð, z]. The voiceless plosives [p, t, k] are moderately less balanced than [s] ($\hat{H}$ = 0.841, 0.866, and 0.832, respectively). The most frequent contrasts with [p] come with the set [f, k, b], all relatively

similar acoustically, while contrasts with [v, ð] are absent and [θ, z] contrast with [p] only once. The voiceless alveolar plosive [t] occurs most frequently in contrast with [b], followed closely by the other voiceless plosives [k, p] and the voiceless alveolar sibilant [s]; missing contrasts are with the set [ʧ, ð], with [z] also occurring only once. Of the three voiceless plosives, [k] is the least balanced in its contrasts, occurring 10 times each with [b, h], 9 times with [s], and 8 times each with [t, d], meaning 5 phones account for 65% of the contrasts with [k] in Experiment 1. Finally, [b] and [h] are highly frequent as target phones, and both occur most frequently in contrast with [k], while the next most common contrasts with [b] and [h] are [t] and [s], respectively. In general, however, [h] is more balanced in its contrast distribution than [b] ($\hat{H} = 0.853$, as compared with 0.835), which more closely aligns with the distribution of contrasts among the other plosives.

The distributions in the sub-experiments 1a and 1b are similar, but given the general sparsity of the contrast matrix there are many notable discrepancies as well that may influence both the replicability of the contrast accuracy results below and the cue integration model in Section 4.4. Excluding differences of 1-2 items, contrasts of greater frequency in Experiment 1a are the following: [b, θ] (5), [k, d] (4), [k, g] (4), [k, s] (3), [p, ʃ] (3), [ʧ, s] (3), and [h, v] (3). In Experiment 1b, the following contrasts are more common than in 1a: [p, ʧ] (5), [f, s] (5), [s, h] (5), and [p, h] (3). See Tables A.2 and A.3 in the appendix for the full Experiment 1a/b distributions.

Word-medial position (VCV). The number of minimal pairs comprising each segmental contrast in VCV position in Experiment 1 are shown in Table 3.10. Given the prevalence of flaps intervocalically, among the most common obstruent contrasts are several with [ɾ]; i.e., [k, ɾ] define 13 minimal pairs (MPs), [z, ɾ] appear in 12 MPs, [b, ɾ] in 10 MPs, and [ʤ, ɾ], [s, ɾ], and [v, ɾ] appear in 9 pairs each. Other frequent contrasts are [p, k] at 14 items, [f, v] at 9 items, and [k, s] and [s, v] at 8 items each. Regarding the contrast distributions among the most frequent target phones in VCV position, [ɾ, p, k, v, s, f, z], as noted above, [ɾ] is frequent enough that it occurs in a contrast with nearly every other obstruent, with only [ʒ, h] absent entirely (excluding [t, d], which are in complementary distribution with [ɾ]), and among the present contrasts only [θ] occurs less

than 3 times. Among the frequent target plosives listed above, the contrast distribution with the voiceless velar [k] is moderately more balanced than with the labial [p] ($\hat{H} = 0.845$, as compared with 0.828), as nearly half of the occurrences of [p] in VCV position (49%) are in contrast with three obstruents: [k], [f], and [ɾ]. By comparison, [k] occurs 14 times with [p], 13 times with [ɾ], 8 times with [s], 6 times each with [f, ʧ], and 5 times each with [b] and [v]. Contrasts absent entirely from the [p] set are [h, ð, ʒ], and those absent from [k] are [d, θ, h].

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**h**	**v**	**ð**	**z**	**ʒ**	**ɾ**
p	1	14	3	3	5	1	1	7	1	3	3	0	5	0	3	0	7
	t	2	4	0	1	0	4	5	0	4	2	0	1	0	2	0	0
		k	5	0	4	6	2	6	0	8	3	0	5	1	5	1	13
			b	1	3	3	0	2	0	0	2	1	2	0	1	0	10
				d	1	2	0	1	0	2	0	0	0	0	1	1	0
					g	0	0	1	1	2	2	0	2	0	1	0	6
						ʧ	1	2	0	5	1	0	1	0	0	1	6
							ʤ	1	1	4	0	0	2	0	3	1	9
								f	1	3	2	0	9	0	1	0	3
									θ	0	0	0	0	0	0	0	1
										s	2	0	8	0	6	2	9
											ʃ	0	3	1	0	1	3
												h	3	0	0	0	0
													v	2	7	0	9
														ð	0	0	4
															z	1	12
																ʒ	0

Table 3.10: Distribution of VCV contrasts in minimal pair stimuli in Experiment 1.

Finally, the fricative set [f, v, s, z] occurs frequently intervocalically in obstruent contrasts in English, and this set is similarly distributed as [p, k] ($\hat{H} \geq 0.83$), with the exception of [z] ($\hat{H} = 0.751$), which primarily occurs with [ɾ] (12 MPs), [v] (7), [s] (6), and [k] (5). The voiceless counterpart to [z], [s], occurs most often in contrast with [ɾ, v, k, z], but occurs more than once with 12 out of 17 obstruents. Finally, the voiceless labiodentals, [f, v], occur most often in contrast

with each other (9 MPs), with [v] contrasting equally often with [ɾ], and also frequently with the alveolar fricatives [s, z] at 8 and 7 items, respectively. The next most common contrasts with [f] (after [v]) come from the voiceless plosive set, with [p] at 7 minimal pairs, [k] at 6, and [t] at 5.

Comparing the distributions in Experiment 1a to 1b, we find the following contrasts notably more frequent in 1a: [v, z] (5), [v, ɾ] (5), [k, ʧ] (4), [ʧ, ɾ] (4), [k, ʃ] (3), [p, f] (3), [p, ɾ] (3), [b, g] (3), and [ʃ, ɾ] (3). In Experiment 1b, the following contrasts occur more frequently than in 1a: [b, ɾ] (6), [s, ɾ] (5), [k, b] (5), [p, b] (3), [p, v] (3), [t, f] (3), and [ʧ, s] (3). See Tables A.4 and A.5 in the appendix for the full VCV contrast distributions in Experiments 1a and 1b.

Word-final position (VC). Table 3.11 shows the distribution of word-final contrasts in Experiment 1. Here, because of the voiced alveolars [d, z], a great number of the most frequent contrasts in VC position occur with one of these two obstruents. For instance, among the contrasts appearing 10 times or more, two-thirds are contrasts with [d] or [z]; namely, the [d, z] contrast is by far the most frequent at 67 items (21% of all word-final contrasts in Exp. 1), followed by [t, d] at 16, [t, z] at 13, and [s, z] at 11. The remaining two highly frequent contrasts are between [t, k] (11 MPs), and [p, t] (10 MPs). In general, however, [d] and [z], though occupying a similar role word-finally as [ɾ] word-medially, are much less balanced in their distributions ($\hat{H}_d = 0.62$, $\hat{H}_z = 0.57$, as compared with 0.86 for [ɾ]), meaning their impact on the word final obstruent system is expected to be far more skewed toward particular contrasts, and the acoustic features that define them.

Beyond [d] and [z], other frequent target phones in VC position are the voiceless plosives [p, t, k], the voiceless alveolar sibilant [s], and the voiced labiodental [v]. The voiceless plosives are fairly balanced in their contrast distributions considering the generally greater asymmetry in VC position ($\hat{H} > 0.75$), with some of their most common contrasts occurring within the voiceless plosive set (item counts between 8 and 11), and with [d, z], and only one non-trivial gap across the set (i.e., excluding highly infrequent sounds such as [ð] and [ʒ]): [p, ʃ]. The voiceless alveolar sibilant [s] is also widely distributed ($\hat{H} = 0.849$), occurring most frequently in contrasts with [z] and [t], and only non-trivially absent from a minimal contrast with [ʧ]. Finally, [v] is quite common

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**v**	**ð**	**z**	**ʒ**
p	10	8	2	1	2	4	1	2	1	3	0	3	1	5	0
	t	11	1	16	6	3	3	5	2	6	3	8	0	13	0
		k	2	8	1	4	1	1	2	5	2	1	0	4	0
			b	2	0	0	1	3	0	2	1	0	0	0	0
				d	1	1	3	3	2	4	5	5	1	67	1
					g	3	1	0	0	2	3	2	0	2	0
						ʧ	2	2	0	0	2	0	0	1	0
							ʤ	1	1	1	0	0	0	1	0
								f	0	2	1	3	0	2	0
									θ	3	2	0	0	1	0
										s	4	3	0	11	0
											ʃ	0	0	0	0
												v	0	9	0
													ð	0	0
														z	1

Table 3.11: Distribution of VC contrasts in minimal pair stimuli in Experiment 1.

in word-final obstruent contrasts at 7th overall, but is much more skewed in its distribution than [p, t, k, s]. Of the 15 potential contrasts in VC position, [v] occurs with only 8, with 50% of its items coming from two contrasts: [v, z] and [v, t]. That is, the role of [v] in word-final position is highly constrained; consequently, we expect the acoustic cues that listeners rely on in identifying [v] word-finally to reflect primarily the narrow requirements of this set: place information in the case of [z]; place, manner, and voicing in the case of [t].

Finally, comparing contrast distributions in the two sub-experiments, the following contrasts were notably more frequent in Experiment 1a than in 1b: [p, ʧ] (4) and [s, z] (3). In Experiment 1b, the following contrasts occurred notably more frequently than in 1a: [p, k] (6), [t, d] (6), [d, z] (5), [t, k] (3), [t, ʧ] (3), [d, v] (3), and [f, v] (3). See Tables A.6 and A.7 in the appendix for the full word-final contrast distributions in Experiments 1a and 1b.

3.3.4.2 Feature distributions

Next we summarize the distribution of obstruent contrasts by distinctive feature transmitted; that is, for each segmental contrast [*a*, *b*] we tabulate whether *a* and *b* differ in *voicing*, *manner*, *place*, or *sibilance*, where each feature is considered independently of the remaining features. For example, both [b, p] and [b, f] are equivalent in transmitting voicing information, despite the former being differentiated only by voicing, while the latter contrasts along both voicing and manner dimensions. This analysis follows that of Miller & Nicely (1955), and is the most closely aligned with the problem of acoustic cue integration in lexical contrast perception, as voicing, manner, place, and sibilance cues remain informative for the identity of the target in contrasts distinguished by two or more features, a set which accounts for 76% of contrasts in Experiment 1.

Word-initial position (CV). Table 3.12 shows the percentage of word-initial contrasts distinguished by voicing, manner, place, and sibilance. Overall, the distribution generally conforms with expectations based on the number of categories comprising each feature; namely, place contrasts (labial, [LOW] coronal, [HIGH] coronal, dorsal, glottal) are the most frequent at 70%, followed by manner contrasts (plosive, affricate, fricative) at 58%, and then the two binary features—voicing and sibilance—the least frequent at 46% and 41%, respectively. Sibilance appears less commonly as a distinguishing feature because it is less balanced than voicing is, resulting in a larger number of contrasts within the nonsibilant class (the larger of the two) than occur within voiceless or voiced sets. Thus, in terms of feature distributions in word-initial contrasts, there are no clear patterns with notable implications either for the accuracy analysis below, or for the general structure of the system.

Word-medial position (VCV). The distribution of featural contrasts intervocalically, shown in Table 3.13, though predictable from the contrast distribution in Table 3.10, does not directly follow from the organization of features as in CV position. That is, place has a greater number of classes than manner by 1—manner = {plosive, affricate, fricative, flap}, place = {labial, [LOW] coronal,

	Voicing	Manner	Place	Sibilance
Exp. 1a	50.0	58.8	83.1	36.9
Exp. 1b	42.5	57.5	80.0	45.0
Total	46.2	58.1	81.6	40.9

Table 3.12: Percentages of word-initial contrasts distinguished by Voicing, Manner, Place, and Sibilance in Experiment 1.

	Voicing	Manner	Place	Sibilance
Exp. 1a	45.6	70.0	75.6	45.6
Exp. 1b	53.4	67.1	75.2	39.1
Total	49.5	68.5	75.4	42.4

Table 3.13: Percentages of word-medial contrasts distinguished by Voicing, Manner, Place, and Sibilance in Experiment 1.

[HIGH] coronal, dorsal, glottal}. However, manner contrasts are nearly equal in frequency, at 69%, as compared with 75% for place contrasts. This is because the most frequent contrasts—those between the flap [ɾ] and other obstruents—amount to a manner distinction, and conversely, with [LOW] coronals being the most common place of articulation, many of these contrasts will not involve a place distinction. As in CV position, both voicing and sibilance are the least frequent featural contrasts intervocalically, but both are slightly more frequent in VCV than in CV position (50% for voicing, as compared with 46% in CV; 42% for sibilance, as compared with 41% in CV). In summary, the featural contrast distribution in Table 3.13 demonstrates the critical role played by manner contrasts in the English obstruent system, and the necessity for acoustic cue integration models to capture those signal characteristics that reliably transmit manner information.

Word-final position (VC). Table 3.14 shows the distribution of VC minimal pairs by contrastive feature, and as with VCV position, the result is unsurprising given the notable asymmetry of VC position, where over 20% of items are distinguished by the [d, z] contrast. Nevertheless, this fact does not undermine the sizable contribution of manner and sibilance features to the system of word-final obstruent contrasts in English. Manner is the most frequent contrastive feature at 64%,

	Voicing	Manner	Place	Sibilance
Exp. 1a	46.3	65.0	56.7	56.9
Exp. 1b	37.7	62.9	52.8	56.0
Total	42.0	64.0	54.9	56.4

Table 3.14: Percentages of word-final contrasts distinguished by Voicing, Manner, Place, and Sibilance in Experiment 1.

followed by sibilance at 56%, both being driven, again, primarily by the prevalence of [d, z] due to English inflectional morphology. As a further result of this asymmetry, place and voicing are less informative word-finally, accounting for 55% and 42% of VC contrasts, respectively: far lower than in CV or VCV positions. Both results—the importance of manner and sibilance contrasts word-finally, and the subordination of voicing and place distinctions—imply a notably distinct acoustic cue hierarchy word-finally than in the other two positions.

3.3.4.3 Phonetic contrast accuracy

As in the target category analysis in Section 3.3.3.3, listener recognition of phonetic contrasts is examined at both *phone* and *feature* levels. The former considers each pair of obstruent phones that mark a distinction between items in the 2AFC task, and contrary to the *undirected* form which is typically adopted for such relations—i.e., in the contrast [a, b], the role of a or b as stimulus or response is ignored—here we treat contrasts as *directed*. To be precise, any given contrast set [a, b] is comprised of two directed contrasts, one with a as the stimulus (written $a \rightarrow b$), and the other with b as the stimulus ($b \rightarrow a$), where $a \rightarrow b :\neq b \rightarrow a$. The utility of examining contrasts in this form is that it can uncover asymmetries in accuracy on a contrast based on the acoustic input (i.e., which member of the contrast is most easily recognizable in the stimulus), and general response biases, though for the latter it must be kept in mind that the data below represent decisions based on contrasts between whole words, and thus if there is a bias toward [t], for example, it must reflect a bias toward [t]-onset words in general, not the [t] category in particular.

Similarly, the analysis of features must also take into account the directionality of feature trans-

mission (e.g., [−voice] → [+voice] vs. [+voice] → [−voice]). This is done by averaging over directed contrasts that match certain feature relations (e.g., [−sibilant] → [+sibilant], [+α place] → [−α place]), where all cases of a feature match between target and competitor are excluded to avoid trivial contributions to feature accuracy. Finally, we note that though the framework of *transmitted information (TI)* analysis is used in the formulation of directed featural accuracy, the formal analysis of TI, as in Miller & Nicely (1955), is not used in the present study because it is a symmetric measure, and thus cannot account for directionality in contrast perception.

We begin with the analysis of lexical contrast accuracy by obstruent phone pair, where as stated above, directionality matters; i.e., $a \rightarrow b :\neq b \rightarrow a$. As before, separate analyses are conducted in CV, VCV, and VC positions. Finally, because contrast distributions when broken down by phone are generally sparse, we will limit discussion to only the accuracy results that derive from multiple items. That is, accuracies deriving from a single minimal pair will be ignored, except as contributions to wider patterns in classes of phones.

Word-initial position (CV). Figure 3.11 shows listener accuracies by phonetic contrast in word-initial position. Because of the combinatoric complexity introduced by the study of ordered pairs among $n = 16$ phones—i.e., $n(n-1) = 240$ potential pairs, though not all combinations are present in the database—contrast accuracies are presented in a confusion matrix layout, where each cell ij indicates the overall accuracy of listeners on the contrast $i \rightarrow j$, with darker blue indicating the highest accuracy quartile ($> 87\%$), dark red the lowest quartile $< 75\%$, and light red/blue the second and third quartiles, respectively. White cells indicate contrasts that are missing from the database, including the non-contrastive identity pairs that form the diagonal of the matrix. Shown alongside the matrix, to the right and below, respectively, are summary diagrams of a hierarchical clustering of the row (stimulus) and column (response) accuracy distributions for each phone. In these *dendrograms*, phones that are closer together—i.e., separated by fewer nodes in the tree—are more similar in distribution, meaning they share more in common in terms of the phones they are accurately distinguished from, and those they are often confused with. For example, the set [p,

ʃ, h] clusters together in the stimulus accuracy tree in the upper right panel of Figure 3.11 due to the fact that all three are commonly confused with the voiceless plosive series, as well as being relatively accurately distinguished from the affricates [ʧ, ʤ].

Several notable patterns are evident in Figure 3.11. First, the voiceless plosives [p, t, k] are commonly misidentified with each other, while their voiced counterparts [b, d, g] are more varied internally. Outside of their internal confusability, however, [p, t, k] exhibit distinct accuracy distributions, which can be seen in their wide separation in the hierarchical clustering of stimulus accuracies (row vectors) in the upper right panel of Figure 3.11. The labial [p], for instance, is generally poorly perceived, particularly when in contrast with [f]; however, recognition is better with more posterior articulations, notably [ʃ, ʧ, ʤ, g]. On the other hand, errors on the alveolar [t] are less widely distributed, and are primarily concentrated among other plosives; that is, with the exception of [ʃ, v], listeners are above average at recognizing [t] when in contrast with affricates and fricatives. Given the greater number of errors on such contrasts when [t] is the competitor, this result could be a response bias due to the generally greater frequency of [t] in the lexicon.[17] Finally, the velar [k] is more starkly manner-constrained in its error patterns than [p] or [t], as it shows frequent errors with all stops but [b, d], and on the other hand, is highly accurate when in contrast with all fricatives except [v], with [s, ʃ, h] contrasts all in the highest accuracy quartile ($> 87\%$).[18] However, as with [t], [k] is highly frequent, and listeners' generally below-average accuracy when [k] is a competitor is also indicative of a response bias toward [k]. For a full listing of bias estimates by contrast, where bias is measured as the proportion of responses given to phone i when contrasted with phone j, see Table 3.15. From this measure, [t] has a very slight bias, at 0.51 on average (though some contrasts approach 0.6), while [k] is much higher at 0.55 on average, with the [k] bias on the [k, ʤ] contrast reaching 0.68.

Considering next the voiced plosives [b, d, g], the lingual subset, [d, g], exhibit similar accuracy patterns (note their close clustering in the dendrogram in Figure 3.11), and are particularly well

[17] Explanations based on markedness and the underspecification of [t], such as in the FUL model of Lahiri & Reetz (2002), could also be derived from this result.

[18] The contrast between [k] and [ð] is also highly accurate, but only represents a single item.

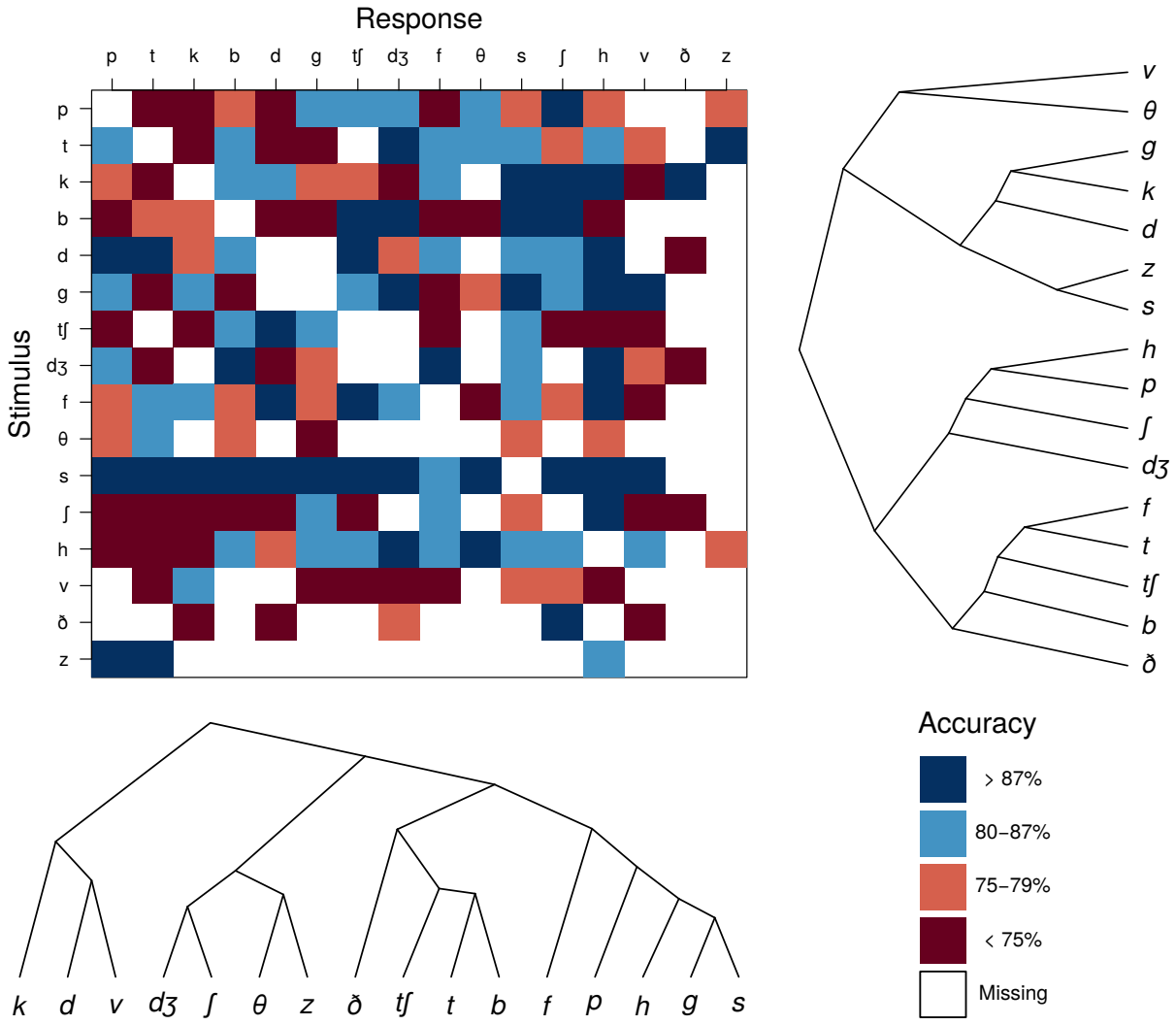

Figure 3.11: Listener accuracies by contrast (CV). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones a and b represented in a greater number of nodes between a and b in the dendrogram.

	p	t	k	b	d	g	ʧ	ʤ	f	θ	s	ʃ	h	v	ð	z	Mean
p	NA	0.46	0.38	0.54	0.44	0.50	0.54	0.50	0.50	0.52	0.44	0.59	0.60	NA	NA	0.42	0.49
t	0.55	NA	0.45	0.55	0.44	0.52	NA	0.58	0.50	0.49	0.46	0.50	0.55	0.56	NA	0.52	0.51
k	0.62	0.55	NA	0.53	0.54	0.48	0.60	0.68	0.49	NA	0.48	0.52	0.60	0.46	0.62	NA	0.55
b	0.46	0.46	0.47	NA	0.43	0.52	0.55	0.48	0.46	0.46	0.46	0.62	0.44	0.50	NA	NA	0.49
d	0.56	0.56	0.46	0.57	NA	NA	0.46	0.54	0.46	NA	0.44	0.55	0.55	NA	0.48	NA	0.51
g	0.50	0.48	0.52	0.48	NA	NA	0.50	0.55	0.48	0.57	0.46	0.50	0.55	0.62	NA	NA	0.52
ʧ	0.46	NA	0.40	0.45	0.54	0.50	NA	NA	0.35	NA	0.45	0.62	0.43	0.44	NA	NA	0.46
ʤ	0.50	0.42	0.32	0.52	0.46	0.45	NA	NA	0.52	NA	0.44	NA	0.45	0.55	0.38	NA	0.46
f	0.50	0.50	0.51	0.55	0.54	0.52	0.65	0.48	NA	0.75	0.50	0.50	0.56	0.56	NA	NA	0.55
θ	0.48	0.51	NA	0.54	NA	0.43	NA	NA	0.25	NA	0.42	NA	0.42	NA	NA	NA	0.43
s	0.56	0.55	0.52	0.55	0.56	0.54	0.55	0.56	0.50	0.58	NA	0.56	0.55	0.56	NA	NA	0.55
ʃ	0.41	0.50	0.48	0.38	0.45	0.50	0.38	NA	0.50	NA	0.44	NA	0.52	0.45	0.42	NA	0.45
h	0.40	0.46	0.40	0.56	0.46	0.45	0.57	0.55	0.44	0.57	0.45	0.48	NA	0.57	NA	0.49	0.49
v	NA	0.44	0.54	0.50	NA	0.38	0.56	0.45	0.44	NA	0.44	0.55	0.42	NA	0.42	NA	0.47
ð	NA	NA	0.38	NA	0.52	NA	NA	0.62	NA	NA	NA	0.57	NA	0.57	NA	NA	0.54
z	0.57	0.48	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.51	NA	NA	NA	0.52

Table 3.15: Response bias on each contrast in CV position in Experiment 1. For each row *i*, which indicates the stimulus phone, the proportion in cell [*i*, *j*] represent the proportion of responses given to phone *i* when contrasted with phone *j*. For example, cell [*p*, *t*] = 0.46 (rounded from 0.455), which derives from a 75% accuracy on the contrast when *p* is the stimulus, and 16% *p* responses when *t* is the stimulus; i.e., (0.75 + 0.16) / 2 = 0.91 / 2 = 0.455.

distinguished from fricatives. The labial, [b], on the other hand, is generally poorly perceived, particularly in contrasts with nonsibilants such as [f, θ, h]. Finally, the voiced plosives show generally lower response biases than the voiceless plosives, at 0.49, 0.51, and 0.52, respectively.

Turning next to the affricates, [ʧ, ʤ], members of this set are quite distinct in their accuracy distributions. Errors on the voiceless affricate [ʧ], for instance, are primarily restricted to other voiceless obstruents, while conversely, [ʤ] is mostly confused with other voiced obstruents. This clear effect of voicing, discussed more precisely later in the analysis of featural accuracy, is not apparent in the results for the other manner classes, which is why the stimulus accuracy dendrogram in Figure 3.11 does not show clear clustering according to voicing. Further, because both [ʧ] and [ʤ] show response biases *away* from themselves and toward their competitors (i.e., both below 0.5, at 0.46), this result is likely acoustic in nature.

Among the fricatives, the dentals, [θ, ð] are highly infrequent and when they do appear are generally poorly recognized. The voiced alveolar, [z] is also infrequent, contrasting only with three phones, [p, t, h], but is accurately perceived in all three instances. However, given the relative rarity [θ, ð, z] in word-initial obstruent contrasts, we will not dwell on them further here.

The voiceless sibilants, [s, ʃ], on the other hand, are frequent and widespread in their appearance in contrasts with other obstruents, but otherwise could not be more distinct in their accuracy distributions. The alveolar, [s], for instance, is above average in accuracy in all contrasts it appears in, with [f] being the only competitor it is not in the highest accuracy range with. On the other hand, [ʃ] is below average in accuracy on all but three contrasts: [ʃ, g], [ʃ, f], [ʃ, h]. Comparing the two on response biases, we see that the bias estimates are consistent with the above results, with [s] at 0.55 and [ʃ] at 0.45, and this relation widens slightly when restricted to just the [s, ʃ] contrast, where [s] comprises 56% of responses. This difference can also be attributed to general acoustic properties of the two sounds, as [s] is notably higher in spectral peak frequency, a characteristic that allows the signal to more clearly stand out from the background noise. The voiceless labiodental, [f], is also generally well-perceived, with the majority of errors occurring when [f] is in contrast with the other labials: [p, b, v].

Finally, [h] and [v] are both quite common and widespread in their contrast distributions, though recalling Figure 3.2, [v] is the most poorly perceived stimulus phone, at below 70%, while [h] is in the middle of the range at between 75 and 80%. Given the poor recognition of [v] overall, the wide range of below-average accuracies in the [v] row of Figure 3.11 is expected, though it does confirm that the aggregate result in Section 3.3.3.3 is robust and not restricted to a narrow set of contrasts. The glottal fricative, on the other hand, is of intermediate accuracy primarily because of listeners' poor performance on contrasts between [h] and the voiceless plosives [p, t, k]. Contrasts between the glottal and other fricatives are generally above average in accuracy. This result lends further support to the mixed classification of [h] in the phonetic and phonological literature, as its noise source is more closely aligned with the aspiration of word-initial plosives than it is with the supraglottal noise sources of other fricatives, which yield notably distinct spectral and amplitudinal characteristics (Stevens, 1971; Shadle, 1985).

Among the patterns noted above, the following are the most consistent across sub-experiments. In both Experiment 1a (Figure A.19) and 1b (Figure A.20) voiced and voiceless plosives remain internally highly confusable, while contrasts between the two sets are moderately more accurate but

not robust. Regarding place of articulation, the labials [p, b] are consistent across sub-experiments in both being poorly perceived in a wide range of obstruent contrasts. Results for the alveolar and velar plosives, however, are more varied in different subsets of the lexicon. Regarding affricates, contrasts with fricatives are generally consistent across Experiments 1a and 1b in showing greater within-voicing errors than between-voicing; however, contrasts between affricates and plosives are more varied in this regard. This result is not surprising given the stronger voicing cues word-initially in fricatives than in plosives (see, for example, the VOI% and LF results in Chapter 2).

In comparison with the stop contrast patterns, the fricative results are much more consistent across subsets of the lexicon. The voiceless alveolar [s] remains robust across contrasts in both Experiment 1a and 1b, while the voiceless postalveolar [ʃ] is consistently misidentified in a wide range of contrasts. Both the dental fricatives [θ, ð] and the voiced labiodental [v] are poorly perceived in most contrasts across sub-experiments, while the contrast accuracies on the voiceless labiodental [f] are less consistent. While [f] is commonly confused with the other labials in Experiment 1b, it is well discriminated from this set in Experiment 1a, suggesting these patterns may be item-specific, or that the strong acoustic similarity between [p, b, f, v] can be easily overriden by other lexical characteristics. Finally, the similarity between [h] and the plosives, particularly the voiceless series, is consistent across sub-experiments, as both Experiment 1a and 1b show high error rates among this set.

Word-medial position (VCV). Intervocalically, the contrast accuracy distributions diverge substantially from those in CV position. Beginning with the voiceless plosives, Figure 3.12 illustrates that the high discriminability of [t, k] from most fricatives in CV position is inverted in VCV, where the majority of such contrasts, at least when [t, k] are presented in the stimulus, are below average in accuracy. Conversely, [p] shows better performance with fricative contrasts in VCV position than in CV, with the voiceless nonsibilants [f, θ] the only set commonly confused with [p] stimuli. Notably, both of the above patterns are direction-specific; intervocalic fricative stimuli are less commonly confused with voiceless plosive competitors. The VCV results are, however,

consistent with CV in showing a high degree of confusability within the [p, t, k] set, and when voiced plosive stimuli are contrasted with voiceless plosive competitors, with accuracy on the reverse relation—[p, t, k] → [b, d, g]—much better. Finally, Table 3.16 shows that there are no clear response biases overall on [p, t, k] in VCV contrasts, though within the voiceless plosives there is a moderate bias toward [p] in [p→k] contrasts (0.54), and toward [t] in [t→p] contrasts (0.55). Further, some of the voiced→voiceless errors may be attributable to response biases, as half of such contrasts show biases of 0.46 or below (i.e., biases away from the target [b, d, g] set and toward [p, t, k]); the labial plosives represent an exception, however, as [b→p] shows a bias of 0.63.

Regarding voiced plosive recognition in fricative contrasts (the presence of contrasts between voiced plosives and affricates in VCV position is sparse and thus cannot be reliably compared to such contrasts in CV position), accuracies are generally lower intervocalically than word-initially. From Figure 3.12, the only notably well-perceived contrast in this set is between [g] and [ʃ] ([g, z] is also highly accurate but only appears in one item), though this result does not exhibit any particularly clear acoustic explanation. Similarly to the voiceless plosive results, when the stimulus-response role is reversed, contrasts between fricatives and voiced plosives, with the exception of [v], are generally well perceived. Finally, all plosives are commonly confused with the alveolar flap, excepting [t, d], which are in complementary distribution with [ɾ].

Examining next the perception of affricates intervocalically, as was shown earlier in the overall accuracy ranks on target phones in Figure 3.3, [ʧ] is quite robust in VCV position, with only the contrast with [f] exhibiting a relatively low accuracy (the affricate voicing contrast, [ʧ, ʤ], only occurs in one item). The voiced affricate, [ʤ], on the other hand, is poorly perceived overall—in the bottom third of obstruents—but from Figure 3.12 we see that this result is primarily driven by contrasts with fricatives and with the alveolar flap, [ɾ], the latter of which, given its frequency, is the most critical for the overall recognition of [ʤ] intervocalically.

Listener accuracy on fricative stimuli in VCV position can generally be divided into two classes based on voicing, with voiceless fricatives above-average in accuracy, and voiced fricatives below-average, with the exception of [z], which is well perceived intervocalically. Among the voiceless

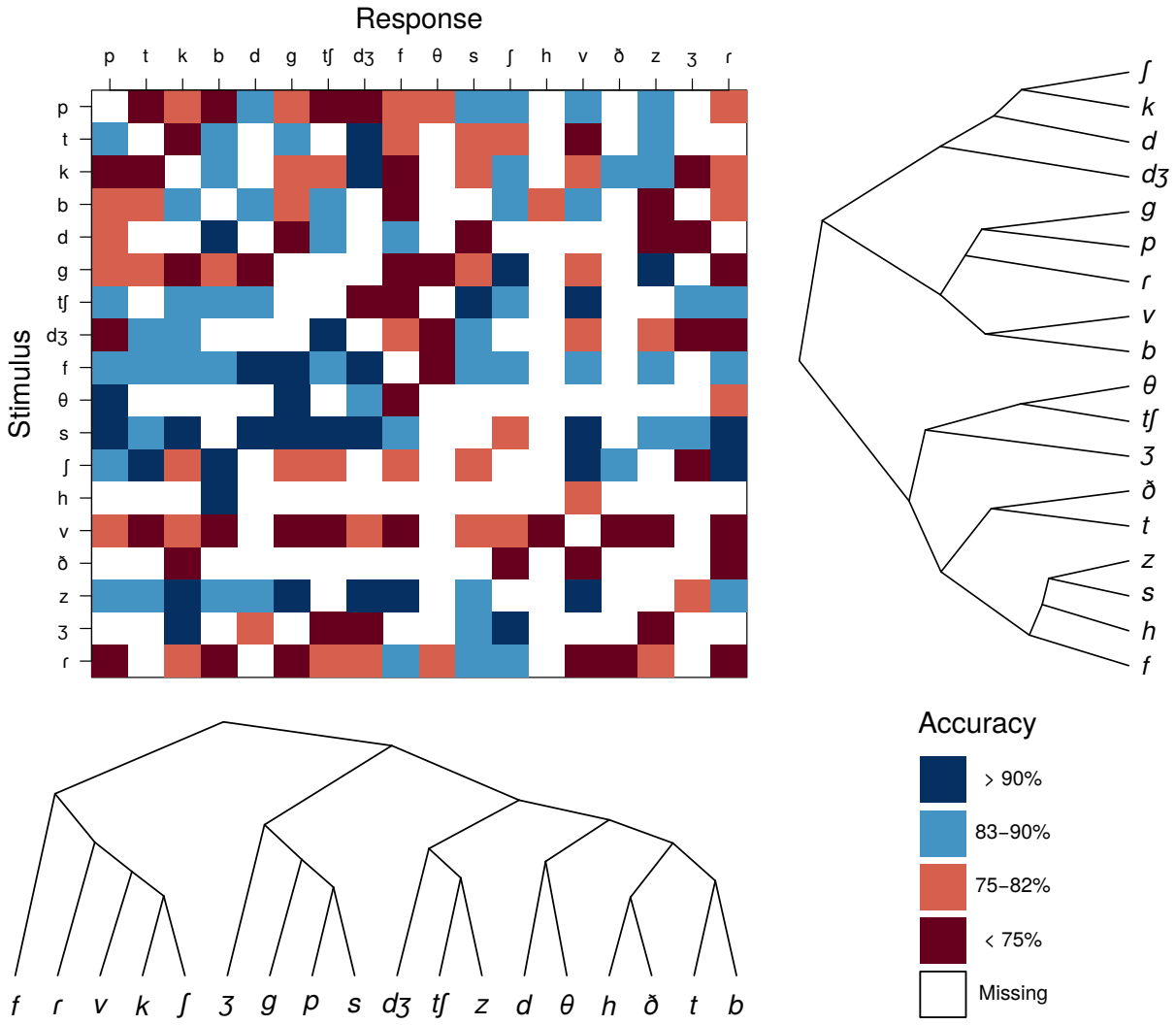

Figure 3.12: Listener accuracies by contrast (VCV). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

	p	**t**	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**h**	**v**	**ð**	**z**	**ʒ**	**ɾ**	**Mean**
p	NA	0.45	0.54	0.37	0.55	0.50	0.42	0.55	0.48	0.40	0.45	0.50	NA	0.55	NA	0.50	NA	0.54	0.49
t	0.55	NA	0.48	0.54	NA	0.55	NA	0.50	0.47	NA	0.48	0.42	NA	0.42	NA	0.48	NA	NA	0.49
k	0.46	0.52	NA	0.50	NA	0.55	0.46	0.54	0.42	NA	0.46	0.52	NA	0.50	0.60	0.47	0.35	0.51	0.49
b	0.63	0.46	0.50	NA	0.48	0.50	0.49	NA	0.44	NA	NA	0.46	0.42	0.62	NA	0.32	NA	0.52	0.49
d	0.45	NA	NA	0.52	NA	0.40	0.51	NA	0.48	NA	0.36	NA	NA	NA	NA	0.42	0.40	NA	0.44
g	0.50	0.45	0.45	0.50	0.60	NA	NA	NA	0.35	0.38	0.42	0.55	NA	0.54	NA	0.48	NA	0.51	0.48
ʧ	0.57	NA	0.55	0.51	0.49	NA	NA	0.30	0.40	NA	0.52	0.52	NA	0.60	NA	NA	0.60	0.54	0.51
ʤ	0.45	0.50	0.46	NA	NA	NA	0.70	NA	0.40	0.32	0.46	NA	NA	0.50	NA	0.45	0.50	0.46	0.47
f	0.52	0.53	0.57	0.56	0.52	0.65	0.60	0.60	NA	0.65	0.48	0.52	NA	0.57	NA	0.45	NA	0.52	0.55
θ	0.60	NA	NA	NA	NA	0.62	NA	0.68	0.35	NA	NA	NA	NA	NA	NA	NA	NA	0.50	0.55
s	0.55	0.52	0.54	NA	0.64	0.58	0.48	0.54	0.52	NA	NA	0.51	NA	0.57	NA	0.50	0.52	0.56	0.54
ʃ	0.50	0.58	0.48	0.54	NA	0.45	0.48	NA	0.48	NA	0.49	NA	NA	0.56	0.57	NA	0.22	0.53	0.49
h	NA	NA	NA	0.57	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.58	NA	NA	NA	NA	0.58
v	0.45	0.57	0.50	0.38	NA	0.46	0.40	0.50	0.43	NA	0.43	0.44	0.42	NA	0.51	0.42	NA	0.53	0.46
ð	NA	NA	0.40	NA	NA	NA	NA	NA	NA	NA	NA	0.42	NA	0.49	NA	NA	NA	0.44	0.44
z	0.50	0.52	0.53	0.68	0.57	0.52	NA	0.55	0.55	NA	0.50	NA	NA	0.58	NA	NA	0.62	0.55	0.56
ʒ	NA	NA	0.65	NA	0.60	NA	0.40	0.50	NA	NA	0.48	0.78	NA	NA	NA	0.38	NA	NA	0.54
ɾ	0.46	NA	0.49	0.48	NA	0.49	0.46	0.55	0.48	0.50	0.44	0.47	NA	0.47	0.56	0.46	NA	0.50	0.49

Table 3.16: Response bias on each contrast in VCV position in Experiment 1. For each row *i*, which indicates the stimulus phone, the proportion in cell [*i*, *j*] represent the proportion of responses given to phone *i* when contrasted with phone *j*. For example, cell [*p*, *t*] = 0.45, which derives from a 75% accuracy on the contrast when *p* is the stimulus, and 15% *p* responses when *t* is the stimulus; i.e., (0.75 + 0.15) / 2 = 0.9 / 2 = 0.45.

fricatives, [ʃ] is the least accurate, particularly in contrast with [g, f, s]. However, listeners are highly accurate at discriminating the frequent [ʃ, ɾ] contrast, and with the exception of the nonsibilant [θ], are above-average on alveolar flap contrasts with the other voiceless fricatives. In fact, aside from the poor perception of [θ, ɾ], confusions on the [f, θ, s] set appears to be restricted primarily to their most acoustically similar counterparts; i.e., [f, θ] and [s, ʃ] are each below-average in accuracy.[19] The distribution of accuracies on contrasts with voiced fricative stimuli is less revealing, as both [v] and [ð] show below-average accuracies across the board, and [ʒ] is also poorly perceived, with the [ʒ, s] contrast the only one that is relatively well perceived and also present in more than one item.

Finally, the alveolar flap, [ɾ], presents a major vulnerability to the perception of intervocalic obstruent contrasts in noise, as it is both one of the least accurately perceived obstruents overall (ranked fourth-to-last in Figure 3.2), and is below-average on all but the more salient voiceless fricatives [f, s, ʃ]. An examination of Table 3.16 reveals that this result is not a consequence of

[19]Note, however, that the [f, θ] contrast only appears in a single item], while there are two [s, ʃ] minimal pairs.

response biases—[ɾ] shows a mean bias of 0.49, and is at its lowest value when contrasted with [s], one of the few well-perceived contrasts with the alveolar flap. Therefore, the poor perception of [ɾ] is likely due to its relatively weak acoustic cues, as its distinctly short duration, though a potentially useful cue (acoustic models in Section 4.3 confirm this expectation), also results in a degradation in the resolution of spectral cues, particularly in the time course over which such cues are available to the listener. And given that we are testing word recognition in noise, shorter durations offer fewer opportunities for spectral cues from the signal to be detected against the background noise; i.e., in such cases there is a greater likelihood that random fluctuations in the multi-talker babble will obscure the critical window in the consonant when spectral cues are available.

Most of the above results are consistent across sub-experiments. In fricative-plosive contrasts, fricatives are more readily identified in the stimulus and distinguished from plosive competitors, while both voiced and voiceless plosives show high error rates in Experiments 1a and 1b as the stimulus member of such contrasts. Within the plosive sets, the asymmetry in greater accuracy on *voiceless* → *voiced* pairs than on contrasts in the opposite direction is also present in both sub-experiments. Finally, with the exception of [p, ɾ] in Experiment 1b, plosives are consistently confused with the alveolar flap, both when presented in the stimulus and when serving as the on-screen competitor.

Regarding affricates, the general robustness of [ʧ] is inconsistent across sub-experiments, showing high accuracies in Experiment 1b, particularly in contrasts with other stops, while in Experiment 1a above-average accuracy rates are primarily restricted to contrasts with fricatives. The poor recognition of its voiced counterpart [ʤ], however, is generally consistent across sub-experiments and is present in the majority of both stop and fricative contrasts.

Among fricative stimuli, the robust distinction between voiceless and voiced sets is reliably present in Experiment 1a and 1b, where voiceless fricatives are accurately perceived across a range of contrasts, and voiced fricatives (excepting [z]) are poorly perceived across the same range. In this regard, recognition of contrasts with the alveolar flap [ɾ] is similarly poor in both sub-experiments, whether [ɾ] is the stimulus or competitor. However, within this set of vulnerable

contrasts, the voiceless fricatives generally exhibit greater stability in recognition, a result which is likely due to the combination of distinct frequency and durational characteristics due to the contrast in voicing and manner of articulation.

Word-final position (VC). Figure 3.13 shows listener accuracies on word-final contrasts in Experiment 1. The most notable patterns in Figure 3.13 are the generally poor recognition of plosives, particularly labials and alveolars, and the relatively accurate perception of fricatives, results which are in greater conformity with those in VCV position than in CV. Beginning with the voiceless plosives, [k] is the most accurate of the three, particularly in contrasts with the lingual stops [d, g, ʧ, ʤ]. The labial [p], on the other hand, is only well-perceived in contrast with its voiced counterpart [b], while [t] is only accurately distinguished from the alveolars [d, s, z]. The voiced plosives are also poorly perceived overall, but are moderately better than their voiceless counterparts. With the exception of [d, ʃ], all three are accurately recognized in voiceless sibilant fricative contrasts, which also happen to appear quite frequently word-finally. Most other voiced plosive contrasts in Figure 3.13 that are above-average in accuracy represent single items, however, though [d, ʤ] is recognized well by listeners and occurs in three minimal pairs. In general, response biases (Table 3.17) appear to favor fricatives over plosives word-finally, with several fricative→plosive contrasts at or above 0.6. These values are much greater than those observed for CV or VCV contrasts, which could be due to the greater amount of decision-constraining information at the end of the word (where the point of contrast is located), the greater distributional and phonotactic asymmetry word-finally, or some combination of the two.

Affricates show a pattern in accuracy distributions that is somewhat intermediate between the CV and VCV results. That is, listeners are more accurate on [ʧ] overall in VC position than in CV, but less than in VCV, while accuracy on [ʤ] is worse in VC than in both CV and VCV positions. The most common confusions with [ʧ] word-finally occur with [k], a result that could be due to similarities between the two in both their release characteristics (the velar plosive being the most likely of the three places to exhibit a release burst word-finally) and in their VC formant transitions

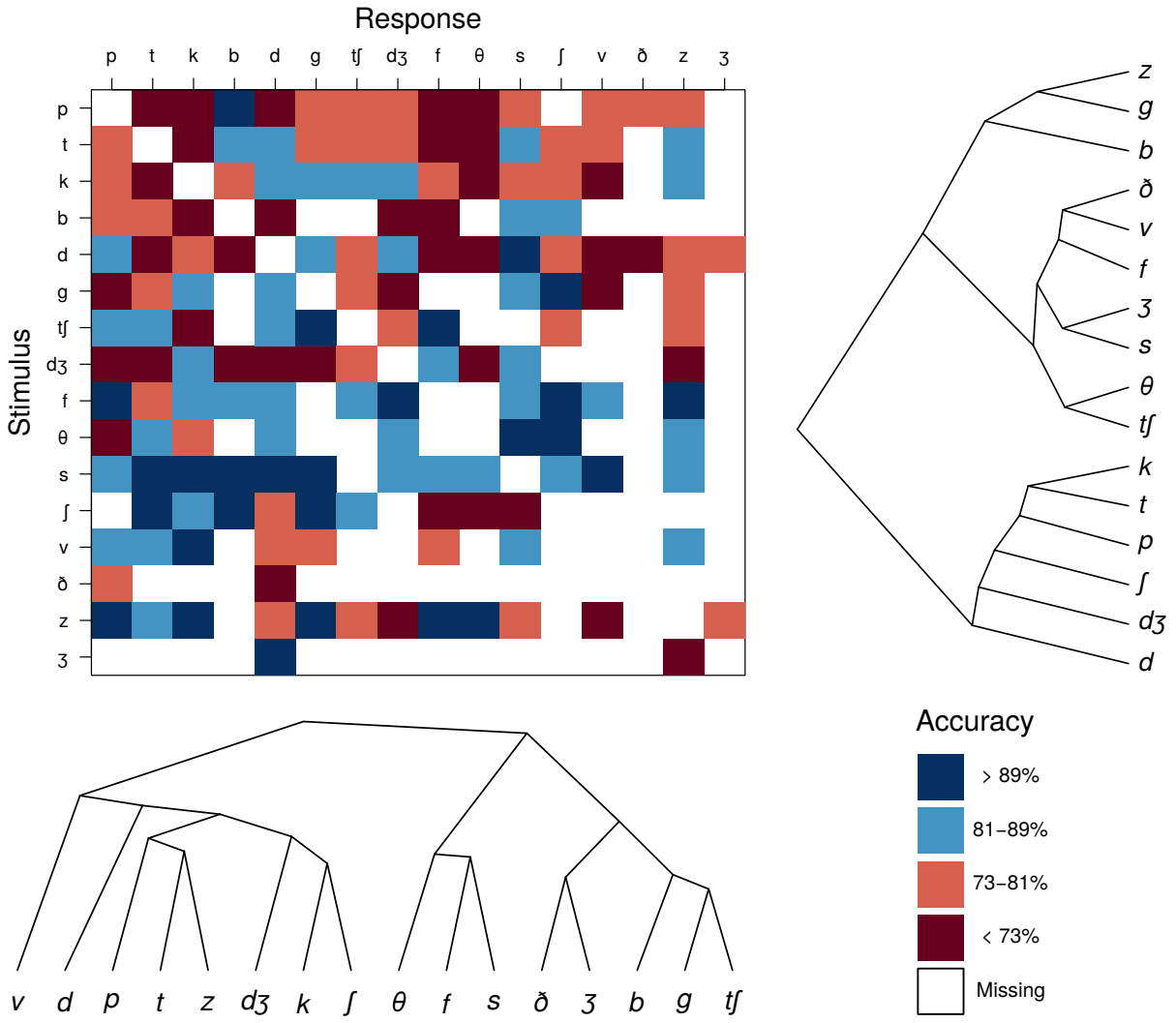

Figure 3.13: Listener accuracies by contrast (VC). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

(both show rapidly a rising second formant into the closure). The voiced affricate, [ʤ], on the other hand, is poorly distinguished from all stops except [k], though the [ʤ, k] contrast only appears in one item, as do the [ʤ, f] and [ʤ, s] contrasts, the two-other contrasts with [ʤ] that were above-average in accuracy.

	p	t	k	b	d	g	ʧ	ʤ	f	θ	s	ʃ	v	ð	z	ʒ	Mean
p	NA	0.48	0.44	0.58	0.40	0.57	0.45	0.55	0.36	0.48	0.47	NA	0.45	0.48	0.45	NA	0.47
t	0.52	NA	0.50	0.57	0.58	0.48	0.46	0.56	0.44	0.38	0.43	0.40	0.47	NA	0.50	NA	0.48
k	0.56	0.50	NA	0.55	0.52	0.50	0.58	0.50	0.45	0.48	0.42	0.48	0.32	NA	0.46	NA	0.49
b	0.42	0.42	0.45	NA	0.46	NA	NA	0.48	0.38	NA	0.44	0.42	NA	NA	NA	NA	0.43
d	0.60	0.42	0.48	0.54	NA	0.52	0.42	0.59	0.43	0.40	0.46	0.48	0.47	0.40	0.50	0.42	0.48
g	0.42	0.52	0.50	NA	0.48	NA	0.42	0.52	NA	NA	0.42	0.48	0.35	NA	0.43	NA	0.46
ʧ	0.55	0.54	0.42	NA	0.57	0.57	NA	0.54	0.52	NA	NA	0.45	NA	NA	0.50	NA	0.52
ʤ	0.45	0.44	0.50	0.52	0.41	0.48	0.46	NA	0.48	0.42	0.52	NA	NA	NA	0.55	NA	0.48
f	0.64	0.56	0.55	0.62	0.57	NA	0.48	0.52	NA	NA	0.48	0.78	0.55	NA	0.48	NA	0.57
θ	0.52	0.62	0.52	NA	0.60	NA	NA	0.57	NA	NA	0.52	0.65	NA	NA	0.48	NA	0.56
s	0.53	0.57	0.58	0.56	0.54	0.57	NA	0.48	0.52	0.48	NA	0.62	0.54	NA	0.54	NA	0.54
ʃ	NA	0.60	0.52	0.57	0.52	0.52	0.55	NA	0.22	0.35	0.38	NA	NA	NA	NA	NA	0.47
v	0.55	0.53	0.68	NA	0.53	0.65	NA	NA	0.45	NA	0.46	NA	NA	NA	0.56	NA	0.55
ð	0.52	NA	NA	NA	0.60	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.56
z	0.55	0.50	0.54	NA	0.50	0.57	0.50	0.45	0.52	0.52	0.46	NA	0.44	NA	NA	0.57	0.51
ʒ	NA	NA	NA	NA	0.57	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.42	NA	0.50

Table 3.17: Response bias on each contrast in VC position in Experiment 1. For each row *i*, which indicates the stimulus phone, the proportion in cell [*i*, *j*] represent the proportion of responses given to phone *i* when contrasted with phone *j*. For example, cell [*p*, *t*] = 0.48, which derives from a 70% accuracy on the contrast when *p* is the stimulus, and 26% *p* responses when *t* is the stimulus; i.e., (0.70 + 0.26) / 2 = 0.96 / 2 = 0.48.

Finally, the voiceless fricatives [f, θ, s, ʃ] are all relatively accurately perceived word-finally. As in CV and VCV positions, [ʃ] is the least accurate of the four, though in VC position errors on [ʃ] are primarily restricted to the other voiceless fricatives [f, θ, s]. Accuracy distributions on contrasts with the voiceless nonsibilant fricatives, [f, θ], also show a predictable acoustic basis, as the greatest number of errors appear in contrasts with voiceless plosives; namely, [f, t], [θ, p], and [θ, k]. Lastly, as in the other two positions, the voiceless alveolar, [s], remains robust across all word-final contrasts.

Regarding the voiced fricatives, results are mixed. Here we focus on the fricatives [v, z], both of which (particularly [z]) appear frequently in word-final contrasts; [ð] and [ʒ], on the other hand, are both quite rare. Errors on [v] are acoustically predictable, and primarily restricted to the voiced plosives [d, g], and its voiceless counterpart [f]. Similarly, the contrast accuracy distribution for [z] is largely predictable from the acoustics. The most accurately perceived contrasts with [z] are the voiceless nonsibilants, and with the exception of [d], the plosives, whereas listeners are below average at recognizing [z] when contrasted with other voiced fricatives or sibilants. Among these results, the most critical is the result for [d, z], which appears in 20% of word-final obstruent

minimal pairs in English, again primarily due to inflectional morphology. Listeners are below average at recognizing this distinction ([d→z] = 76%, [z→d] = 75%), but with no evidence of response biases. This means that the morphological contrast between [d] and [z] either is not acoustically robust, or listeners are not able to pick up on the right cues when stimuli are presented in isolation, perhaps because for morphological contrasts like [d, z], syntactic cues are typically available that reduce the weight on bottom-up information from the acoustic signal.

Finally, considering the robustness of the above patterns among VC contrasts in different item and participant sets, there were several notable commonalities and discrepancies between Experiment 1a and 1b. First, both voiced and voiceless plosives are generally poorly recognized, particularly labials, though the poor overall recognition of [t] reported above is not replicated in contrasts with other plosives in Experiment 1a. Further, the greater accuracy on [k] is primarily restricted to Experiment 1a, as [k] is below-average in contrasts with all obstruents but the voiced alveolar sibilant [z]. Thus, overall, word-final plosive recognition is quite unstable and represents a potential vulnerability for the system, especially considering the frequent occurrence of such contrasts in the English lexicon. Results for affricates are generally consistent across sub-experiments, with [ʤ] poorly perceived in most contrasts in both Experiment 1a and 1b, while [ʧ] is more accurate overall, though it is consistently poorly perceived in contrasts with [k] and [ʃ], two obstruents of notable acoustic similarity with [ʧ] word-finally.

Fricatives, on the other hand, are generally accurate across a range of contrasts, though the voiceless nonsibilants [f, θ] are less consistent in this regard. The frequent errors between [f, θ] and the voiceless plosives, for example, primarily occur in Experiment 1b. The voiceless alveolar sibilant [s], however, is robust across most contrasts in both sub-experiments, while [ʃ] is consistently poorly perceived in contrasts with other fricatives (contrasts between [ʃ] and stop consonants are less consistent). Among the voiced fricatives only [v] and [z] participate widely in word-final contrasts, and while the contrast recognition patterns with [v] are quite consistent in Experiments 1a and 1b, showing good recognition when in contrast with fricatives and with voiceless plosives, but poor recognition in contrasts with voiced plosives, [z] is less consistent in this regard. Listeners

in Experiment 1b were much poorer at detecting [z] across a range of contrasts than in Experiment 1a. However, on some of the most frequent contrasts in this set—namely, [z, d], [z, s], and [z, v]—listeners were below-average in accuracy in both Experiment 1a and 1b. As we have repeatedly emphasized throughout the dissertation, this result for [d, z] is particularly notable given that such contrasts comprise approximately 20% of the word-final obstruent contrasts in English, at least in whole-word models of lexical contrast where distinctions between morphologically complex forms are incorporated into the system.

3.3.4.4 Featural contrast accuracy

Next we examine listener perception of featural contrasts by measuring their accuracy at identifying the *voicing*, *manner*, *place*, or *sibilance* of the target obstruent. This analysis amounts to a subsetting of minimal pairs by featural distinction presented to the listener, and then measuring their accuracy as a function of the feature class presented. For example, in the analysis of voicing contrast accuracy, we measure listener accuracy on voiceless stimuli when in contrast with voiced competitors, and vice versa for voiced stimuli. Similarly, in the analysis of manner contrast perception, we examine separately plosives, affricates, and fricatives when presented in contrast with obstruents of an opposing manner class. As before, featural contrast accuracy is analyzed as a function of noise level, word length, and absolute/relative target frequency, with results for CV, VCV, and VC positions considered separately within each analysis.

Accuracy by Noise Level. As in the analysis of stimulus feature accuracy by noise level, here we are interested in how backround noise affects feature transmission and the extent to which transmission is symmetric or asymmetric between the classes that compose each feature. More precisely, symmetric systems do not depend on which class (e.g., *voiceless* versus *voiced*) is in the stimulus and which is in the competitor, and asymmetric systems show differential transmission rates as a function of the stimulus and competitor class.

Word-initial position (CV). Overall, listeners were similarly accurate on the four featural contrasts, ranging at +2 dB between 87% for manner and place distinctions to 89% for sibilance distinctions, while accuracies at −2 dB ranged from 74% in the case of place contrasts to 78% for sibilance contrasts. However, in a mixed-effects logistic regression predicting Accuracy from Feature and Noise Level, with Listener mean differences modeled as a random intercept, there was a significant overall effect of Feature ($\chi^2(6) = 22.3$, $p = 0.001$) wherein there is a moderate but significant advantage for contrasts differing in sibilance (ps < 0.05), while no significant differences between the remaining three features were observed (ps > 0.1). Further, the only effect of noise level on this relation was the loss of a significant distinction between voicing and sibilance at +2 dB, while all other distinctions were constant across SNRs. Thus, with the exception of obstruent sibilance, the presence of a given featural distinction has little impact on listener accuracy, and therefore the analysis below focuses primarily on the transmission of each feature as a function of which value is present in the stimulus.

Figure 3.14 shows word-initial featural contrast accuracies by noise level, and largely agrees with the overall accuracy patterns by feature class in Figure 3.5. Namely, voiceless obstruents remain more recognizable than voiced obstruents, particularly at a lower SNR, affricates are more poorly recognized than plosives and fricatives, and [LOW] coronals and dorsals consistently transmit place information more reliably than labials, [HIGH] coronals, and glottals. The one notable difference is in the transmission of sibilance, which is symmetric between sibilant and nonsibilant stimuli. That is, in terms of overall accuracy, listeners are more accurate on sibilant stimuli than nonsibilant stimuli, but when the contrast is restricted to cases where the two obstruents differ in sibilance, listeners are as good at recognizing a sibilant onset as they are at recognizing its absence. This result is consistent with the phonetic contrast analysis above in that nonsibilants are primarily confused with other nonsibilants, and thus their poor recognition does not depend on the transmission of sibilance, but rather the acoustic characteristics of sibilants that make them poor at transmitting other featural information like place and manner.

Each the above relations were significant in a mixed-effects logistic regression on featural

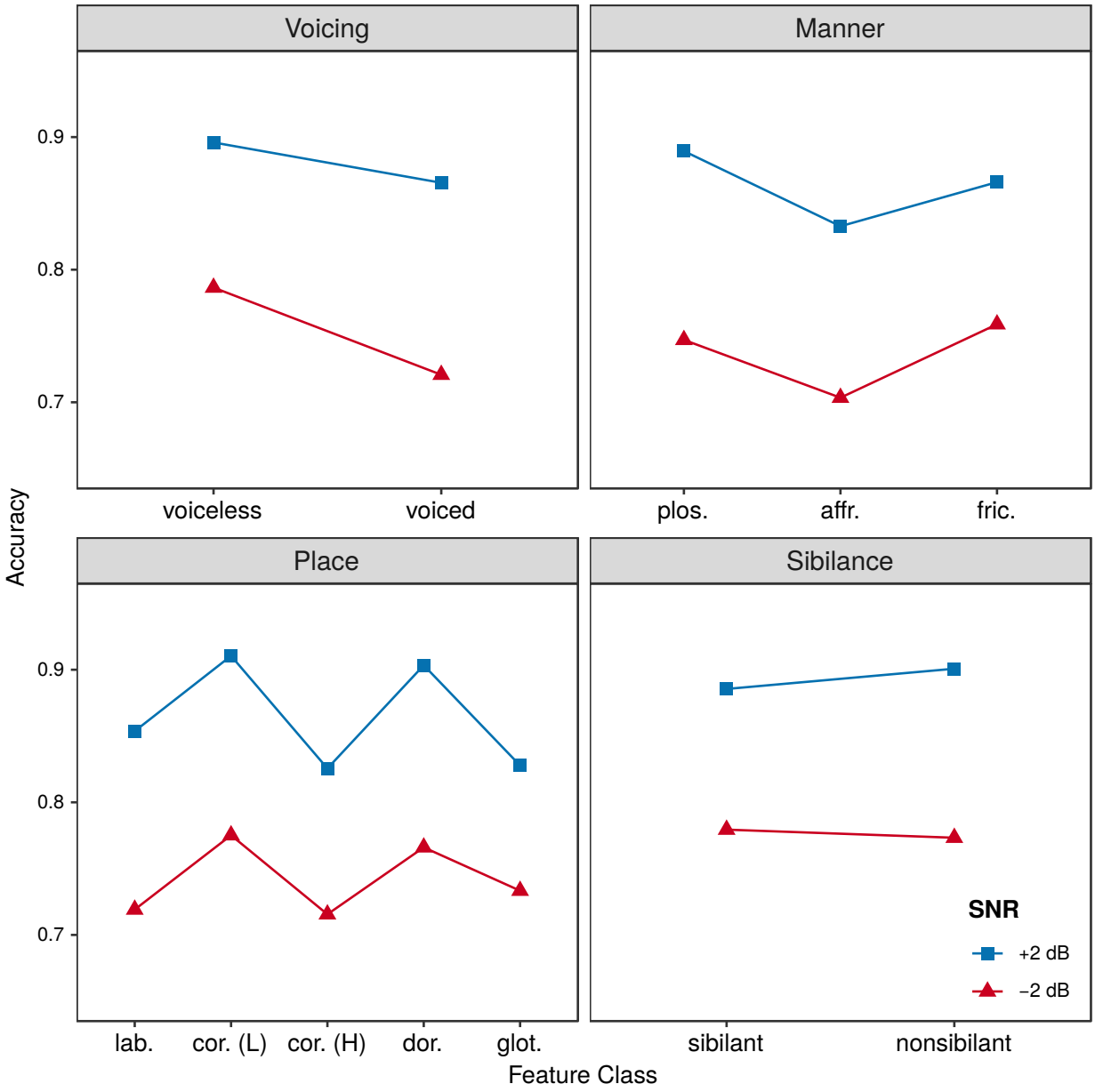

Figure 3.14: Featural contrast accuracies by target feature and SNR in CV position in Experiment 1.

accuracy with the Target Feature Class and Noise Level as fixed effects, and Listener as a random intercept. Regarding voicing perception, *voiced* stimuli are lower in accuracy than *voiceless* stimuli at both +2 dB ($\beta = -0.297$, $z = -2.573$, $p = 0.010$) and −2 dB SNR ($\beta = -0.346$, $z = -3.942$, $p < 0.001$). Regarding manner perception, the *affricate* < *plosive* pattern is consistent across SNRs (+2 dB: $\beta_{p-a} = 0.472$, $z = 3.422$, $p < 0.001$; −2 dB: $\beta_{p-a} = 0.228$, $z = 2.086$, $p = 0.037$), while affricates are worse at transmitting manner information than fricatives at −2 dB ($\beta_{a-f} = -0.281$,

$z = -2.546$, $p = 0.011$), but are not statistically different at +2 dB ($\beta_{a-f} = -0.241$, $z = -1.777$, $p = 0.076$). The place effect is largely driven by a robust separation between the following two sets: {[LOW] coronals, dorsals} > {labials, [HIGH] coronals, glottals}. The only statistical exception to this pattern is the relation between dorsals and glottals at −2 dB, which was not significant ($\beta_{d-g} = 0.193$, $z = 1.510$, $p = 0.131$). Finally, the novel observation that sibilance perception is symmetric—i.e., ([+sibilant] → [−sibilant]) ≈ ([−sibilant] → [+sibilant])—is statistically robust (ps > 0.1). Finally, these general patterns are largely consistent across sub-experiments, though there is some variation in the size of each effect. See Figures A.25 and A.26 in the appendix for further details.

Word-medial position (VCV). Intervocalically, obstruent sibilance is again the most accurately transmitted, at 92% at +2 dB and 79% at −2 dB, followed by voicing (+2 dB: 90%, −2 dB: 77%), manner (+2 dB: 89%, −2 dB: 75%), and place (+2 dB: 88%, −2 dB: 75%). As in CV position, there was a significant overall effect of Feature ($\chi^2(6) = 44.3$, $p < 0.001$) based on the following relation: *manner/place* < *voicing* < *sibilance* (ps < 0.05). Further, this relation did not vary significantly overall as a function of background noise level, though the distinction between voicing and manner was restricted to −2 dB, and the distinction between voicing and sibilance was restricted to +2 dB. All other effects were constant across SNRs.

When broken down by stimulus feature class, listeners' featural accuracy is nearly identical to their overall accuracy. That is, Figure 3.15 exhibits most of the critical patterns in Figure 3.6, such as the robust voicing (*voiced* < *voiceless*) and sibilance (*nonsibilant* < *sibilant*) effects, though the latter is primarily restricted to the lower SNR. Further, the more accurate transmission of manner information from fricative stimuli, and correspondingly low accuracies on manner contrasts with plosive or flap stimuli, follows that in the overall accuracy results in Figure 3.6, where effects are greatest at −2 dB, though at the higher SNR the fundamental *stop/flap* < *fricative* relation remains. Finally, in the transmission of place information, few asymmetries are observed at +2 dB, while at −2 dB there is a notable advantage of coronals over labials/dorsals, consistent with the overall

accuracy results in Figure 3.6.

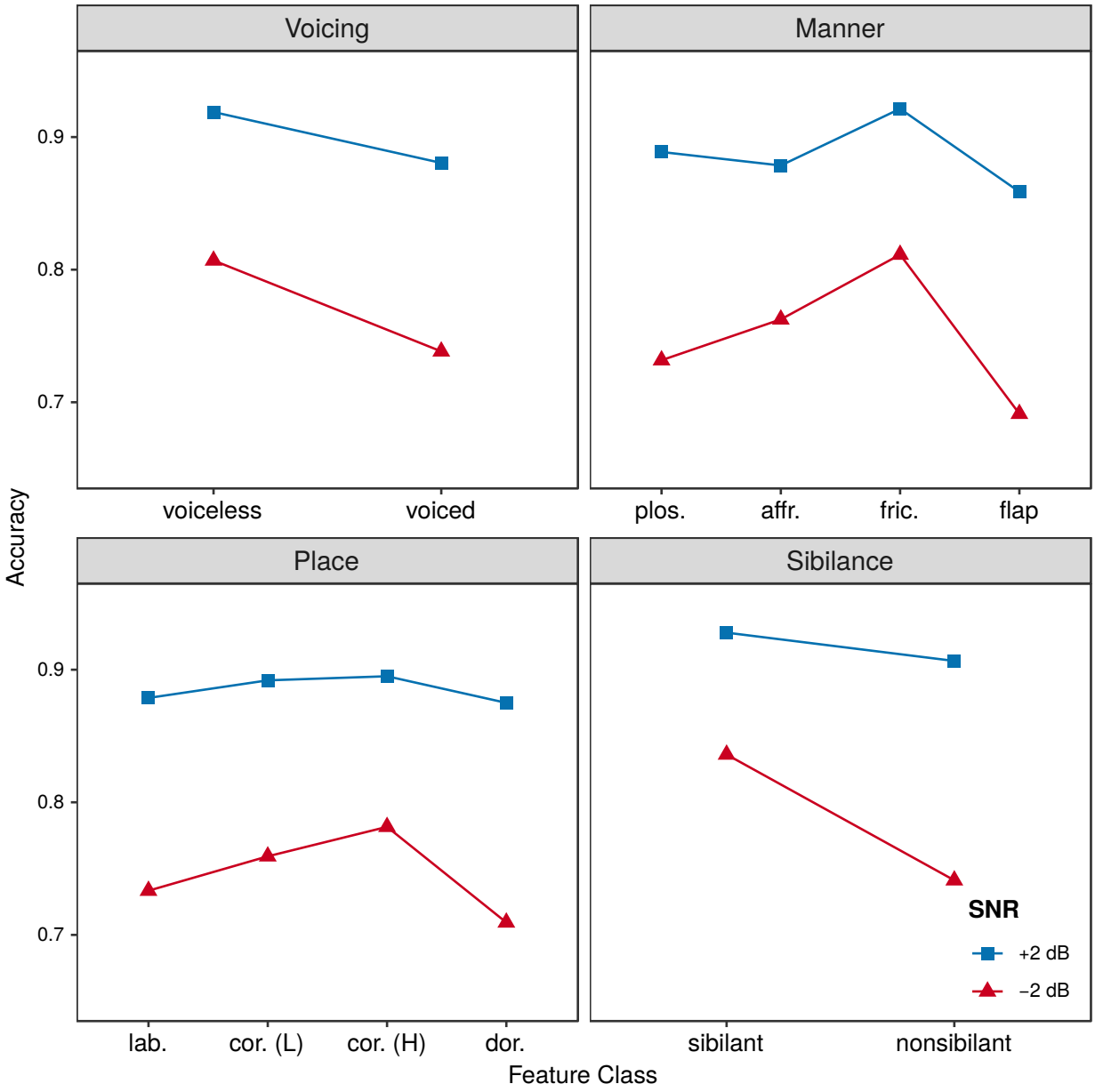

Figure 3.15: Featural contrast accuracies by target feature and SNR in VCV position in Experiment 1. The glottal fricative has been omitted from the place of articulation results due to sparsity of data.

Both voicing and sibilance effects were confirmed statistically at both SNRs (voicing: ps $<$ 0.001; sibilance: $\beta_{+2} = 0.279$, $z = 1.975$, $p = 0.048$; $\beta_{-2} = 0.585$, $z = 5.985$, $p < 0.001$). Further, the poor transmission of manner information from intervocalic stops and flaps relative to fricatives is robust across SNRs (ps < 0.01), while flaps and plosives are relatively even in manner contrast

accuracy, though flaps are slightly less accurate ($\beta_{+2} = 0.269$, $z = 2.105$, $p = 0.035$; $\beta_{-2} = 0.181$, $z = 1.904$, $p = 0.057$). Finally, the place effects noted above are statistically robust in that no differences suggesting a break from symmetry were obtained at $+2$ dB (ps > 0.05), while at -2 dB the disadvantage for labial and dorsal stimuli was generally confirmed (ps < 0.05), though no significant distinction between labials and [LOW] coronals was obtained ($\beta_{lc-l} = 0.162$, $z = 1.898$, $p = 0.058$). These effects were largely consistent across sub-experiments, with the one exception being place of articulation, which showed a coronal advantage at $+2$ dB in Experiment 1a that is neither present in Experiment 1b, nor overall, though this pattern is partly consistent with the place asymmetries at -2 dB. See Figures A.27 and A.28 in the appendix for full featural contrast results by sub-experiment.

Word-final position (VC). The relative patterning of overall featural accuracy in word-final contrasts differs somewhat from that in CV and VCV positions. At both $+2$ dB and -2 dB, voicing information is the most effectively transmitted at 88% and 79%, respectively, followed closely by sibilance at 87% and 76%, and finally manner/place at 85% at $+2$ dB and 75/74% at -2 dB. As with word-initial and word-medial contrasts, this *manner/place* $<$ *sibilance* $<$ *voicing* relation was significant overall ($\chi^2(6) = 36.3$, $p < 0.001$), with the only impact of background noise level occurring in the loss of a significant difference between manner and sibilance at -2 dB, and between voicing and sibilance at $+2$ dB.

Figure 3.16 shows featural contrast accuracies in word-final position, and largely conforms with the overall accuracy results in Figure 3.7. Among the patterns that replicate across sub-experiments (see Figures A.29 and A.30 in the appendix for details) are the greater accuracy of voicing transmission from voiceless stimuli (ps < 0.001), and the greater accuracy at perceiving manner of articulation from fricative stimuli than from plosives (ps < 0.001). Place and sibilance, on the other hand, are relatively symmetric across stimulus classes. The former result was obtained previously in the overall accuracy results, but the latter represents a notable departure that is consistent with the patterns observed in CV and VCV position. That is, listeners are just as good

at detecting the presence of a sibilant in the stimulus, which is predicted given the many salient acoustic properties of sibilants, as they are at detecting nonsibilants when contrasted with sibilant-offset competitors ($ps > 0.1$). The discrepancy between this result and the consistently higher overall accuracy on sibilant stimuli has been interpreted as evidence for the perceptual salience of the sibilance feature, as listeners are just as accurate at detecting the absence of the [+sibilant] feature value in the stimulus as they are at detecting its presence.

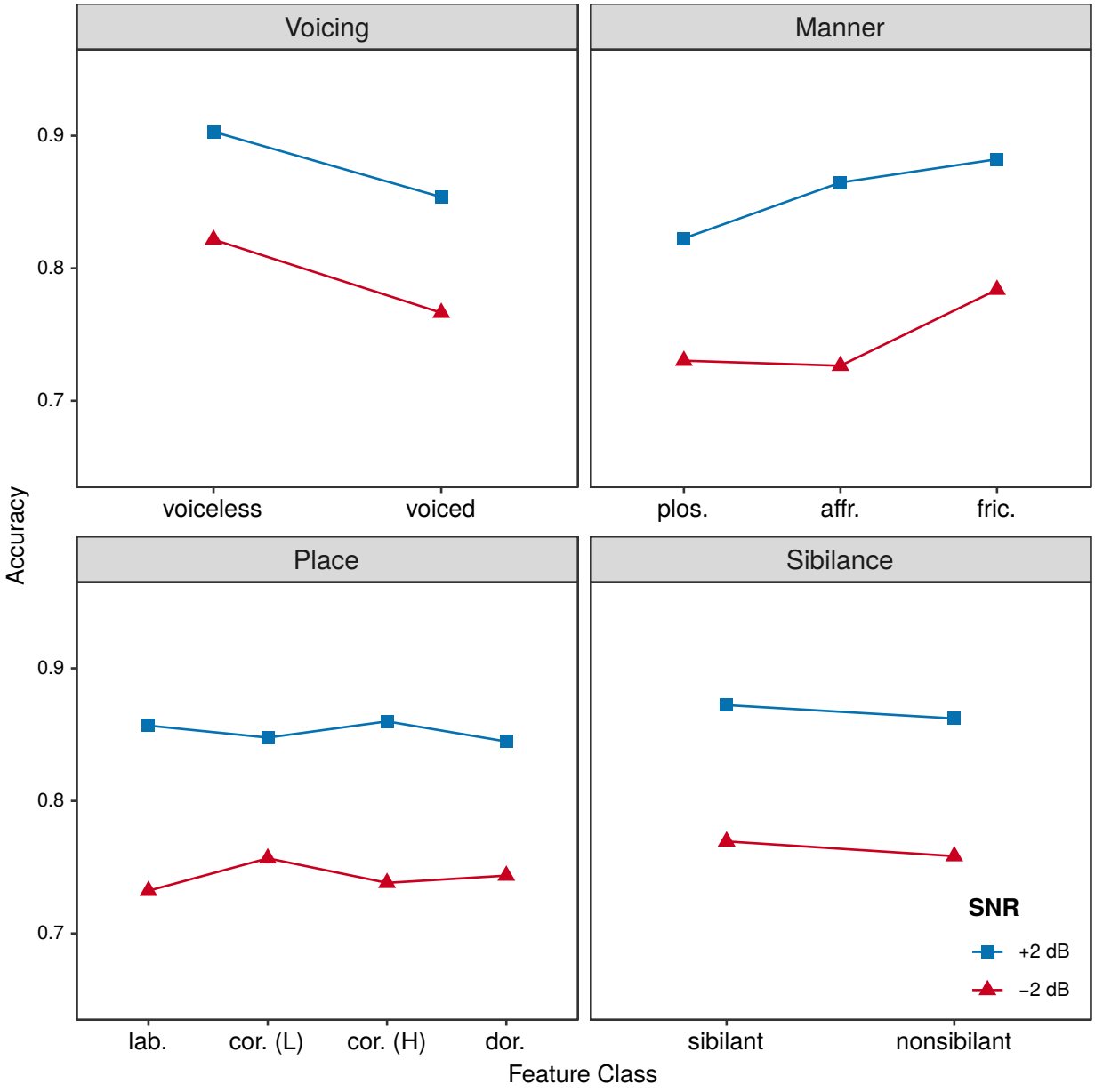

Figure 3.16: Featural contrast accuracies by target feature and SNR in VC position in Experiment 1.

Accuracy by Word Length and Frequency. Next we examine what effects, if any, additional factors such as word length and word frequency have on feature transmission in minimal-pair contrasts, where again the emphasis is on assessing symmetries and asymmetries in feature transmission as a function of lexical characteristics that are beyond the immediate scope of the contrast.

Word-initial position (CV). We begin with word-initial contrasts, wherein the sibilance advantage is largely robust to variation in word length and frequency. Regarding word length, the *voicing/manner/place* $<$ *sibilance* relation reported above is significant overall ($\chi^2(6) = 25.8$, $p < 0.001$), though the distinctions between sibilance and voicing, and sibilance and place, are not significant in polysyllabic items due to their reduced item number and therefore increased variance. Similarly, neither absolute nor relative target frequency interacts significantly with Feature (ps > 0.1), though there is a significant increase in the *sibilance* $>$ *voicing* advantage as a function of absolute frequency ($\beta_x = 0.105$, $z = 2.008$, $p = 0.045$).

Figure 3.17 shows word-initial featural contrast accuracies as a function of stimulus feature, word length, and word frequency. As in the overall accuracy results, voicing perception is consistently more robust with voiceless stimuli than with voiced stimuli, both across word lengths and absolute/relative frequency levels (ps > 0.1), and across sub-experiments (see Figures A.31 and A.32 in the appendix for details). Further, the symmetry in sibilance perception, shown in the equality between sibilants and nonsibilants in Figure 3.14, is preserved across word lengths (ps > 0.1), though both absolute and relative target frequency interact significantly with sibilance perception in yielding at reduced accuracy on sibilant targets relative to nonsibilants as the frequency of the target decreases. More precisely, in predicting listener accuracy on minimal pairs contrasting in sibilance, the sibilance of the target (sibilant [ref], nonsibilant) interacts negatively with absolute frequency ($\beta = -0.220$, $z = -2.820$, $p = 0.005$) and relative frequency ($\beta = -0.289$, $z = -3.691$, $p < 0.001$), though there is no significant three-way interaction between target sibilance, absolute frequency, and relative frequency.

This result is indeed counterintuitive given the acoustic salience of sibilants, which should

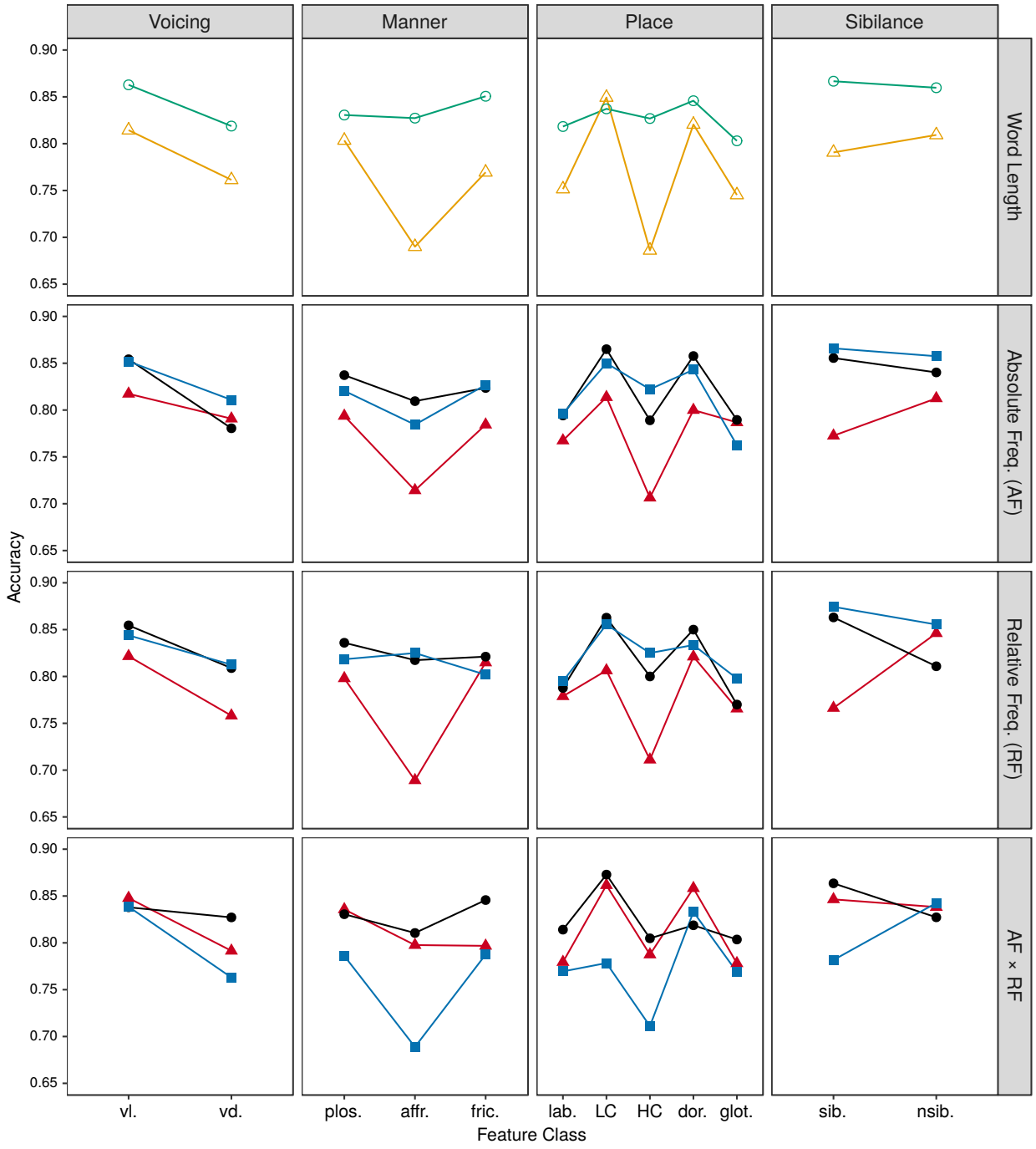

Figure 3.17: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, RF×AF) in CV position in Experiment 1. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

make them less susceptible to top-down biases such as those from the word frequency (see the discussion of target feature class results in Figure 3.8, for example). Upon closer examination of word-initial sibilance contrasts in Experiment 1, this result does appear to be an artifact of the overall frequency distributions in sibilant-onset words versus nonsibilant-onset words, as sibilant targets occupy a much greater portion of the low-frequency range than do nonsibilants ($\mu_{sib} = 2.123$, $CI_{sib} = [2.04, 2.20]$; $\mu_{nsib} = 2.350$, $CI_{nsib} = [2.27, 2.43]$). This difference in frequency distributions is not observed overall, which is why the *sibilant* > *nonsibilant* relation emerged in the overall accuracy analysis but not in the analysis restricted to sibilance contrasts. Therefore, the previous discussion of symmetric sibilance perception may simply be due to frequency biases disproportionately aiding recognition of the acoustically weaker nonsibilant-onset words. However, this result remains interesting and important for the study of the distribution of acoustic information in the lexicon, as it is an example of the leveling out of information from bottom-up and top-down sources that provides for a more robust system of contrast in the lexicon.

Finally, regarding place and manner of articulation, the featural accuracy patterns in Figure 3.17 largely conform with the overall accuracies by target feature class in Figure 3.8. Namely, there are much greater distinctions between target obstruents in polysyllabic items than in monosyllables (manner: affricate < fricative < plosive, ps < 0.05; place: [HIGH] coronal < labial/glottal < dorsal / [LOW] coronal, ps < 0.05), and while manner effects are primarily restricted to low-frequency items, the {labial, [HIGH] coronal, glottal} < {[LOW] coronal, dorsal} relation is robust across the absolute and relative target frequency range. Recalling the sibilance results above, the substantially lower accuracy of affricates relative to plosives and fricatives does coincide with lower mean frequencies of affricate-onset words (1.667, as compared with 2.367 for plosives and 2.377 for fricatives); however, baseline frequency differences do not account for the place effects in Figure 3.17. Both place and manner effects are largely replicated across sub-experiments, though Experiment 1b does show somewhat greater variability in manner and place effects as a function of word frequency.

Word-medial position (VCV). Among intervocalic contrasts, feature transmission significantly depends on word length ($\chi^2(3) = 19.4$, $p < 0.001$), as in disyllables we observe the following relation—*manner/place* < *voicing/sibilance* (ps < 0.001)—whereas in trisyllables voicing perception is the least accurate (ps < 0.05), followed by place, manner, and sibilance, though no distinctions among the latter three are significant. Regarding modulation of feature transmission by word frequency, neither absolute nor relative target frequency interact significantly with the feature transmitted (ps > 0.1). Taken together, we may assume that the sibilance advantage is generally robust to changes in item length/frequency, that manner and place are typically less accurately transmitted, though not always significantly so, and that the transmission of voicing intervocalically is relatively more variable, particularly as a function of the length of the word.

Figure 3.18 shows word-medial featural contrast accuracies by word length and frequency. As in CV position, the results when isolated to the featural contrast in question largely conform with the overall accuracy results in Figure 3.9. That is, voicing perception is more robust for voiceless-onset words across both word lengths and frequencies (see Figures A.33 and A.34 in the appendix for similar patterns in Experiments 1a and 1b, respectively). Despite these consistent trends, however, there are effects of word length and frequency on voicing perception; namely, though target voicing does not interact significantly with word length ($\beta_x = 0.222$, $z = 1.285$, $p = 0.199$), the *voiceless* > *voiced* relation is only significant among disyllables ($\beta = 0.458$, $z = 5.851$, $p < 0.001$) due in part to the high variance in the much smaller trisyllabic item set ($\beta = 0.236$, $z = 1.537$, $p = 0.124$). The significant interactions between target voicing and word frequency, both absolute ($\beta_x = -0.228$, $z = -3.266$, $p = 0.001$) and relative ($\beta_x = -0.219$, $z = -3.038$, $p = 0.002$), reflect the fact that voicing perception becomes more symmetric (i.e., a smaller discrepancy between voiceless and voiced targets) at lower absolute/relative target frequencies. Further, there are no notable baseline differences in the frequencies of voiceless-onset words versus voiced-onset words, so this effect is not an artifact, but rather reflects the fact that the greater acoustic salience of voiceless obstruents can nevertheless be counteracted by top-down frequency biases, though the result appears to be a pull toward the mean recognition rate (affecting voiceless more than voiced

given their higher starting accuracy) rather than a constant reduction across all items.

The perception of sibilance is similar to that for voicing, and is consistent with the overall accuracy results in Figure 3.9. Namely, sibilance contrasts are better perceived when the stimulus phone is a sibilant than when it is a nonsibilant, an effect which generally holds across items, though it declines in trisyllables ($\beta_x = 0.414$, $z = 2.093$, $p = 0.036$) and at lower absolute/relative target frequencies (AF: $\beta_x = -0.144$, $z = -1.769$, $p = 0.077$; RF: $\beta_x = -0.246$, $z = -3.167$, $p = 0.002$). Both sub-experiments replicate this pattern (see Figures A.33 and A.34 in the appendix for details). Finally, while there remains a difference in baseline frequencies of sibilant and nonsibilant targets, it is reduced relative to CV position, which explains the intermediate sibilance effect between that in CV position and that in the overall results in Figure 3.9.

Regarding place and manner of articulation, there are no notable deviations between Figure 3.18 and Figure 3.9 in terms of the effect of word length and frequency on featural contrast perception. Namely, fricatives maintain their stimulus advantage over plosives across word lengths, and over flaps in disyllables, and are generally better perceived across most of the frequency range, though at high absolute and relative frequencies affricates and fricatives are similarly accurate. Flaps, on the other hand, show little improvement in the transmission of place information with increases in word frequency, while plosives show some effect of frequency but maintain their disadvantage relative to fricatives and affricates. Further, each of these patterns are broadly replicated across sub-experiments, though the results for plosives and affricates are more variable than those for fricatives and flaps.

As in the overall accuracy results, Place and Word Length do not interact significantly ($\chi^2(3) = 7.24$, $p = 0.065$), though there is an apparent reversal in accuracy on dorsal and [LOW] coronal obstruents relative to [HIGH] coronals between di- and tri-syllables. Nevertheless, these distinctions are small, as all four stimulus classes show correct place transmission rates between 80 and 85%. The significant interaction between word frequency and place of articulation (AF: $\chi^2(3) = 30.2$, $p < 0.001$; RF: $\chi^2(3) = 9.0$, $p = 0.029$) primarily reflects a relatively greater impact of absolute/relative target frequency on the perception of plosives and [HIGH] coronals than on [LOW]

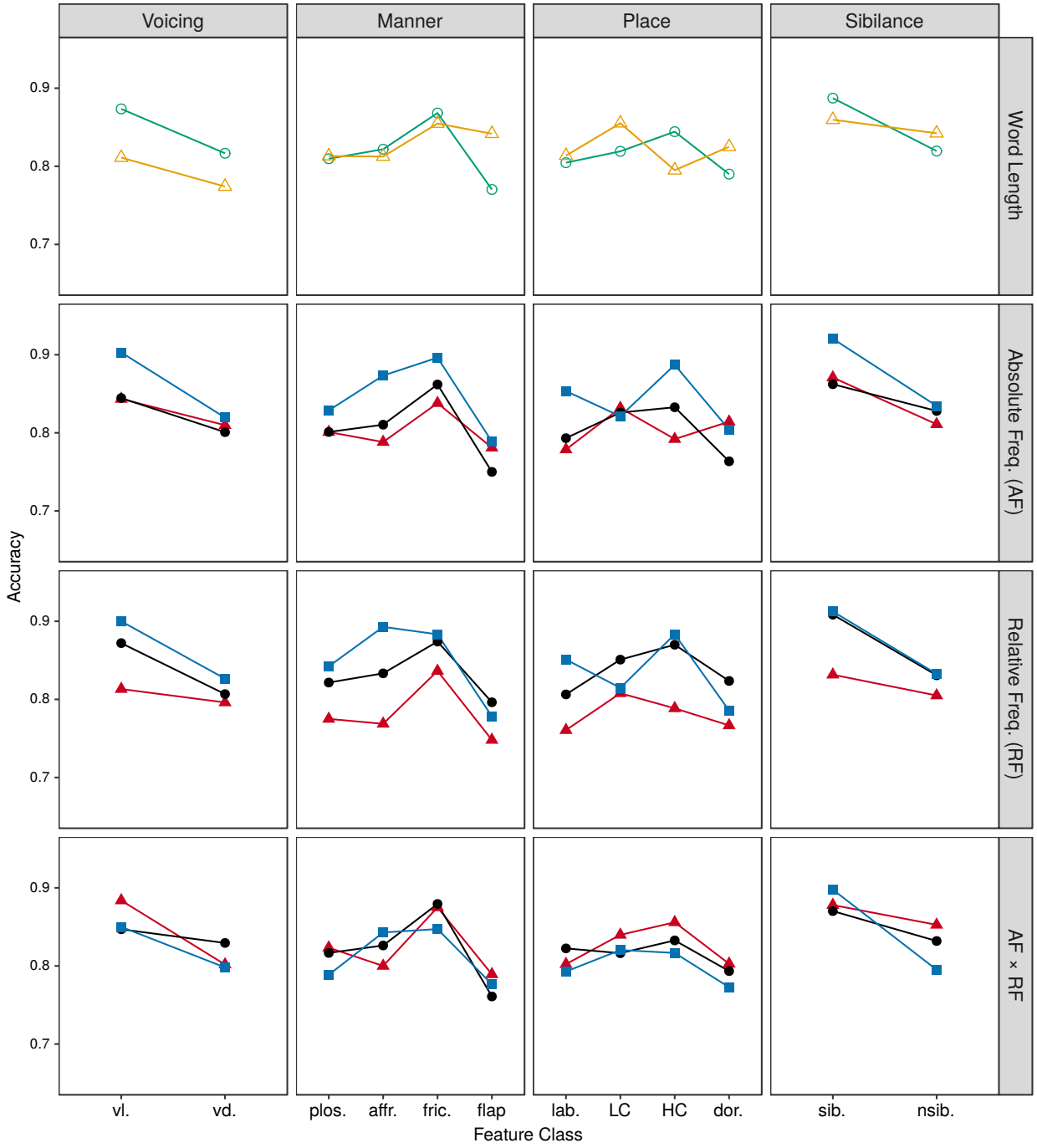

Figure 3.18: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, RF×AF) in VCV position in Experiment 1. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively. Glottals have been omitted from the place analysis due to their sparsity in the data.

coronals and dorsals. Both of these results can be explained by the baseline frequency ranges of each stimulus class in place contrasts, as [LOW] coronals tend to occur in relatively higher-frequency words, while dorsals occupy the lower frequency range to a greater degree than the other three classes, with labials and [HIGH] coronals intermediate between the two. Frequency differences do not entirely account for the relative differences between stimulus classes, however, as models of place perception from absolute/relative frequency alone are significantly poorer than those that include the place of articulation of the stimulus (AF: $\chi^2(6) = 43.6,\ p < 0.001$; RF: $\chi^2(6) = 23.8,\ p < 0.001$). See Figures A.33 and A.34 in the appendix for comparable patterns in Experiments 1a and 1b.

Word-final position (VC). Neither word length nor frequency significantly modulate feature transmission word-finally (ps > 0.1), though the *place* $<$ *manner* $<$ *voicing/sibilance* relation generally reduces in polysyllabic contrasts due to their greater sparsity in the lexicon. Figure 3.19 shows word-final featural contrast accuracies by word length and absolute/relative target frequency. As in CV and VCV positions, the results when restricted to minimal pairs contrasting in a certain feature show little difference from the overall accuracy results by stimulus feature class. The transmission of voicing, for instance, is asymmetric in showing an advantage for voiceless-offset stimuli over their voiced counterparts that is robust across word lengths and frequencies; i.e., interactions with Word Length, Absolute Frequency, and Relative Frequency are not significant (ps > 0.1), though there is a significant three-way interaction between target voicing, AF, and RF ($\beta_x = 0.167$, $z = 2.762,\ p = 0.006$) indicating greater symmetry between voiceless and voiced targets at the highest and lowest absolute/relative word frequencies. These effects benefit in part from greater baseline frequencies of words ending in voiceless obstruents (AF = 2.678 as compared with 2.487 for voiced), but are also consistent with the generally greater acoustic salience of word-final voiceless obstruents. Nevetheless, it is important to note that the overall accuracy results in Figure 3.10 benefit from a similar baseline difference in word frequencies. Thus, here we have a case unlike the sibilance results in CV position where both bottom-up and top-down information benefit the

same class of a given feature. See Figures A.35 and A.36 in the appendix for similar results in Experiments 1a and 1b.

The perception of sibilance is also consistent with the overall accuracy results in Figure 3.10 and across sub-experiments in showing a relatively minor overall advantage for sibilant-offset stimuli that is sensitive to changes in both word length ($\beta_x = 0.407$, $z = 3.282$, $p = 0.001$), though no significant interactions were obtained with absolute or relative frequency, or the interaction between the two (ps > 0.05). Regarding word length, difference between sibilant and nonsibilant targets is significant in monosyllables ($\beta = 0.284$, $z = 3.276$, $p = 0.001$) but not in polysyllables ($\beta = -0.123$, $z = -1.383$, $p = 0.167$). Finally, as in CV position sibilant and nonsibilant targets differ in their baseline frequencies, with nonsibilants significantly greater in absolute frequency ($\mu_{sib} = 2.272$, $CI_{sib} = [2.2, 2.3]$; $\mu_{nsib} = 2.556$, $CI_{nsib} = [2.5, 2.6]$). Therefore, despite the fact that nonsibilants benefit on average from a top-down bias, the greater acoustic salience of sibilants appears to outweigh this lexical advantage.

As in CV and VCV positions, the manner and place effects word-finally are more complex. Beginning with manner of articulation, the distinction between plosives and fricatives significantly varies as a function of word length ($\beta_x = -0.614$, $z = -5.033$, $p < 0.001$), with plosives $<$ fricatives in monosyllables ($\beta = -0.693$, $z = -8.247$, $p < 0.001$) but not in polysyllables ($\beta = -0.080$, $z = -0.899$, $p = 0.368$). Affricates are somewhat intermediate between the two but have been excluded due to their general sparsity in word-final contrasts. Word frequency has a relatively minor effect on place transmission that is approximately equal between plosive and fricative stimuli (ps > 0.1), meaning that the fricative advantage is quite robust to shifts in top-down information.

Contrary to manner of articulation, the transmission of place information is relatively constant across stimulus classes. There is a significant interaction between stimulus place and word length ($\chi^2(3) = 50.9$, $p < 0.001$), but the only significant distinction is between labials and the remaining places, where labials exhibit the lowest accuracy in monosyllables (ps < 0.05) and the highest accuracy in polysyllables (ps < 0.001). Regarding word frequency, while Figure 3.19 shows some variability in place transmission by stimulus class across the absolute and relative frequency ranges,

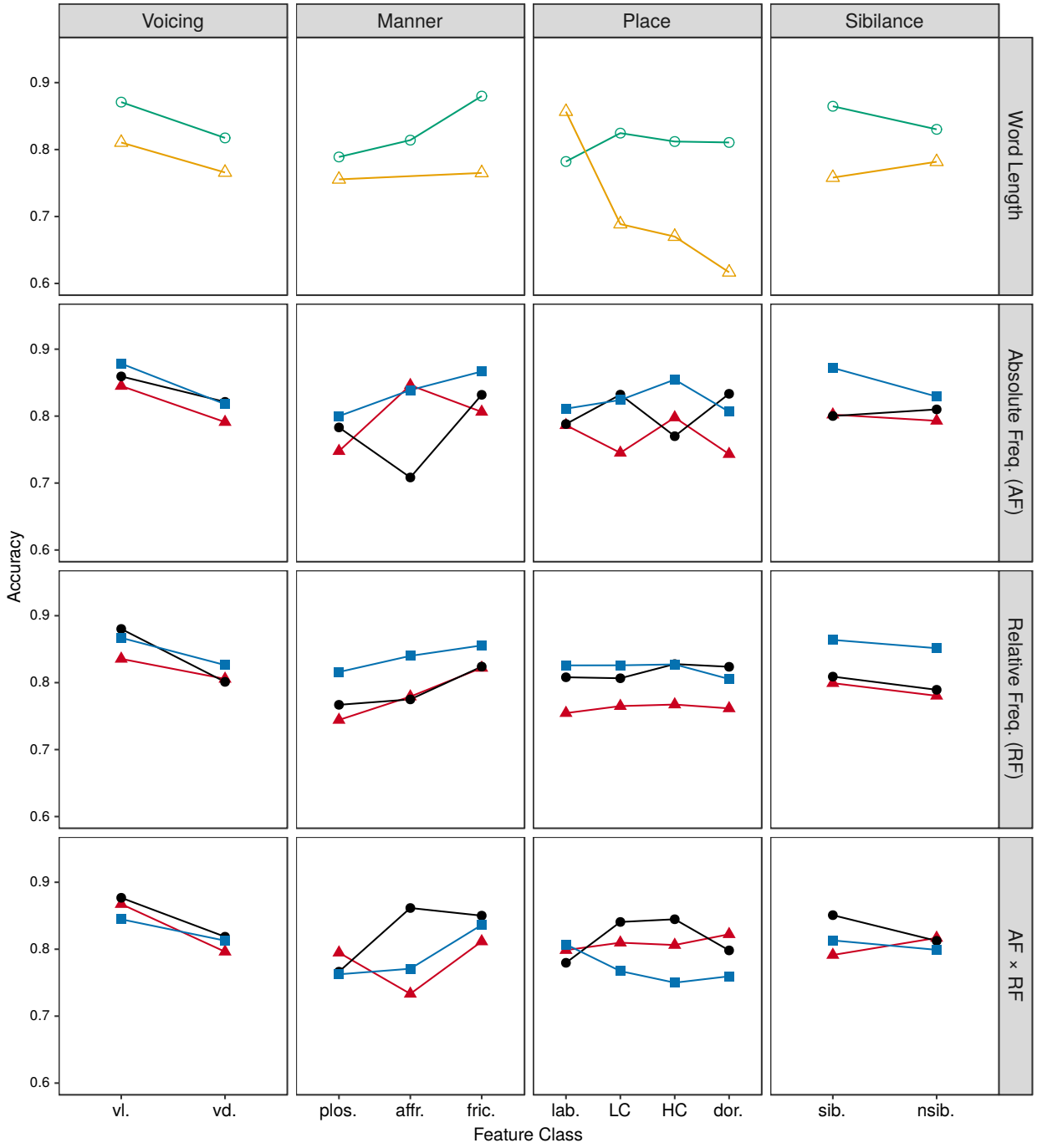

Figure 3.19: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, RF×AF) in VC position in Experiment 1. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively. Affricates in polysyllabic items have been omitted from the manner analysis due to their sparsity in the data.

no significant effects emerged with AF, RF, or AF×RF ($ps > 0.1$). In general, the assessment of smaller feature classes such as those comprising obstruent place and manner is less stable word-finally than in CV or VCV positions because of the high degree of imbalance among word-final obstruent contrasts (see Table 3.11 for details).

3.3.4.5 Summary of contrast recognition results

Several key findings may be taken away from the analysis of contrast recognition presented above. First, just as the analysis of category distributions found obstruents to be highly uneven in their frequency of occurrence in minimal pairs in the lexicon, the distribution of items by contrast is even more asymmetric across the range of possible contrasts among obstruents. Moreover, the degree of imbalance in contrast distributions increases as the position of the contrast progresses further into the word (i.e., from CV to VCV to VC). Across positions the most frequent contrasts are those involving the voiceless plosives and the voiceless alveolar sibilant [s], though most contrast distributions show significant positional dependencies. For instance, given the prevalence of the alveolar flap intervocalically, most of the highly frequent contrasts in VCV position involve [ɾ], while in VC position the majority of contrasts involve alveolar obstruents; namely, *d–z*, *t–d*, *t–z*, *s–z*, *p–t*, and *t–k*. Word-initially and word-medially, contrasts involving place and manner distinctions are the most prevalent, while word-finally, manner and sibilance play a relatively greater role, with place and voicing less informative in VC contrasts. Ultimately, the most important conclusion from the analysis of contrast distributions is that multi-feature contrasts—comprising over 75% of contrasts overall, 74% in word-initial and word-final positions and 79% intervocalically—far outweigh the single-feature contrasts that form the basis for much of the analysis of speech acoustics and perception in the phonetic literature. And thus, as we will address in the next section, the assessment of listeners' accuracy on different obstruent contrasts in the lexicon is only partially informative as to the ultimate role of each contrast in preventing or allowing misperceptions in speech communication. The distribution of such contrasts is also critical, as it governs the ultimate likelihood that randomly chosen minimal pair is correctly distinguished.

Turning next to contrast accuracy, as was anticipated in the introduction to this section, listeners' recognition of different obstruent phones depends in part on the phones they are in competition with, though there are also many obstruents that are fairly consistent across contrasts in being either well or poorly perceived by listeners. The voiceless sibilant fricative [s], for instance, is robust across most contrasts in all three positions, while the labial plosives [p, b] and voiced nonsibilant fricatives [v, ð] are generally poorly perceived across CV, VCV, and VC positions. The high internal confusability of voiceless plosives is also quite consistent across positions, whereas voiced plosives are more generally vulnerable to misperception in a wide range of contrasts. Errors on affricate stimuli largely fall within homorganic voicing classes, particularly word-initially where errors on [ʧ] are primarily restricted to voiceless fricatives, and errors on [ʤ] occur most often in contrasts with voiced fricatives. Word-medially and word-finally, [ʤ] is poorly perceived overall, while [ʧ] is generally well perceived intervocalically, and word-finally is primarily confused with two acoustically similar voiceless obstruents: [k] and [ʃ].

Finally, fricatives tend to be the most accurately perceived obstruents among the three manner classes, particularly voiceless fricatives. Among the voiceless fricatives, [f] and [s] are the most accurate across positions, while [θ] is poorly recognized word-initially, though accuracy on [θ] is much higher in VCV and VC positions. The glottal fricative [h] is also poorly recognized word-initially, but errors on [h] are more constrained, occurring primarily with the similarly aspirated voiceless plosive series. The voiceless postalveolar sibilant [ʃ], on the other hand, is generally poorly perceived, particularly in contrasts with other voiceless obstruents. Among the voiced fricatives, [z] is generally well-perceived, with the most frequent errors occurring word-finally in contrasts with other alveolar obstruents. All other voiced fricatives exhibit high error rates in most contrasts, though the perception of [v] word-finally is more constrained and primarily is confused with voiced plosives. Thus, overall we find both featurally constrained error patterns, which are both acoustically and phonologically motivated, and general error rates indicative of obstruents that are broadly less salient, and therefore more sensitive to noise masking.

When broken down by featural contrast, obstruent sibilance, and to a lesser extent voicing, is

generally more accurately transmitted across positions than is manner or place information. Further, this relation is largely consistent across changes in background noise level, word length, and word frequency. Sibilance perception is asymmetric, however, in that word-medially and word-finally listeners are more accurate when the stimulus phone is a sibilant than when it is a nonsibilant, though this distinction is lost word-initially (and reduced word-finally) due to nonsibilants benefitting from a top-down bias due to baseline differences in word frequency. Obstruent voicing is also asymmetric in exhibiting more [+voice] $\rightarrow$ [−voice] errors than [−voice] $\rightarrow$ [+voice], a pattern which is also generally robust across SNRs, word lengths, and word frequencies.

Manner and place perception, however, is more variable. Regarding manner of articulation, fricative stimuli are generally more accurately perceived than plosives, affricates, or flaps, a pattern which is consistent with the overall accuracy results in the previous section. Plosives, on the other hand, are generally poor at transmitting manner information in word-medial and word-final contrasts, though their accuracy word-initially is comparable to fricatives. The perception of plosives is also more sensitive to changes in noise level and word frequency, a result which is also obtained for affricates but derives more from their relatively sparse distribution in the lexicon. Lastly, flaps are consistently poorly perceived across a range of modulating stimulus characteristics.

The perception of place of articulation is the most complex, and given the large number of constituent classes it is also more sensitive to artifacts from non-phonetic characteristics of the items comprising each contrast. In general, labials are the least effective at transmitting place information, while [LOW] coronals are typically the most accurate stimuli, with glottals, dorsals, and [HIGH] coronals intermediate between the two and more dependent on contrast position. That is, dorsal obstruents are similar to [LOW] coronals word-initially, but are relatively less accurate in VCV position, while the opposite pattern is obtained for [HIGH] coronals. Among VC contrasts, however, the perception of place is more symmetric across stimulus classes. Finally, these relations, though generally consistent across SNRs, word lengths, and word frequencies, tend to be enhanced at higher noise levels and lower frequencies.

While the above patterns have largely been described in phonetic terms, it is important to

emphasize that by examining lexical contrasts we are admitting an array of non-phonetic factors such as word frequency, morphological composition, and syntactic/semantic biases, that may each play a role in listener perception of a given contrast. However, this lack of control over confounding higher-order factors is by design, and requires a reorientation of our perspective to thinking about functions and outcomes in addition to causes. For example, frequent errors between [ʧ] and [k] word-finally may reflect acoustic similarities between the two phones, such as their similar noise amplitude profiles and vowel-offset F2 transitions, but they may also reflect item-based response biases, as the [k]-offset word in most of the [k, ʧ] minimal pairs is the more frequent of the two. Nevertheless, this confounding effect of word frequency does not negate the ultimate outcome that in VC position, the *ʧ*→*k* contrast is vulnerable to errors in the lexicon. Cases such as these have broader implications for the structure of the obstruent system in English, and the role of both acoustic and non-acoustic cues in the functioning of that system in encoding meaning in the signal.

3.3.5 Cumulative error contribution

The analysis of listener accuracy by category/contrast, while informative as to the relative discriminability of different stimuli, does not account for asymmetries in their lexical distribution, such as those presented in Sections 3.3.3.1–3.3.4.2. In this section we look at the degree to which a given category or contrast contributes to the overall number of recognition errors committed in the experiment, and by extension, the lexicon which the item set has been sampled to represent. Such an analysis leads ultimately to estimates of components of the phonetic system that represent points of robustness versus those which constitute potential vulnerabilities in the maintenance of lexical contrast. Further, as in the accuracy analyses in the previous sections, potential modulating effects of noise level, word length, and word frequency are considered in assessing the structure of the cumulative error distribution.

3.3.5.1 Phonetic category errors

Beginning with the obstruent phone / feature class of the target item in each minimal pair, in the sections below we examine the cumulative error contributions of each category in word-initial, word-medial, and word-final contrasts.

Word-initial position (CV). Figure 3.20 displays the total proportion of errors in Experiment 1 that each phone accounts for as a function of how frequently that phone occurs as the stimulus member of a minimal pair contrast. Phones above the median error line indicate sounds responsible for more than their expected share of listener errors, whereas those below the line indicate sounds that listeners are more accurate at identifying words beginning with such sounds than expected. The phones of high occurrence frequency (rightward along the x-axis) are of particular importance because they contribute the most to overall word recognition performance, with sounds above the line (notably [b] and [p]) representing the greatest vulnerabilities in the system, while high-frequency phones below the line (notably [s] and [k]) represent points of critical stability. Overall, the relative position of phones above or below the median line is generally consistent at both $+2$ dB and -2 dB, though at the lower SNR, due to the greater acoustic uncertainty there is much less variability in error proportions relative to expectations. Finally, while these results are aggregated across Experiments 1a and 1b, the patterns described above are consistent across sub-experiments, suggesting they are robust and likely to be replicated throughout the lexicon. See Figures A.37 and A.38 in the appendix for the full CV error distributions in Experiments 1a and 1b.

Considering next the proportion of total word recognition errors each phone accounts for as a function of word length, Figure 3.21 shows that while polysyllabic items exhibit greater overall variance, many of the critical phones play a similar role in both sets; namely, [s] and [k] are robust in both mono- and polysyllabic items, while the labial plosives [p, b] represent notable vulnerabilities across word lengths. The key differences between the two sets are the disproportionate contributions of [h] to errors among monosyllabic contrasts, and [v] to errors on polysyllabic contrasts. Further, [f] represents a point of robustness among monosyllables, while [d], and to a

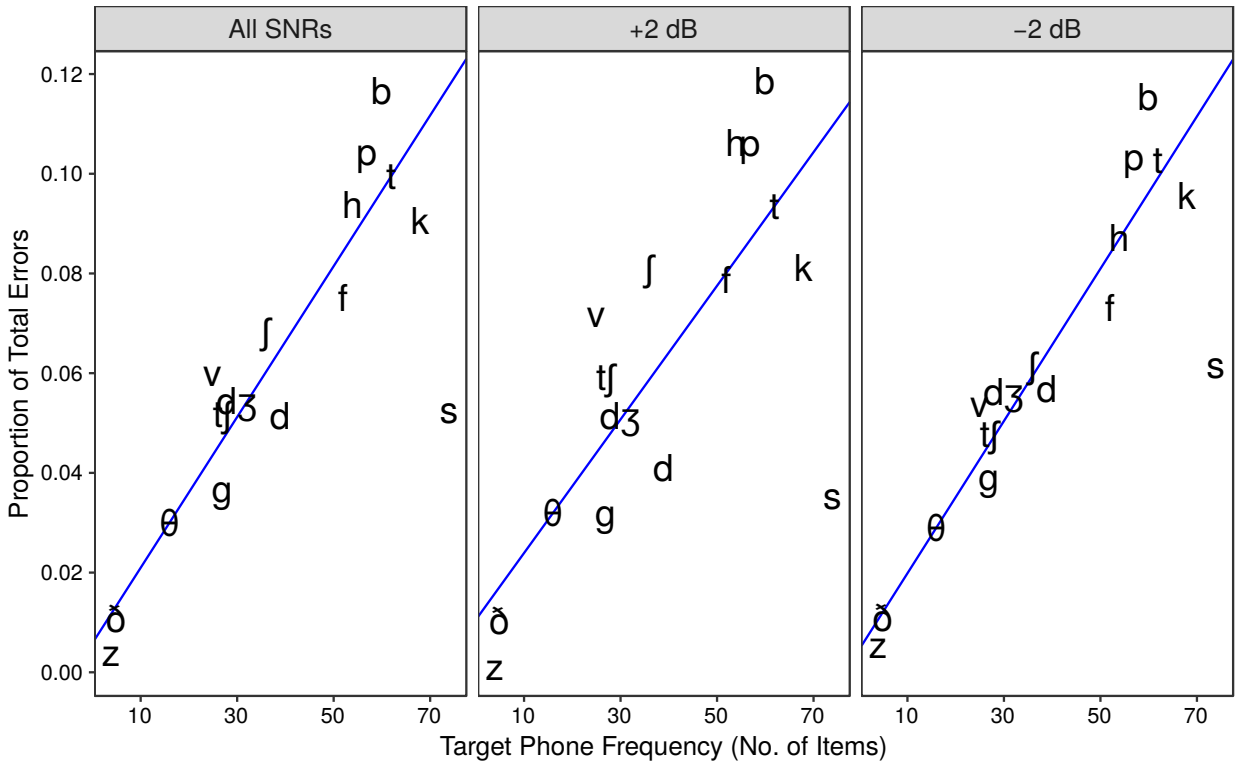

Figure 3.20: Proportion of errors in CV position in Experiment 1 (overall and by SNR) attributable to each target phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

lesser extent [g],[20] play a similar role in polysyllables. These results are largely consistent across sub-experiments, though the robustness of [f] is primarily restricted to Experiment 1a, while the vulnerability of [p] is primarily restricted to Experiment 1b (see Figures A.43 and A.44 in the appendix for details).

Finally, we assess the role of target word frequency in the contribution of each phone to the cumulative error distribution. This analysis is further informative as word frequency is ultimately a proxy for the token frequency of each contrast more broadly in the language, while till now we have primarily been concerned with type distributions in the lexicon. Figure 3.22 shows cumulative error contributions of each target phone in high- and low-frequency items based on a median split of the absolute frequency of the target word. Across both sets, [b] is responsible for a disproportionately high number of errors, while [s] is responsible for much fewer errors than expected, both consistent

[20]The voiceless dental [θ] is also well below the median, but as it occurs fewer than 5 times it is not expected to contribute much to the overall stability of the system.

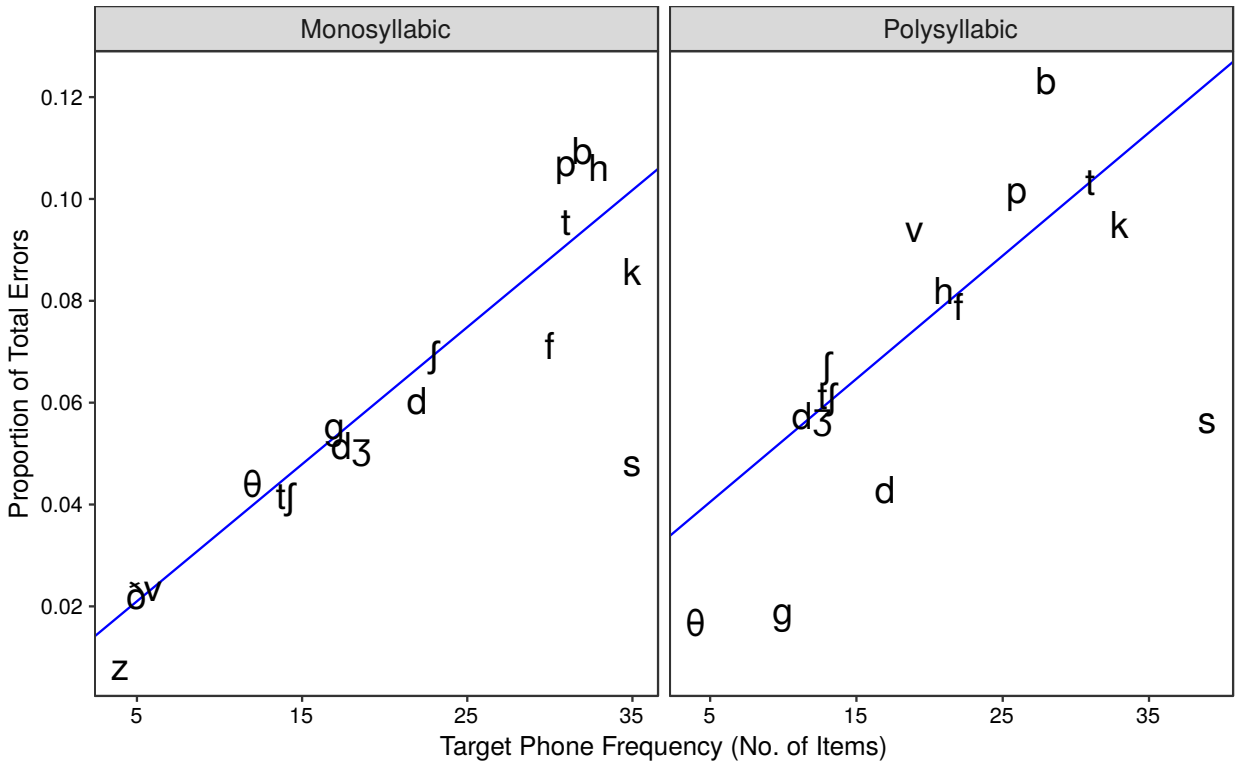

Figure 3.21: Proportion of errors in CV position in Experiment 1 attributable to each target phone as a function of word length. Lines indicate median regression fits.

with the previous results across SNRs and word lengths. Among high-frequency words, [h], and to a lesser extent [p], contribute disproportionately to the errors in Experiment 1, whereas [t] and [k] are relatively lower in error rates, though not nearly as robust as [s]. Phone contributions to cumulative errors are more varied among low-frequency words, with the set [v, ʃ] relatively higher in errors, and a large set of obstruents, [h, f, k, d, g] below-average in error rates.

These effects are largely consistent across sub-experiments (see Figures A.49 and A.50 in the appendix for details), with the role of [h] in low-frequency words the only notable discrepancy ([h] is above expectations in errors in Exp. 1a and below expectations in Exp. 1b). Overall, then, we find that while it is in low-frequency words where differences in acoustic salience and discriminability of obstruent consonants emerge, the high-frequency items are the ones most likely to impact the stability of the system in speech communication, and thus points of robustness, such as [s, t], and points of vulnerability, such as [b, h], within the high-frequency set are the most

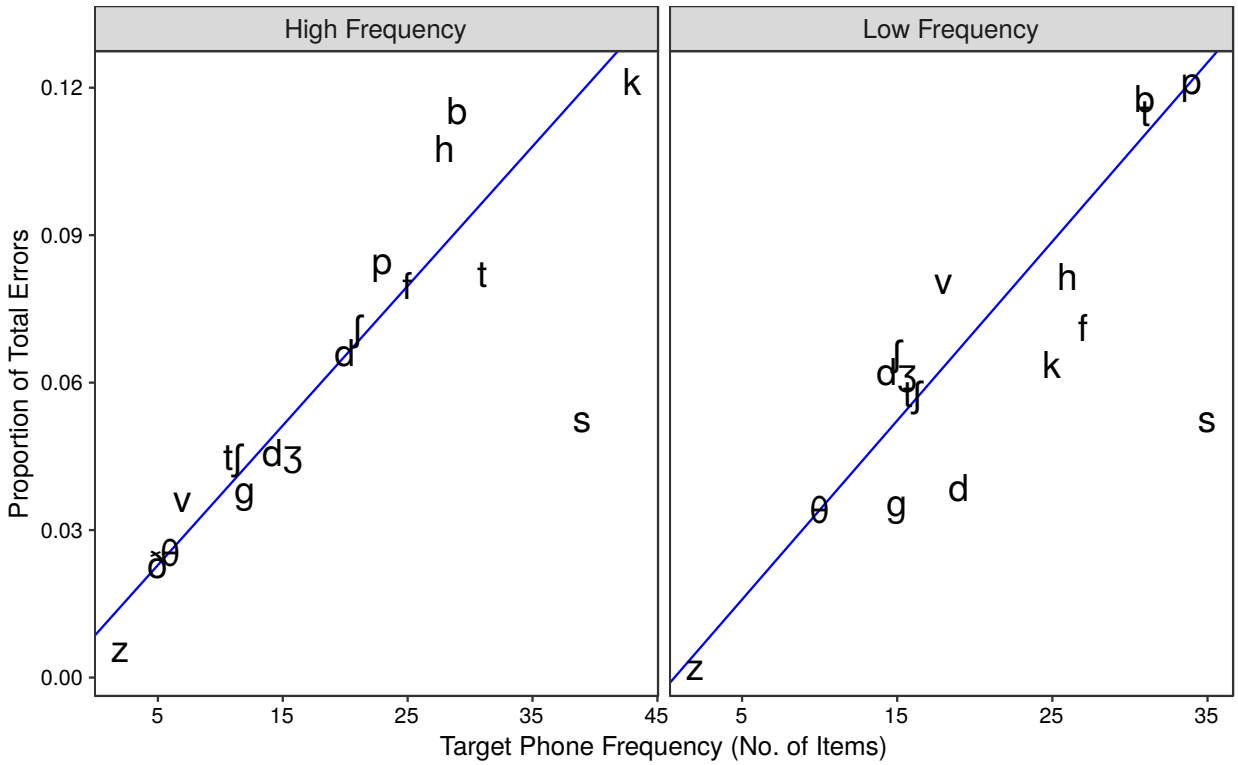

Figure 3.22: Proportion of errors in CV position in Experiment 1 attributable to each target phone as a function of target word frequency. Lines indicate median regression fits.

critical results obtained from this analysis.

Word-medial position (VCV). Figure 3.23 shows the proportion of listener errors on VCV minimal pairs as a function of the number of experimental items in which a given sound occurs contrastively. As noted earlier, the alveolar flap [ɾ] dominates the distribution, occurring in nearly one-fifth of all VCV stimuli. Further, listeners commit more errors on flaps than is expected based on median error rates among the 18 target phones (i.e., the point in Figure 3.23 is above the median regression line). This effect is especially pronounced at +2 dB SNR, suggesting the discrimination of flaps from other obstruent phones in English represents a serious vulnerability in the system when perception is perturbed by background noise. Similarly, [v] contributes a disproportionate amount of errors, while remaining quite frequent as a member of VCV contrasts in the lexicon. On the other hand, the alveolar sibilants [s, z], as well as the voiceless labiodental fricative [f], contribute disproportionately to the maintenance of contrast, being both highly frequent and lower

than average in error rates. Finally, in addition to these patterns being consistent across noise levels, they are also consistent across sub-experiments. See Figures A.39 and A.40 in the appendix for the full error distributions by target phone in Experiments 1a and 1b.

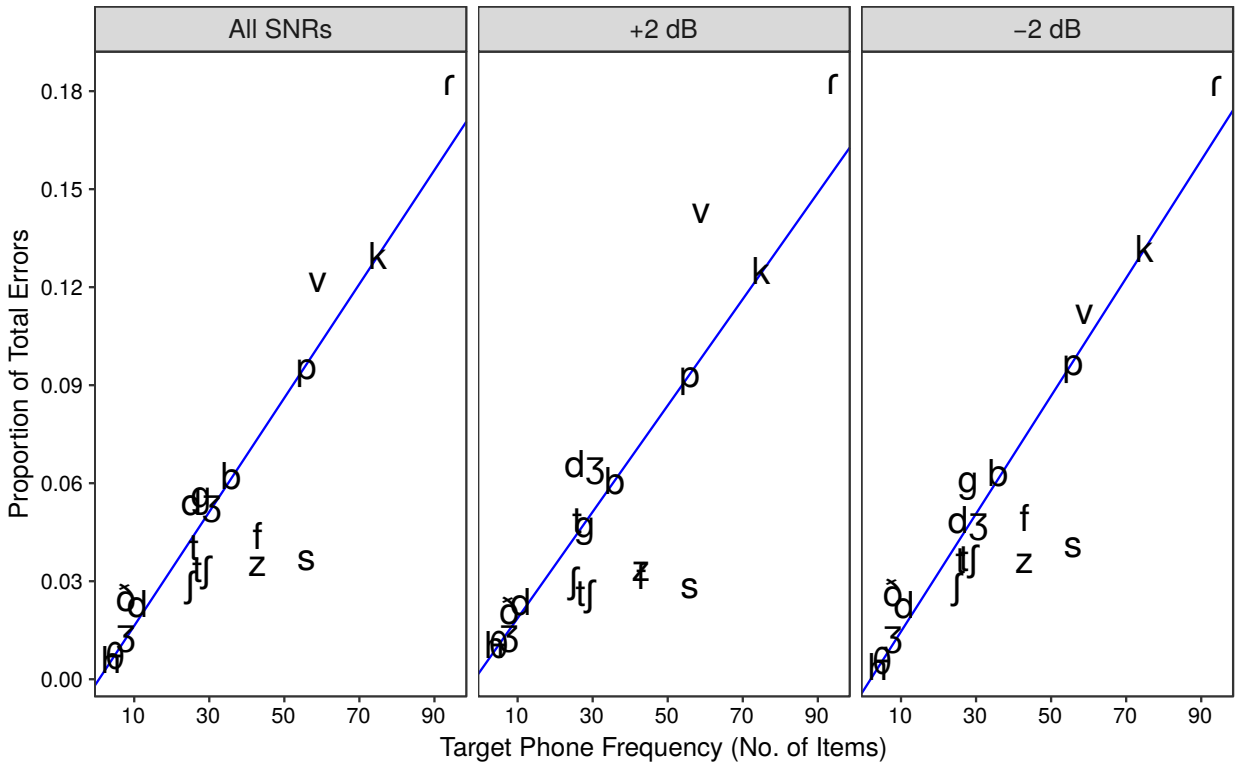

Figure 3.23: Proportion of errors in VCV position in Experiment 1 (overall and by SNR) attributable to each target phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

Regarding the distribution of errors among target phones as a function of word length, Figure 3.24 shows cumulative error contributions from each phone in disyllabic and trisyllabic target words. Given the much greater number of disyllabic minimal pairs, the overall error distribution in Figure 3.23 largely reflects patterns among disyllables; however, some results, such as the below-expectation error rates for [f, s], and the above-expectation error rate for [v], are consistent across word lengths. Further, these results are consistent across sub-experiments, suggesting the relative role of different phones in intervocalic obstruent contrasts generalizes well to the lexicon as a whole. See Figures A.45 and A.46 in the appendix for details.

Finally, Figure 3.25 shows the cumulative contribution of each target phone to listener errors

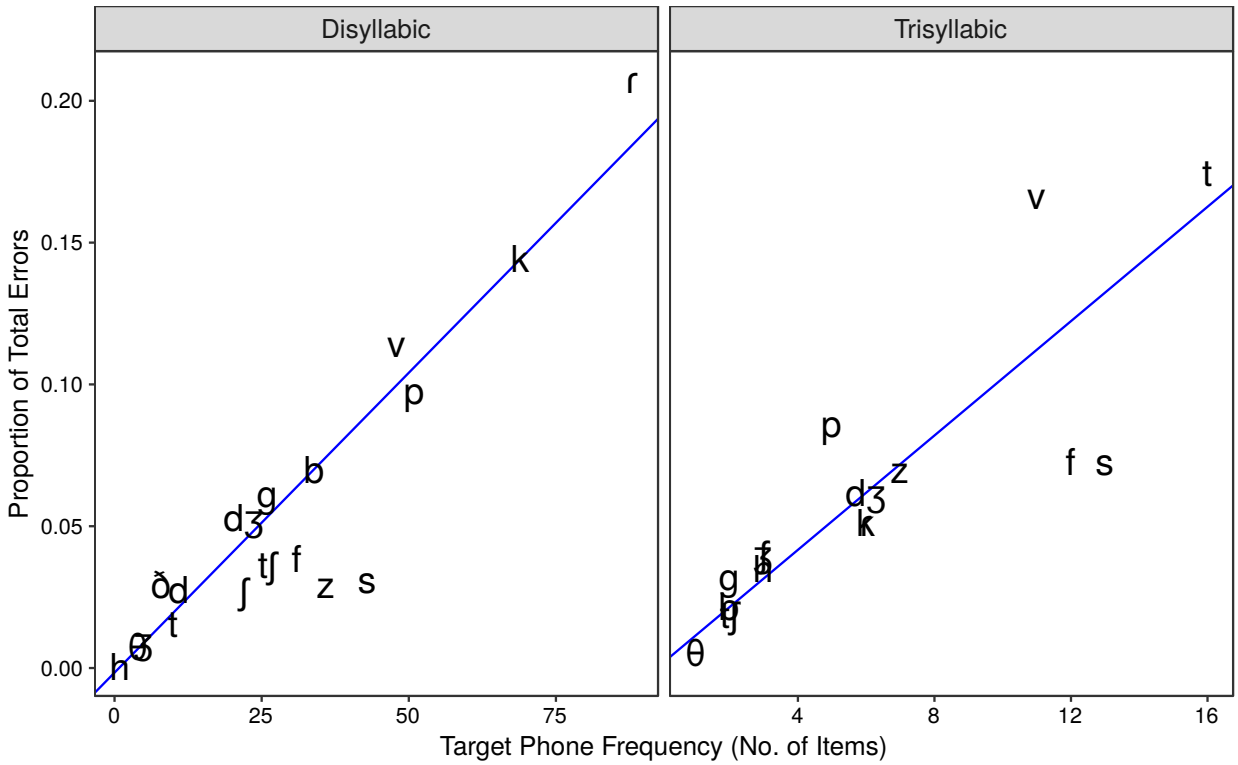

Figure 3.24: Proportion of errors in VCV position in Experiment 1 attributable to each target phone as a function of word length. Lines indicate median regression fits.

in high- and low-frequency words. The error distributions in Figure 3.25 are revealing in two key ways. The first is that the fricatives [s, z, f] are robust across word frequencies, suggesting their role in the lexicon depends to a large degree on their relative acoustic salience, and thus is less sensitive to changes in word frequency (or word length, as described above). Regarding phones that play the converse role of contributing disproportionately to word recognition errors, the phones [ɾ] and [v] are largely frequency-constrained in their patterning. The alveolar flap, for instance, is notably vulnerable in high-frequency words, suggesting that the perception of [ɾ] in obstruent contrasts, at least when listening in a noisy environment, is particularly problematic for communication because such items where it is disproportionately poorly perceived occur often. On the other hand, the disproportionate errors attributable to [v] occur much more in low-frequency words, a result which could reflect the lower salience of [v] allowing for a greater impact of top-down biases that favor the competitor word in a given contrast.

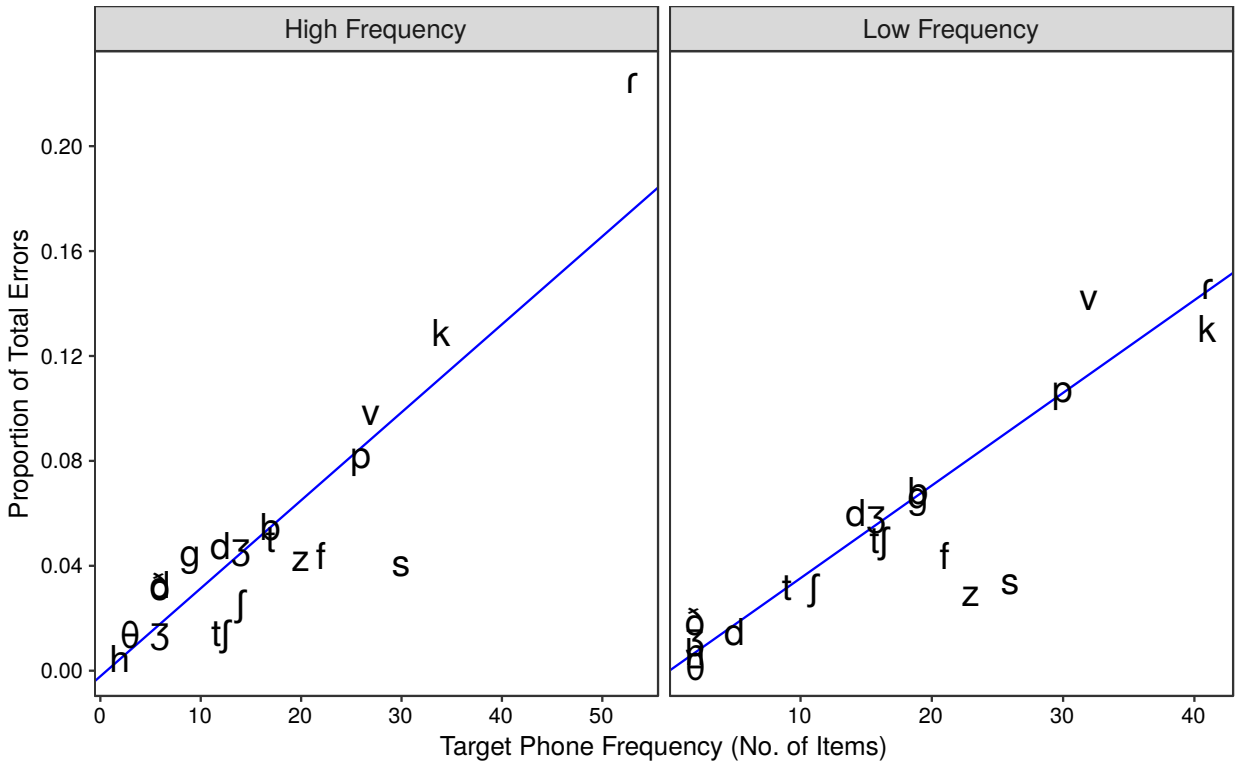

Figure 3.25: Proportion of errors in VCV position in Experiment 1 attributable to each target phone as a function of target word frequency. Lines indicate median regression fits.

Given that these results are consistent across sub-experiments (see Figures A.51 and A.52 for details), we can conclude that the most critical phone responsible for word-recognition errors intervocalically is [ɾ], whereas the sibilant fricatives and the voiceless labiodental [f] play a critical role in maintaining lexical contrast in VCV position. In this regard a crucial question to be addressed in future research is the distribution of flaps in contrasts with non-obstruent phones, as the relative frequency of such contrasts in the lexicon will impact both the overall role of [ɾ] in English, as well as theoretical implications of the underspecification of the [CORONAL] feature in speech perception.

Word-final position (VC). As Figure 3.26 shows, the distribution of errors among target phones in word-final contrasts is much more skewed toward a few obstruents that play an outsized role in the lexicon. As discussed in the prior distributional analysis these phones are the voiceless plosives [p, t, k], the voiced alveolar plosive [d], the sibilant fricatives [s, z], and the voiced labiodental

fricative [v]. Among this set, the majority contribute an expected proportion of errors given their frequency in the lexicon. The only notable deviations are the disproportionately high error rate of [d] and the relatively low error rate for [s]; [p] and [v] play similar roles, respectively, as points of vulnerability and robustness, but are much closer to the expected error count than [d] or [s]. Further, though reduced at −2 dB, these effects are consistent across SNRs and sub-experiments (see Figures A.41 and A.42 in the appendix for details).

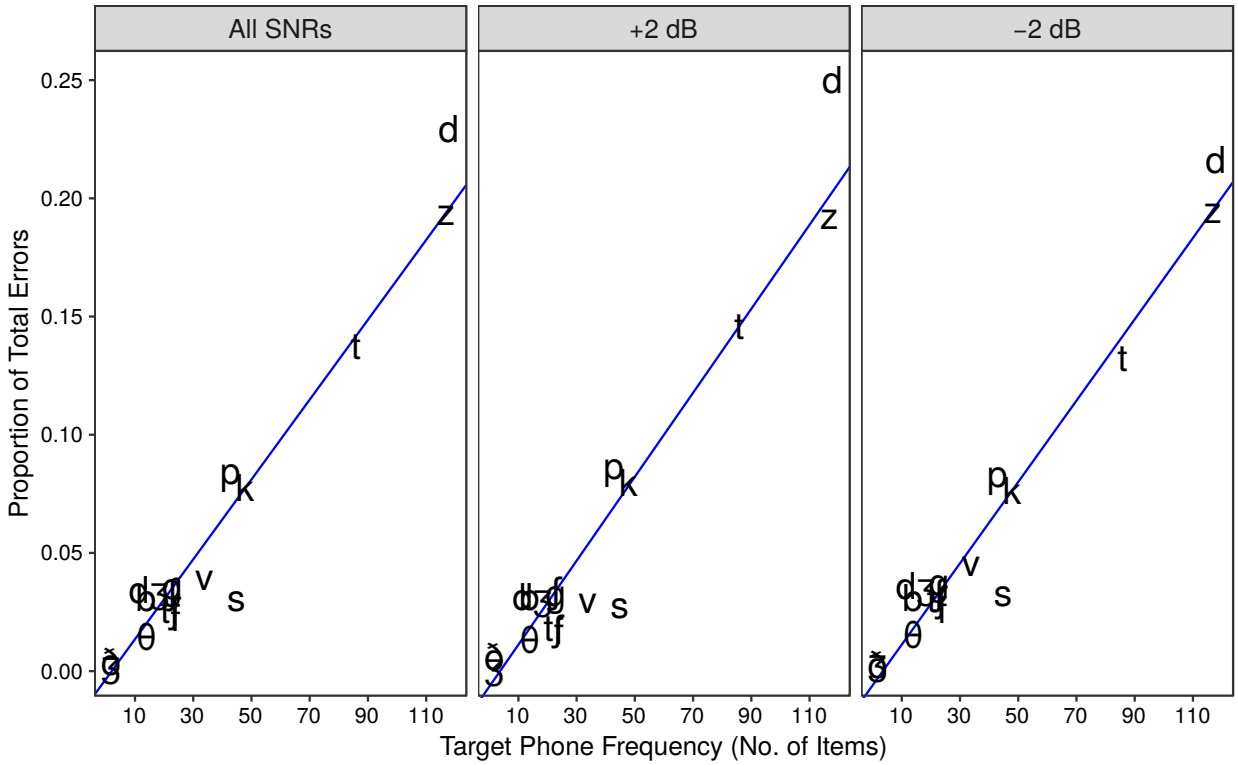

Figure 3.26: Proportion of errors in VC position in Experiment 1 (overall and by SNR) attributable to each target phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

Regarding the role of word length in word-final contrasts, in monosyllabic items we find a notably distinct pattern from that in polysyllables. Figure 3.27 shows that [p] and [d] contribute disproportionately to listener errors word-finally in monosyllables, while neither phone exhibits the same vulnerability in polysyllabic items. The two phones that play this role in polysyllables are [k] and [z], though both are relatively minor effects and are also inconsistent across sub-experiments (see Figures A.47 and A.48 in the appendix for details). Turning to the phones that are notably

robust word-finally, in monosyllables [s] and [z] consistently play this role overall and across sub-experiments, while in polysyllables [s] and [v] are the disproportionately robust target phones. Overall, we see that while [d] and [z], which primarily occur in contrast with each other in polysyllabic items due to English inflectional morphology, are largely consistent with expectations in terms of their contribution to listener errors word-finally. We will revisit this issue in the next section where the cumulative error contributions of contrasts such as *d–z* are directly examined.

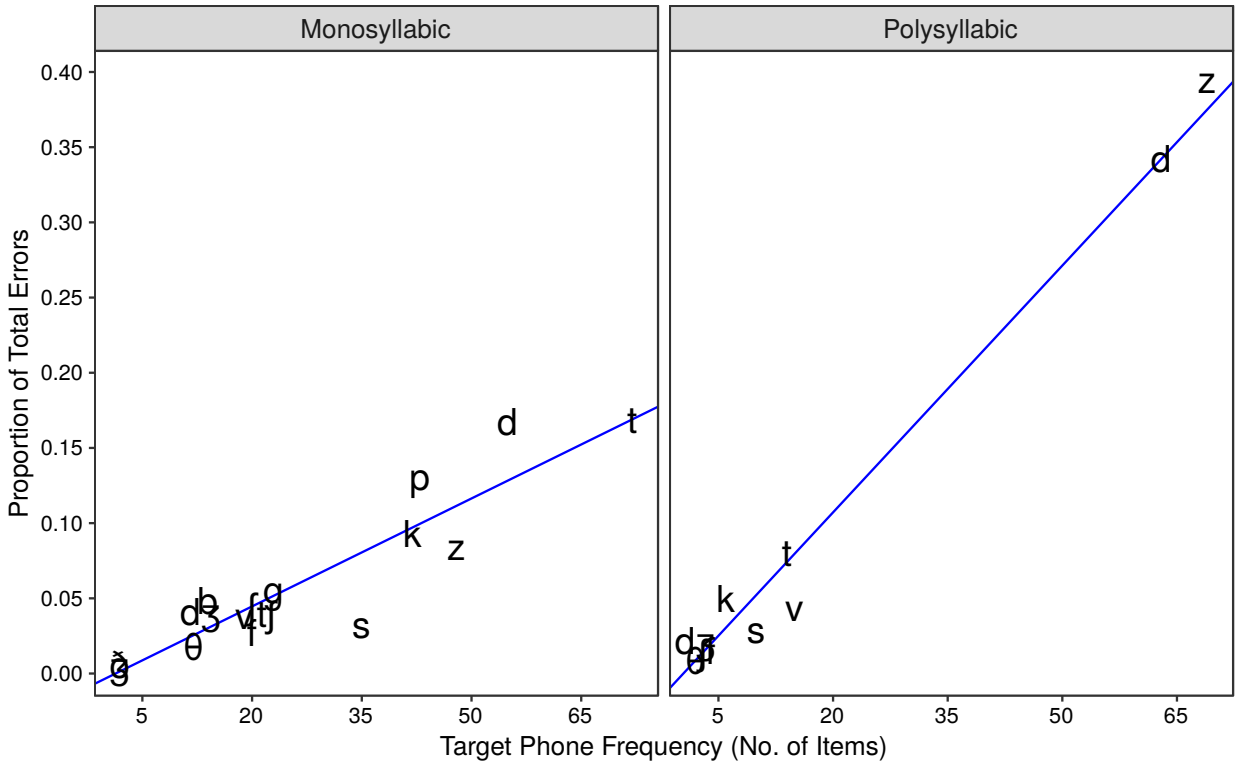

Figure 3.27: Proportion of errors in VC position in Experiment 1 attributable to each target phone as a function of word length. Lines indicate median regression fits.

Finally, Figure 3.28 shows the distribution of listener errors on word-final contrasts as a function of word frequency and the item frequency of the target phones in those contrasts. Unlike in CV and VCV positions, there is generally greater variance in target phone contributions to listener errors in high-frequency items than in low-frequency items, though among the more frequent phones this effect is largely diminished. For instance, [p] is a particularly large source of errors in high-frequency words, but not in low-frequency words, while [d] and [s] represent consistent

points of vulnerability and robustness, respectively, across word frequencies. Further, these results are largely consistent across sub-experiments (see Figures A.53 and A.54 in the appendix for details), suggesting that with the exception of [p], the type distribution of errors among target phones in word-final contrasts is generalizable to token distributions that may more accurately reflect the likelihood of a given contrast playing a role in communication.

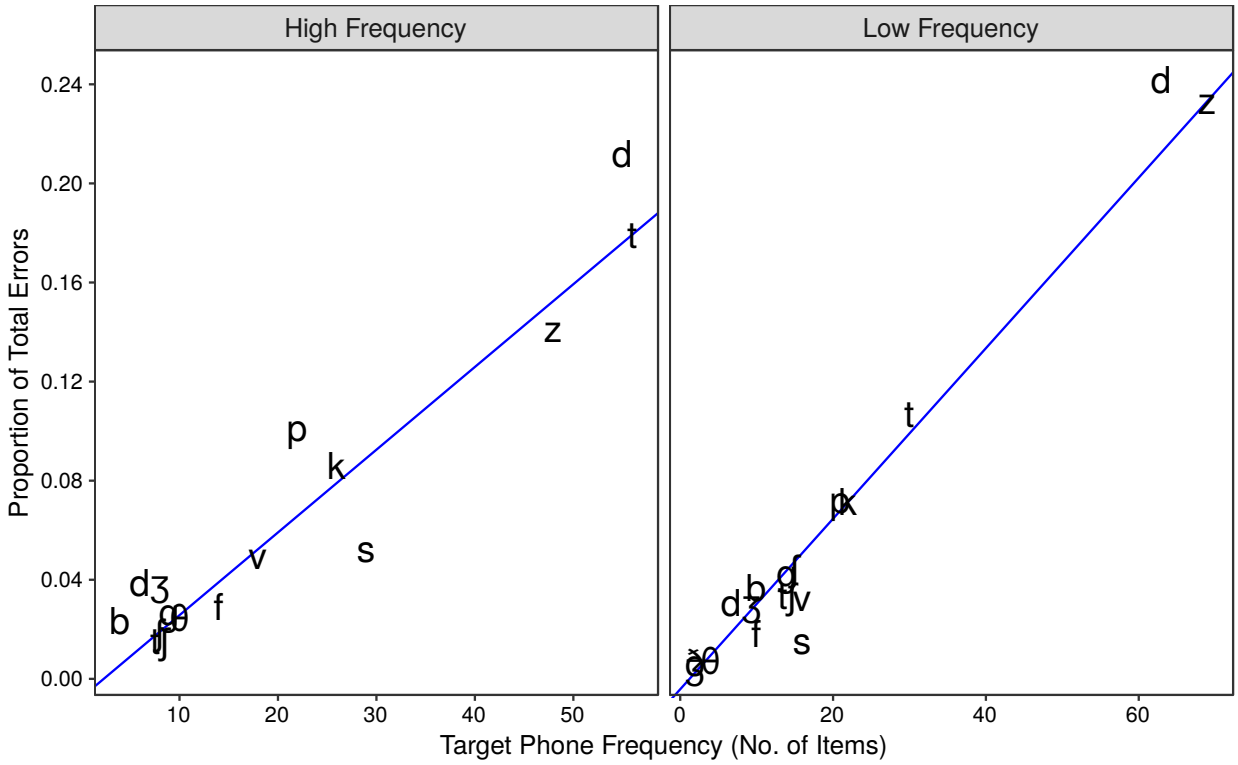

Figure 3.28: Proportion of errors in VC position in Experiment 1 attributable to each target phone as a function of target word frequency. Lines indicate median regression fits.

3.3.5.2 Phonetic contrast errors

As in the analysis of listener accuracy, our next question is how listener errors on specific obstruent contrasts contribute to the cumulative error patterns among target phones presented above. This analysis is further critical to the study of the role of the obstruent system in the lexicon, as it connects directly to the maintenance of contrast in the lexicon. In that regard, the analysis below is closely connected to the study of system structure under acoustic perturbation in Chapter 5.

Word-initial position (CV). Figure 3.29 shows the cumulative error contributions from word-initial obstruent contrasts both overall and by noise level, where the contrasts which are further from the expectation based on a median regression fit are displayed in larger font sizes to highlight points of relative vulnerability and robustness in the system. Recall that the most vulnerable target phones from Figure 3.20 were the labial plosives [b, p], while the voiceless velar plosive [k] and the voiceless alveolar sibilant [s] were notably robust in showing lower-than-expected error rates while also being highly frequent members of lexical contrasts. From Figure 3.29 we see that errors on [p] largely involve the labials [b, f] and the voiceless velar [k], while errors on [b] are more confined to the labial set; i.e., [p, f, v], though the *b–v* contrast occurs far less frequently than *b–p* or *b–f*. The other two contrasts in Figure 3.29 with notably higher error rates than expected, *k–t* and *k–h*, both involve [k] and thus the robustness of [k] as a target phone appears to largely reflect asymmetries in its perception relative to other acoustically similar obstruents.

Among the most robust contrasts in Figure 3.29 are several contrasts involving [s]: *k–s*, *h–s*, *s–t*, *b–s*, and *ʤ–s*. Thus, the role of [s] in the lexicon, as reviewed previously in the contrast accuracy results, is fairly consistent across contrasts. Other robust contrasts which are closer to their expected error rates are *b–k* and *d–k*, and thus the role of [k] is relatively evenly split between more robust and more vulnerable contrasts. All of the above patterns are consistent across sub-experiments and noise levels, though as Figure 3.29 illustrates, the variance in contrast error contributions is greatly reduced at -2 dB relative to $+2$ dB. See Figures A.55 and A.56 in the appendix for the cumulative error distributions in Experiments 1a and 1b.

Figure 3.30 shows the cumulative contribution of each contrast to listener errors as a function of word length. Here we see that the distribution of contrasts among mono- and poly-syllabic items varies considerably, but the relative role of each contrast in increasing or reducing listener word-recognition errors is fairly consistent. The most notable exception is the contrast between [b] and [k], which is notably more robust in polysyllables than in monosyllables. Results also vary across sub-experiments, though again primarily in terms of the distribution of contrasts in each experiment rather than in their relative role as points of vulnerability or robustness in the system

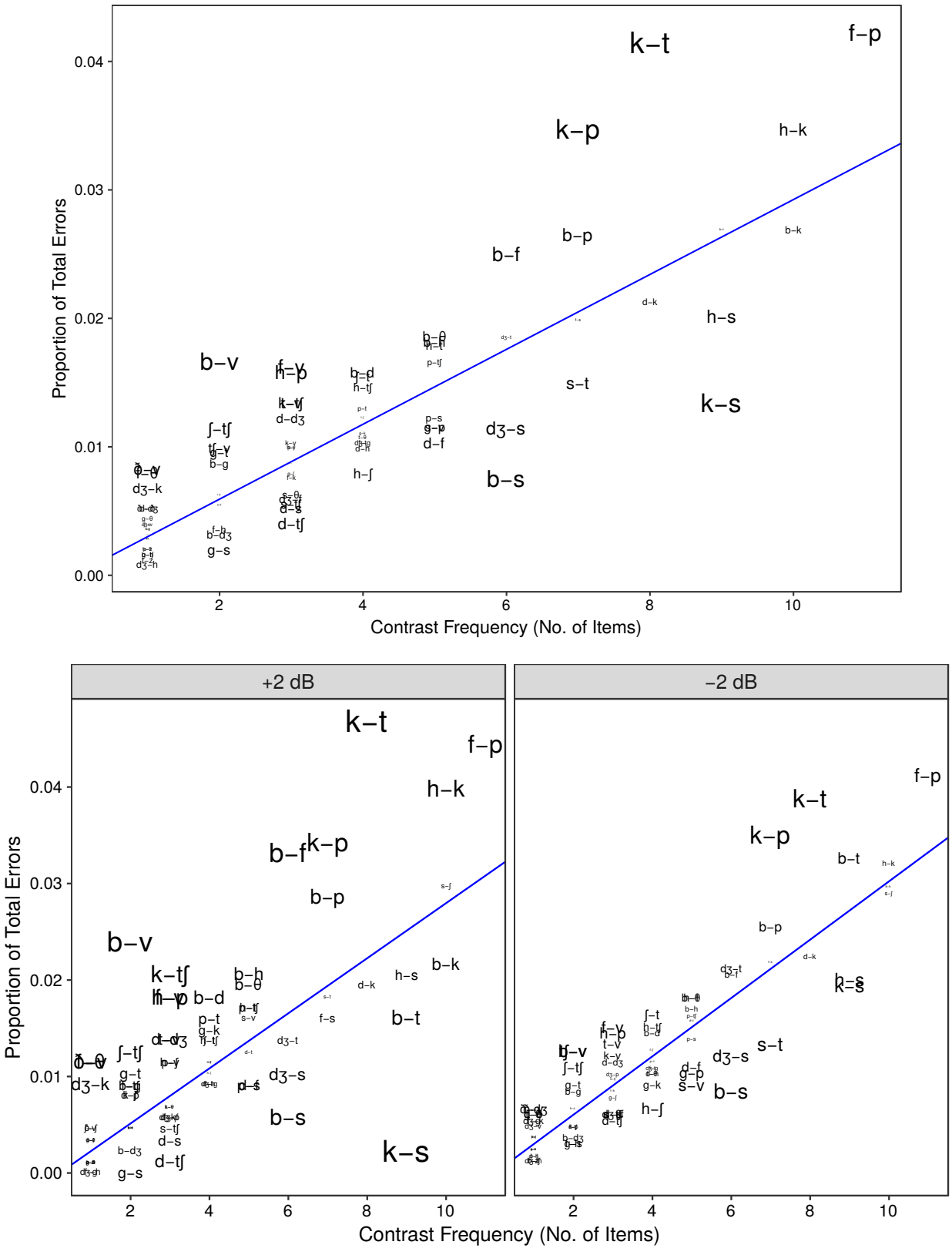

Figure 3.29: Proportion of errors in CV position in Exp. 1 attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

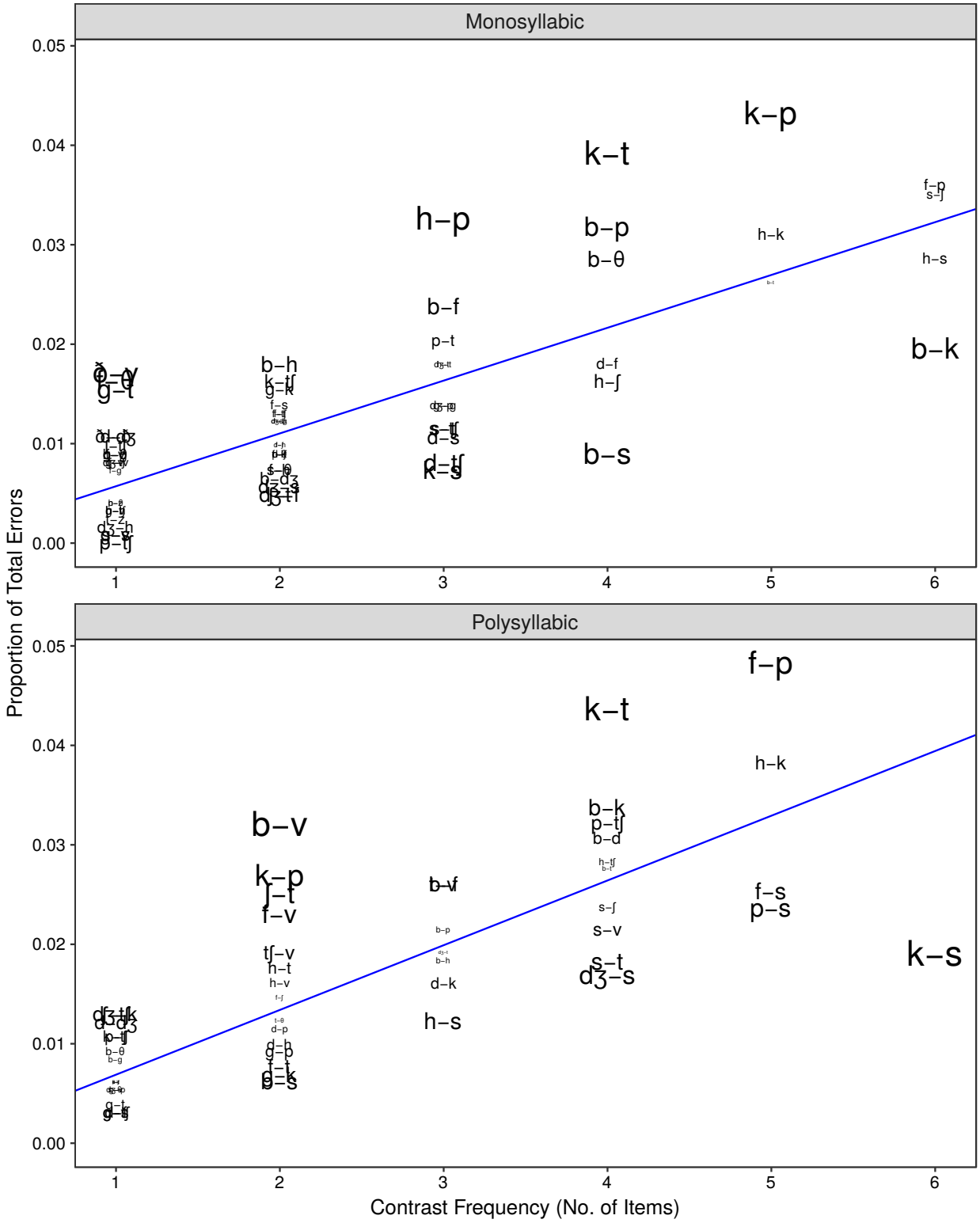

Figure 3.30: Proportion of errors in CV position in Exp. 1 attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

(see Figures A.57 and A.58 in the appendix for details).

Finally, Figure 3.31 shows cumulative error contributions from each contrast within low- and high-frequency items. As in the analysis of target phone errors in the previous section, the role of contrasts among high-frequency words is the most critical as it more closely reflects the probability a given contrast will be misperceived in communication. Among the more vulnerable contrasts in this set are *f–p* and *k–p*, and *h–k*, while *k–s* and *b–s* represent notable points of robustness. The distribution of contrasts in the low-frequency set is generally more variable, consistent with their greater reliance on bottom-up information from the acoustic signal. The most robust contrasts in this set are *h–s*, *b–k*, *s–t*, *ʤ–s*, and *d–k*, while *k–t*, *b–f*, *b–p*, and *k–p* represent notable points of vulnerability. Combining these frequency-dependent distributions with the overall patterns in Figure 3.29, we can identify the most robust contrasts as *k–s*, *b–s*, and *h–s*, while the most vulnerable are *f–p* and *k–p*, the latter of which is robust in both low- and high-frequency sets. These results are generally consistent across sub-experiments, though given the differences in item distributions the extent to which a given contrast contributes disproportionately to listener errors varies in each experiment. See Figures A.59 and A.60 for the complete error distributions by word frequency in Experiments 1a and 1b.

Word-medial position (VCV). The cumulative contribution of intervocalic contrasts to listener errors in Experiment 1 is shown in Figure 3.32, and as expected from the analysis of category and contrast distributions in VCV position, the majority of the critical contrasts—i.e., points of potential vulnerability/robustness—occur with the alveolar flap [ɾ]. The one notable exception to this pattern is the voiceless plosive contrast *k–p*, which is both highly frequent and a much greater source of listener errors than expected based on the median error rate in VCV position. Other contrasts in this vulnerable set are the contrasts between the alveolar flap and several voiced obstruents, [b, g, ʤ, v, ð], while frequent contrasts between [ɾ] and the voiceless plosives [p, k] are closer to the expected error rate.

Among the notable points of robustness in Figure 3.32 are contrasts with both voiced and voice-

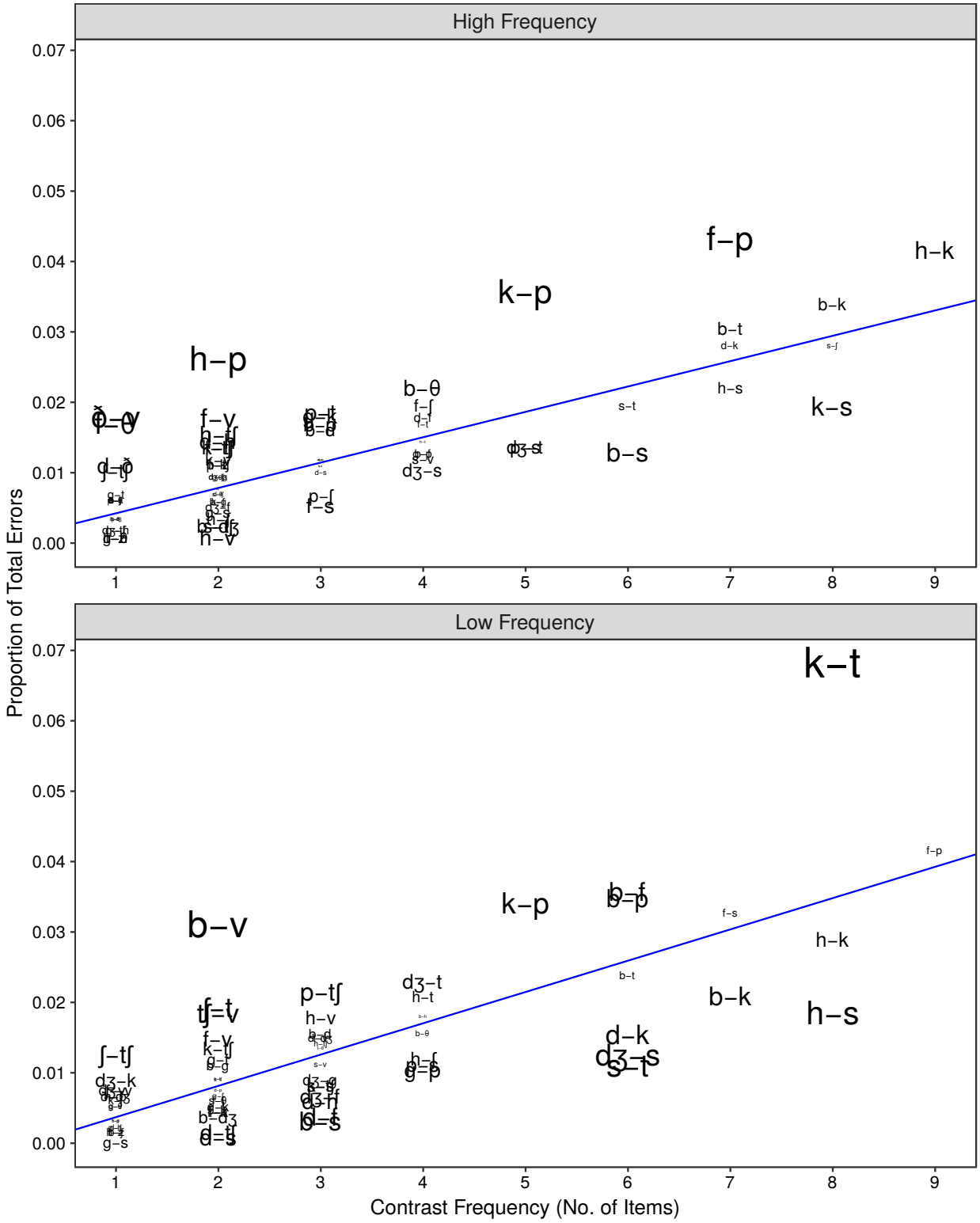

Figure 3.31: Proportion of errors in CV position in Exp. 1 attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

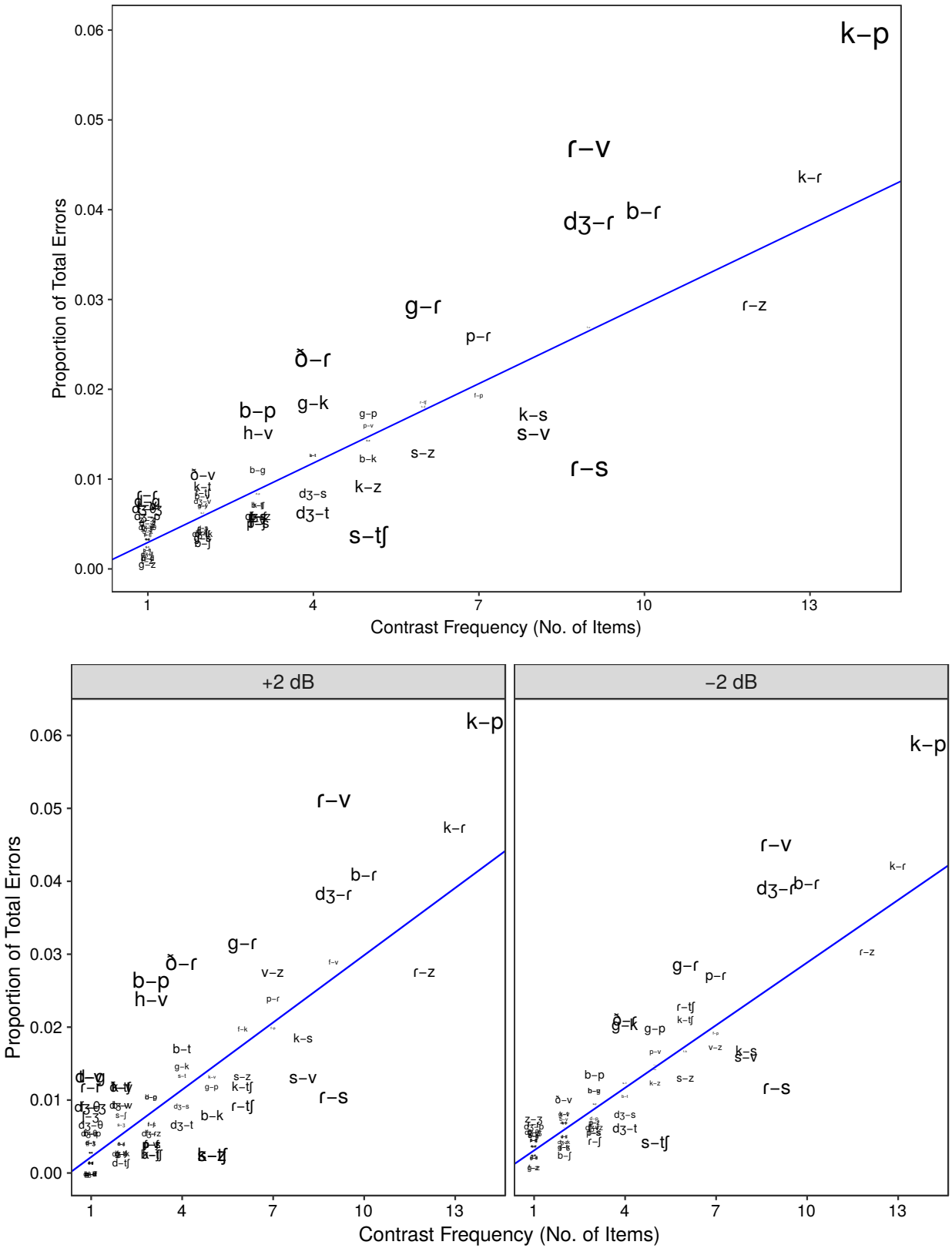

Figure 3.32: Proportion of errors in VCV position in Exp. 1 attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

less sibilants, *ɹ–z* being the most frequent contrast in this set that is below the median error rate, and *ɹ–s* exhibiting the greatest reduction in error relative to expectations. Other contrasts in this set are *s–v*, *k–s*, and *s–ʧ*, each of which can be considered to play a key role in preserving lexical distinctions intervocalically. These results are further consistent across SNRs, with the variability about the median line primarily increasing in magnitude at +2 dB, with few changes in directionality. The greatest of these changes in magnitude are an increase in the relative vulnerability of *b–p*, *ð–ɹ*, and *h–v* at +2 dB, and an increase in the relative robustness of sibilance contrasts at −2 dB. Experiments 1a and 1b are largely consistent in replicating these patterns, though Experiment 1b shows greater variability overall in contrast error rates; i.e., points of vulnerability and robustness are relatively greater compared to Experiment 1a. See Figures A.61 and A.62 in the appendix for the full error distributions by sub-experiment.

When broken down by word length, we find that the patterns noted above are primarily due to disyllabic items, a result again of their much greater frequency in the lexicon relative to trisyllables. Thus, much of the new information provided in Figure 3.33 is about the relative role of contrasts in trisyllabic items, though due to their low type frequency it is harder to distinguish identify *critical* items in this set, as a difference between 4 and 2 items, for example, is comparatively insignificant compared to the 10+ item differences in the disyllabic set. Nevertheless, among trisyllables the contrasts exhibiting the highest error rates are *h–v*, *b–p*, *f–t*, and *s–t*, while with the exception of *s–t* much of the remaining sibilance distinctions are relatively robust in perception and thus exhibit lower-than-expected error rates. Because of the general sparsity of trisyllabic contrasts in the lexicon, these results do not generalize well across sub-experiments, though the disyllabic results (at least in terms of directionality) are largely consistent between Experiments 1a and 1b (see Figures A.63 and A.64 in the appendix for details).

Finally, in Figure 3.34 we examine the relative contribution of intervocalic contrasts to listener errors as a function of target word frequency, where again the primary focus is on high-frequency items, as error distributions in this set most closely relate to the relative likelihood a given contrast will contribute to accurate word recognition in speech communication. Many of the more vulner-

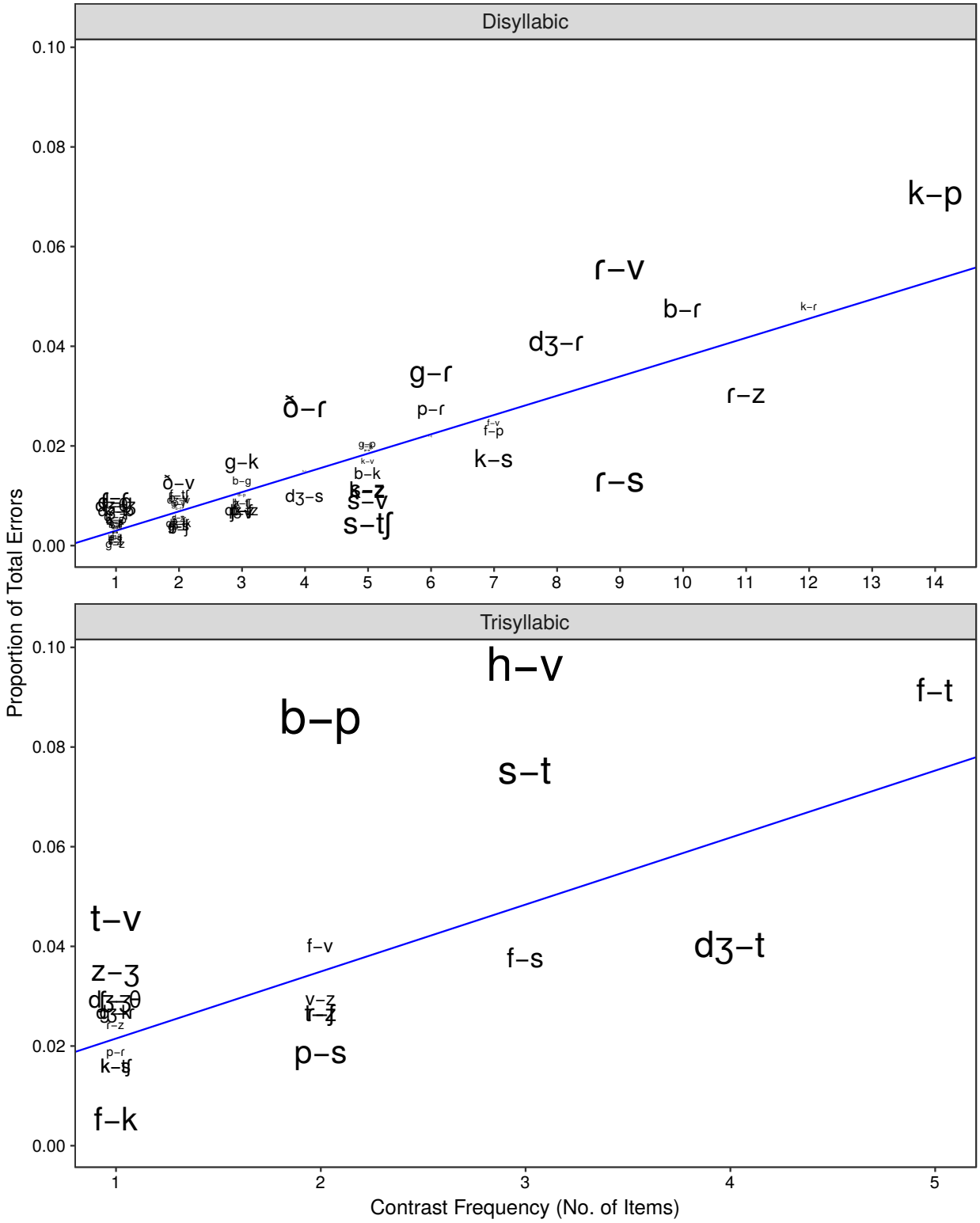

Figure 3.33: Proportion of errors in VCV position in Exp. 1 attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

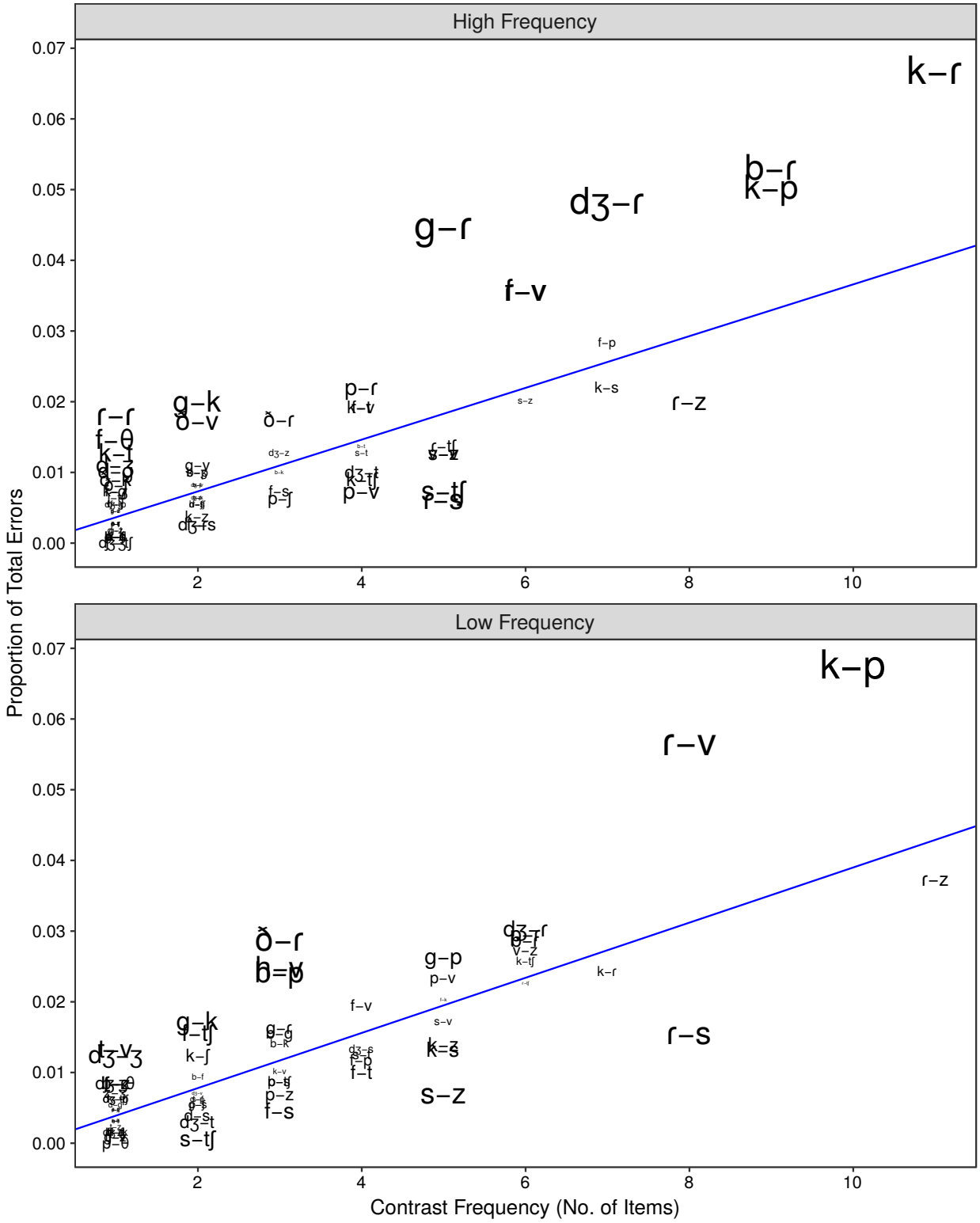

Figure 3.34: Proportion of errors in VCV position in Exp. 1 attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

able contrasts between voiced obstruents and flaps that emerged in the aggregate results, such as *b–ɾ*, *ɟ͡ʒ–ɾ*, and *g–ɾ* primarily occur among relatively high-frequency words, while others such as *k–ɾ* and *f–v* are much more robust in low-frequency items and thus appear closer to the expected error rate when aggregated across low- and high-frequency sets. The voiceless plosive contrast *k–p*, on the other hand, is consistently vulnerable in contrasts involving both high- and low-frequency words. The general characteristics of the points of robustness in VCV contrasts, however, are relatively consistent across word frequencies, as nearly all such contrasts involve either a sibilance distinction, or a distinction between sibilants, all of which are highly salient intervocalically (see the accuracy analysis in Section 3.3.4.3 for details). The directionality of these effects is generally consistent across sub-experiments, though due to differences in contrast and item distributions, Experiments 1a and 1b do vary in the role of such contrasts (item count) and in the size of the increase/decrease in error rate relative to expectations (positive/negative residuals). See Figures A.65 and A.66 in the appendix for details.

Word-final position (VC). Figure 3.35 shows the cumulative error contribution from word-final contrasts in Experiment 1. For clarity of presentation, the contrast between [d] and [z] which accounts for over 20% of word-final contrasts in Experiment 1 has been excluded from the Figure 3.35, though this does not impact our assessment of relative points of vulnerability/robustness as the number of errors on *d–z* is consistent with expectations. Among the most vulnerable contrasts word-finally are those between voiceless plosives (namely, *k–t* and *p–t*), while points of robustness primarily involve sibilance and voicing contrasts, two features which are perceptually robust in VC position. These patterns are further consistent across noise levels, as well replicating in general across sub-experiments, though *k–t* is more variable in this regard, exhibiting much higher error rates than expected in Experiment 1b relative to Experiment 1a (see Figures A.67 and A.68 in the appendix for details).

Regarding the impact of word length on the relative role of word-final contrasts in promoting or inhibiting listener recognition, Figure 3.36 shows that while the role of voiceless plosive contrasts

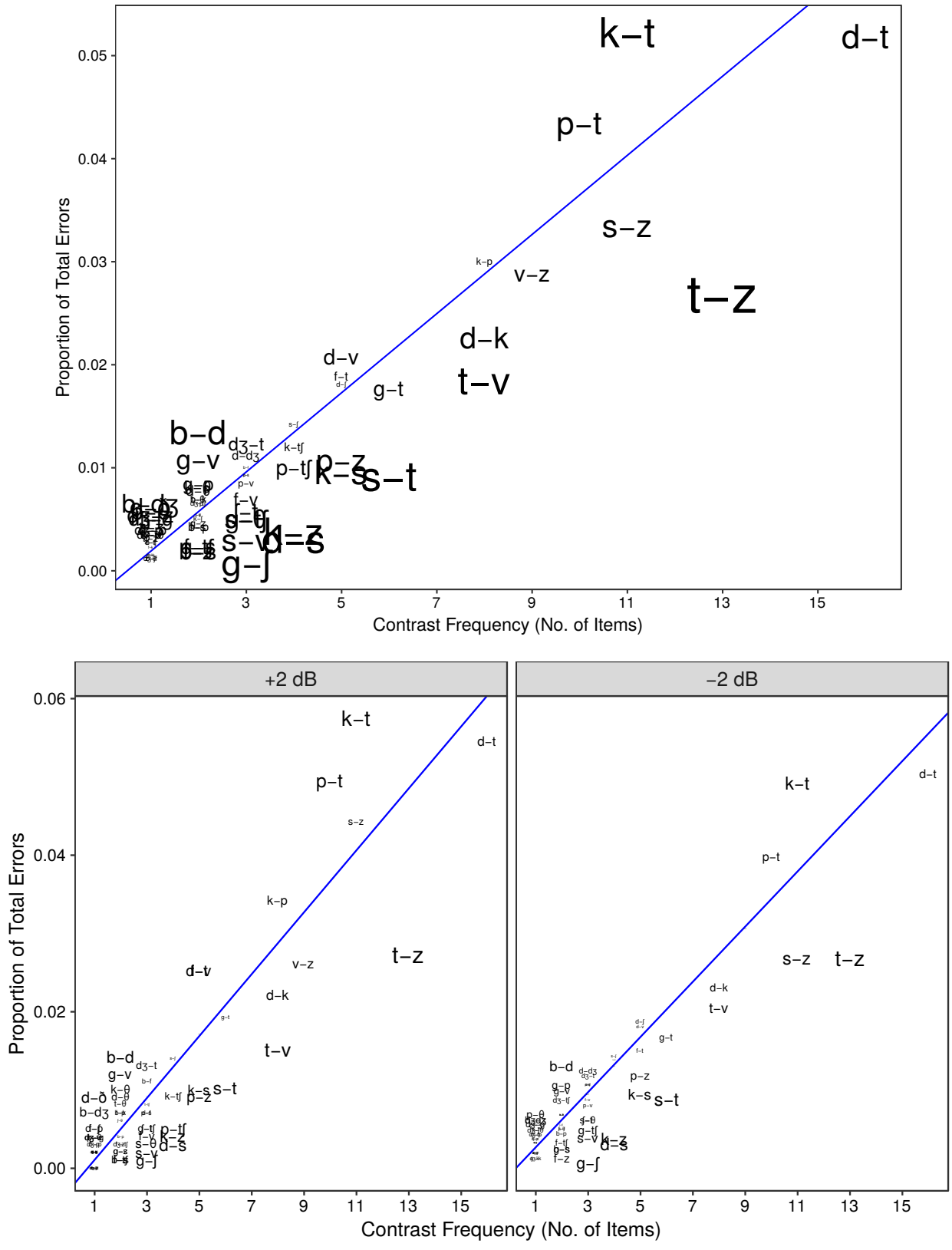

Figure 3.35: Proportion of errors in VC position in Exp. 1 attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits. The contrast *d–z* has been left out of the plot for clarity purposes, but lies along the median line at $x = 67$.

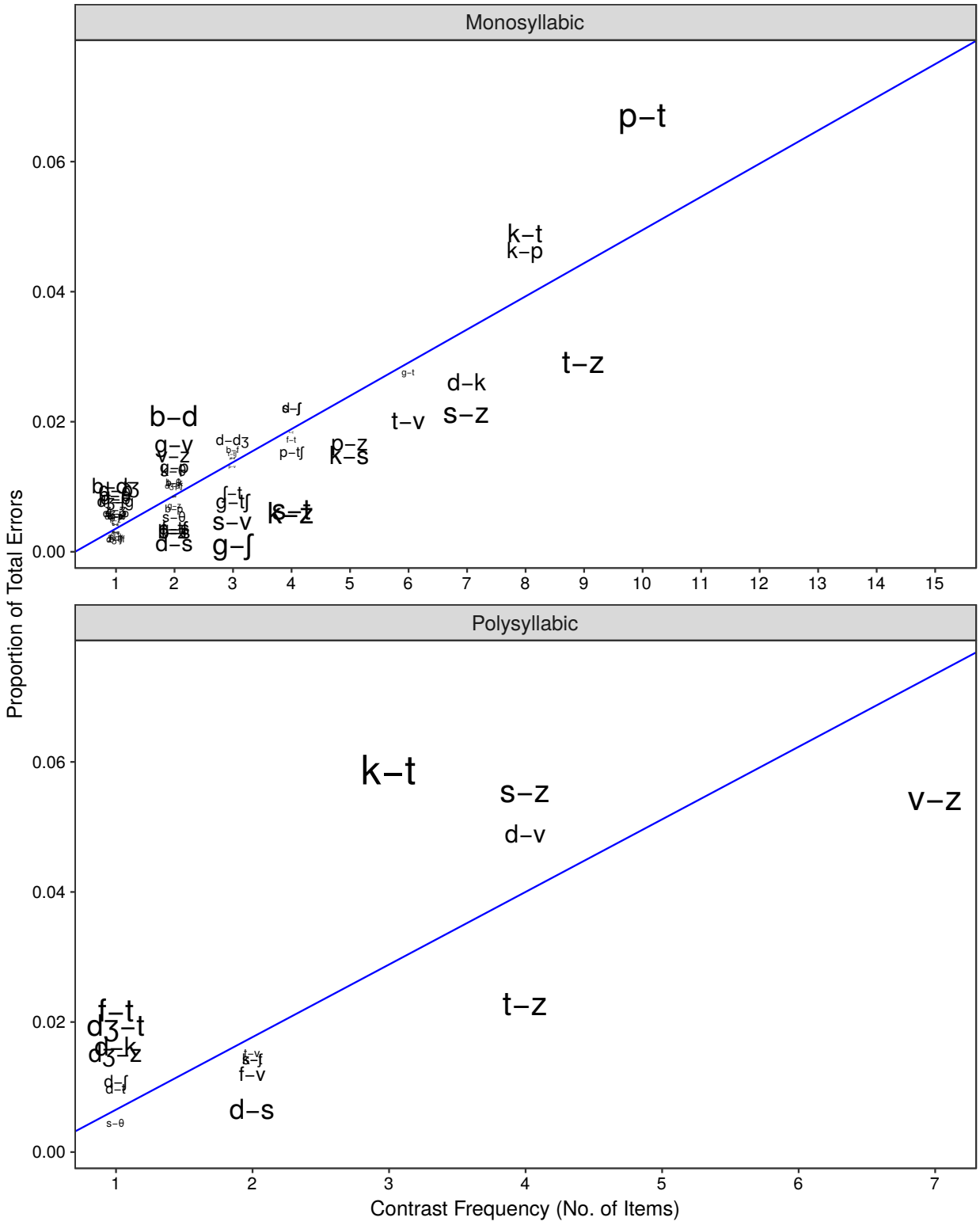

Figure 3.36: Proportion of errors in VC position in Exp. 1 attributable to each contrast as a function of item count and word length. Lines indicate median regression fits. The contrast *d–z* has been left out of the plot for clarity, but is consistent with the expected error rate.

and sibilance contrasts word-finally is consistent across word lengths, the relative discriminability of the voicing contrast *s–z* varies significantly as a function of word length, being notably robust in monosyllables but vulnerable in polysyllables. This result replicates across Experiments 1a and 1b (see Figures A.69 and A.70 in the appendix for details), and is consistent with the greater reliance of sibilant voicing distinctions on duration in VC position, given that [z] shows considerable devoicing word-finally, as durational distinctions generally reduce both with increasing word length and as a function of word stress, both characteristics that are detrimental for the perception of word-final voicing in polysyllabic contrasts.

Finally, Figure 3.37 shows the cumulative distribution of errors among word-final obstruent contrasts as a function of word frequency. In general the relative vulnerability of voiceless plosive contrasts, and the relative robustness of sibilance contrasts, is consistent across word frequencies, though the distribution of contrasts within each set varies between low- and high-frequency sets. Among high-frequency items, voiceless plosive contrasts with the labial [p] are most vulnerable, while voicing and sibilance contrasts are notably robust. In the low-frequency set *k–t* disproportionately contributes to listener errors, which is expected given that the distinction between [k] and [t] relies heavily on characteristics of the noise spectrum which are either reduced in salience word-finally, or absent entirely in cases where such stops are unreleased. The directionality of these effects is consistent across sub-experiments, though Experiments 1a and 1b differ in the relative frequency of each contrast, and thereby their role as points of critical vulnerability or robustness in the lexicon. See Figures A.71 and A.72 in the appendix for details.

3.3.5.3 Summary of cumulative error results

When the cumulative contribution of each phone/contrast to listener errors is studied, we get a clearer picture of the relative role each component of the obstruent system plays in the discrimination of items in word recognition. At one end of the spectrum are phones and contrasts that both occur in many minimal pairs and are responsible for considerably more errors than expected based on listeners' typical error rate on obstruent distinctions. This set represents a potential vulnerabil-

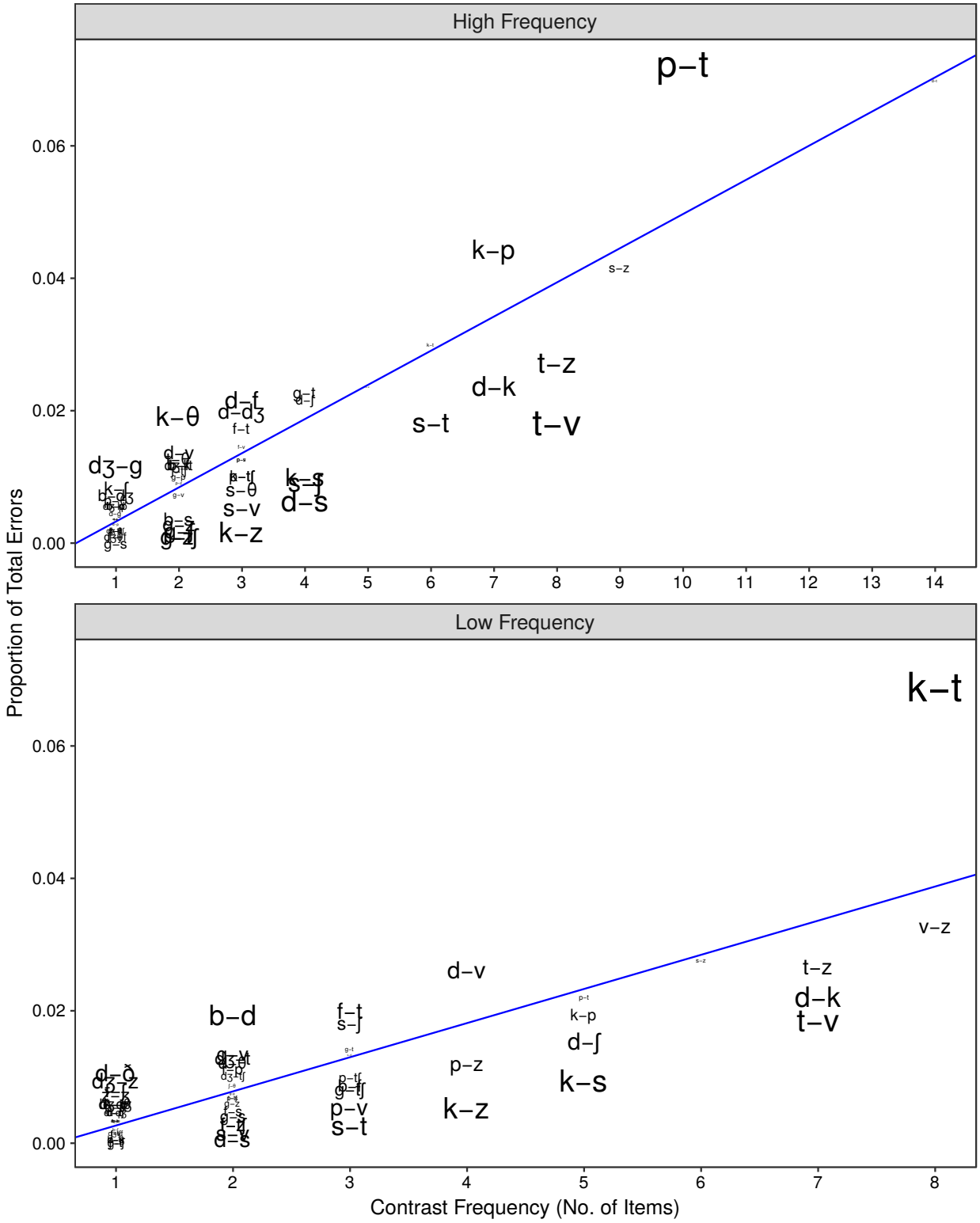

Figure 3.37: Proportion of errors in VC position in Exp. 1 attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

ity in the system, as listeners are both more likely to encounter such items in communication, and when presented in the stimulus they are relatively more likely to misperceive the target. Across contrast positions, distinctions among the voiceless plosives [p, t, k] represent substantial vulnerabilities in the system, being both highly frequent in the lexicon and easily confusable in word recognition in noise. By comparison, sibilance contrasts, particularly those involving the voiceless alveolar sibilant [s], represent notable points of robustness, being highly frequent in the lexicon and more accurate than nearly all other contrasts.

Moving forward to the study of cue integration in Chapter 4 and system structure in Chapter 5, error distributions such as these provide a critical window on the core perceptual patterns that drive the overall success of listeners in word recognition. Such patterns not only determine the relative weight attached to different cues in perception, but also have implications for the response of the system of contrasts in the lexicon to uncertainty in the acoustic signal, be it from overall perturbation by background noise, or specific cue loss/degradation.

3.3.6 Discussion

In the analysis of lexical contrast perception in Experiment 1, our focus has been on providing a detailed description of the relative identifiability of different obstruent phones and feature classes in the stimulus, the relative discriminability of phonetic/featural contrasts, and the aggregate role of such units in contributing to or preventing listener errors in word recognition.

Regarding target phone/feature perception, sibilant fricatives are robust across positions, noise levels, word lengths, and generally word frequencies, whereas labials are consistently poorly perceived (excepting [f, v] in VC position), and are easily affected by variation in noise level and word length. As a class, voiceless obstruents are generally more accurately perceived than their voiced counterparts, particularly in VCV position, and fricatives are more accurately perceived on average than other manners of articulation, and are generally robust to changes in noise level, word length, and word frequency.

When broken down by specific phonetic/featural contrasts, the degree of asymmetry evident

both in the phonological distributions in the lexicon and in listener perception increases substantially. Voiceless plosives and the alveolar sibilant [s] dominate the contrast distributions across all three positions, while intervocalically contrasts with the alveolar flap [ɾ] occupy a significant portion of the system, and word-finally contrasts primarily involve alveolar obstruents, particularly the morphological contrast between [d] and [z]. Further, at a featural level, the majority of contrasts are distinguished by two or more features, a result which runs counter to the tradition in the phonetic literature of isolating single-feature distinctions, but which will likely determine many of the cue-weighting outcomes in Chapter 4. In terms of accuracy, several phones are consistently perceptually robust (*s*, *f*) or weak (*p*, *b*, *v*, *ð*) across phonetic contrasts and positional contexts. Others are more featurally localized, such as the grouping of [ʧ] errors among voiceless obstruents and [ʤ] among voiced obstruents, and the grouping of [h] errors with the voiceless aspirated series [p, t, k]. Overall, sibilance is the most accurately transmitted featural contrast, though it is asymmetric in being much more accurate when the target is a sibilant than when it is a nonsibilant. Voicing transmission is similarly asymmetric, with listeners more accurate at perceiving voiceless obstruents against voiced competitors than vice versa. The most notable results for manner and place transmission are the greater salience of fricatives and [LOW] coronal obstruents.

Finally, in the previous section we examined the cumulative contribution of each phone/contrast to listener errors in Experiment 1. This analysis revealed quite consistent points of robustness (sibilance contrasts, primarily with [s]) and vulnerability (place distinctions among voiceless plosives and nonsibilant contrasts with labial plosives in general) across positions. Here we have the clearest example of the impact of the warping of phonological distributions provided by the lexicon, as it appears that the majority of listeners' success or failure in word recognition depends on a small set of distinctions. In the next chapter the impact of these asymmetries on cue integration will be explored, while Chapter 5 will examine the extent to which the system of obstruent contrasts in the lexicon depends on this set in its response to perturbation by background noise and by the loss in contrastive information from a given acoustic cue.

3.4 General discussion

In this chapter we have presented a comprehensive description of the perception of obstruent contrasts in a wide range of minimal pairs from a model lexicon of a single native English speaker. Just as Chapter 2 provided the acoustic basis for the study of lexically dependent cue weighting in Chapter 4, the present chapter provided the perceptual basis for such models. However, beyond simply providing a thorough description of the data that will serve as a baseline for the cue-integration models to track, the present chapter demonstrates many new considerations that emerge in the study of phonetic systems as a function of the lexical distinctions they encode. For instance, word length and particularly word frequency are not typically factors that are assessed in the study of the fundamental perceptual characteristics of the phonetic system. They may be examined *post hoc*—e.g., in studies of speech rate or in models of spoken word recognition and lexical access—but the inventory assumption does not directly necessitate their involvement at a phonetic level. But in a framework where the fundamental units of analysis are relations between words, such lexical characteristics directly emerge as possible modulating factors in perception. This means that in accounting for more naturalistic speech perception contexts, there is less of an information gap to bridge, and less of a need for external mechanisms, such as the lexical feedback procedure in TRACE (McClelland & Elman, 1986), to resolve misalignments in bottom-up expectations that may simply be due to listeners not attending to a particular cue or feature because they have learned over time that that cue in a given contrast and context is uninformative, either due to variability in the lexical items in which it occurs, or due to the reliable presence of other lexical information that may be more stable.

Chapter 4

Cue integration

Outline

This chapter presents results of several statistical models of cue integration designed to test the relative compatibility between inventory- and lexicon-based systems of phonetic contrast with regard to the relative weight assigned to acoustic cues in speech perception. Two main classes of models are introduced in Sections 4.3 and 4.4. The first examines cue integration under the assumption of ideal recognition, and thus presents cue weights that represent the information available to listeners in the signal. The second models cue integration in the prediction of actual listener recognition behavior, and thus captures the relative weight assigned to each cue dimension in perception. Within each class of models three models are compared which vary in their degree of dependence on the lexicon: an *inventory model* (controlled syllable acoustics, balanced contrast representation), a *weighted inventory model* (syllable acoustics, contrasts representative of lexical distribution), and a *lexicon model* (lexical acoustics and contrast distributions). Finally, in Section 4.5 we present results Experiment 2, which serves as a verification that the cue-integration models are causally valid by testing whether model-predicted changes in recognition accuracy due to cross-splicing are reflected in listener responses.

4.1 Introduction

4.2 Modeling methodology
The methodology for modeling cue integration in the prediction of both ideal discrimination and listener behavior-based contrast recognition is outlined.

4.3 Cue integration in ideal recognition
Statistical models integrating the acoustic cues from Chapter 2 in the discrimination of contrasts in the lexicon and inventory are presented. Because the model aims for maximal separation of contrasts it provides estimates of the relative contribution of each cue to the information that is available to the listener in the acoustic signal.

4.4 Cue integration in listener recognition
Statistical models integrating acoustic cues from Chapter 2 in the prediction of listener recognition behavior are presented. These models complement the models in Section 4.3 by tracking listeners' actual signal-parsing behavior, both points of success and failure in word recognition.

4.5 Experiment 2: Cross-splicing validation
A subset of items from Experiment 1 are cross-spliced such that their predicted recognition probability from the cue-integration models in Section 4.4 either increases or decreases. This experiment serves as a test of the causal validity of such models in speech perception.

4.6 General discussion

4.1 Introduction

The analysis of obstruent acoustics and perception in Chapters 2 and 3 provides two critical components of the project of linking the acoustic-phonetic system to the higher-order structure of form distinctions in the lexicon that enables the encoding of meaning in speech communication. The present chapter provides the first direct link between the two in the form of statistical models of cue integration, where the primary focus is on the utility of each cue in the lexicon, and the degree to which such cue weights are predictable from an independent inventory of contrasts, which largely focuses on the acoustic properties of controlled syllables balanced in the distribution of both obstruent contrasts and vowel contexts. The latter is further subdivided into a model retaining this balance—the inventory model—and a *weighted* inventory wherein both contrast and vowel-context distributions are scaled to match those in the lexicon. From patterns of agreement and disagreement between the three models we are able to describe both the general scalability of acoustic cues between different architectures of the phonetic system—i.e., between one that assumes an independent inventory of speech sounds and one that is embedded in the ensemble of form distinctions in the lexicon—and the source of discrepancies where they arise; e.g., is a cue upweighted or downweighted in the inventory relative to its role in the lexicon because of fundamental differences in the acoustics of the two data sets, or because of differences in contrast distributions that fail to emphasize the characteristics of the contrasts which play the greatest role in the lexicon?

Each such cue-weighting analysis is performed for two main classes of model: one that examines cue integration under ideal recognition (Section 4.3), seeking to find the cue structure that maximally distinguishes contrastive pairs from non-contrastive pairs (i.e., detects the presence/absence of contrast); and one that is designed to track patterns in listener recognition (Section 4.4), and thus aims to uncover the relative weight assigned to acoustic cues in speech perception. Both models shed light on the multivariate acoustic structure of the obstruent system, and by tracking for each model the mechanisms by which that structure scales between inventory and lexicon we are able to provide an empirical assessment of the impact of the inventory assumption on our understanding of the acoustic encoding of speech, and the degree to which the shift toward a more

lexicon-centric approach, as this thesis advocates, is necessary.

Finally, in Section 4.5 results of Experiment 2 are presented, which replicates the 2AFC design in Experiment 1, but with manipulated stimuli that serve to provide causal evidence of listener sensitivity to different acoustic cues, as well as to validate in general the closed-class model predictions. Experiment 1, in addition to providing data on category and contrast recognition patterns, is used in Section 4.4 to build a statistical model of listener behavior based on acoustic and lexical cues. From this model we obtain the relative weight of each acoustic parameter based on the overall role that parameter plays in predicting listener recognition behavior. Yet these results are correlational and model-dependent. In Experiment 2, we elicit causal evidence for particular cues by replacing the target consonant and adjacent vowel (i.e., the target diphone) with that from another item in the database whose values along a particular acoustic dimension are either enhanced (more distinct from values associated with the other member of the obstruent contrast) or reduced (less distinct from the competitor obstruent).[1] For example, the word *sip* might be enhanced in spectral peak frequency relative to *ship* by cross-splicing onto *sip* an [s] from another word whose peak frequency is higher and therefore more distinct from [ʃ]. The extent to which listeners are more accurate on *enhanced* stimuli than *reduced* stimuli will serve as evidence for (a) the causal role of different acoustic cues in lexical contrast perception, and (b) the validity of the statistical model from which such predicted enhancements and reductions were derived.

4.2 Modeling methodology

In Sections 4.3 and 4.4 we use the acoustic parameter set described in Chapter 2 to predict the presence/absence of a contrast in the former, and the relative discriminability of contrasts by listeners in the latter, referred to hereafter as the *ideal perceiver model* and *listener model*, respectively. In the ideal perceiver model, contrast presence, **y**, is a dichotomous variable (0 = same, 1 = different), and each acoustic cue in the predictor matrix **X** is included as the absolute difference between the cue value in one member of the contrast and the cue value in the other; i.e., for the contrast between

[1] To be clear, diphones are cross-spliced to preserve consonant-vowel transition properties.

bit and *pit*, $\Delta\text{VOT} = |\text{VOT}_{bit} - \text{VOT}_{pit}|$. These are the same Δ-parameters used to report contrast effects in Figures 2.2–2.90.

In the *listener model* this set of contrast parameters is supplemented with the corresponding raw parameters from the target word in each minimal pair—referred to hereafter as the *target* parameters—as well as the absolute frequency of the target word and the frequency of the target relative to the competitor. Word frequency is not included in the parameter set for the inventory model, as inventory-based cue weights are assumed to be independent of higher-order information such as that from word frequency. The response $\mathbf{y}$ in the listener model is the probit-transformed listener accuracy on each item, where accuracies in the lexicon and weighted inventory models are drawn from the results of the 2AFC task in Experiment 1, and accuracies for the inventory data are drawn from published data in two prior studies of obstruent recognition in controlled syllables: Woods et al. (2010) and Cooke & Scharenborg (2008). Finally, because the latter two experiments employed an n-alternative forced choice design, where n is the number of target consonants examined in the study, responses were converted to pairwise contrast accuracies via the Luce choice rule (Luce, 1959). Here we must emphasize that the relationship between contrast accuracies obtained in this manner and those obtained directly via a 2AFC task is not perfect. Nevertheless, in the absence of 2AFC experiments on a comprehensive set of obstruent-contrastive syllables, this approximation is necessary until such a study is run in future work on this project.

The statistical model employed in mapping the predictor set $\mathbf{X}$ onto the outcome $\mathbf{y}$ is a Bayesian Additive Regression Tree (BART) with the following general form:

$$\mathbf{y} = f\left(\sum_{i=1}^{k} \mathcal{T}_i^{\mathcal{M}}(\mathbf{X})\right) + \varepsilon \ , \ \varepsilon \sim N(0, \sigma^2)$$

where $\mathcal{T}_i^{\mathcal{M}}(\mathbf{X})$ is the ith decision tree in a $k = 200$-tree ensemble, $\mathcal{M}$ is the set of terminal node parameters, $\mathbf{X}$ is the $n \times p$ matrix of predictor variables, $\mathbf{y}$ is the $n \times 1$ vector of responses, and f is a linking function. In the *ideal perceiver model*, where the outcome is a dichotomous variable, a probit link is used; i.e., $f = \Phi$, the Gaussian cumulative distribution function. In the *listener model*, the identity function is used, as $\mathbf{y}$ is a continuous variable in that model. Finally, the prior

on the error variance in the listener model is set at $\sigma = 0.5/2\sqrt{k}$, while $\sigma = 3/2\sqrt{k}$ in the ideal perceiver model. Both values are based on the recommendations in Chipman et al. (2010), which were further tested via cross-validation.

Each tree $\mathcal{T}_i$ is composed of a set of splitting rules formulated as $x_j < c$, where x_j is a variable in $\mathbf{X}$ and c is a splitting value. For example, a splitting rule could be the following: $\Delta\text{VOT} < 20$, where observations meeting this inequality might be predicted to be recognized less accurately (lower probability that $y = 1$) than those where the VOT difference between target and competitor is greater than 20 ms. Each observation descends through a tree according to these splitting rules until it arrives at a terminal node where a partial predicted value is assigned. Final predicted values are obtained by summing these terminal node predictions across all trees in the ensemble. By summing across trees, bias from any given decision tree is reduced, resulting in a model that is both more reliable in generating out-of-sample predictions, and whose parameter estimates are more robust and generalizable beyond the training data. Finally, as a Bayesian model the predicted response for a given observation is not a single point value but a posterior distribution which we can use to obtain both central estimates of the prediction, via posterior means or medians, and confidence intervals from upper and lower quantiles of the posterior distribution (e.g., 0.975 and 0.025 for a 95% confidence interval, 0.75 and 0.25 for an inter-quartile/50% range). All models were fit using the `bartMachine` package in R (Kapelner & Bleich, 2016).

The BART model was chosen for two reasons. First, as a tree-based model it is relatively less sensitive to multicollinearity in the predictors than are linear models such as logistic regression and linear discriminant analysis. Second, the ensemble learning method used in BART, where predictions are derived from the aggregation of several simpler models, reduces bias from any one model in a manner similar to the Random Forest algorithm, but with fewer problems with overfitting given the Bayesian method of model fitting based on weakly informative priors.

Finally, separate models were fit to CV, VCV, and VC contrasts, where the parameters included in each model are only those which are defined for that position; i.e., all of the parameters described above can be measured for intervocalic contrasts, but pre-consonantal parameters (in-

cluding parameters like consonant duration, which is equivalent to noise duration in voiceless obstruents in CV position) are undefined for word-initial contrasts, and post-consonantal parameters are undefined for word-final contrasts. From each model, cue weights were determined as follows. First, the partial dependence of $\mathbf{y}$ on each cue x_j was estimated, which is computed by measuring, for a given fixed value of x_j, the average (probit-transformed) value of $\hat{\mathbf{y}}$ as $\mathbf{X}_{-j}$ varies over its marginal distribution. By iterating this measurement over a range of x_j values, such as the quantiles $(0.05,\ 0.15,\ ...,\ 0.95)$, we can measure the range of variation in $\hat{\mathbf{y}}$ due to x_j while controlling for effects of other variables on the response. We then derive an aggregate cue weight by averaging the piecewise change in $\hat{\mathbf{y}}$ quantile-to-quantile. This measure reduces to the linear slope of the partial dependence function in cases where x and y are linearly related, but can also account for nonlinearities in cases where the variation in y is restricted to a narrower range of x. Finally, in the case of target cue weights—i.e., the raw parameter values in the target, not the Δ-parameters comparing both target and competitor—where the partial dependence function is not expected to increase monotonically, the absolute value of each piecewise slope is taken prior to averaging.

From each cue weight we then measure the rank of the cue based on the median of the posterior of the partial dependence function, with variance in cue weights accounted for by measuring cue weights at the 25th and 75th percentiles of the posterior and noting any rank shifts due to overlap in the cue weight distributions. The reason we focus on ranks and not the raw cue weights is that it is difficult to compare weights across models, and further, given that weights vary based on both the experimental and statistical methodology, the literature on cue-weighting primarily focuses on the *relative* importance of a given cue rather than its *absolute* utility in the acoustics or perception. Next we review the three models—the *inventory model*, the *weighted inventory model*, and the *lexicon model*—that are used to relate systems of contrast in the inventory and lexicon, as well as to determine, where cue-weight discrepancies arise, what the primary cause of the discrepancy is.

Inventory model. In what we refer to as the *inventory model*, the input matrix, $\mathbf{X}$, is the set of acoustic parameters from controlled syllable productions from the target speaker (referred to else-

where as the *inventory database*), which includes only contrast parameters in the ideal perceiver model, and both target and contrast parameters in the listener model. The response, **y**, in the ideal perceiver model is the contrast presence/absence (0/1) vector where contrasts were created by systematically pairing each obstruent in the 17-phone set (18 in VCV), controlling for position and vowel context, and equivalence relations (within-category comparisons) were created by pairing the two repetitions of the same item. Thus, the within-category pairs represent the same distinctions that were used to illustrate approximate chance ranges for contrast effects in the individual parameter results above. Finally, the within-category ($y = 0$) set was oversampled to balance the response variable for model estimation purposes. In the listener model, **y** is the probit-transformed listener accuracy drawn from the CaST database (Woods et al., 2010) for CV and VC contrasts, and from the CC08 database (Cooke & Scharenborg, 2008) for VCV contrasts. Further, models of the CaST and CC08 responses based on the acoustics of stimuli used in each experiment are provided for reference to ensure that the inventory model, which is based on the acoustics of the target speaker, adequately fits the perception data.

Weighted inventory model. In attempting to link the acoustic properties of English obstruents to the system of contrasts in the lexicon, one possible alternative to the direct study of acoustic distinctions between words is to use a *weighted inventory* approach. That is, much like in the traditional methodological division between phonetics and phonology, in the weighted inventory model the fundamental acoustic properties of English obstruents are held to be independent of their phonological distribution in the lexicon. With this assumption, however, we can still study how the lexicon might warp cue weights based on which information is more likely to play a role in listeners' perception of distinctions between real words. This is done by sampling the acoustic measurements from controlled syllables—i.e., drawing observations from the input matrix **X** in the inventory model—such that the proportion of occurrence of each obstruent contrast in a given position and vowel context matches that in the lexicon.

In the ideal perceiver model, the lexicon is modeled via a subset of the MALD database that

includes the most frequent words comprising 95% of the tokens in each of several corpora—COCA (written and spoken; Davies, 2009), SUBTLEX-US (Brysbaert et al., 2012), and Google Unigram (Michel et al., 2011). This criterion reduces the 26,793-word MALD database—16,021 words, excluding lexical hermits (words with no minimal-pair neighbors)—to a core subset of 3,406 words participating in 11,972 minimal-pair contrasts.[2] Within this set 2,501 minimal pairs (from 1,649 words) involve contrasts between obstruents. We will refer to this subset of the MALD database, which is used in the lexicon model below and in the perturbation analysis in Chapter 5, as the Lex95 set. In the listener model the set of 960 minimal pairs presented in Experiment 1 is used.

From the minimal pairs in the Lex95 data, obstruent contrasts in CV, VCV, and VC positions in words of up to three syllables in length were identified (matching the stimulus criteria in Experiments 1 and 2 in Chapter 3) and used as the basis for both contrastive ($y = 1$) and within-category ($y = 0$) items, where the latter were drawn from CV, VCV, and VC portions of the non-contrastive interval of each minimal pair; e.g., in the pair *sip–sit*, the word-offset distinction between [p] and [t] forms the contrast, and the word-onset [s, s] pair forms the within-category item. In the listener model no within-category pairs are used, and thus the Experiment 1 set of contrasts forms the complete data set. Thus, the only difference between the inventory model and the weighted inventory model is in the relative frequency of each contrast.

Lexicon model. Finally, the *lexicon model*, which is the focus of the present study, aims to predict the presence/absence of contrast in the Lex95 set in the ideal perceiver model, and the recognition probability of minimal-pair contrasts by listeners in the listener model, directly from the acoustic characteristics of the words participating in such contrasts. That is, the lexicon and weighted inventory models are identical in their contrast disributions, but differ in the source of the acoustic data, the former coming from real words and the latter from controlled syllables. This distinction, however, has far greater theoretical and methodological implications than a simple comparison of word and nonword production characteristics. Since the acoustic measurements for the lexicon model are made from the minimal pairs themselves, and the outcome predicted by the

[2]By comparison, there are 40,312 minimal pairs in the full MALD database.

model is based on contrasts rather than specific obstruent categories, the lexicon model operates on a relaxed assumption regarding the phonological composition of each word, as it only knows that two intervals in a word pair are contrastive (i.e., are acoustically distinct in a linguistically meaningful way) and that they fit within the broad featural class of [−sonorant]. By comparison, the weighted inventory model must still assume (1) a precise set of obstruent phones and vowel contexts that are manipulated in the controlled syllables that comprise the inventory database, and (2) a mapping of phone sequences in the inventory onto phonological forms in the lexicon.

4.3 Cue integration in ideal recognition

Given the results for individual parameters in Chapter 2, the next question to address is: *what is the relative weight that should be assigned to each cue when information from all cues is integrated in a general discrimination problem?* In this section we examine the integration of cues in a statistical model of contrast discrimination that aims to approximate the information available to the listener under ideal recognition conditions. Later, in Section 4.4, we will explore a similar model where contrast discrimination is not assumed to be perfect, but rather matches listener behavior. Finally, we should emphasize that by *ideal recognition* we mean the acoustic information that is available in the signal, not accounting for auditory or linguistic factors that may impact the parsing of that information. For the latter we would require information on perception in 'clean' listening conditions (i.e., without background noise) that is gradient enough to detect listener uncertainty prior to a correct decision. Methods using eye-tracking and pupillometry are the most promising for acquiring such data, and will be explored in future research.

Below we review the relative weights of acoustic cues in the *lexicon*, *inventory*, and *weighted inventory* models of word-initial, word-medial, and word-final contrasts. In each analysis, we examine several cues in greater depth that represent key points of similarity and difference between the three models, where for the latter there are three major kinds of disagreement investigated, termed *distributional*, *acoustic*, and *composite*.

Distributional effects correspond to cue weight discrepancies which arise due to differences

in the distribution of contrasts in the lexicon and inventory, where the lexicon model is expected to place greater weight on the cues distinguishing frequent contrasts and lesser weight on cues whose utility is primarily restricted to contrasts that are more sparse in the lexicon. Acoustic effects arise where the acoustic characteristics of real words diverge from controlled syllables to a large enough degree that estimates derived from the latter do not scale to the former and therefore are problematic as a means of understanding the acoustic structure of the lexicon and the general uptake of acoustic information in speech communication. Finally, composite effects arise in cases where the both distributional and acoustic sources of disagreement appear to be involved.

The weighted inventory model is used as a means of detecting both distributional and acoustic effects, as cases of good agreement between the weighted inventory model and the lexicon model, but otherwise poor agreement between the lexicon and inventory, suggest that the distortion of contrast distributions in the balanced inventory is the primary cause of the cue-weight discrepancy. On the other hand, if the cue ranking in the weighted inventory model diverges notably from that obtained in the inventory and lexicon, this result suggests that the latter two models are achieving similar cue weights via different means. That is, given that the acoustic characteristics of each obstruent contrast are identical between the two inventory models, if the acoustics of the inventory and lexicon were similar then a warping of contrast distributions to match that in the lexicon should result in *greater*, not lesser agreement with the lexicon model.

Finally, if the lexicon model disagrees with both inventory and weighted inventory models, this result indicates that cue weight discrepancies between the lexicon and inventory cannot be resolved merely by retuning the model to attend to contrasts in a manner consistent with their distribution in the lexicon. Rather, a composite effect of both distributional and acoustic differences between the two models must be responsible for such a result. All three effects—distributional, acoustic, and composite—as well as cases of agreement between the three models, will be discussed below in separate subsections within each position-specific analysis.

	Predicted					
	Inventory		W. Inv.		Lexicon	
Observed	0	1	0	1	0	1
0	50	0	32	2	33	2
1	2	48	0	66	3	62
Accuracy:		98%		98%		95%
Precision:		1		0.97		0.97
Recall:		0.96		1		0.95
$F1$ Score:		0.98		0.98		0.96

Table 4.1: Confusion matrices and model fit statistics for inventory, weighted inventory, and lexicon models of word-initial contrast presence/absence (0/1).

4.3.1 Word-initial position (CV)

Confusion matrices and fit statistics for inventory, weighted inventory, and lexicon models of word-initial contrasts are presented in Table 4.1. Overall, each model was highly accurate and exhibited minimal bias (lower $F1$ scores), neither toward false positives (lower *precision*) nor false negatives (lower *recall*). Thus, in the analysis of cue weights below we can assume that the predictive power of each parameter in the model is a reasonable approximation to the general utility of that parameter in the discrimination of contrasts in ideal recognition.

Figure 4.1 shows cue rankings in the lexicon, inventory, and weighted inventory models of word-initial obstruent contrasts. Parameters are ordered in descending rank in the lexicon, with inventory and weighted inventory ranks matched for comparison. Ranks were derived from the median cue weight following the analysis of partial dependence functions described above. Error bars indicate the range of fluctuation in cue ranks based on overlap between the IQRs of each cue weight. Parameters with negative cue weights indicative of a correlative relation with the outcome that is inconsistent with the use of that parameter as a cue to contrast,[3] are shown as translucent bars in Figure 4.1, and should not be compared internally given the uncertainty in their interpretation.

In the lexicon model, among the more influential cues are vowel-onset F2 ($F2_{CV}$), release

[3]Such weights suggest that as two signals become more similar along a certain dimension they are more likely to be distinguished. This is counterintuitive and rather reflects the presence of a confounding variable in the relation.

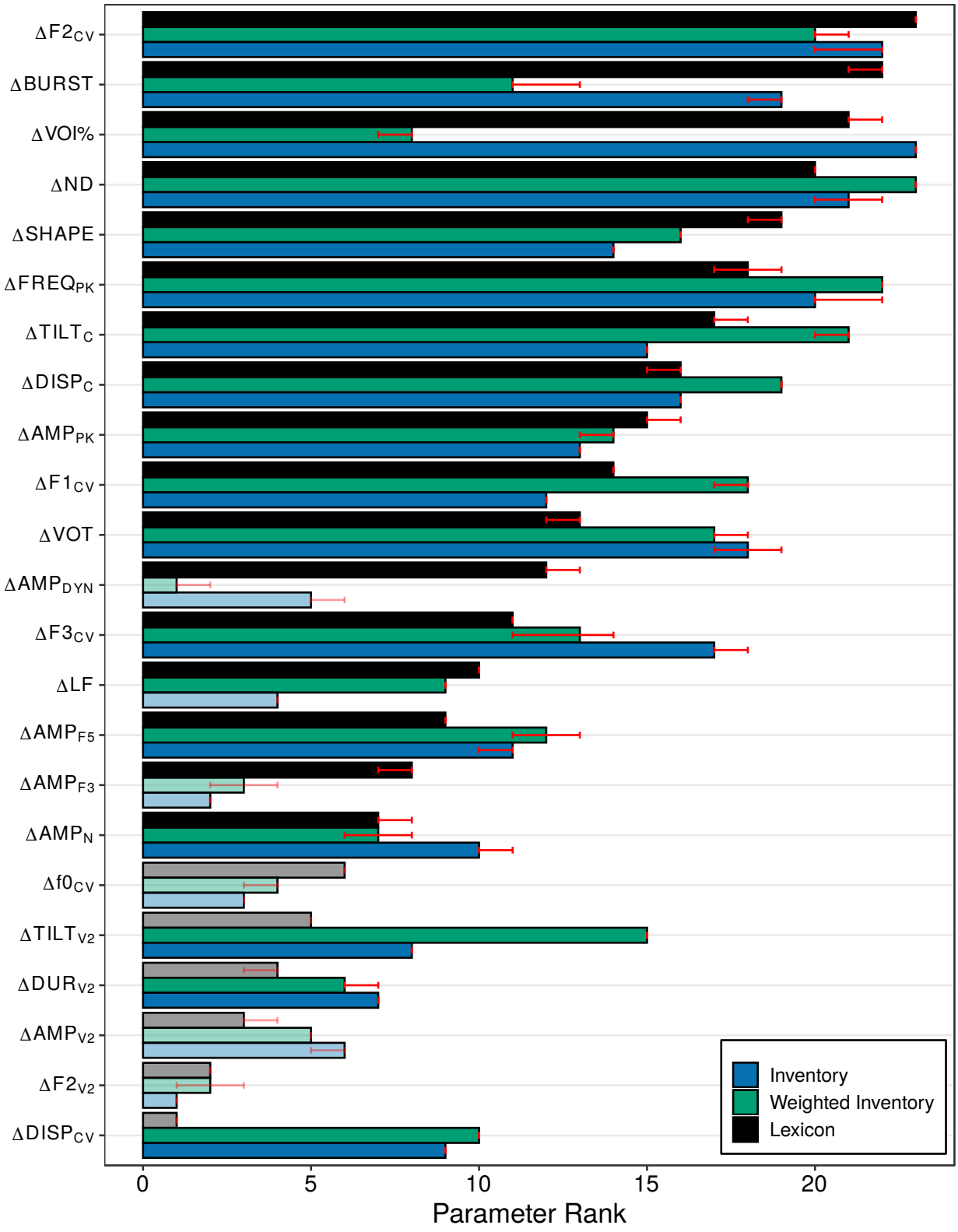

Figure 4.1: Acoustic parameter ranks in lexicon, inventory, and weighted inventory models of word-initial contrasts under the assumption of ideal recognition. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

burst presence (BURST), consonant voicing percentage (VOI%), noise duration (ND), and spectral shape (SHAPE). With the exception of spectral shape, which primarily isolates postalveolars from other obstruents, all of these parameters serve in a diverse array of contrasts, as $F2_{CV}$ indexes place and to a lesser extent voicing and manner, BURST and ND index manner of articulation, and VOI% indexes voicing and manner. Conversely, many of the low-weighted cues capture vowel characteristics that were shown to be independently less discriminative in the preceding sections.

The inventory model largely agrees with the lexicon in both higher- and lower-ranked cue sets, showing high ranks for $F2_{CV}$, BURST, VOI%, and ND, and relatively low ranks for the vowel parameters $TILT_{V2}$, DUR_{V2}, AMP_{V2}, and $F2_{V2}$, though spectral dispersion at vowel onset is of much greater utility in the inventory than in the lexicon. The weighted inventory model is similar in this regard, though with a much higher rank for $TILT_{V2}$, while among the highly ranked cues in the lexicon, BURST and VOI% are of comparatively less utility in the weighted inventory than in the inventory or lexicon. Overall, the highest ranked cues in the weighted inventory are ND, $FREQ_{PK}$, $TILT_{C}$, $F2_{CV}$, and $DISP_{C}$, while in the inventory the set {VOI%, $F2_{CV}$, ND, $FREQ_{PK}$, BURST} exhibits the highest ranking.

Figures 4.2 and 4.3 provide a more direct display of these model relations in the form of correlations between cue ranks (equivalent to Spearman's rank correlation between cue weights) and the distribution of cues according to rank differences between the lexicon model and the inventory models. From Figure 4.2 we see that the cue ranks in both inventory models are highly correlated with those in the lexicon, though the correlation is higher between the inventory and lexicon, as parameters such as AMP_{DYN}, VOI%, BURST, $TILT_{V2}$, and $DISP_{CV}$ are quite distinct in the weighted inventory relative to the lexicon. Comparing the cue-weight discrepancies in Figure 4.3 we see that cues in the lower left quadrant, particularly those such as AMP_{DYN} that are far from both dashed axes (indicating equivalence rank equivalence between a pair of models), are notably underestimated in the inventory models relative to their utility in the lexicon, while cues in the upper right quadrant, such as $DISP_{CV}$, are overestimated in this regard. Cues that lie far from the origin along the *x*-axis, such as low-frequency energy (LF), indicate potential distributional sources

of disagreement in cue weights between the inventory and lexicon, while those lying far out along the *y*-axis, such as VOI%, indicate potential acoustic discrepancies. Finally, cues toward the center of the figure, such as $F2_{CV/V2}$ and AMP_{PK}, are quite consistent between all three models.

From these relations we have identified four cues to examine in greater detail—vowel-onset F2 ($F2_{CV}$), low-frequency energy (LF), consonant voicing percentage (VOI%), and dynamic amplitude (AMP_{DYN})—representing points of *agreement*, *distributional disagreement*, *acoustic disagreement*, and *composite disagreement*, respectively. We begin with $F2_{CV}$, a highly weighted cue that shows remarkable agreement between the three models.

4.3.1.1 Cue Agreement: Vowel-onset F2

Figure 4.4 shows the partial dependence functions for $F2_{CV}$ in the inventory, weighted inventory, and lexicon models, as well as the $F2_{CV}$ distributions in the data input to each model. All three models in Figure 4.4 exhibit substantial overlap in both the distribution of $\Delta F2_{CV}$ among items in their input data, and in the relationship between $\Delta F2_{CV}$ and the model-predicted likelihood of contrast presence. Results are further summarized in Figure 4.5, which shows the relationhip between phonetic contrast means along the $F2_{CV}$ dimension in the inventory and lexicon. From Figure 4.5 we see the source of the close agreement between the three models. Not only do both the inventory and lexicon exhibit a wide range of $\Delta F2_{CV}$ values in contrastive pairs, but the majority of this range lies above the 75th percentile of the within-category values in both data sets. Further, the mean $\Delta F2_{CV}$ values obtained for contrasts in the lexicon and inventory are highly correlated, particularly the contrasts which are the most frequent in the lexicon (shown with larger text sizes in the figure), meaning that when such contrasts receive greater emphasis in the weighted inventory model, the conformity in their acoustics allows for a similar fit.

Examining Figure 4.5 more closely, we find clear evidence of vowel-onset F2 indexing place of articulation, both within and across manner of articulation. At the upper end of the $\Delta F2_{CV}$ range are frequently occurring contrasts such as *b–d*, *d–f*, and *d–h*, along with several contrasts involving the voiceless sibilant fricatives [s, ʃ]. In fact, one of the most robust distinctions is that between

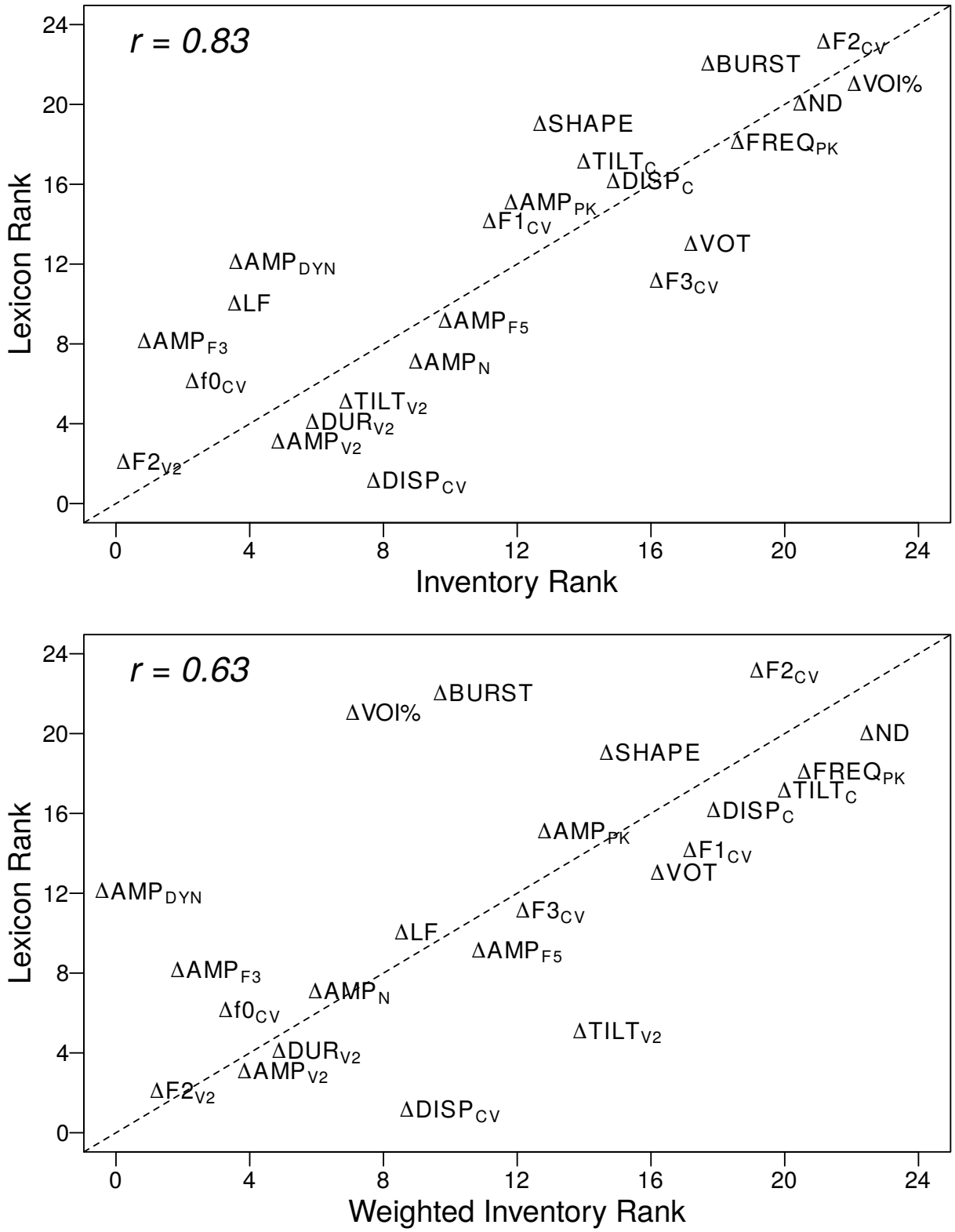

Figure 4.2: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in CV position. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

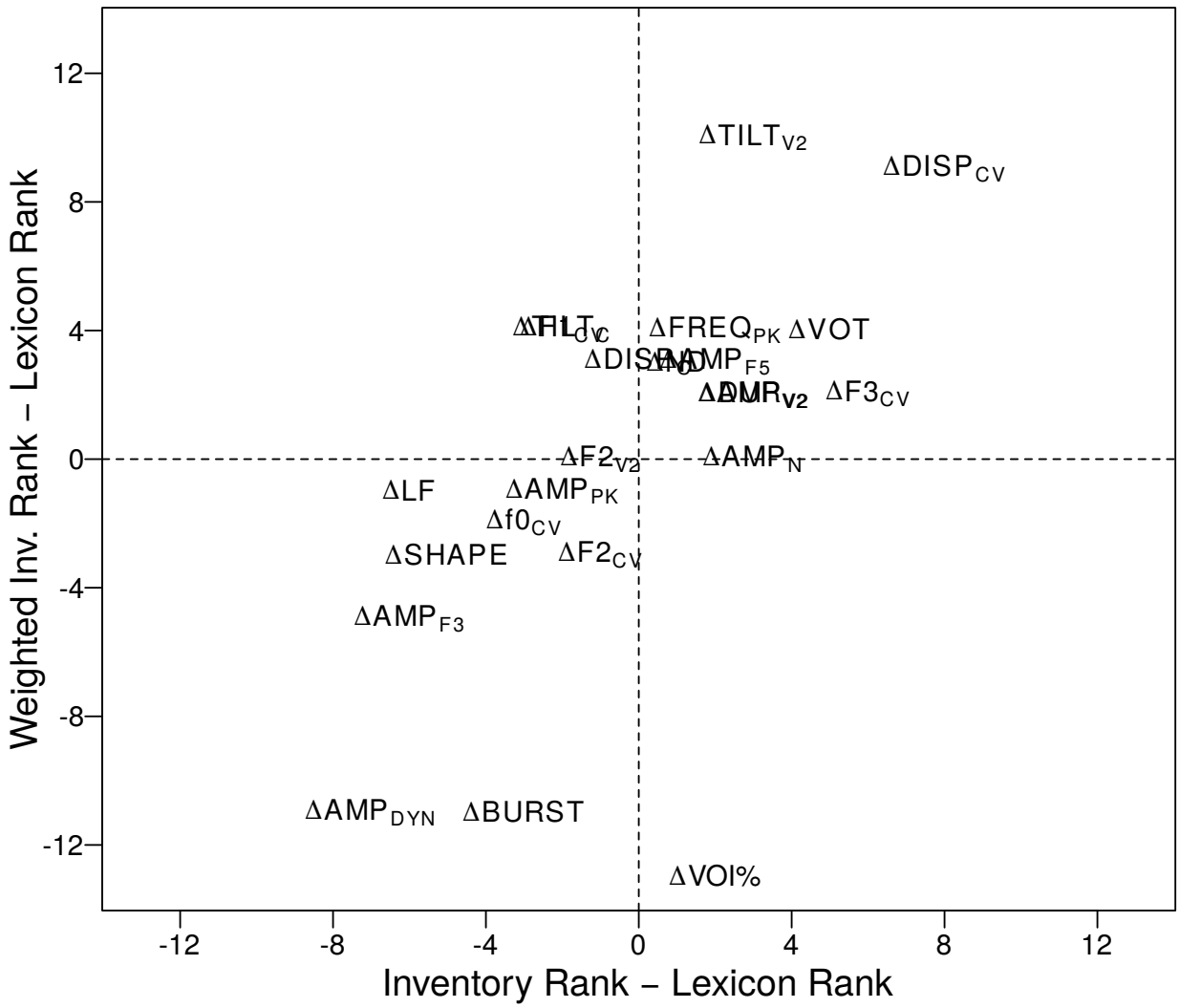

Figure 4.3: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in CV position. Dashed lines indicate equivalence relations between each pair of models.

postalveolar and non-postalveolar obstruents, as the set [ʃ, ʧ, ʤ, ʒ] exhibits consistent F2 onsets in the 1.5–2 kHz range that are well above the onset F2 frequencies of most other obstruents (see Figure 2.79 in Chapter 2 for details), though many of these contrasts are not labeled in Figure 4.5 because they are below the item frequency threshold adopted for visual clarity.

At the lower end of the $\Delta F2_{CV}$ range are contrasts such as *b–p*, *b–f*, and *f–p* that overlap largely with the within-category range due to their shared place of articulation (other voicing contrasts such as *k–g*, *t–d* are more distinct but are not labeled in Figure 4.5 because they are relatively less frequent). Place contrasts between voiceless obstruents (excepting the aforementioned contrasts

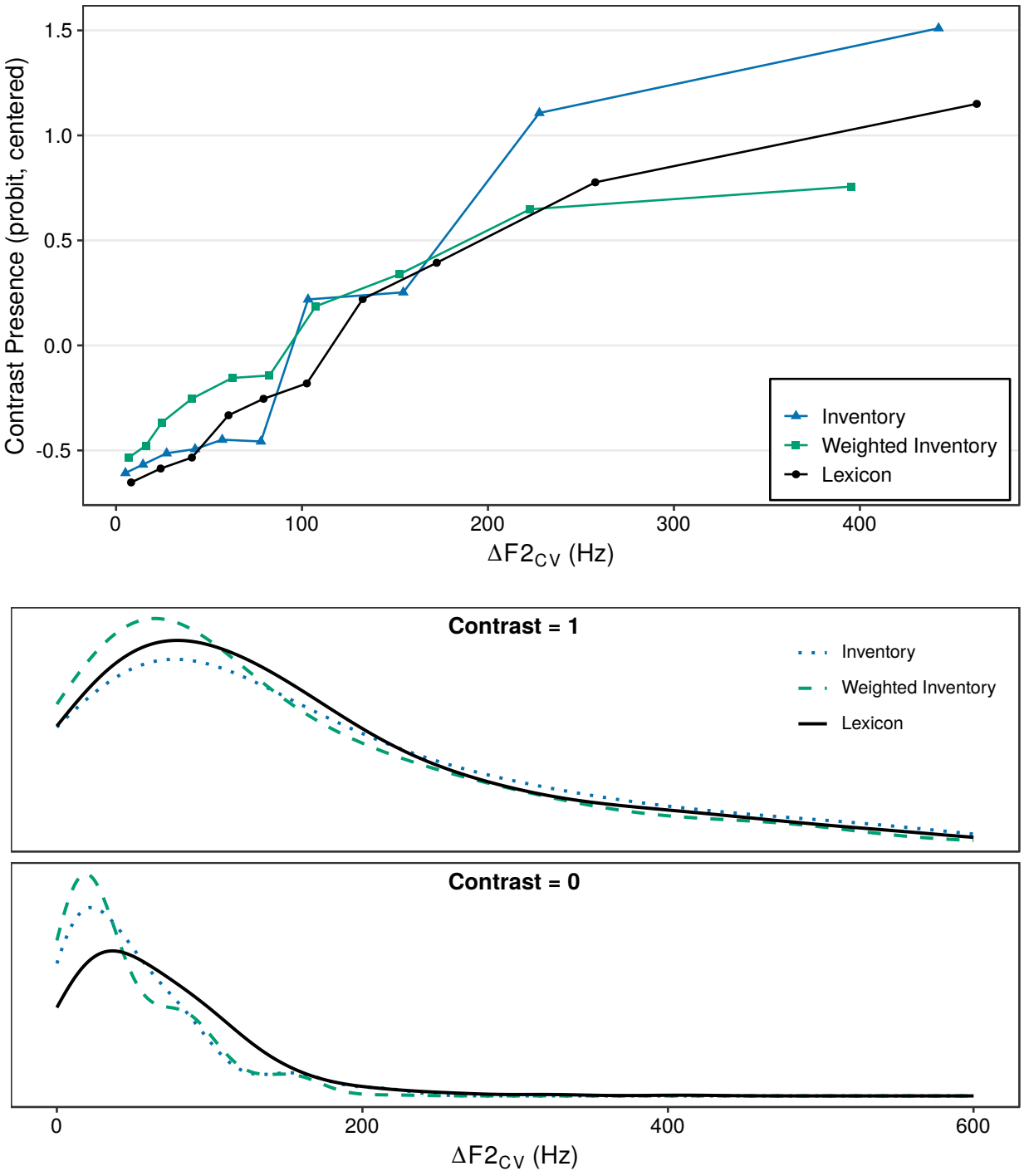

Figure 4.4: Partial dependence functions for $F2_{CV}$ (top panel) and $F2_{CV}$ distributions (bottom panels) in the inventory, weighted inventory, and lexicon models of ideal recognition in word-initial position.

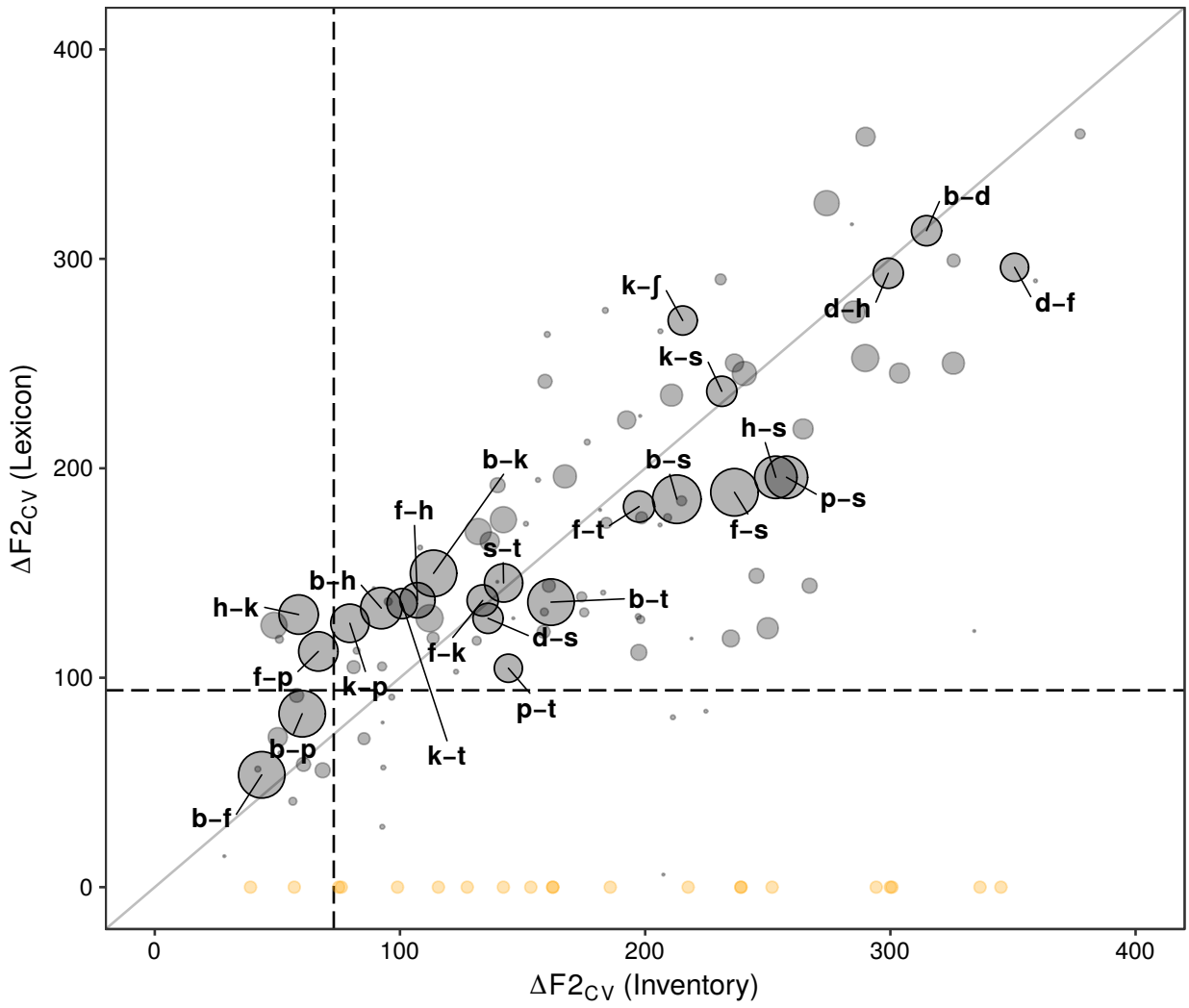

Figure 4.5: Relationship between $\Delta F2_{CV}$ means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-initial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 50% of items are labeled. Contrasts absent from the lexicon are shown in orange. Several infrequent contrasts at the upper end of the $\Delta F2_{CV}$ range have been excluded for clarity.

involving sibilants) are also relatively less distinct in F2 onsets compared to voiced contrasts or contrasts between voiced and voiceless obstruents (due to $F2_{CV}$ partially reflecting voicing differences). However, as noted earlier, the general picture from the distributions of $F2_{CV}$ in Figures 4.4 and 4.5 is that for the vast majority of contrasts the cross-category distinction in F2 onset is greater than the distinctions that arise due to within-category variability. For this reason $F2_{CV}$ is a robust predictor of contrastiveness in both the lexicon and inventory, as well as in the weighted inventory

model of lexical contrasts.

4.3.1.2 Distributional Disagreement: Low-frequency Energy

Overall, there are few cues in the ideal perceiver model of word-initial contrasts that show that by weighting the inventory contrasts according to their frequency in the lexicon greater conformity is brought about with the use of that cue in the lexicon model. Figure 4.3, for example, shows a largely vertical distribution of cue-rank differences, suggesting greater disagreement between the lexicon and weighted inventory models as compared with the balanced inventory model. The low-frequency energy in the consonant noise spectrum, however, is one cue that receives a relatively low weight in the inventory, but which is brought into closer agreement with the lexicon by accounting for the lexical distribution of obstruent contrasts. Figure 4.6 shows that this agreement is achieved largely by downweighting the role of relatively infrequent contrasts that show similar ΔLF values to the within-category items, and upweighting contrasts exhibiting greater LF distinctions.

From Figure 4.7 we see that cases of the latter include several contrasts with labial obstruents, such as *b–f*, *d–f*, *f–p*, *b–s*, *f–t*, *f–k*, and *p–s*, as well as contrasts between sibilant fricatives and plosives, such as *s–t*, *d–s*, *k–ʃ*, and *k–s*, all of which occur frequently in the lexicon and are well-above the within-category range in both the inventory and lexicon. The contrasts that overlap with the within-category range are dominated by contrasts between plosives, a result which is consistent with the fact that the word-initial plosive voicing distinction in English is one of the degree of aspiration rather than a true voicing distinction, and thus both sets exhibit similar low-frequency energy profiles. In summary, because of the reduction in true voicing distinctions among word-initial obstruents in English, LF primarily indexes manner distinctions, which due to their relatively high frequency in the lexicon allows LF to remain useful in distinguishing minimal pairs, a result that is diminished when contrasts are balanced out in the inventory.

Put another way, Figure 4.7 is similar to Figure 4.5 in showing a generally close correlation between the low-frequency energy distribution in controlled syllable contrasts and that in minimal-pair contrasts in the lexicon. However, unlike with $F2_{CV}$, there is greater overlap between the

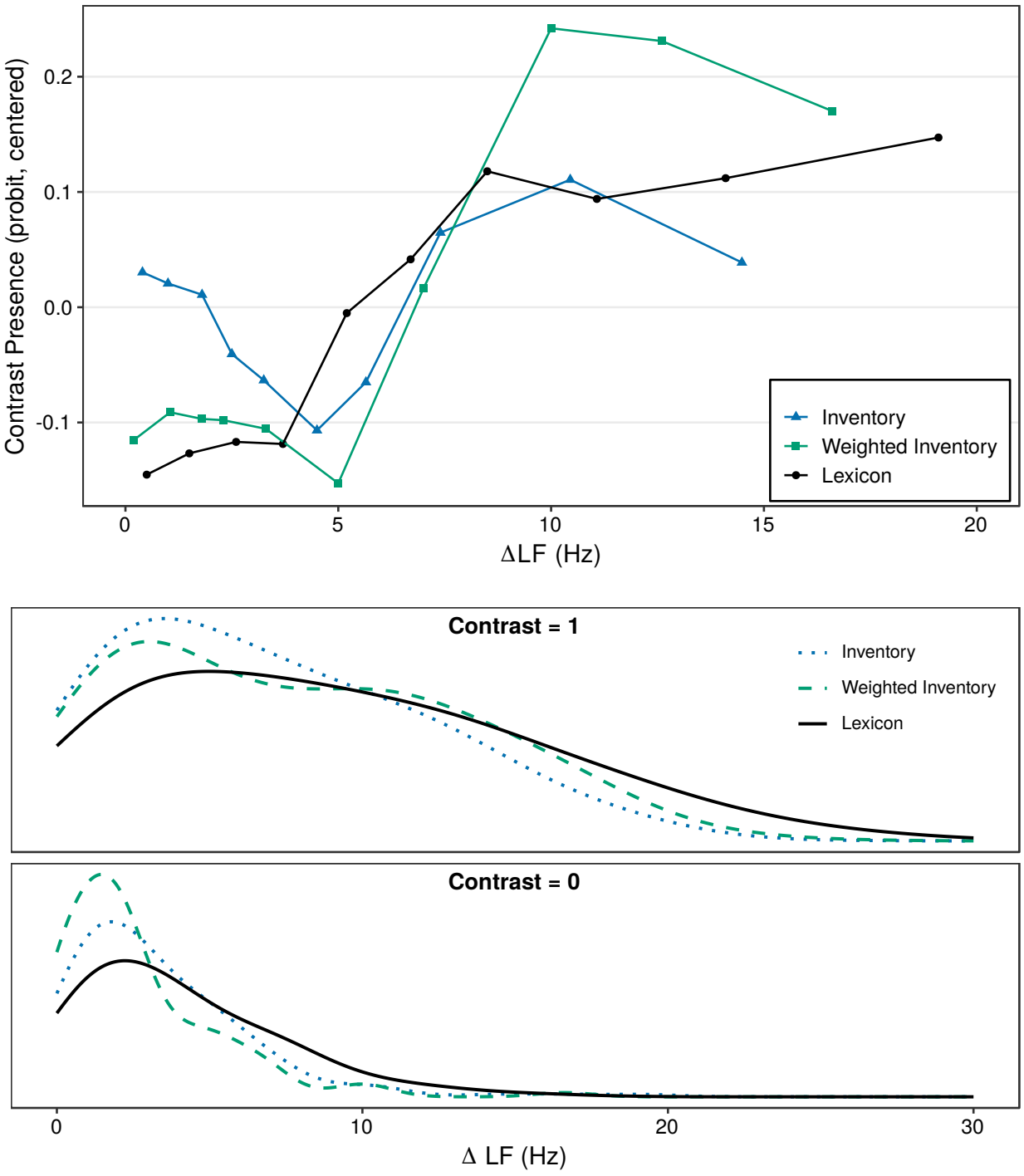

Figure 4.6: Partial dependence functions for LF (top panel) and LF distributions (bottom panels) in the inventory, weighted inventory, and lexicon models of ideal recognition in word-initial position.

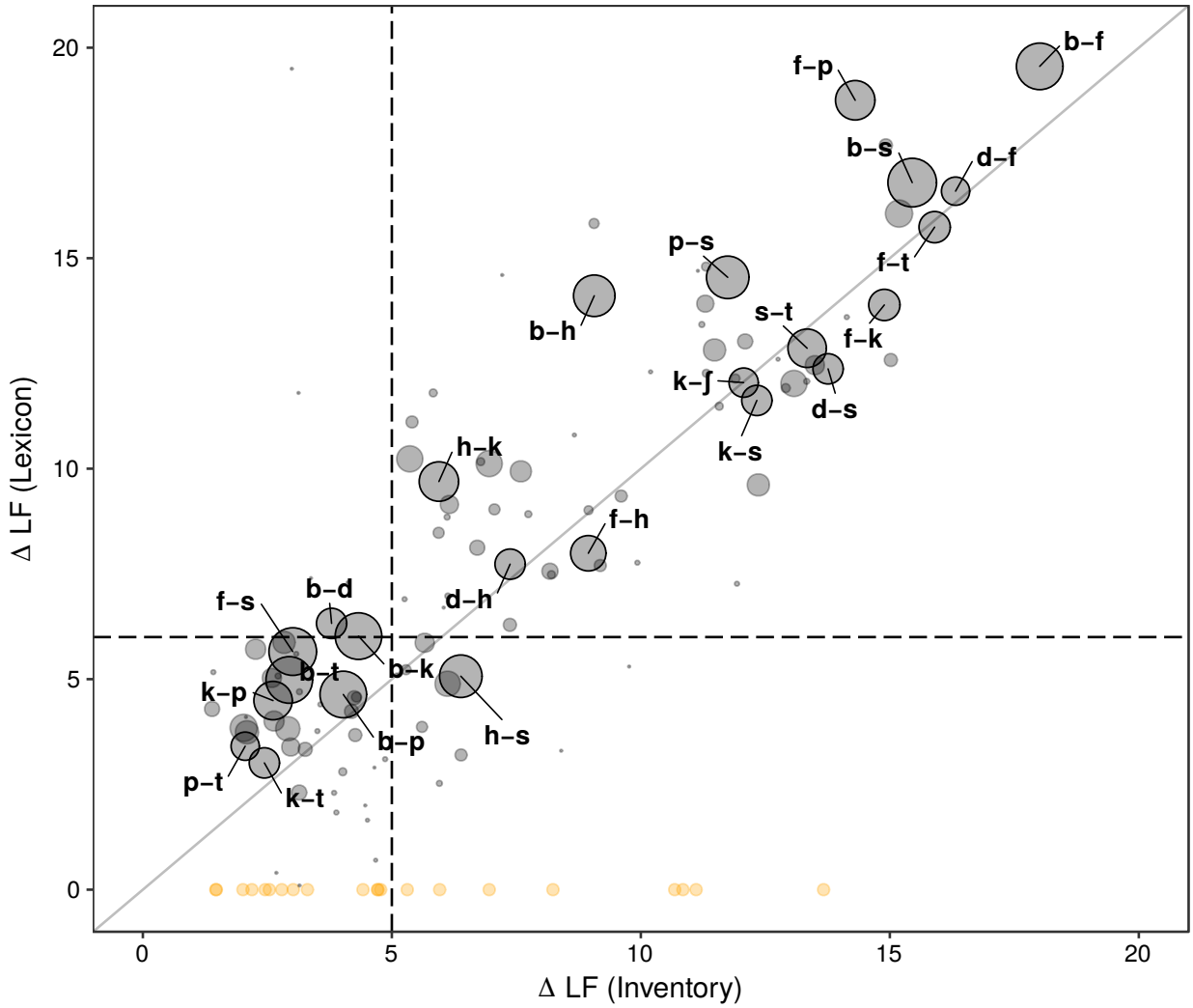

Figure 4.7: Relationship between ΔLF means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-initial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 50% of items are labeled. Contrasts absent from the lexicon are shown in orange.

contrastive and non-contrastive pairs in both sets, but given that most such contrasts are relatively infrequent in the lexicon this result does not lead to the spurious negative correlation in Figure 4.6 between ΔLF and contrast presence below 5 dB that is observed in the inventory where that same set plays a relatively larger role. This difference can also be seen in the distribution of points in the partial dependence function, where the inventory range below 5 dB accounts for 50% of the distribution, whereas in the lexicon only 30% of minimal pairs are less than 5 dB apart. This does not mean that the relative difference in low-frequency energy is not still useful in the inventory, just

that its utility is restricted to a narrower range of contrasts than those which occur in the lexicon.

4.3.1.3 Acoustic Disagreement: Consonant Voicing Percentage

Figure 4.8 shows a quite distinct pattern relative to that in Figures 4.4 and 4.6, as the distinction between contrastive and non-contrastive items occurs over a relatively narrow range of ΔVOI% values: generally below 25%. Further, Figure 4.9 shows that mean ΔVOI% values per contrast are notably less correlated. For instance, excluding the contrasts at the highest and lowest ends of the ΔVOI% dimensionı,e,, focusing on those primarily in the 10–50% range—there is no clear relation between the parameter values in the inventory and those in the lexicon. Nevertheless, within both sets there is a sizeable separation between the contrastive and non-contrastive ranges, leading ultimately to an aggregate cue weight that is the highest ranking in the inventory and third-highest in the lexicon.

This lack of acoustic agreement between the VOI% values measured from balanced contrasts in controlled syllables and those measured from minimal pairs results in a downweighting of ΔVOI% in the weighted inventory model because the acoustic measurements from one database scale poorly to the other and are generally less distinct in the inventory than in the lexicon. More precisely, the reason VOI% is downweighted in the weighted inventory model relative to the lexicon and inventory models is that many of the more frequent contrasts in the lexicon that lie outside of the within-category range, primarily manner/voicing distinctions involving the voiced plosives [b] and [d], exhibit lower voicing percentage differences in controlled syllable data than in the minimal-pair contrasts that comprise the lexicon (shown in Figure 4.9 in the location of these contrasts above the gray identity line, indicating greater distinction in the lexicon than in the inventory). This means that the result of sampling the inventory to be representative of the lexicon has the aggregate effect of reducing the distinction between contrastive and non-contrastive ΔVOI% values, shifting the distribution toward narrower VOI% distinctions (see the density plots in Figure 4.8) and thereby weakening the overall role of VOI% in the weighted inventory model.

Again, as in the case of the role of low-frequency energy in each model, this weakening does

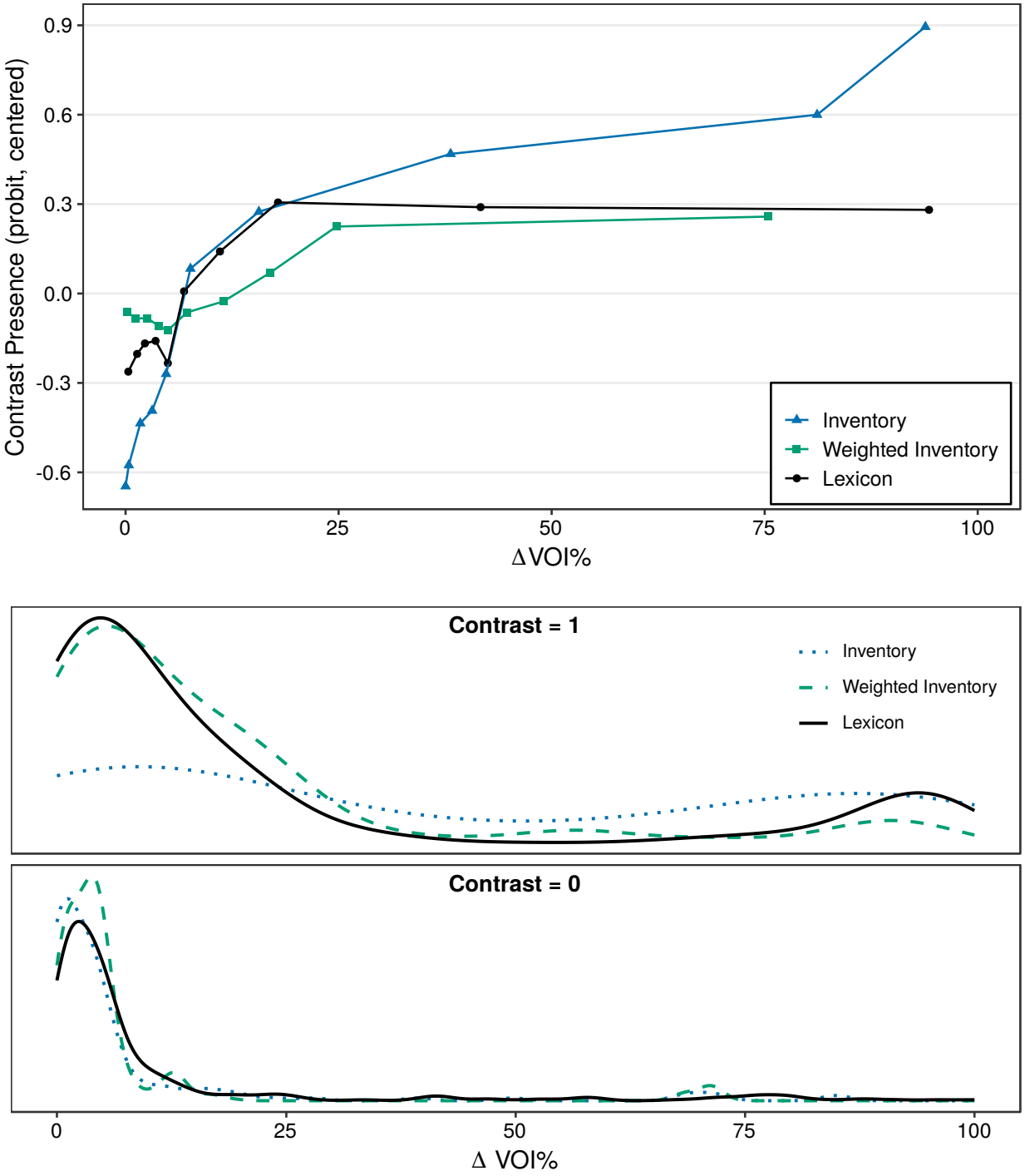

Figure 4.8: Partial dependence functions for VOI% (top panel) and VOI% distributions (bottom panels) in the inventory, weighted inventory, and lexicon models of ideal recognition in word-initial position.

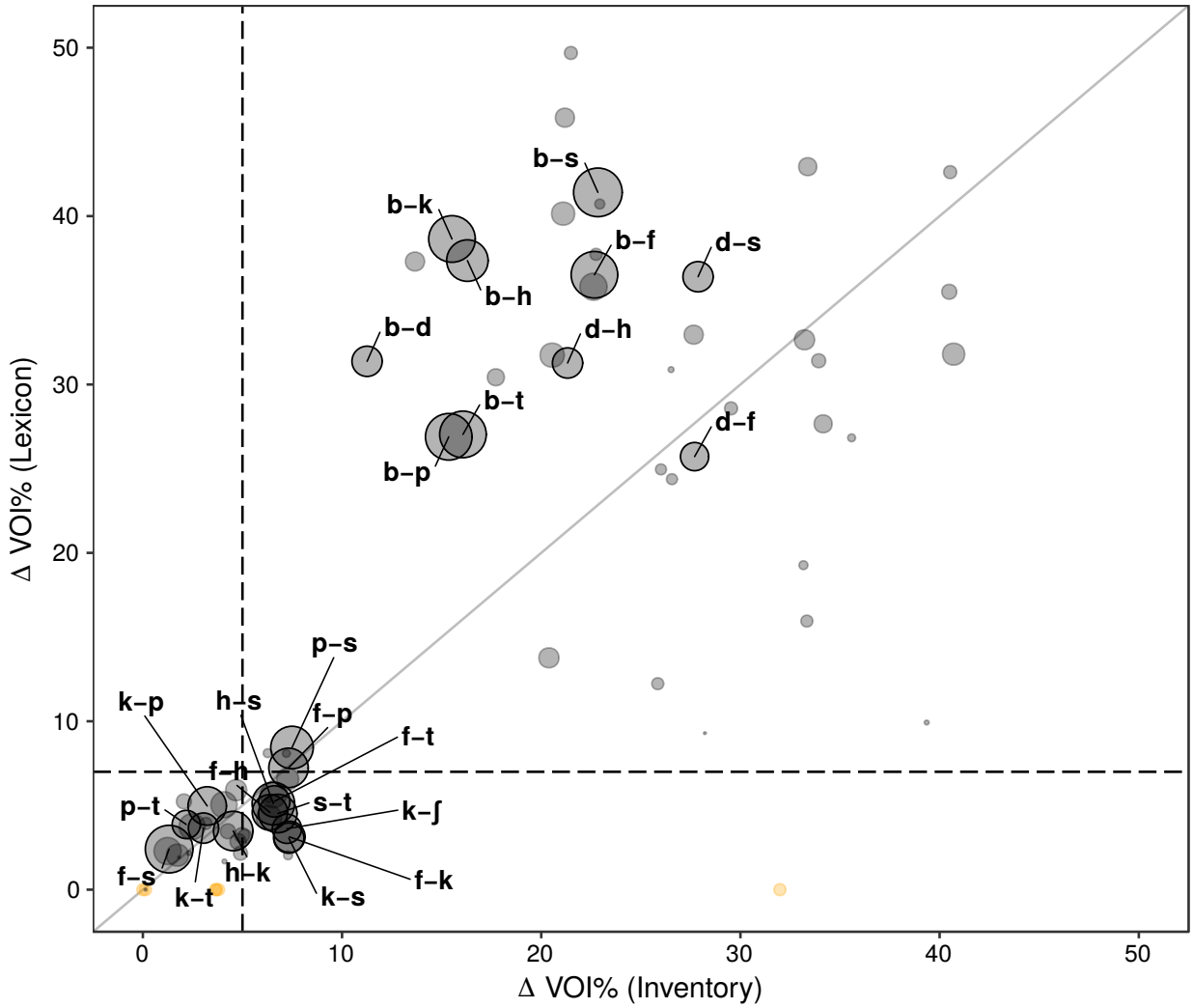

Figure 4.9: Relationship between ΔVOI% means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-initial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 50% of items are labeled. Contrasts absent from the lexicon are shown in orange. Several contrasts with ΔVOI% values above 50 were excluded to clarify distinctions among the majority of contrasts which lie below 50.

not remove the role of VOI% in the weighted inventory model, it simply shows that both inventory and lexical approaches can arrive at similar aggregate cue weights while being inconsistent in the way a cue is distributed among the different phonetic contrasts that comprise each system. Such discrepancies are of limited consequence in ideal perceiver models which do not differentiate among contrasts, only seeking to maximally separate contrastive from non-contrastive items, but in the listener models in Section 4.4 where contrasts differ in these discrepancies will pose a much

greater problem for scaling estimates between the inventory and lexicon, and could lead researchers working in the former paradigm to misidentify how a cue will operate in word recognition.

4.3.1.4 Composite Disagreement: Dynamic Amplitude

Finally, in AMP_{DYN} we have a case where both acoustic and distributional discrepancies between the inventory and lexicon conspire to yield cue weights in the inventory models that notably underestimate the role of dynamic amplitude in the lexicon. Figure 4.10 shows the role of AMP_{DYN} in the three models. In the lexicon model, we see a notable increase in contrast presence likelihood starting at around 5 dB, where most contrasts in the lexicon are well above the this threshold, whereas in the inventory a much greater proportion of contrastive pairs are within the non-contrastive range and thus the overall role of AMP_{DYN} in the inventory is negligible. This overlap between contrastive and non-contrastive ΔAMP_{DYN} values in the inventory and weighted inventory models can also be seen in the leftward shift of the contrastive distributions in Figure 4.10 relative to the lexical distribution, and in the corresponding rightward shift of the non-contrastive distributions in the inventory models, both of which reduce the overall utility of dynamic amplitude relative to its role in the lexicon.

From Figure 4.11 we see that in the weighted inventory model, while there are several highly frequent contrasts that exhibit sizeable differences in dynamic amplitude (well above the within-category range), there are also a large number of contrasts such as *b–p*, *b–k*, and *b–t* that are acoustically less distinct in the inventory than in the lexicon, and so the upweighting of cues from these contrasts does not improve the discriminability of contrastive and non-contrastive pairs. Further, the majority of the dynamic amplitude distinctions in the inventory are smaller than those in the lexicon, particularly among higher-frequency contrasts (shown in Figure 4.11 as points above the gray identity line). Overall, while there is a generally higher correlation between controlled-syllable and real-word measurements of dynamic amplitude than in the corresponding VOI% relation, the correlation between the two databases is lower in Figure 4.11 than in Figures 4.5 and 4.7 for $F2_{CV}$ and LF, respectively, both cases where the inventory acoustics scale well to the lexicon. Thus, there

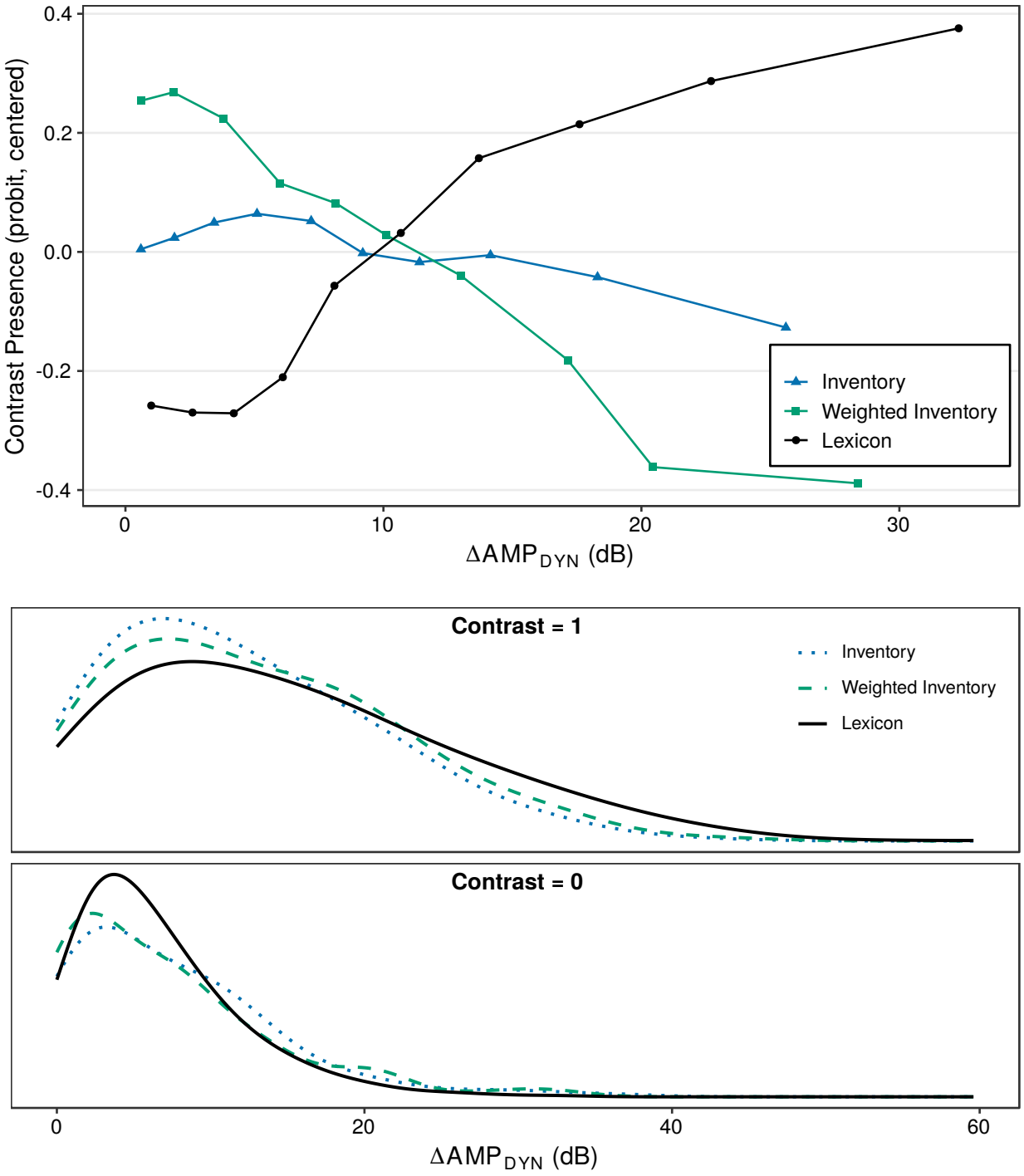

Figure 4.10: Partial dependence functions for AMP_{DYN} (top panel) and AMP_{DYN} distributions (bottom panels) in the inventory, weighted inventory, and lexicon models of ideal recognition in word-initial position.

is evidence in Figure 4.11 of both distributional and acoustic factors in the underestimation of the role of dynamic amplitude in the lexicon from both inventory models.

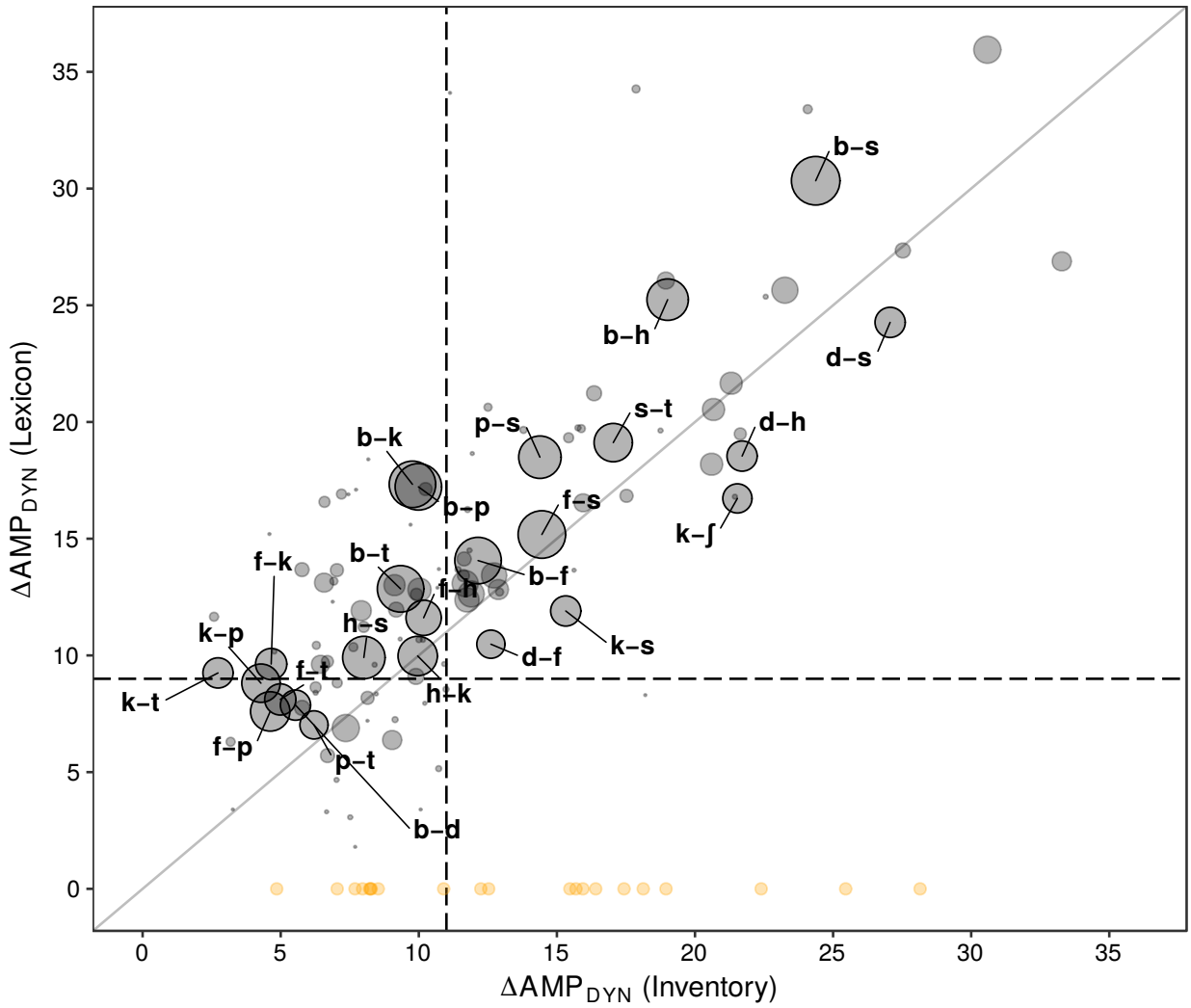

Figure 4.11: Relationship between ΔAMP_{DYN} means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-initial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 50% of items are labeled. Contrasts absent from the lexicon are shown in orange.

Finally, as discussed in the introduction of this section, negative cue weights can also arise due to confounding relations between the cue of interest and other cues in the model. Therefore, while the distributional and acoustic disagreements between the inventory models and the lexicon models are contributing factors to the ultimate difference in aggregate cue weights, they are not the comprehensive in accounting for the scaling problem. Confounding relations between cues

in such a high-dimensional space are difficult to unpack, but in the case of dynamic amplitude we must assume that there are other features which index sibilance and voicing, such as spectral tilt ($TILT_C$), spectral dispersion ($DISP_C$) and spectral peak amplitude (AMP_{PK}), which are more directly related to contrastiveness in the inventory models than dynamic amplitude.

4.3.2 Word-medial position (VCV)

As in word-initial position, all three models of intervocalic contrasts yield excellent fits to the data, both in terms of overall accuracy, and high precision/recall (Table 4.2). Figure 4.13 shows cue ranks in lexicon, inventory, and weighted inventory models of intervocalic contrasts. Among the highest-ranked cues in the lexicon are spectral peak frequency ($FREQ_{PK}$), spectral tilt of the consonant ($TILT_C$), spectral shape (SHAPE), consonant voicing percentage (VOI%), voice cessation time (VCT), and vowel-onset F2 ($F2_{CV}$). This set primarily distinguishes consonant place and voicing, but parameters such as spectral tilt and spectral shape also exhibit notable manner distinctions that expand the overall discriminative power of the set. Several of these parameters, such as $FREQ_{PK}$, VOI%, and $F2_{CV}$ are also highly ranked in the inventory and weighted inventory models, while other cues such as noise duration (ND), spectral tilt at V2 onset ($TILT_{V2}$), and relative F3 amplitude (AMP_{F3}) are of much greater utility in the inventory models than in the lexion. Finally, among the higher-weighted cues in each model, $TILT_C$ and VCT are each highly ranked in the inventory, as in the lexicon, though both cues receive lower weight in the weighted inventory model. No cues in this set exhibit the converse pattern of showing closer agreement between the lexicon and weighted inventory.

Among the lowest-ranked cues in the lexicon are vowel-offset/onset f0 ($f0_{VC/CV}$), preceding/-following vowel duration ($DUR_{V1/V2}$), following vowel amplitude (AMP_{V2}), F2 and spectral tilt of V1 ($F2_{V1}$, $TILT_{V1}$), and spectral dispersion at the VC transition ($DISP_{VC}$). Thus, with the exception of F1 and F2 at V1 offset ($F1_{VC}$, $F2_{VC}$), cues from the preceding vowel appear to be less useful than those from the following vowel, both in the lexicon and inventory. Other low-ranked cues in the inventory that receive relatively higher weight in the lexicon are noise amplitude (AMP_N),

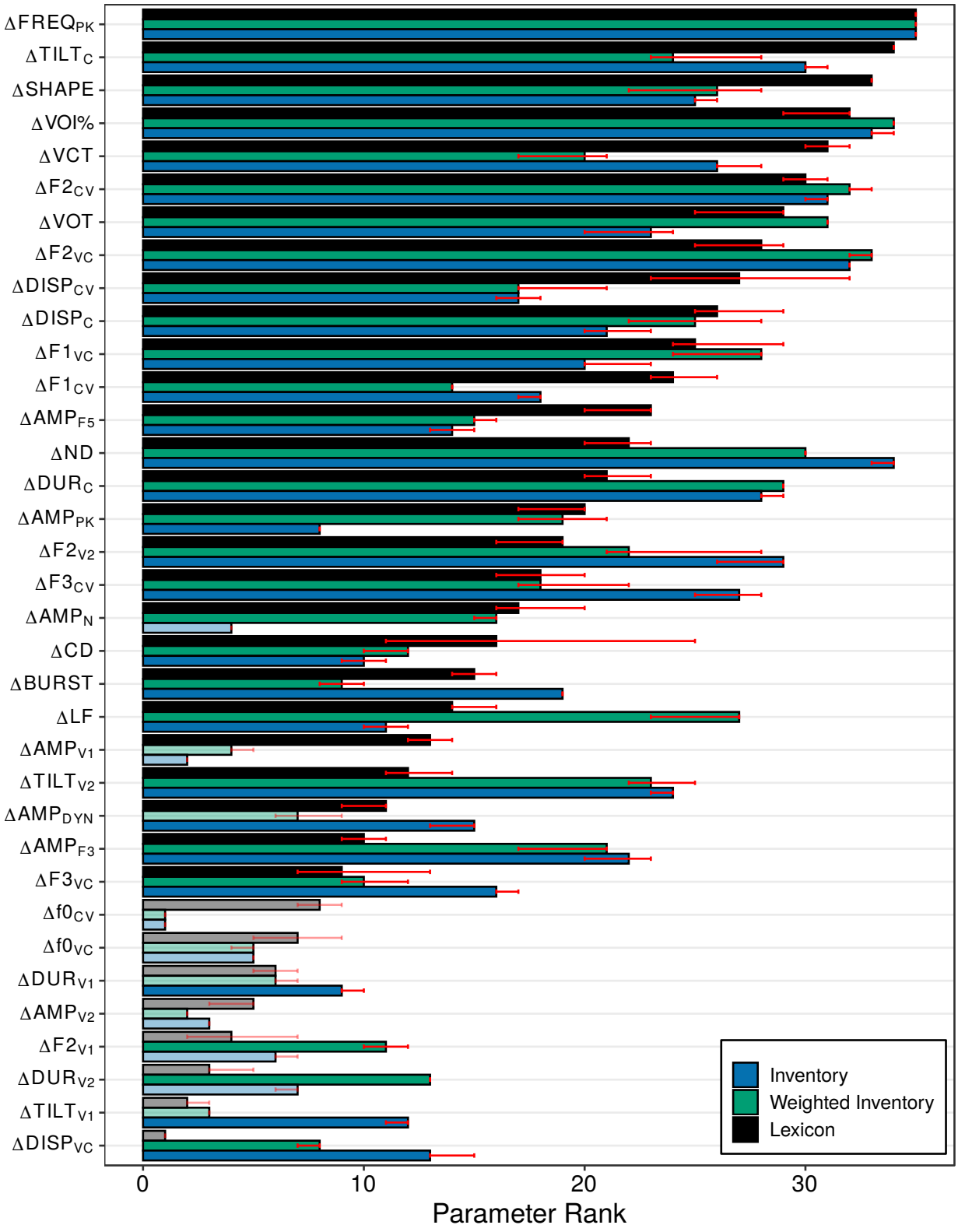

Figure 4.12: Acoustic parameter ranks in lexicon, inventory, and weighted inventory models of word-medial contrasts under the assumption of ideal recognition. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

	Predicted					
	Inventory		W. Inv.		Lexicon	
Observed	0	1	0	1	0	1
0	46	0	36	0	38	2
1	0	54	0	64	0	59
Accuracy:	100%		100%		98%	
Precision:	1		1		0.97	
Recall:	1		1		1	
$F1$ Score:	1		1		0.98	

Table 4.2: Confusion matrices and model fit statistics for inventory, weighted inventory, and lexicon models of word-medial contrast presence/absence (0/1).

relative F5 amplitude (AMP_{F5}), and spectral dispersion at the CV transition ($DISP_{CV}$), while in addition to these cues, F1 at V2 onset ($F1_{CV}$) is of relatively low rank in the weighted inventory model. Finally, the sets {ND, DUR_C, $TILT_{V2}$, $F2_{V2}$, $F3_{CV}$, $TILT_{V2}$, AMP_{F3}, $F3_{VC}$} in the inventory, and {ND, DUR_C, LF, $TILT_{V2}$, AMP_{F3}} in the weighted inventory, receive much higher weights than in the lexicon.

The general relationship between cue ranks in each model is further summarized in Figures 4.13 and 4.14, which show respectively the correlation between cue ranks in the lexicon and those in the inventory / weighted inventory, and the relationship between rank differences between the lexicon model and each inventory model. Unlike in the models of word-initial contrasts, in VCV position the cue weights in the weighted inventory are relatively more correlated with those in the lexicon than are the weights in the balanced inventory, though both correlations are quite high at 0.78 and 0.72, respectively. In Figure 4.14 we see this difference in a greater horizontal spread of cue differences, reflecting greater disagreement between the lexicon and inventory, and thus greater agreement between with the weighted inventory, but also we find that many cues exhibit potential composite sources of disagreement (distributional and acoustic) in lying far from the origin along the $y = x$ axis (in the upper-right and lower-left quadrants). These cues include cues such as $DISP_{CV}$, AMP_{V1}, AMP_{F5}, $f0_{CV}$, SHAPE, and $F1_{CV}$, whose role in the lexicon is underestimated by both inventory models, and cues such as $TILT_{V2}$, AMP_{F3}, ND, DUR_C, and $DISP_{VC}$, whose

lexical role is overestimated by models based on controlled syllable acoustics.

Among the cues exhibiting points of disagreement between the inventory and lexicon that appear to primarily result from differences in the distribution of contrasts in the homogeneous inventory and the fundamentally heterogeneous lexicon—i.e., *distributional* discrepancies—are the amplitudes of the spectral peak (AMP_{PK}) and overall noise interval (AMP_{N}), which are notably underestimated in the inventory model, and F3 at vowel offset/onset ($F3_{VC/CV}$) and the spectral tilt of the preceding vowel ($TILT_{V1}$), which are notably overestimated in the inventory analysis. Finally, while there is generally less agreement between the three models in VCV position relative to CV position, there remain several cues that closely agree in cue ranks and are of relatively high utility in each model. These cues include spectral peak frequency ($FREQ_{PK}$), F2 at vowel onset ($F2_{CV}$), and consonant voicing percentage (VOI%). From among these four classes of cues exhibiting points of agreement and disagreement of different types, four cues were chosen to exemplify the behavior of each class in word-medial contrast discrimination: spectral peak frequency ($FREQ_{PK}$; cue rank *agreement*), spectral peak amplitude (AMP_{PK}; *distributional disagreement*), voice cessation time (VCT; *acoustic disagreement*), and relative F3 amplitude (AMP_{F3}; *composite disagreement*). We begin with spectral peak frequency, which is the highest-ranked cue in all three models.

4.3.2.1 Cue Agreement: Spectral Peak Frequency

In all three models—inventory, weighted inventory, and lexicon—the peak frequency of the obstruent noise spectrum is highly discriminative, as shown in Figure 4.15, where most of the area in the contrastive distributions is to the right (higher $\Delta FREQ_{PK}$ values) of the corresponding distributions for within-category distinctions, and all three partial dependence functions show a sizeable increase in contrast presence likelihood between approximately 1 to 3 kHz. However, despite the close agreement there are some key differences in the behavior of spectral peak frequency in the lexicon and inventory. First, the discrimination point between contrastive and non-contrastive $\Delta FREQ_{PK}$ values is approximately 500 Hz lower in the lexicon than in the inventory, while in the

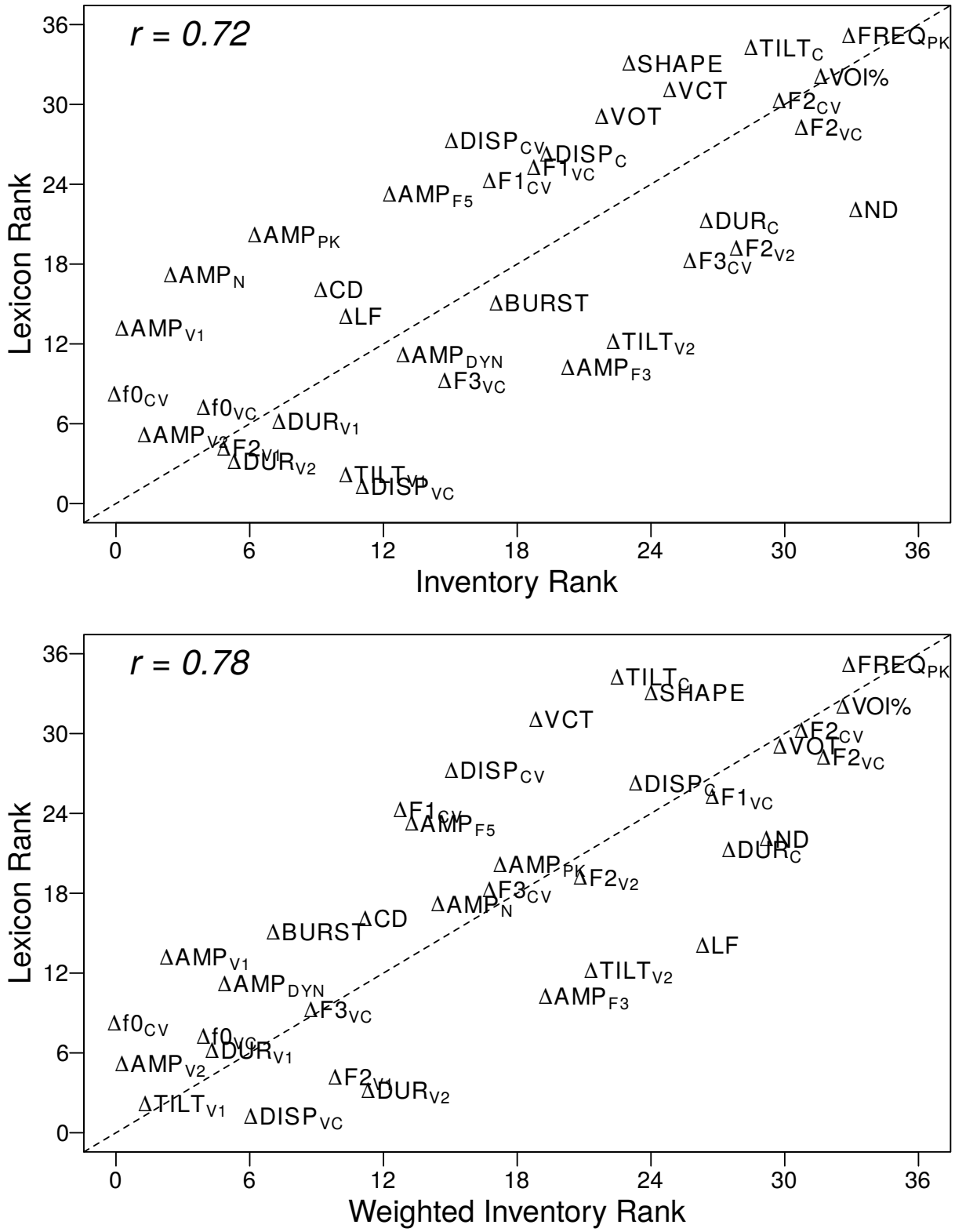

Figure 4.13: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VCV position. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

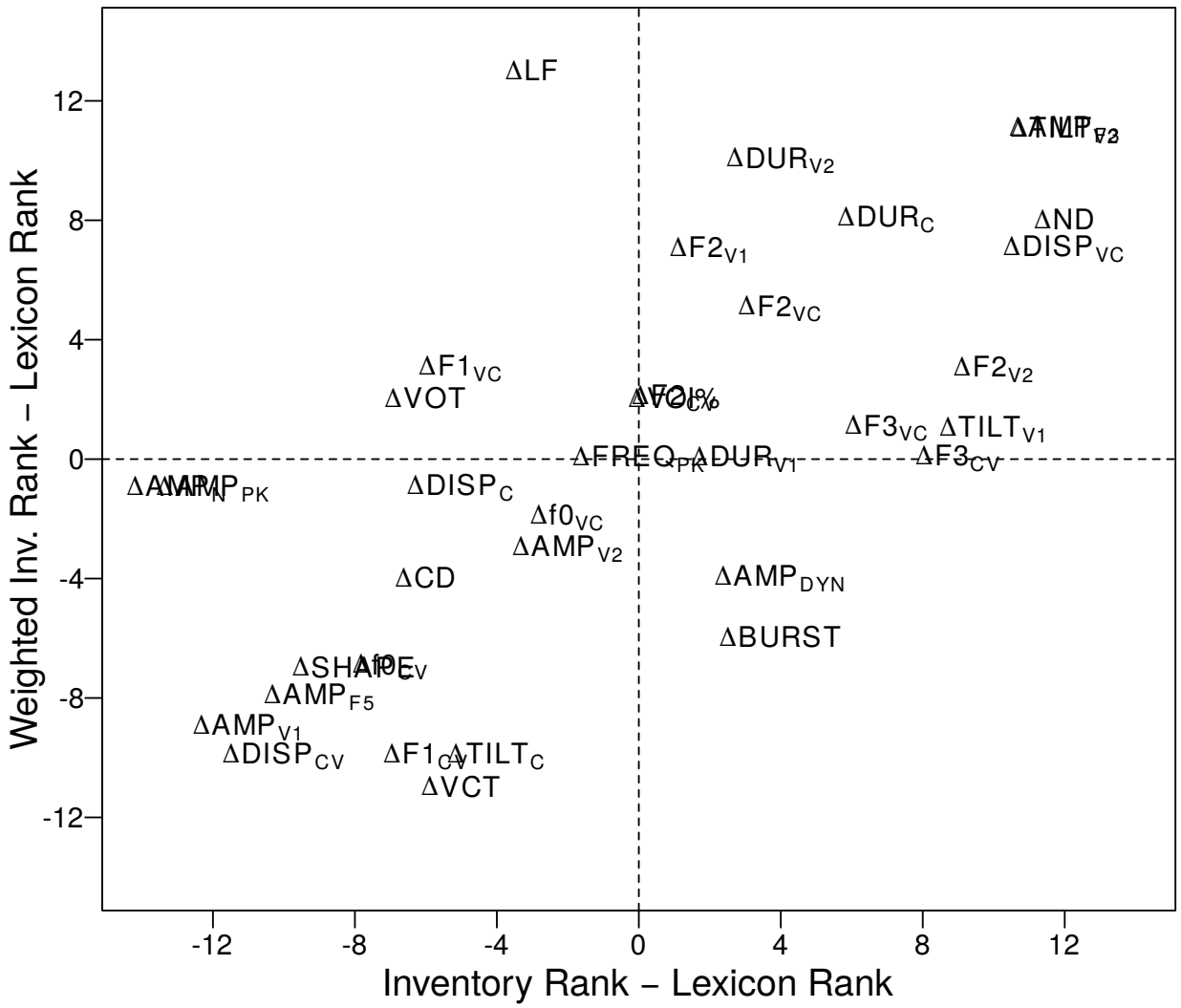

Figure 4.14: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VCV position. Dashed lines indicate equivalence relations between each pair of models.

weighted inventory model this point is further lowered by approximately 300 Hz, meaning that spectral peak frequency shows greater overlap between within- and cross-category distinctions in the inventory than in the lexicon.

Examining the relation between FREQ_{PK} differences in the lexicon and inventory by phonetic contrast (Figure 4.16), we see that many of the more frequent contrasts such as those between the plosives [p, t, k, d, g] and the alveolar flap [ɾ] are more distinct in the lexicon (*y*-axis) than in the inventory. Further, many of the contrasts between sibilants and nonsibilants, which as expected from Chapter 2 exhibit the greatest distinctions in spectral peak frequency, play a greater

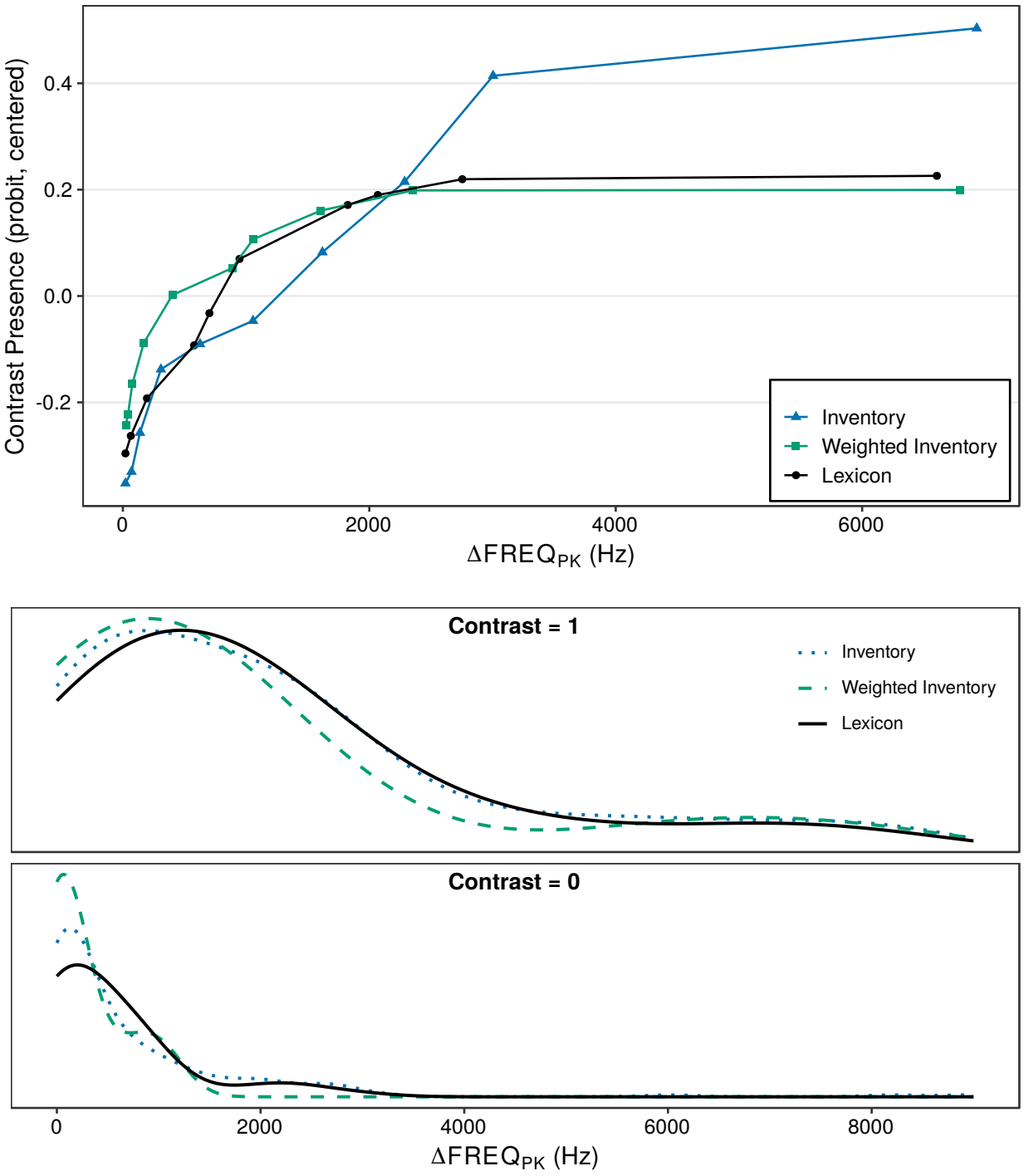

Figure 4.15: Partial dependence functions for FREQ_{PK} (top panel) and FREQ_{PK} distributions (bottom panels) in the inventory, weighted inventory, and lexicon models of ideal recognition in word-medial position.

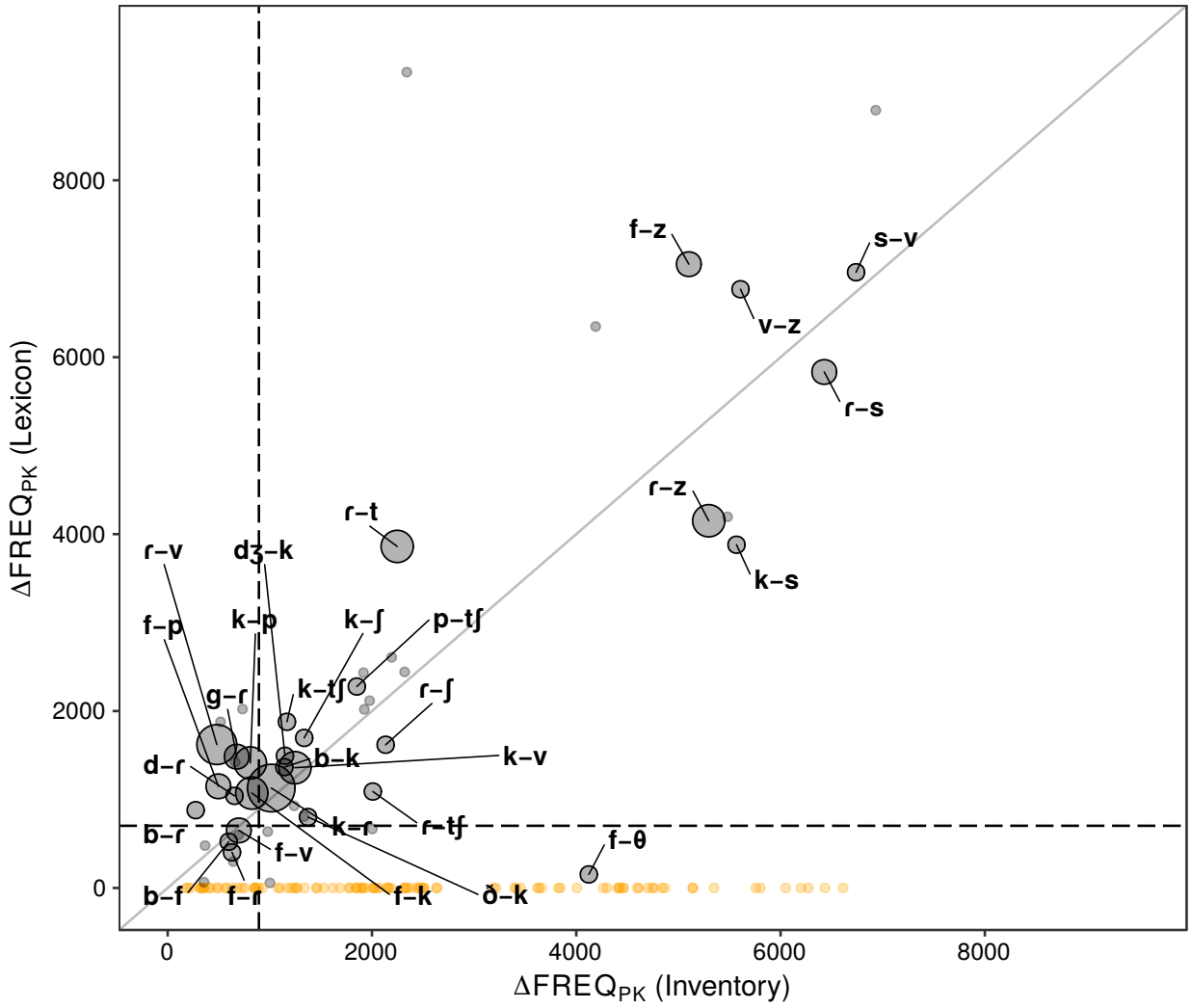

Figure 4.16: Relationship between $\Delta FREQ_{PK}$ means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-medial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 75% of items are labeled. Contrasts absent from the lexicon are shown in orange.

role in the lexicon than in the inventory. Here it is important to draw the reader's attention to the contrasts in Figure 4.16 that are present in the inventory but not in the lexicon (the orange points). In the analysis of word-initial contrasts, most contrasts occurring in the inventory were also present in minimal pairs in the lexicon, but the intervocalic contrast set in the lexicon is much more sparse. Therefore, while *sibilant–nonsibilant* contrasts are similarly distinct in the inventory, they comprise a relatively smaller proportion of the total contrast set, thus shifting the distribution of $\Delta FREQ_{PK}$ toward less distinct, within-place contrasts (see the relation between contrastive and

non-contrastive distributions in Figure 4.15 for details).

For this reason, and the fact that among the contrasts present in both the lexicon and inventory most show close correlations between the measurements from the two data sets, the loss of less-distinct contrasts in the weighted inventory model results in greater agreement with the lexicon model, though in aggregate spectral peak frequency is highly discriminative in all three models.

4.3.2.2 Distributional Disagreement: Spectral Peak Amplitude

One cue that is ranked relatively low in the inventory model, but which in the weighted inventory is brought into complete agreement with the lexicon ranking, is spectral peak amplitude. Figure 4.17 shows the partial dependence functions and contrastive/non-contrastive distributions for AMP_{PK} in each model. In the inventory model we see that the predicted contrast likelihood is relatively flat between 0 and 8 dB, only showing a notable increase between 8 and 11 dB, with a more slight increase over the final quantile between 11 and 15 dB. The lexicon model, on the other hand, increases linearly between 0 and 10 dB, with diminishing gains above 10 dB similar to that in the inventory model, while in the weighted inventory we see a somewhat intermediate result of a sharp increase in contrast likelihood between 4 and 7 dB and a relatively shallow increase over the remainder of the ΔAMP_{PK} range.

This shift toward the lexicon result that is achieved by weighting contrasts in the inventory according to their lexical distribution is also shown in Figure 4.17 in the rightward shift of the weighted inventory distribution relative to the inventory, though there remains greater area in the tails of the lexicon distribution. Therefore, while accounting for the distributional disagreement between the lexicon and inventory eliminates the aggregate cue rank difference between the two models, the full behavior of spectral peak amplitude is not completely aligned between models operating on acoustic properties of controlled syllables and those derived from real words. This is why we describe points of distributional disagreement as characterizing the *primary* source of the discrepancy, not the complete cause.

In Figure 4.18 we get a better picture of which contrasts contribute to the shape of the AMP_{PK}

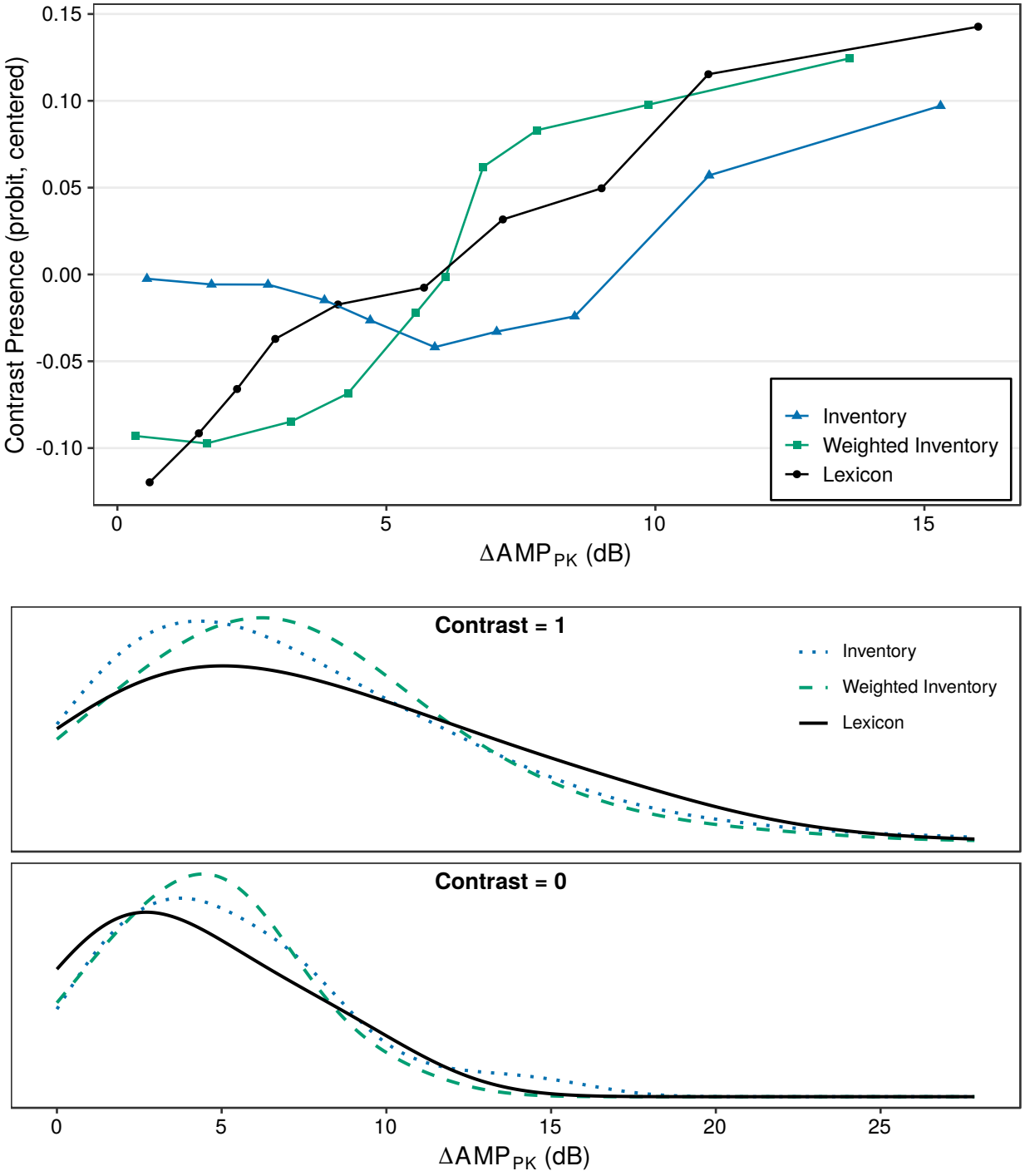

Figure 4.17: Partial dependence functions for AMP_{PK} (top panel) and AMP_{PK} distributions (bottom panels) in the inventory, weighted inventory, and lexicon models of ideal recognition in word-medial position.

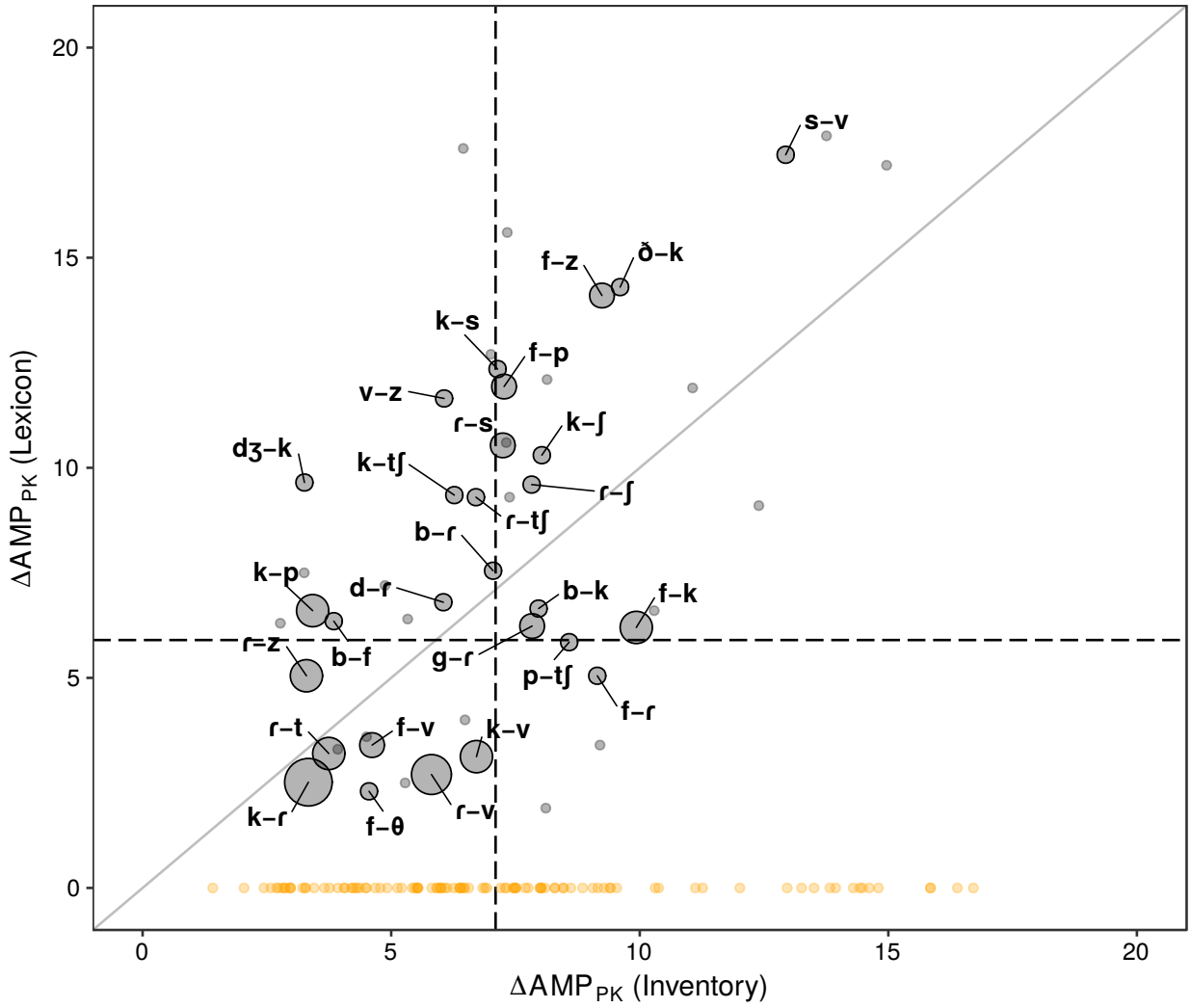

Figure 4.18: Relationship between ΔAMP_{PK} means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-medial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 75% of items are labeled. Contrasts absent from the lexicon are shown in orange.

distributions and partial dependence functions in Figure 4.18. Overall, there is a modest but significant correlation between the ΔAMP_{PK} values among contrasts in the inventory and those in the lexicon ($r = 0.564$, $p < 0.001$), though not to the degree observed for spectral peak frequency ($r = 0.792$, $p < 0.001$). At the lower end of the range are contrasts between nonsibilant fricatives and contrasts between voiceless plosives and the alveolar flap [ɾ], which though differing in the location of the spectral peak are similar in peak amplitude.[4] At the upper end of the range are

[4]Note that minimal pairs between [t] and [ɾ] are inconsistent with the phonological description of the two phones

contrasts between sibilant and nonsibilant obstruents, and between nonsibilant fricatives, which exhibit relatively low spectral peak amplitudes, and voiceless plosives (e.g., *ð–k*, *f–p*, and *f–k*), whose burst spectra are generally louder and more defined.

The problem for the use of spectral peak amplitude in the inventory is that well over half of the contrasts (i.e., including those not present in the lexicon) exhibit AMP_{PK} distinctions that are not any larger than the natural within-category variation in peak amplitude. Thus, by restricting the weighted inventory model to only those contrasts present in minimal pairs in the lexicon, many such items are excluded and a greater proportion of items lie above the non-contrastive range (seen in Figure 4.18 in the gray points above 6 dB along the *x*-axis). Again, the acoustic agreement between the two datasets is much lower than that observed for low-frequency energy (LF) among word-initial contrasts, and there is some redistribution of the roles of different contrasts in the weighted inventory model (e.g., *f–k* yields more distinct values when drawn from the inventory, while *k–p* is less distinct in controlled syllables than in real-word contrasts), but the end result is that spectral peak amplitude is of much higher utility when accounting for the lexical distribution than when every contrast in the inventory is of equal weight.

4.3.2.3 Acoustic Disagreement: Voice Cessation Time

Turning next to a cue which is ranked relatively high in both inventory and lexicon models, but which appears to behave differently in each model due to differences in the acoustics of its constituent contrasts, is voice cessation time (VCT). Figure 4.19 shows the partial dependence functions and underlying cue distributions for VCT in the lexicon, inventory, and weighted inventory models, and illustrates a general weakening of the relation between ΔVCT and contrast likelihood in the weighted inventory relative to the inventory and lexicon models. Further, the acoustic disagreement between the inventory and lexicon can be seen in both the rightward shift of the partial dependence inflection point in the inventory relative to the lexicon (∼15 ms) and in the bimodal distribution of contrastive VCT distinctions in the inventory that is absent from the lexicon.

being in complementary distribution, and such items were avoided in the design of Experiment 1; nevertheless they do occur on occasion when the alveolar plosive is hyperarticulated.

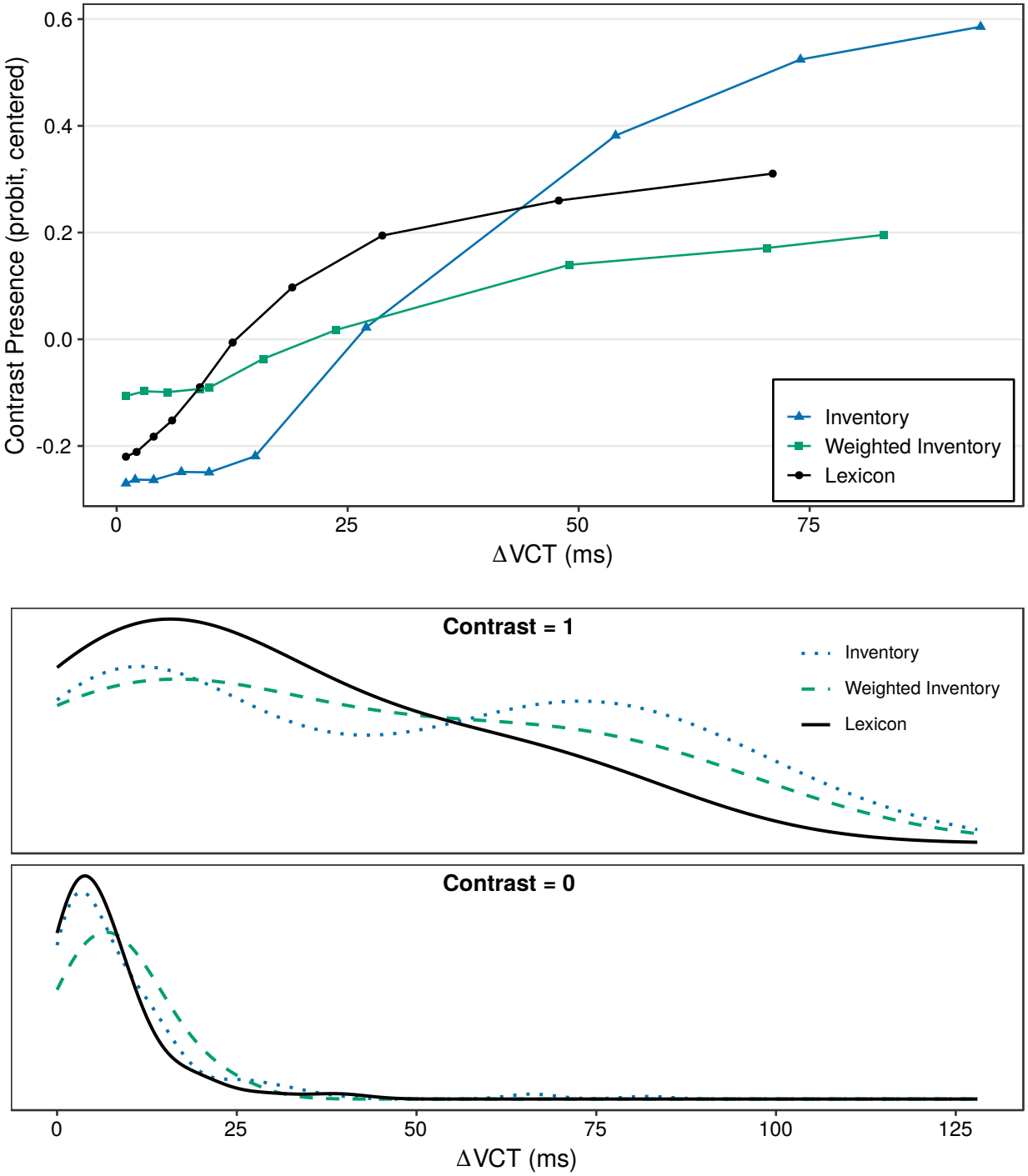

Figure 4.19: Partial dependence functions (top panel) and distributions (bottom panels) of VCT in the inventory, weighted inventory, and lexicon models of ideal recognition in word-medial position.

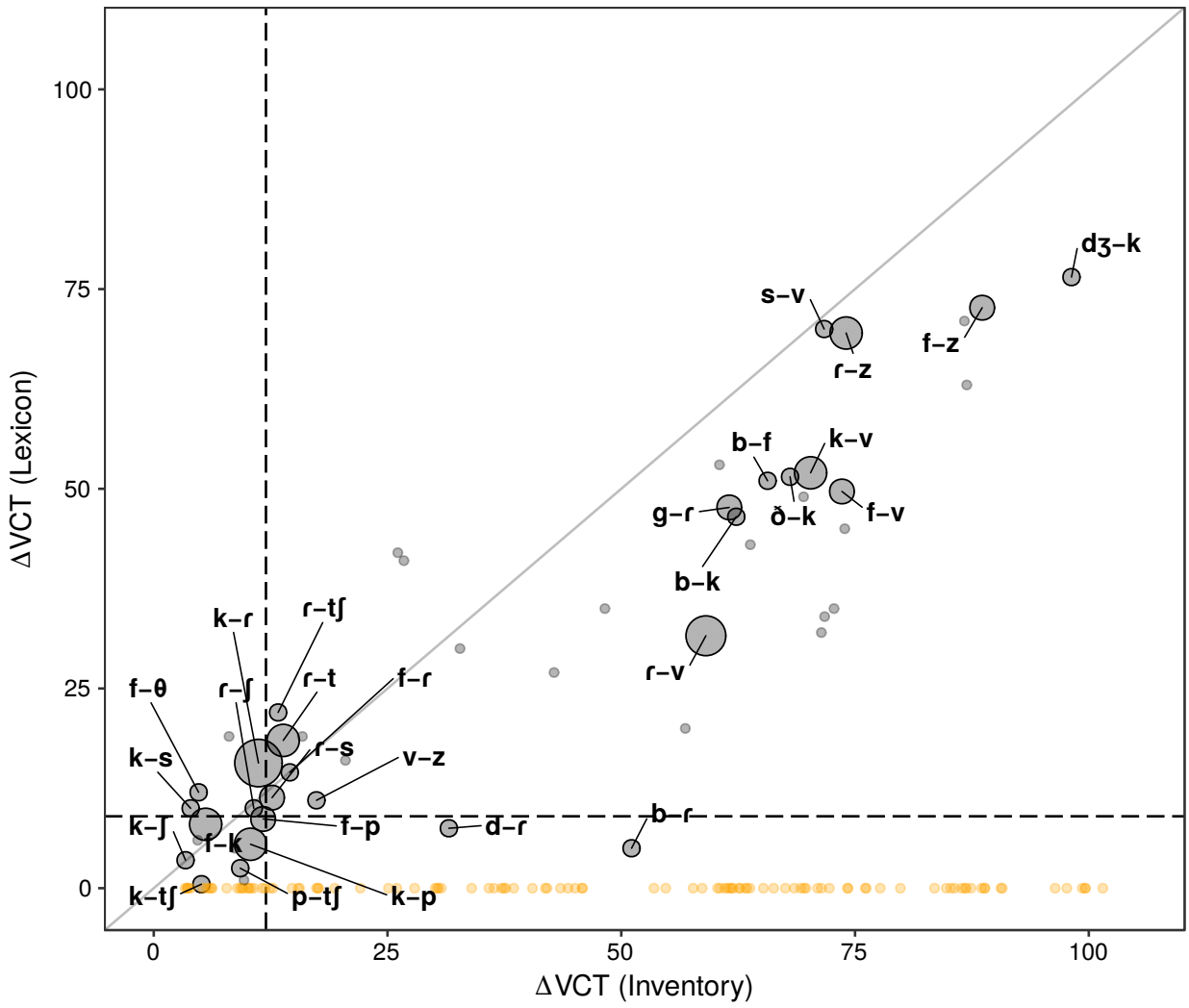

Figure 4.20: Relationship between ΔVCT means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-medial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 75% of items are labeled. Contrasts absent from the lexicon are shown in orange.

The result of drawing acoustic values from the inventory and mapping them onto the lexical distribution via the weighted inventory model is then to weaken the mode corresponding to lower ΔVCT contrasts in the lexicon that contributes to the sharp distinction between contrastive and non-contrastive pairs around 10 dB. In the upper ΔVCT range, primarily driven by distinctions in voicing and to a limited extent manner,[5] contrasts are enhanced in the weighted inventory model

[5]Many such contrasts involve the alveolar flap [ɾ], which exhibits a distinctly intermediate VCT distribution that is longer in duration than most voiceless obstruents, but shorter than most voiced obstruents given that the VCT of fully voiced stops is defined to be equal to their total duration, which leaves flaps shorter than other voiced obstruents.

(seen also in the alignment of contrasts below the gray identity line in Figure 4.20), but given that such contrasts were already well above the within-category range in the lexicon this enhancement does little to change their role in the weighted inventory. Further inspection of Figure 4.20 reveals that among the more frequent contrasts whose VCT distinctions are more narrowly differentiated from the within-category set—primarily contrasts between voiceless obstruents and the alveolar flap—the relation between the lexical and inventory acoustics is in the opposite direction of that in the higher ΔVCT range: namely, smaller distinctions in the inventory than the lexicon.

Combining these results with the distribution of ΔVCT values among contrasts absent from the lexicon, which show substantial overlap with the within-category range for low ΔVCT as well as a wide range of contrasts where voice cessation times are notably distinct (between 25 and 100 ms apart), we have a case where a cue is highly discriminative in two systems despite exhibiting divergent underlying acoustic distributions. Of course, we cannot say from the present data whether the acoustic properties of contrasts absent from the Lex95 database would agree with those in the inventory if they were present. This question is left for future research modeling larger samples of the lexicon. Nevertheless, cases such as these are instructive for the general problem of scaling cue estimates from the inventory to the lexicon, because they illustrate how aggregate cue weights can obscure details in the behavior of a particular cue across a range of contrasts, as well as demonstrating how a model-based re-weighting of contrasts from controlled syllable data to match distributions in the lexicon can provide misleading estimates of the role of certain cues in the recognition of real words.

4.3.2.4 Composite Disagreement: Relative F3 Amplitude

Finally, we examine the relative amplitude of the consonant noise spectrum in the F3 region (AMP_{F3}) as a case of *composite* disagreement, where both acoustic and distributional discrepancies appear to be involved in the cue-weight difference between inventory-based and lexicon-based models of ideal cue integration. Figure 4.21 shows partial dependence functions and distributions of ΔAMP_{F3} in the lexicon, inventory, and weighted inventory. The inventory model shows a

sigmoidal partial dependence curve with a sharp division between predicted contrastive and non-contrastive items at around 7 dB, while the relationship between F3 amplitude and contrastiveness is relatively flat in the lexicon, with the weighted inventory intermediate between the two and approximately linear between 0 and 10 dB.

The ΔAMP_{F3} distributions in Figure 4.21 provide further information on how such discrepancies arise from relations between contrastive and non-contrastive AMP_{F3} distinctions in the three models, where the inventory model shows a sharp decline in the non-contrastive distribution above 7 dB, leading to the corresponding inflection in contrast likelihood at that point in the partial dependence function. The lexicon model, on the other hand, shows the opposite pattern in exhibiting a relatively shallow non-contrastive distribution and a more sharply skewed contrastive distribution, meaning the two overlap for the majority of the ΔAMP_{F3} range and thus AMP_{F3} is not at all predictive of contrast presence. Finally, the weighted inventory model is intermediate between the two in showing a skewed non-contrastive distribution similar to that in the inventory, while the contrastive distribution in the weighted inventory is relativel shallower with a mode that is further to the right than both inventory and lexicon distributions, resulting in a partial dependence function where contrast likelihood begins to increase at a much earlier point in the ΔAMP_{F3} range, but which is also shallower in slope than the inventory model.

Figure 4.22 shows the relation between AMP_{F3} distinctions in phonetic contrasts derived from the controlled syllable data in the inventory model and those derived from minimal pairs in the lexicon. Overall the two do not correlate ($r = 0.137$, $p > 0.1$), and over half of contrasts in the inventory lie beyond the 75th percentile of the within-category range (the x dimension), very the vast majority of contrasts lie within the corresponding range in the lexicon. Further, the contrasts in the lexicon which do show notable distinctions in relative F3 amplitude, such as those between postalveolar and non-postalveolar obstruents, are relatively infrequent. Finally, among the more frequent contrasts in the lexicon, the majority exhibit relatively greater AMP_{F3} distinctions in the controlled syllable data than in real-word minimal pairs (shown in the greater number of contrasts below the gray identity line in Figure 4.22), which is why the role of AMP_{F3} in the weighted

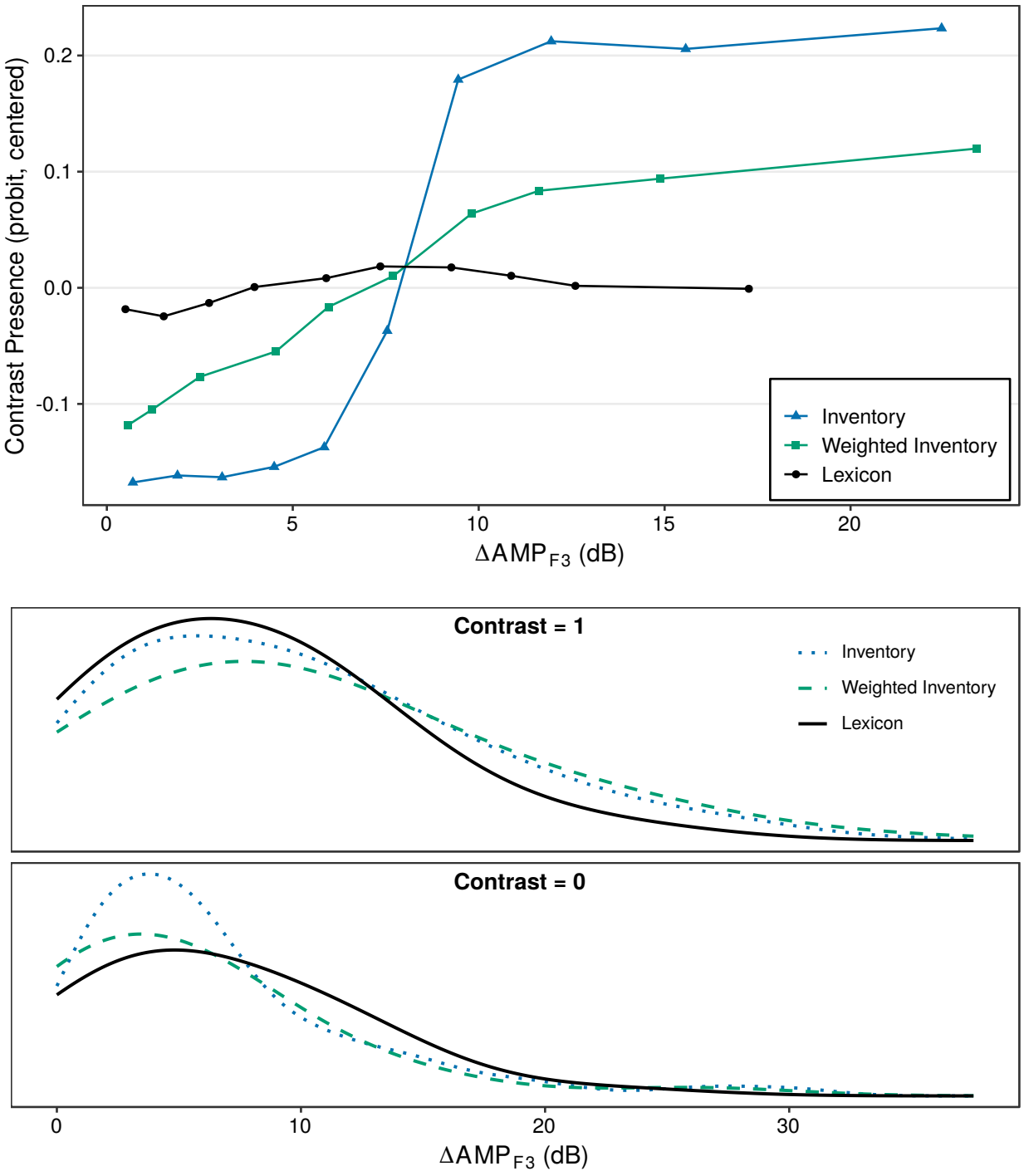

Figure 4.21: Partial dependence functions (top panel) and distributions (bottom panels) of AMP_{F3} in the inventory, weighted inventory, and lexicon models of ideal recognition in word-medial position.

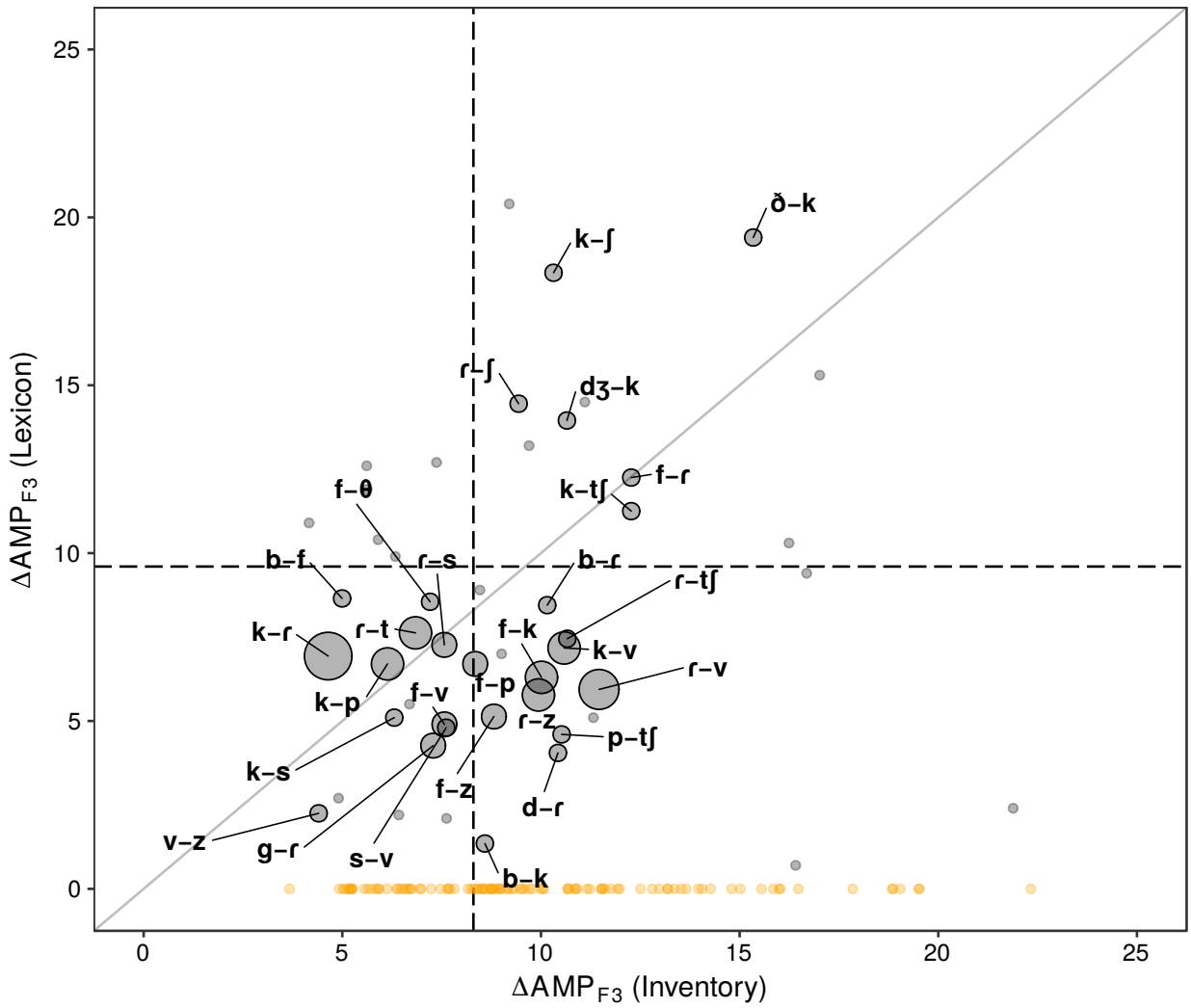

Figure 4.22: Relationship between ΔAMP_{F3} means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-medial position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 75% of items are labeled. Contrasts absent from the lexicon are shown in orange.

inventory is enhanced relative to the lexicon, though when the contrasts which are not present in the lexicon are accounted for AMP_{F3} remains marginally more discriminative in the inventory model. Thus, here we have a case where both the lack of acoustic agreement in AMP_{F3} between the two data sources, and the sizeable difference in contrast distributions in each system, leads to a fundamental discrepancy in cue weights where the role of AMP_{F3} in the lexicon cannot be inferred from a database of controlled syllables designed to capture the relevant phonological and acoustic features of the system, but which requires direct access to contrasts as they occur in the lexicon.

	Predicted					
	Inventory		W. Inv.		Lexicon	
Observed	0	1	0	1	0	1
0	37	11	38	0	36	3
1	1	51	1	61	3	59
Accuracy:		88%		99%		95%
Precision:		0.82		1		0.95
Recall:		0.98		0.98		0.95
$F1$ Score:		0.89		0.99		0.95

Table 4.3: Confusion matrices and model fit statistics for inventory, weighted inventory, and lexicon models of word-final contrast presence/absence (0/1).

4.3.3 Word-final position (VC)

Table 4.3 shows the confusion matrices and fit statistics for the ideal perceiver models of word-final contrasts in the inventory, weighted inventory, and lexicon. Overall, the three models show good fits to the data, though the inventory model is moderately lower in accuracy due to a bias toward contrast presence (precision = 0.82 as compared with 0.95 in the lexicon and 1 in the weighted inventory), which is the first case of a notable bias in any of the ideal perceiver models. Nevertheless, this deviation in model fit remains relatively small and should not pose a problem for the evaluation of cue weights below.

Figure 4.23 shows parameter ranks in the three models, and with the exception of spectral dispersion at the VC transition (DISP_{VC}) shows good agreement between the highest ranked cues in each model. These cues include noise duration (ND), F2 at vowel offset (F2_{VC}), spectral tilt of the consonant noise spectrum (TILT_{C}), spectral peak frequency (FREQ_{PK}), and consonant voicing percentage (VOI%), though the latter is ranked relatively higher in the inventory and weighted inventory than in the lexicon. Among the lowest-ranked cues in the lexicon are noise amplitude (AMP_{N}), burst presence (BURST), low-frequency energy (LF), F2 of the preceding vowel (F2_{V1}), f0 at vowel offset (f0_{VC}), and spectral tilt at vowel offset (TILT_{V1}). These cue ranks generally agree with those in the inventory, while the role of AMP_{N}, BURST, and F2_{V1} in the weighted inventory is relatively greater, though only AMP_{N} is ranked in the top half of cues in the weighted

inventory model. Thus, the greatest discrepancies between the three models occur over the middle half of cue ranks in the lexicon. The most notable differences that emerge in this set are the high ranking of spectral shape (SHAPE) in both inventory models despite SHAPE playing a relatively minor role in the lexicon, and the low ranking of consonant duration (DUR_C) and preceding vowel amplitude (AMP_{V1}) in the inventory models relative to the lexicon.

These relations are further summarized in Figures 4.24 and 4.25, where in the former we see that the cue ranks in the inventory are highly correlated with those in the lexicon, at 0.81, while the correlation between the lexicon and weighted inventory ranks is much lower at 0.57, though this result is primarily driven by a limited set of nine cues—$DISP_{VC}$, AMP_{V1}, DUR_C, AMP_{PK}, $F2_{V1}$, BURST, VCT, AMP_N, and SHAPE—that exhibit sizeable deviations between the two models. The remainder of the cues show an almost perfect identity relation between the weighted inventory and lexicon ranks. The inventory model, on the other hand, exhibits relatively fewer outliers but shows slight rank differences for most cues.

Given these patterns, the distribution of cues in Figure 4.25 according to rank differences between the lexicon and the two inventory models shows the expected greater variance along the y-axis (weighted inventory rank $-$ lexicon rank) than the x-axis, though there are also several cues aligned along the $y = x$ diagonal that are indicative of composite disagreements in cue weight. This set includes spectral shape, which as noted earlier is highly overestimated by both inventory models in terms of its role in the lexicon, and AMP_{V1}, DUR_C, and $DISP_{VC}$ (excluded from the figure for visual clarity), which are each notably underestimated. Consonant voicing percentage (VOI%) and dynamic amplitude (AMP_{DYN}) show similar composite effects, though to a much smaller degree than VOI%, AMP_{V1}, DUR_C, and $DISP_{VC}$. Among the cues that represent potential points of distributional disagreement, only DUR_{V1} is of note, as AMP_{F5} is ranked relatively low in all three models. By comparison, there are many more points of potential acoustic disagreement, the greatest being noise amplitude (AMP_N) and spectral peak amplitude (AMP_{PK}), and to a lesser extent voice cessation time (VCT), burst presence (BURST), and F2 of the preceding vowel ($F2_{V1}$). Finally, as noted in the discussion above, there are several cues which show close agreement between

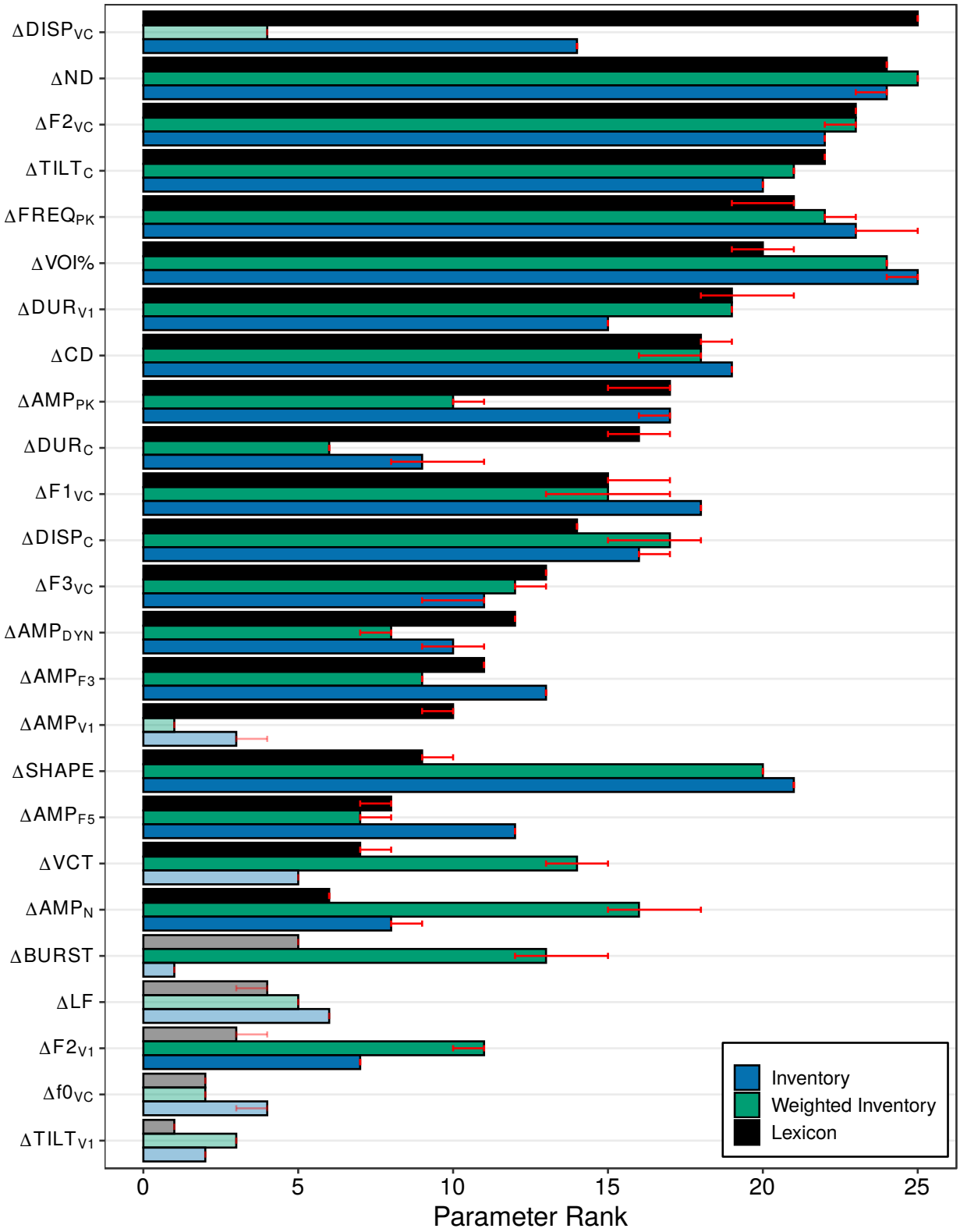

Figure 4.23: Acoustic parameter ranks in lexicon, inventory, and weighted inventory models of word-final contrasts under the assumption of ideal recognition. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

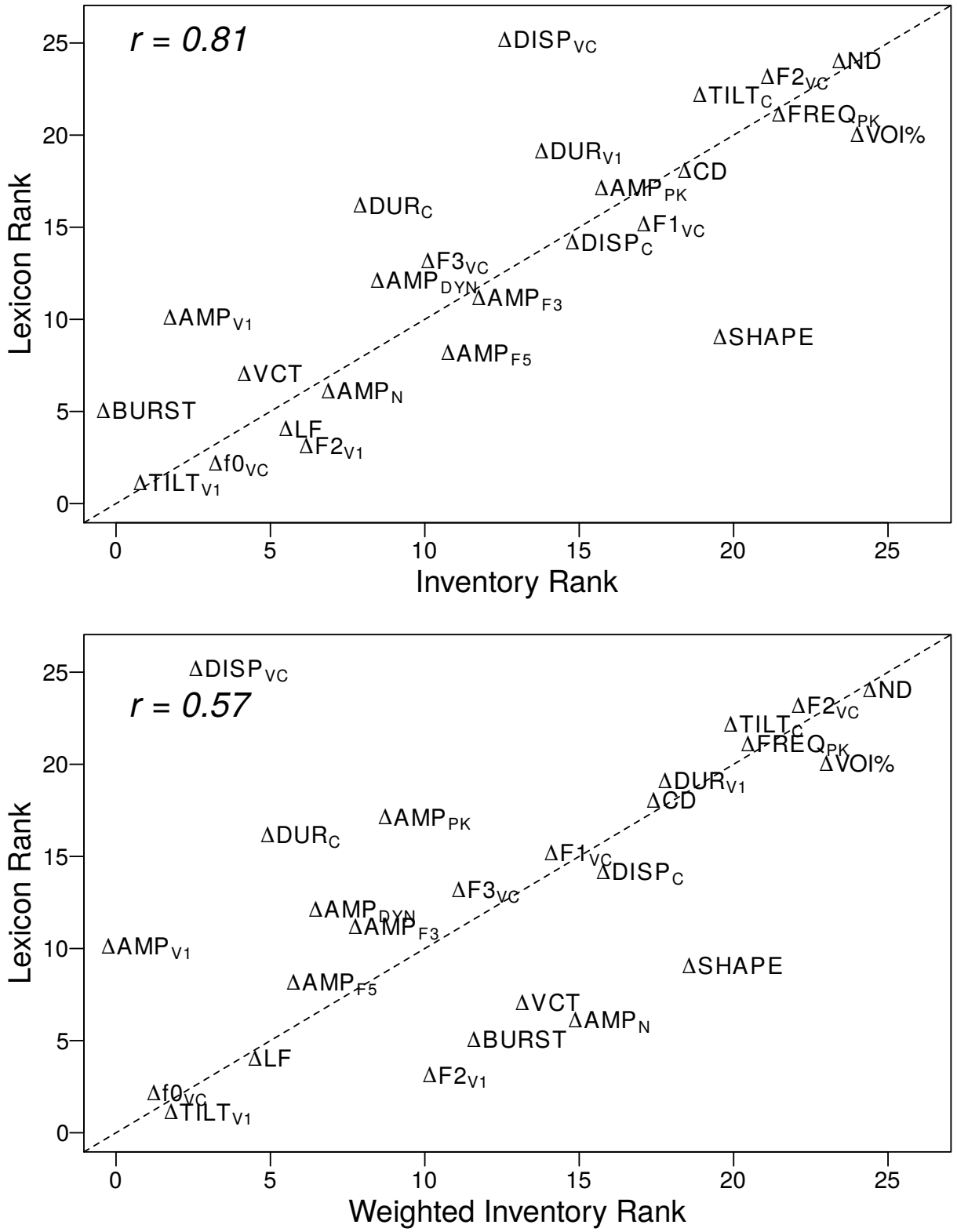

Figure 4.24: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VC position. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

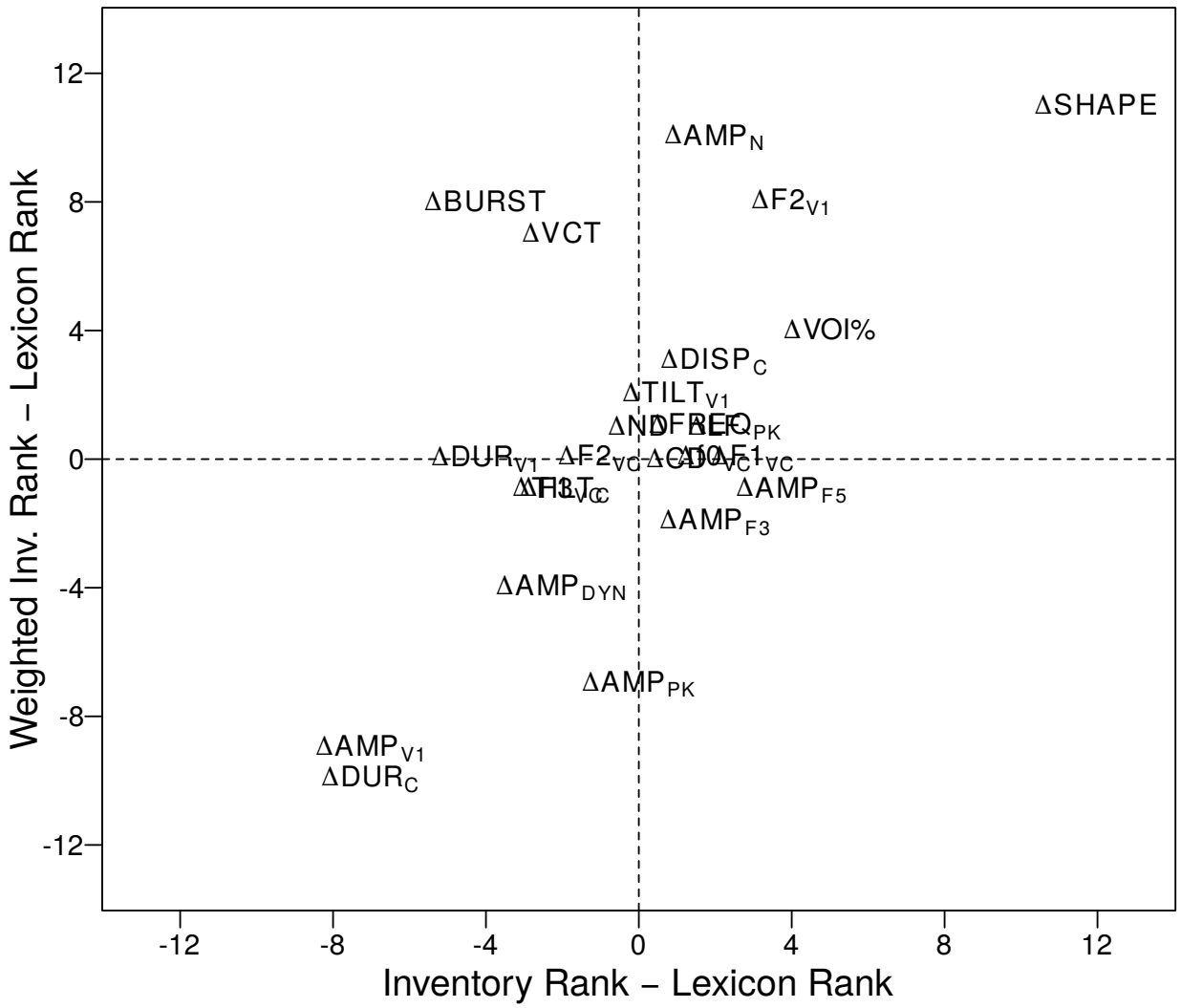

Figure 4.25: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VC position. Dashed lines indicate equivalence relations between each pair of models. The parameter DISP_{VC} has been excluded from the plot for clarity purposes; its coordinates are $(-11, -21)$.

all three models, including noise duration (ND), closure duration (CD), F2 and F3 at vowel offset (F2_{VC}, F3_{VC}), spectral tilt (TILT_{C}, TILT_{V1}), and spectral peak frequency (FREQ_{PK}). Next we review in detail exemplars of each cue type, adopting noise duration (ND) as a prominent case of cue *agreement*, preceding vowel duration (DUR_{V1}) as a case of *distributional disagreement*, noise amplitude (AMP_{N}) as a case of *acoustic disagreement*, and spectral shape (SHAPE) as a case of *composite disagreement*. We begin with noise duration, which is similarly highly weighted in all three models.

4.3.3.1 Cue Agreement: Noise Duration

Figure 4.26 shows partial dependence functions and distributions of noise duration distinctions (ΔND) in the inventory, weighted inventory, and lexicon models of word-final obstruent discrimination. All three models show a monotonic increase in contrast likelihood over the full ΔND range, though the greatest rate of change in the partial dependence functions occurs between 40 and 100 ms. This is the point at which within-category differences in noise duration rapidly decrease in likelihood while ΔND distributions among obstruent contrasts reach their modal value, though the contrastive distributions in Figure 4.26 do not exhibit a prominent mode, but rather remain relatively flat until approximately 100 ms.

In Figure 4.27 we see that there is a large number of manner contrasts, as well as fricative voicing contrasts such as *s–z*, that exhibit distinct differences in noise duration that are closely correlated between the inventory and lexical data. The contrasts showing greater overlap with the within-category range are dominated by plosive distinctions in the lexicon, while all other contrasts in this range are relatively infrequent, and the proportion of contrasts not present in the lexicon that fall within this range is also relatively small compared to those with ND distinctions above 50 ms. Thus overall, in noise duration we have a case of a cue whose weight in the lexicon can be accurately derived from a balanced inventory of contrasts in controlled syllables, which is due both to the high degree of consistency in ΔND values observed in both databases, and to the highly productive set of featural contrasts distinguished by noise duration, which allows for noise duration to exhibit good coverage over the notably asymmetric contrast distribution in the lexicon.

4.3.3.2 Distributional Disagreement: Preceding Vowel Duration

The role of preceding vowel duration is substantial in all three models, with only a modest discrepancy in cue ranking between the inventory and lexicon. Nevertheless, by accounting for the distribution of items in the lexicon this distinction is largely eliminated as the weighted inventory model is brought into complete agreement with the lexicon in terms of aggregate cue weight on $\mathrm{DUR}_{\mathrm{V1}}$. Figure 4.28 shows the partial dependence functions and distributions of preceding vowel

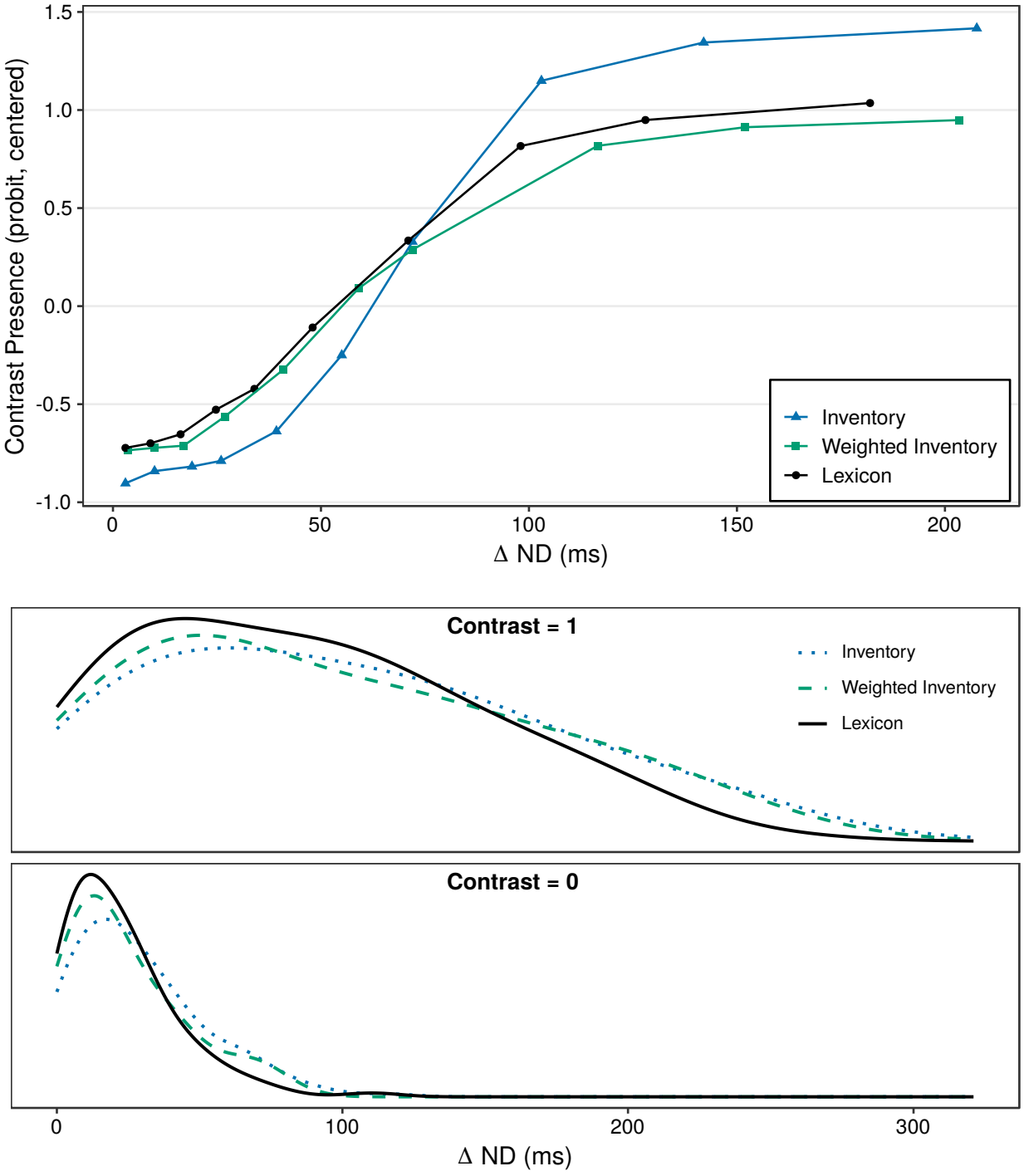

Figure 4.26: Partial dependence functions (top panel) and distributions (bottom panels) of ND in the inventory, weighted inventory, and lexicon models of ideal recognition in word-final position.

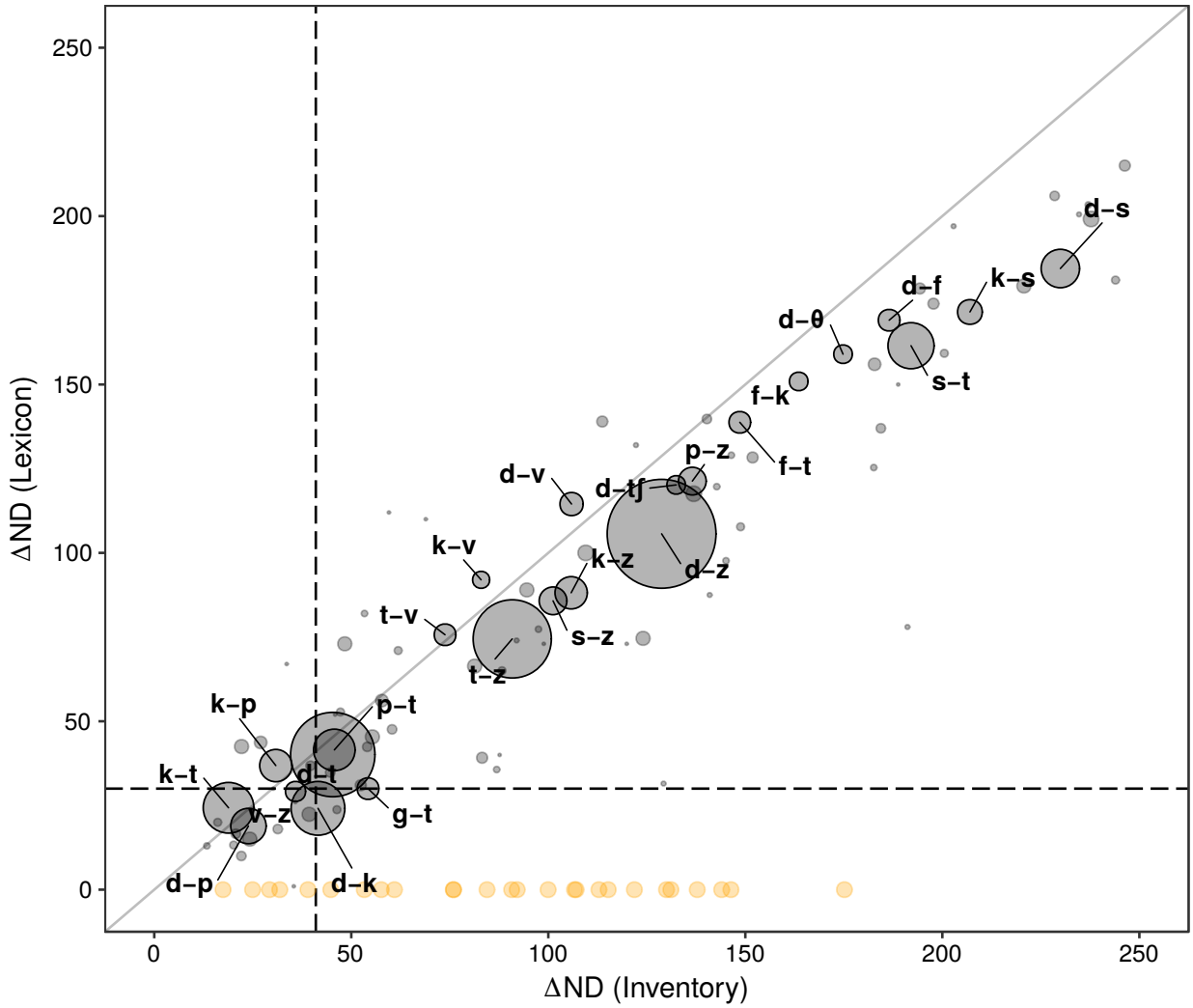

Figure 4.27: Relationship between ΔND means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-final position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 63% of items are labeled. Contrasts absent from the lexicon are shown in orange.

duration in each model. Overall, while all three models show the majority of the increase in contrast likelihood within the 30–80 ms range, the slope of the inventory model is slightly shallower than in the lexicon and weighted inventory models. The latter two models differ, however, in the location of their inflection points, occurring around 40 ms in the weighted inventory and around 60 ms in the lexicon. Further, the partial dependence function in the lexicon model continues to increase over the 75–130 ms range, while increases in contrast likelihood in the weighted inventory level out at around 65 ms.

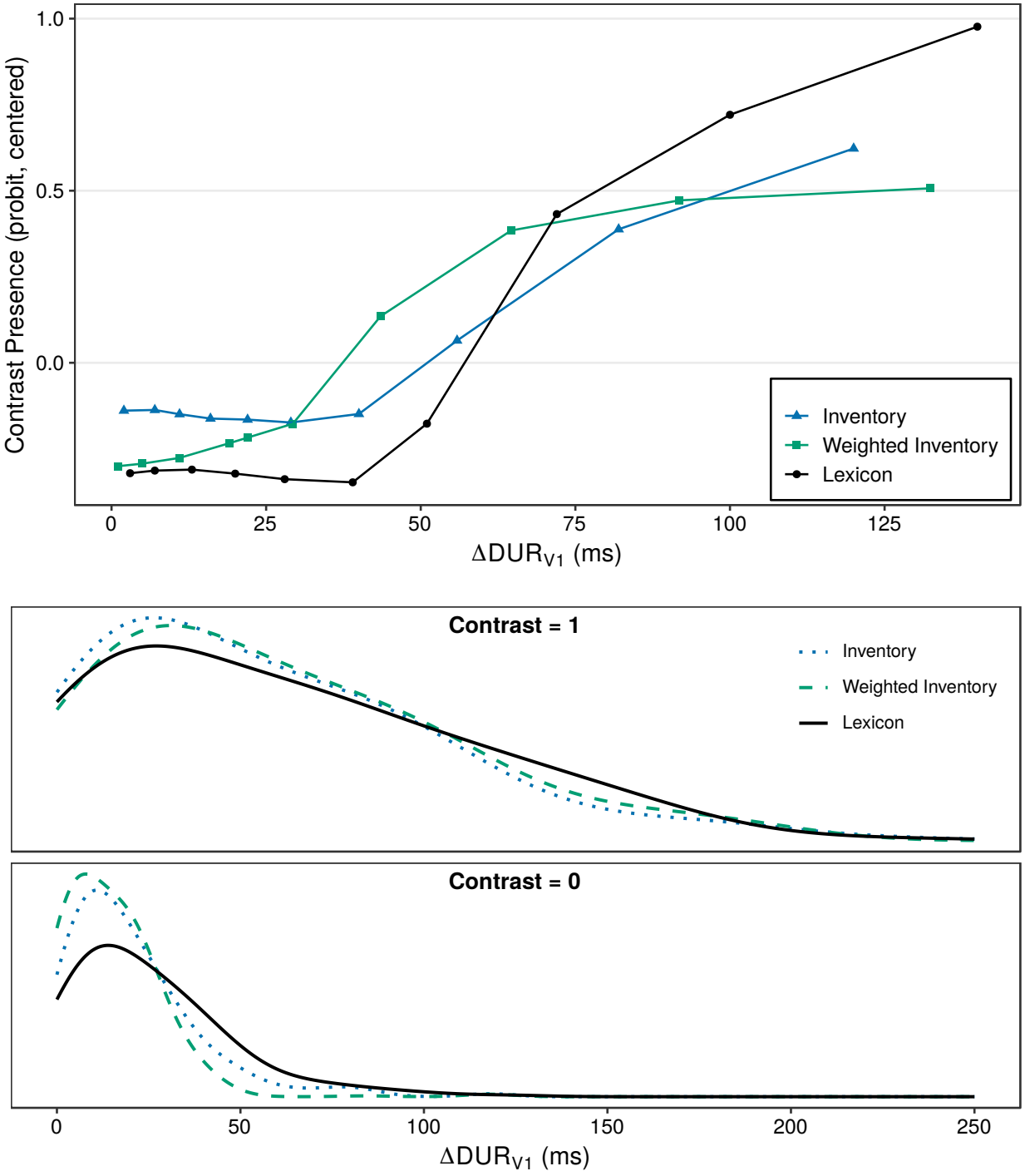

Figure 4.28: Partial dependence functions (top panel) and distributions (bottom panels) of DUR_{V1} in the inventory, weighted inventory, and lexicon models of ideal recognition in word-final position.

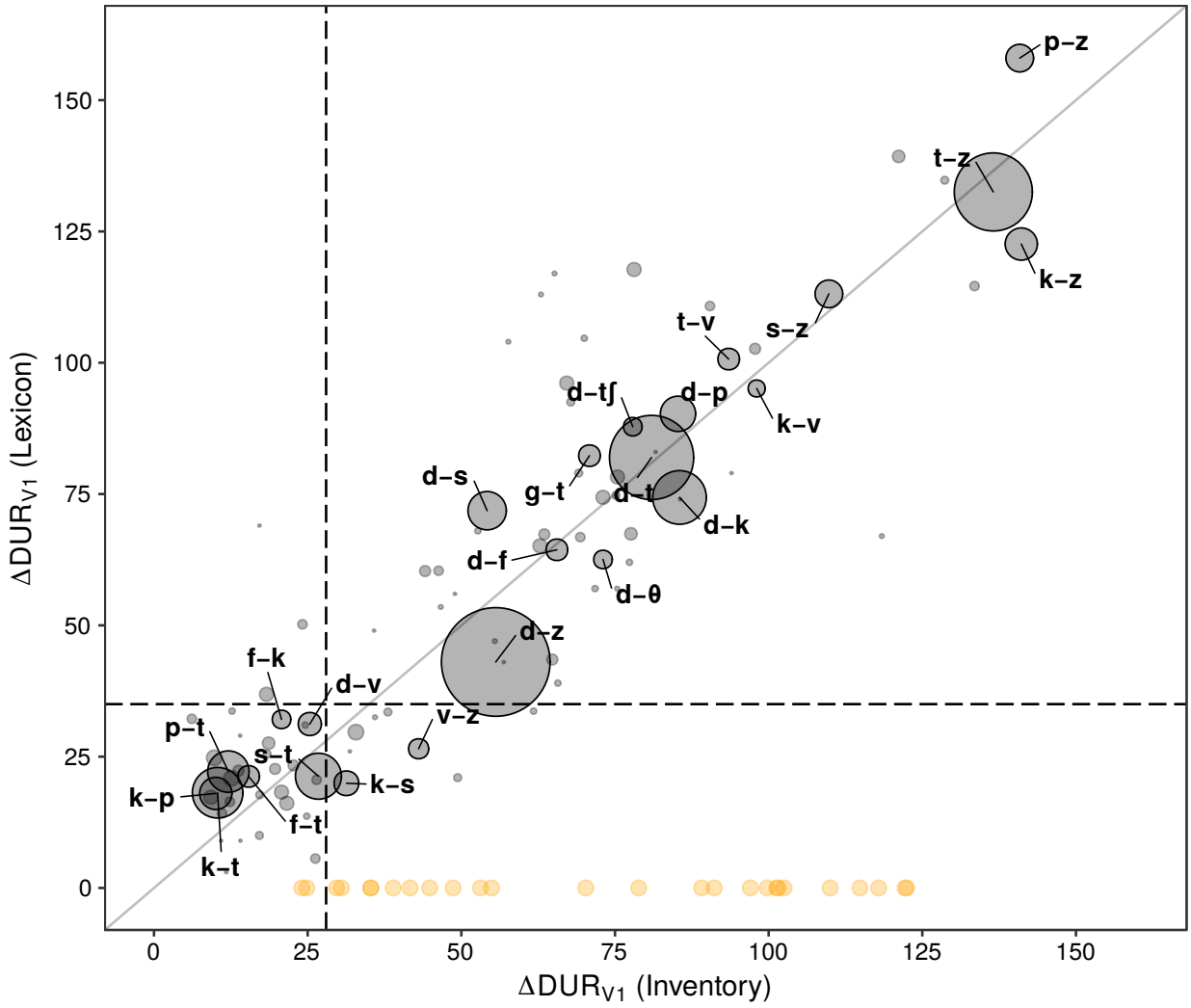

Figure 4.29: Relationship between ΔDUR_{V1} means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-final position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 63% of items are labeled. Contrasts absent from the lexicon are shown in orange.

The ΔDUR_{V1} distributions in Figure 4.28 further clarify this relationship in showing for contrastive items a slight rightward shift in the weighted inventory distribution relative to the balanced inventory, as well as a slight leftward shift in within-category DUR_{V1} distinctions in the weighted inventory. This relative enhancement of contrast in the weighted inventory is due to two results shown in Figure 4.29. First, among the most frequent contrasts, *d–z* and *t–z* (and to a lesser extent *d–k*), vowel duration differences are slightly larger in controlled syllables than in the lexicon, meaning that when such contrasts are given greater weight in the weighted inventory model the

end result is an increase in the distinction between contrastive and non-contrastive vowel duration differences relative to the inventory model. More critically for the interpretation of vowel duration as a cue to word-final voicing, however, is the relatively diminished role of within-voicing contrasts in the lexicon, as with the exception of *d–z*, most such contrasts are well within the range of within-category variation and thus offer no role for $\mathrm{DUR_{V1}}$ as a cue for discrimination. Since these contrasts are equally represented in the inventory the overall role of preceding vowel duration is downweighted relative to its role in the lexicon.

4.3.3.3 Acoustic Disagreement: Noise Amplitude

Turning next to poor scaling between the inventory and lexicon due to acoustic discrepancies between contrasts as they occur in controlled syllables and those present in real-word distinctions, noise amplitude is one such cue. Figure 4.30 shows the partial dependence functions and distributions of $\Delta\mathrm{AMP_N}$ in each model. Both inventory and lexicon models are show a minimal effect of noise amplitude distinctions on contrast likelihood, while the weighted inventory model shows a sizeable increase in contrast likelihood over much of the range, but particularly between 3 and 7 dB. The lexicon model also shows a consistent increase between 5 and 10 db, but the relation is relatively shallow compared to that in the weighted inventory. The distributions in Figure 4.30 illustrate that this erroneous upweighting of $\mathrm{AMP_N}$ in the weighted inventory model derives both from a narrowing of the within-category distribution relative to the inventory and lexicon data, and an increase in the proportion of contrasts between 5 and 7 dB, the critical threshold between contrastive and non-contrastive $\Delta\mathrm{AMP_N}$ ranges.

Figure 4.31 shows that noise amplitude is indeed poorly correlated between the inventory and lexical data, though it is not immediately clear how the distribution of $\Delta\mathrm{AMP_N}$ values among phonetic contrasts results in a relative upweighting of noise amplitude in the weighted inventory model. The majority of the contrasts in the lexicon are more distinct (i.e., above the gray identity line) in the lexicon than in the inventory, meaning that the discriminative power of noise amplitude in the weighted inventory should be *reduced* rather than enhanced. The only explanation then

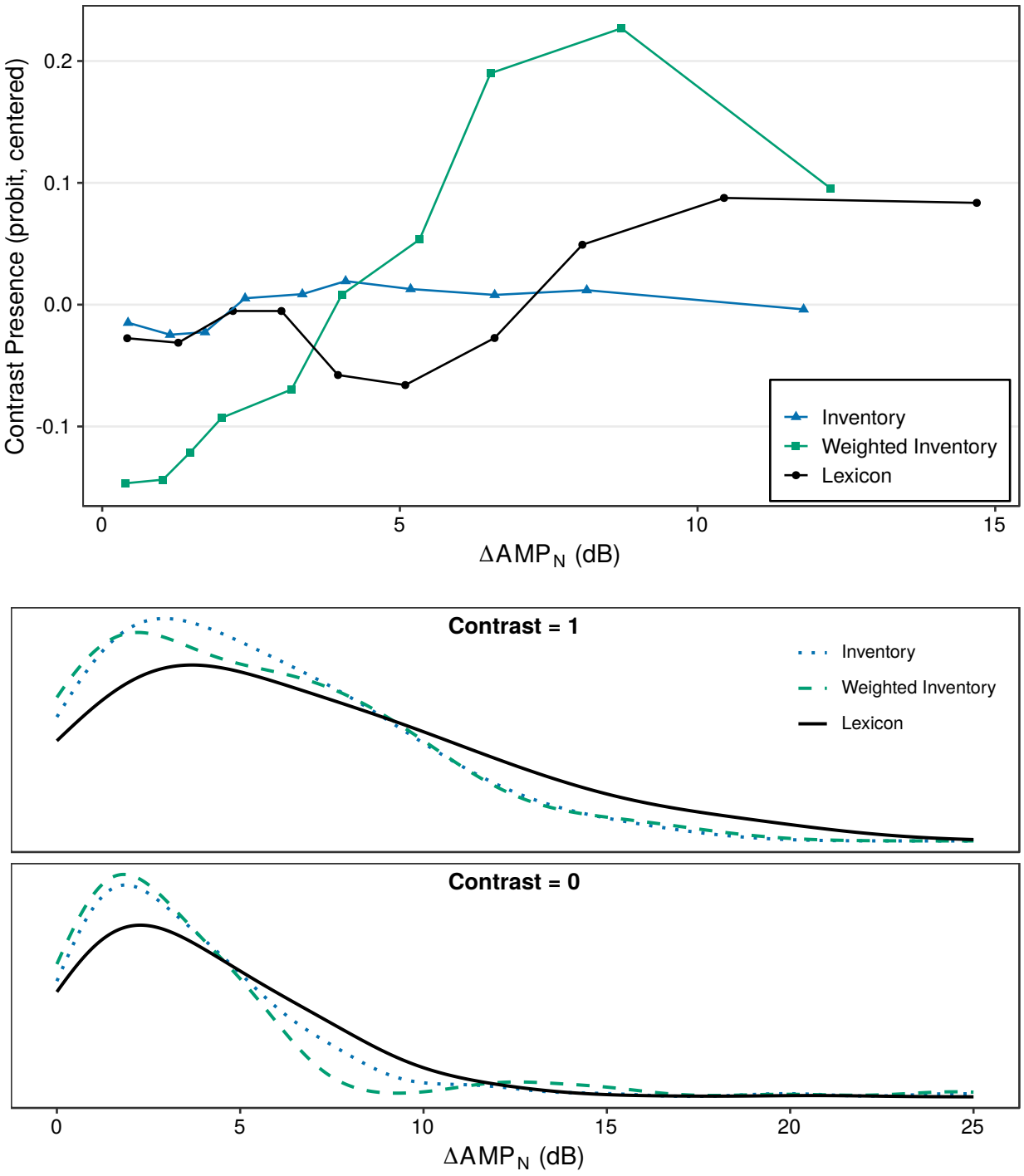

Figure 4.30: Partial dependence functions (top panel) and distributions (bottom panels) of AMP_N in the inventory, weighted inventory, and lexicon models of ideal recognition in word-final position.

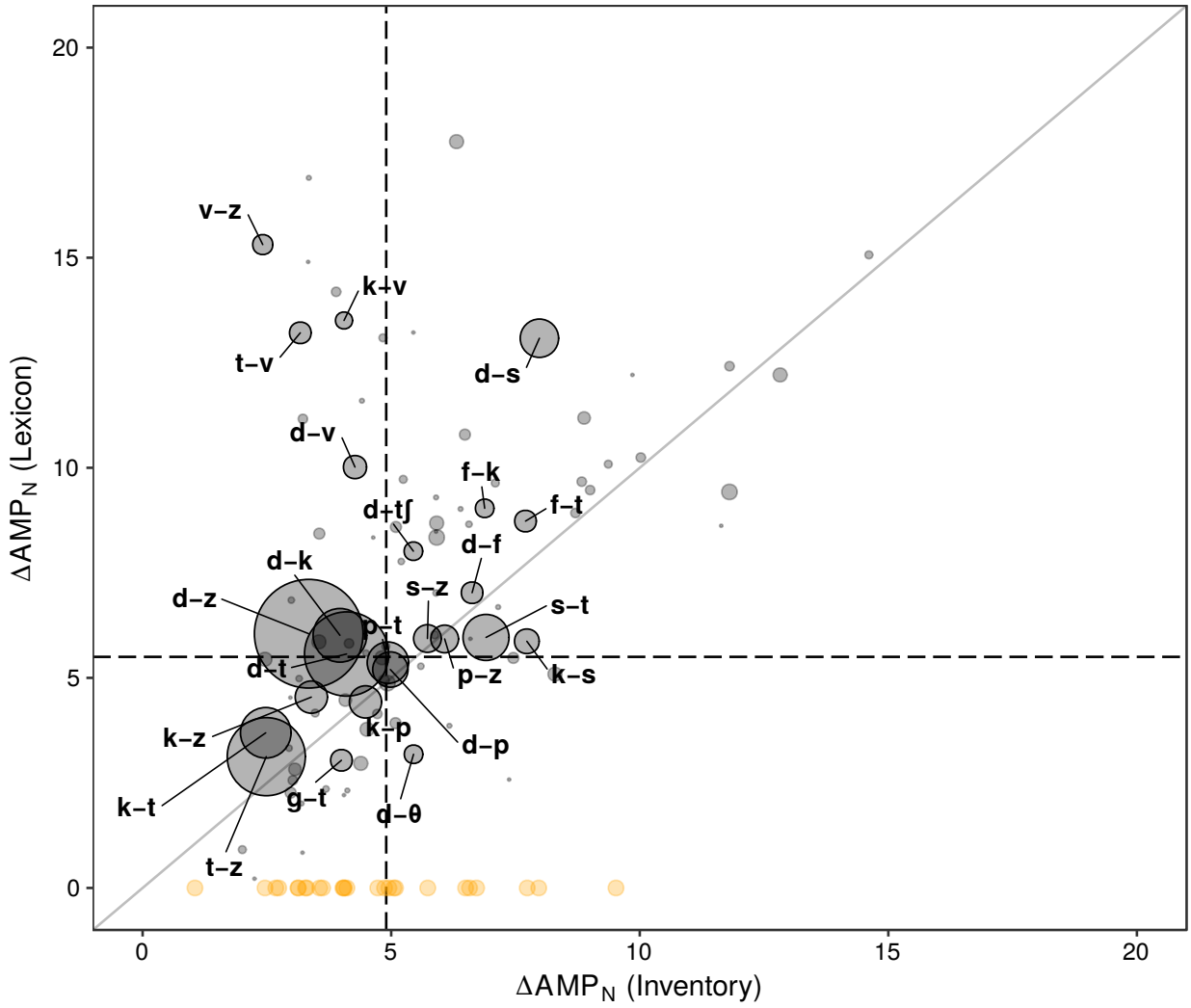

Figure 4.31: Relationship between ΔAMP_N means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-final position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 63% of items are labeled. Contrasts absent from the lexicon are shown in orange.

is that the change in cue weights between the inventory and weighted inventory is due rather to the reduction of within-category variance from the controlled syllable data, which creates a lower boundary between contrastive and non-contrastive items. This shift, combined with the narrowing of the contrastive range (note there is much greater variance along the *y*-axis than the *x*-axis), results in a more rapid jump in contrast likelihood around 5–7 dB in comparison with the more gradient shift in the lexicon. However, this is just one possibility, and given the variability in the mapping between inventory and lexicon, as well as the generally poor separation of items provided

by AMP_N, it is also possible that the weighted inventory result is driven by a confounding relation with another cue. Nevertheless, both explanations are consistent with the source of instability in weighted inventory estimates deriving in part from poor acoustic alignment between the two databases. Were the acoustic characteristics of noise amplitude distinctions in controlled syllables more similar to those in real words, the integration of noise amplitude in the weighted inventory model would have better matched that in the lexicon and yielded a lower aggregate cue ranking.

4.3.3.4 Composite Disagreement: Spectral Shape

Finally, we examine a case of composite disagreement wherein the role of spectral shape in the lexicon is substantially overestimated by both inventory models, indicating that both distributional and acoustic factors appear to be involved in generating this discrepancy. Figure 4.32 shows partial dependence functions for spectral shape in each model, as well as the ΔSHAPE distributions in the data input to each model. The inventory model shows the greatest effect of spectral shape on contrast likelihood among the three, increasing approximately linearly between 2 and 10 dB, while the weighted inventory exhibits a sharp increase between 2 and 2.5 dB, with a more gradual asymptotic increase over the remainder of the range. The relation between spectral shape distinctions and contrastiveness in the lexicon, on the other hand, is much shallower, increasing linearly between approximately 2 and 10 dB but over a much narrower range of *y* values. The source of this discrepancy can be seen in part in the distributions in Figure 4.32, wherein the contrastive set in the inventory has a much heavier tail indicative of many contrasts whose ΔSHAPE values are well above the range of within-category variance. Both lexicon and weighted inventory models, on the other hand, are more skewed, with most of their contrasts lying between 0 and 15 dB; however, the within-category range for the weighted inventory is much more concentrated at low ΔSHAPE values than the lexicon, which accounts for spectral shape retaining a role in the weighted inventory model despite the loss of many high ΔSHAPE contrasts from the inventory.

Figure 4.33 further clarifies how each aggregate distribution arises from spectral shape distinctions among particular phonetic contrasts. Recall that in Chapter 2 we showed that spectral shape

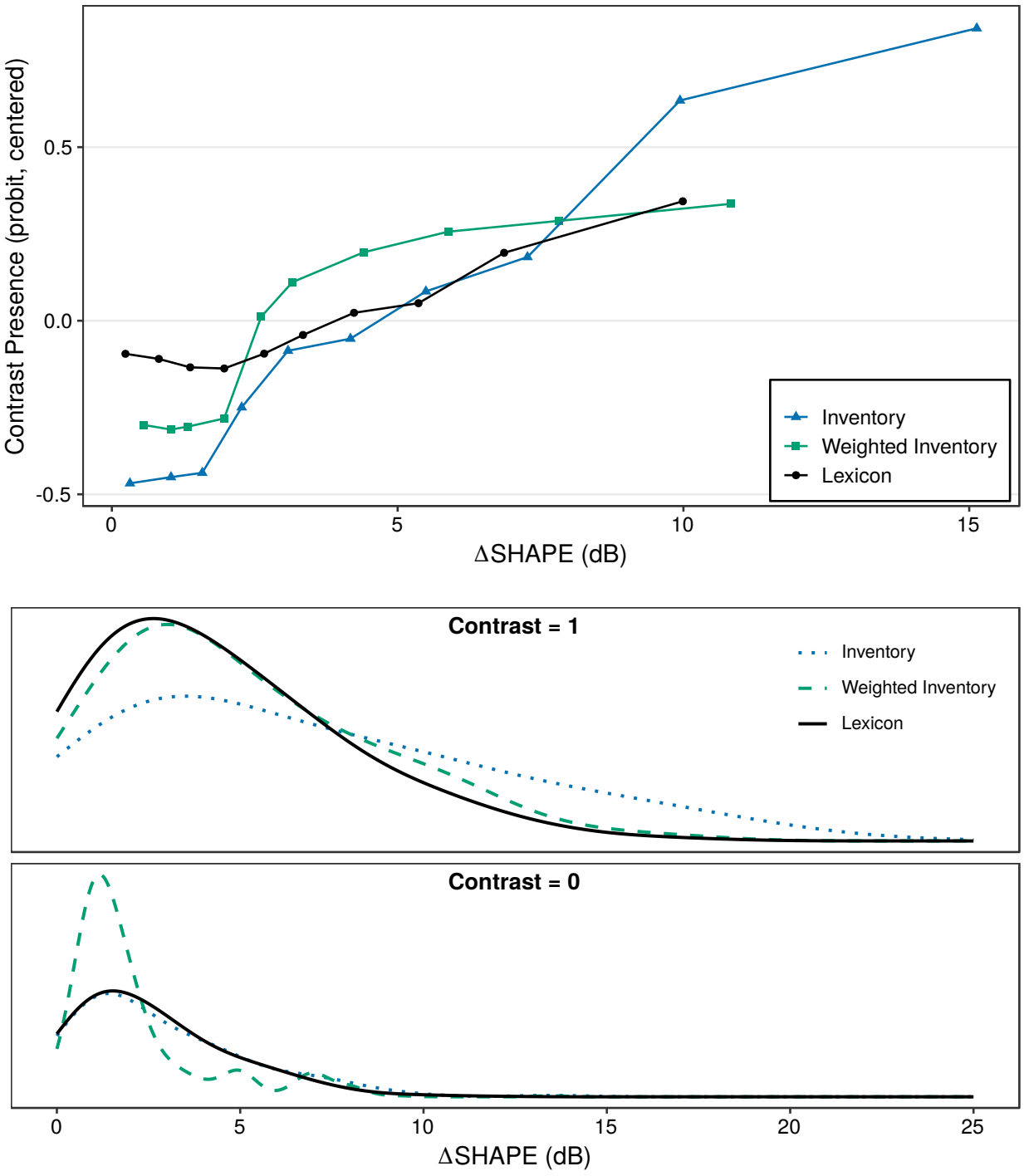

Figure 4.32: Partial dependence functions (top panel) and distributions (bottom panels) of SHAPE in the inventory, weighted inventory, and lexicon models of ideal recognition in word-final position.

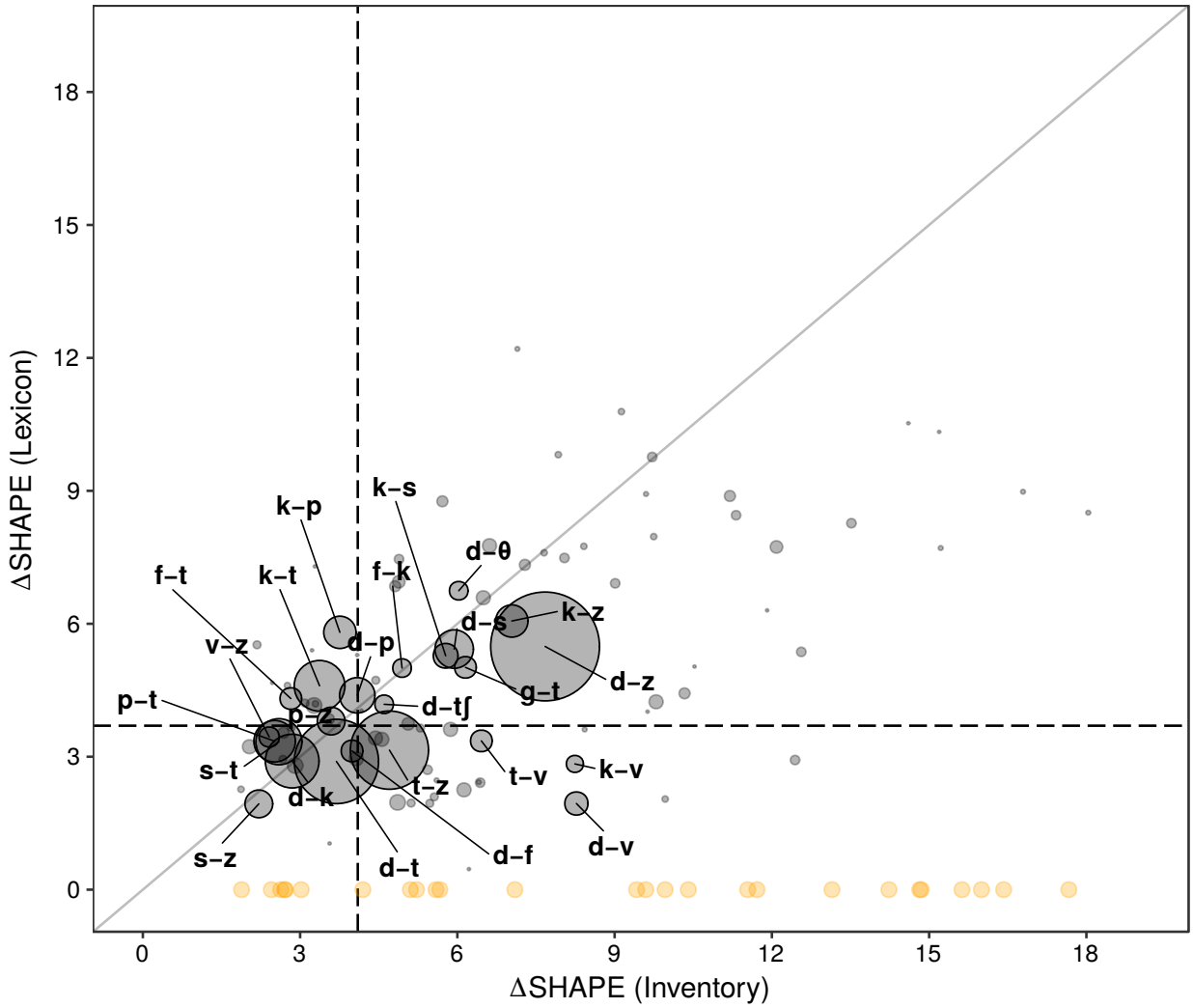

Figure 4.33: Relationship between ΔSHAPE means by phonetic contrast in the inventory and lexicon models of ideal recognition in word-final position. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Contrasts comprising the top 63% of items are labeled. Contrasts absent from the lexicon are shown in orange.

is a robust discriminator of postalveolar obstruents from non-postalveolars, a result that was partly anticipated in its original proposal in (Evers et al., 1998) as a cue to the *s–ʃ* distinction. Many such contrasts, however, are absent from the lexicon and can be seen in the distribution of orange points between above 10 dB. Further, among the contrasts which are present in the lexicon, particularly the highly frequent *d–z*, *t–z*, and *d–t* contrasts, the SHAPE cue is relatively less distinct in real words than in controlled syllables, meaning that even when the lexical distribution is accounted for in the weighted inventory model, the role of spectral shape in the lexicon is overestimated.

4.3.4 Discussion

In analyzing cue integration within an ideal perceiver framework, we were able to explore the encoding potential of a wide range of cues when optimally weighted for the discrimination of obstruent consonants within a given system. Across positions, the most influential cues in the lexicon were F2 at vowel onset/offset, consonant voicing percentage, consonantal spectral tilt, and spectral peak frequency. Many of these cues were also highly weighted in the inventory model, while cues such as noise duration and spectral shape were more consistently upweighted relative to the lexicon model. Vocalic cues, on the other hand, particularly f0, F3, and vowel-midpoint F2, were generally of little utility, either in the lexicon or inventory systems.

In each position-specific model, points of agreement and disagreement between the inventory, weighted inventory, and lexicon models were identified, with the latter then classified into *distributional*, *acoustic*, and *composite* disagreement types. Distributional cue disagreements included the use of low-frequency energy word-initially, spectral peak amplitude word-medially, and preceding vowel duration word-finally, and reflected the relatively greater prevalence in the lexicon of voicing contrasts word-initially, sibilance contrasts word-medially, and voicing/manner distinctions word-finally. In all such cases the acoustics of the inventory and lexicon data are closely correlated, but the contrasts in which they occur are relatively better discriminated in the lexicon than the inventory, leading to an aggregate underestimation by the inventory model of cue utility in the lexicon which can largely be overcome in the weighted inventory model by sampling the inventory data to match the distribution of contrasts in the lexicon.

Acoustic sources of disagreement had less to do with distributional differences in the two systems, and more to do with the relative compatibility between inventory and lexicon acoustics. As a result, cue ranks in the weighted inventory model are brought out of alignment with the lexicon. For example, word-initially, consonant voicing percentages are poorly correlated between the inventory and lexicon, showing generally greater voicing differences in the lexicon than in the inventory, but largely uncorrelated by contrast. This leads to a relative downweighting of VOI% in the weighted inventory model. Word-medially, a similar result is obtained for voice cessation time

(VCT), which is relatively less discriminative of voicing contrasts with the alveolar flap [ɾ] in the inventory data than in the lexicon. Finally, among VC contrasts noise amplitude distinctions are poorly correlated between the inventory and lexicon, and thus cue estimates scale poorly via the weighted inventory model, overestimating the role of AMP_N and the range over which it is most discriminative.

Finally, several points of composite disagreement were identified, including dynamic amplitude, relative F3 amplitude, and spectral shape. In all such cases there are both distributional and acoustic discrepancies between the two systems that result in poor scaling relations. This taxonomy of sources of cue disagreement serves primarily to account for how the behavior of the system depends on both acoustic and phonological distributions, and to provide some means of linking the inventory and lexicon models in an explanatory way. This procedure is critical to understanding how the inventory assumption impacts our understanding of cue integration, and how previous findings from the canonical study of phonetic systems in the literature can be incorporated into this new lexical framework in the future. However, one problem with the ideal perceiver models is that they can only distinguish between contrastive and noncontrastive items. Distinctions in acoustic salience within the contrastive set can only be distinguished by incorporating data from listener perception, which we examine in the next section.

4.4 Cue integration in listener recognition

In Section 4.3 we sought to predict contrast discrimination under ideal recognition conditions; i.e., assuming contrasts are perfectly distinguished and thereby estimating the relative utility of each acoustic cue in distinguishing that set. In the present section the goal is to use the same acoustic data to predict listener recognition behavior, and thereby arrive at estimates of cue weights that more closely reflect the integration of acoustic information in speech perception. Further, one of the challenges noted in the discussion of the ideal perceiver models was the fact that in many instances the model was unable to distinguish between contrasts that are well beyond the within- vs. cross-category threshold along a particular cue dimension. However, the listener models, in

tracking the relative perceptibility of each contrast, can capture these differences and thus provide more gradient information on the role of each cue in the obstruent system. Finally, as with the ideal perceiver models, cue-integration results for the listener models are presented separately by contrast position (CV, VCV, VC), where within each position general cue ranking distributions are first reviewed (for both *target* and *contrast* parameters), followed by correlations and cue-ranking differences between the lexicon and inventory/weighted-inventory models, and lastly detailed analyses of points of agreement and disagreement between the three models in terms of the relative ranking of cues based on their contrast parameterization (i.e., the Δ-parameters as in Section 4.3), where the latter is divided into *distributional*, *acoustic*, and *composite* types.

4.4.1 Word-initial position (CV)

Beginning with models of cue integration at word/syllable onset, Figure 4.34 shows the model fit for the inventory, weighted inventory, and lexicon models, as well as the reference model based on acoustic measurements from stimuli in the Woods et al. (2010) study from which estimates of controlled syllable recognition patterns were derived for use the inventory model. Here we see that all models overestimate listener accuracy, a consequence of the greater concentration of data in the upper accuracy range than in the lower range. Nevertheless, each model shows a generally monotonic relationship over most of the accuracy range, with the accuracy of the fit generally declining for lower accuracy items. Overall, the root-mean-squared error (RMSE) of the models (measured on the probit scale) was 0.258, 0.309, 0.419, and 0.355 in the inventory, reference, weighted inventory, and lexicon models. These fits indicate that all three models are able to account for the general rank-ordering of contrast in terms of their relative discriminability by listeners, and therefore the parameter weights derived from such models can be held to be a viable approximation to the relative importance of different acoustic cues in speech perception. Further, the fact that the inventory model provides as good of a fit to the perception data from Woods et al. (2010) as the reference model based on the stimuli used in that study, as well as the significant correlations in both target ($r = 0.563$, $p = 0.005$) and contrast ($r = 0.628$, $p = 0.001$)

parameter ranks, validates our use of the controlled syllable productions from the target speaker in the evaluation of cue integration in the inventory model.

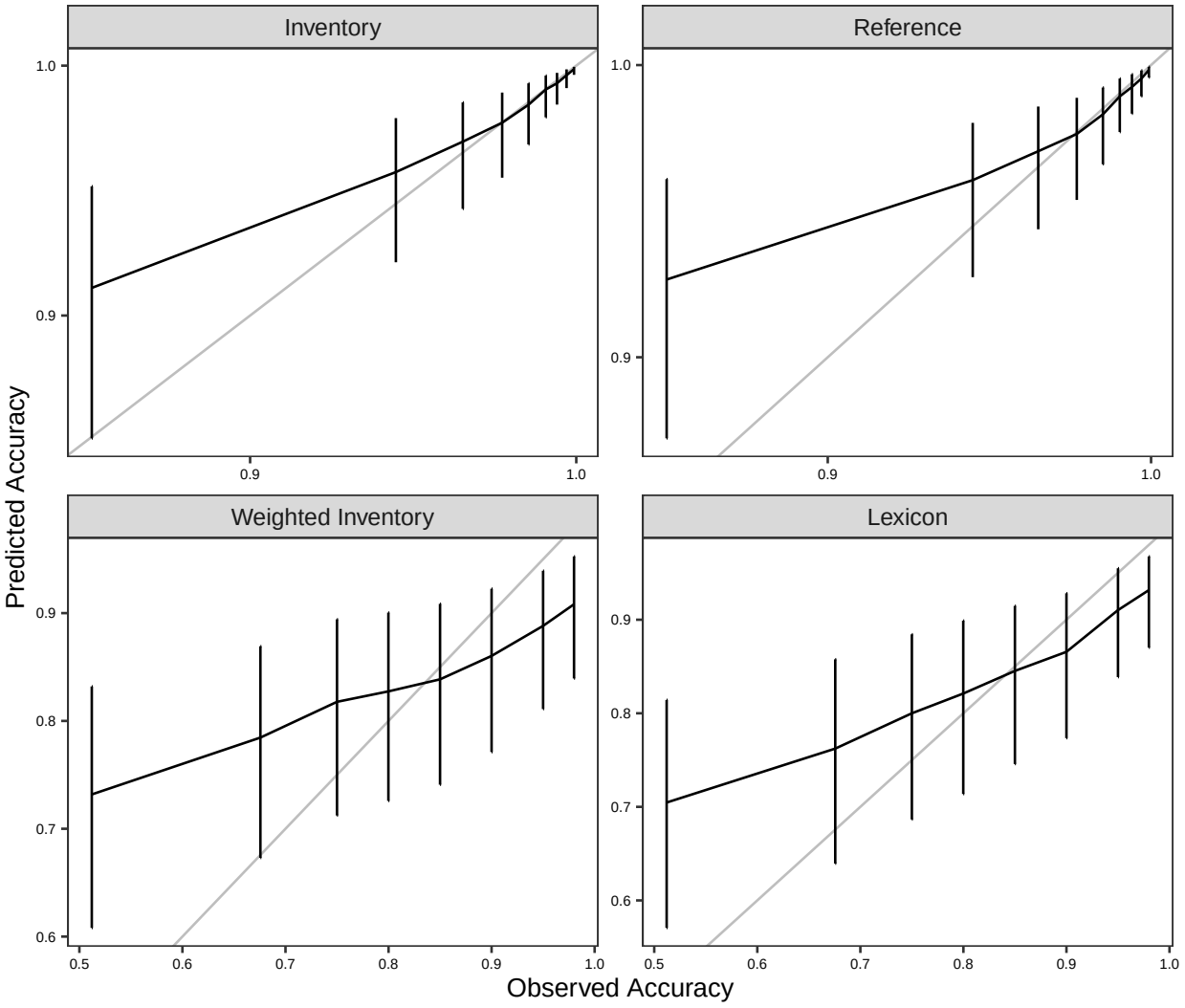

Figure 4.34: Listener model fit in the inventory, reference, weighted inventory, and lexicon models of word-initial contrasts. Lines indicate posterior medians averaged at the observed quantiles (0, 0.1, ..., 1). Error bars indicate the middle 90% of the posterior distribution. Solid grey lines show the identity function.

Given the general monotonic form of the predictions in each model, a meta-modeling approach was pursued wherein a secondary linear model was fit between the predictions of the BART model, $\hat{y_0}$, and the observed outcome y; i.e., $\hat{y_1} = \beta + \alpha\hat{y_0}$. This approach has the effect of transforming the predicted values to account for the fact that the cue-integration model tends to overestimate listener accuracy, particularly at the lower end of the accuracy range, as well as underestimating listener

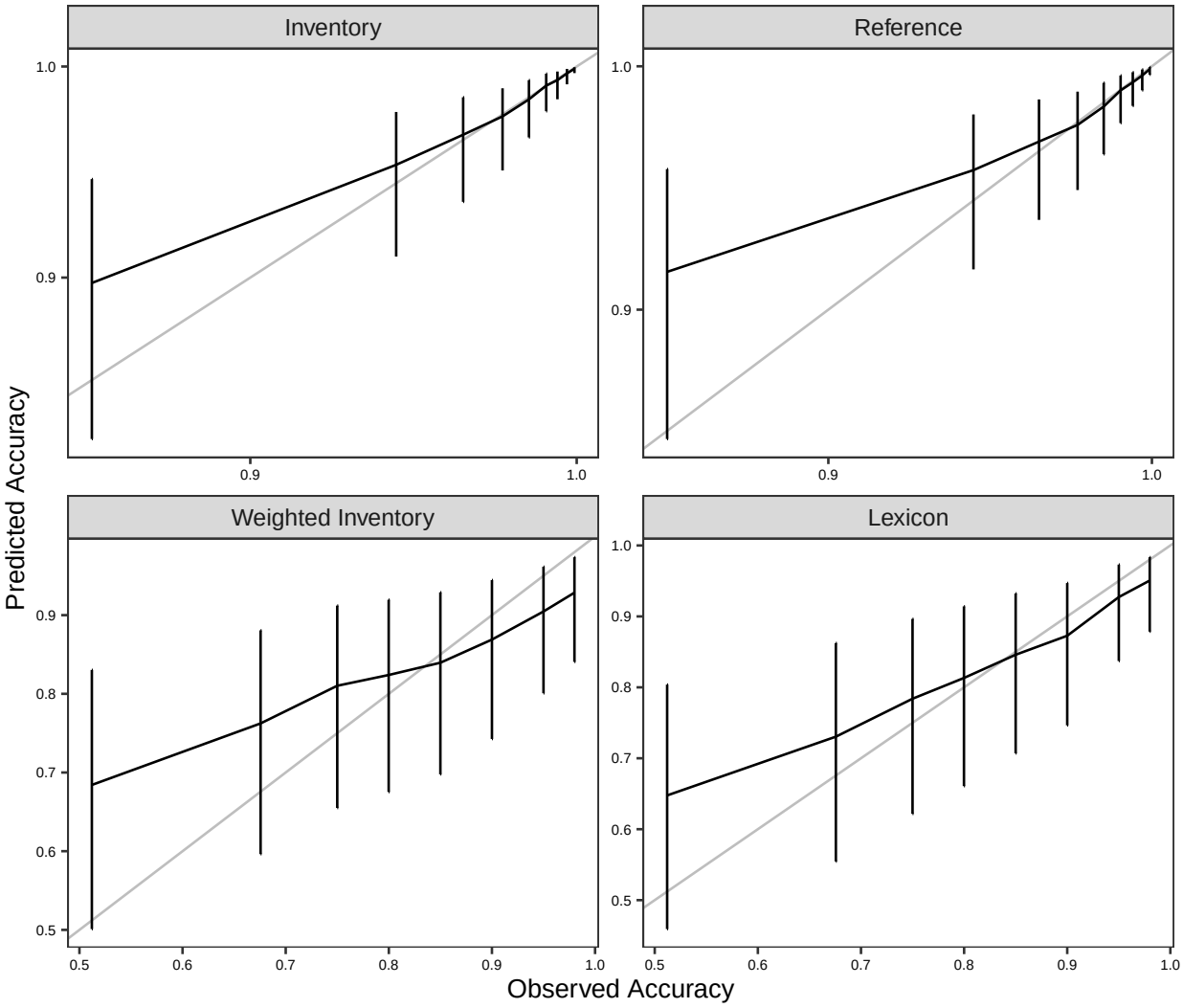

Figure 4.35: Transformed listener model fit based on a meta-model of the inventory, reference, weighted inventory, and lexicon predictions of word-initial contrasts. Lines indicate posterior medians averaged at the observed quantiles (0, 0.1, ..., 1). Error bars indicate the middle 90% of the posterior distribution. Solid grey lines show the identity function.

accuracy for items approaching ceiling performance in the weighted inventory and lexicon models (see Figure 4.35 for the meta-model fits). By applying this transformation we do not affect any cue weights, but we improve the accuracy of the model predictions (by approximately 3% in the inventory models and over 5% in the lexicon model), predictions which are used in the simulation of cue perturbation in Chapter 5.

Acoustic cue ranks from the three models of CV contrast recognition are shown in Figures 4.36 and 4.37, where Figure 4.36 shows results for the target cues, and Figure 4.37 results for

the contrast cues. In the detailed analysis of cue agreement and disagreement between the three models we focus on the contrast parameters (just as in the analysis of ideal perceiver models in the previous section), but first we assess the relative utility of different acoustic characteristics of the target item in order to discern which properties of the acoustic signal generally correspond with relatively good or poor recognition. Among the most critical target cues in the lexicon model in Figure 4.36 that are consistently ranked in the top third of cues in both Experiment 1a and 1b (see Figures A.73 and A.74 in the appendix for details) are noise amplitude (AMP_N), spectral peak frequency ($FREQ_{PK}$), consonant voicing percentage (VOI%), spectral peak amplitude (AMP_{PK}), and spectral shape (SHAPE).

The two amplitudinal parameters, $AMP_{N/PK}$, perform as expected in yielding higher predicted discrimination accuracy at higher amplitudes, where the threshold for noise amplitude after which recognition rates begin to change is around 56 dB, and the threshold for spectral peak amplitude is around 23 dB. Both cues are similarly highly ranked in the inventory, but with the opposite directionality, while the weighted inventory model mirrors the lexical pattern for both noise and spectral peak amplitude, though the latter shows much closer agreement between the two. This result in the inventory model—that louder target obstruents are more poorly recognized than comparatively quieter ones—appears to be counterintuitive, but closer examination of the weighted inventory results and the acoustic description of each cue in Figures 2.26 and 2.38 in Chapter 2 reveals that this result is not due to acoustic differences between the two datasets, but rather reflects differences in the configuration of contrasts in the two distributions. The threshold for noise amplitude, which is also the point at which predicted accuracy begins to decline in the inventory model, largely separates sibilants and voiced obstruents from voiceless plosives and voiceless nonsibilant fricatives. Each of these sets exhibit errors in their respective perception data which primarily occur within the set, which means that in a balanced inventory of contrasts the number of within-set distinctions is larger in the former than in the latter; i.e., there are fewer voiceless plosives and voiceless nonsibilant fricatives than sibilant/voiced obstruents. However, in the lexicon the frequency of voiceless plosive contrasts increases, while the frequency of contrasts between sibilants, as well as the over-

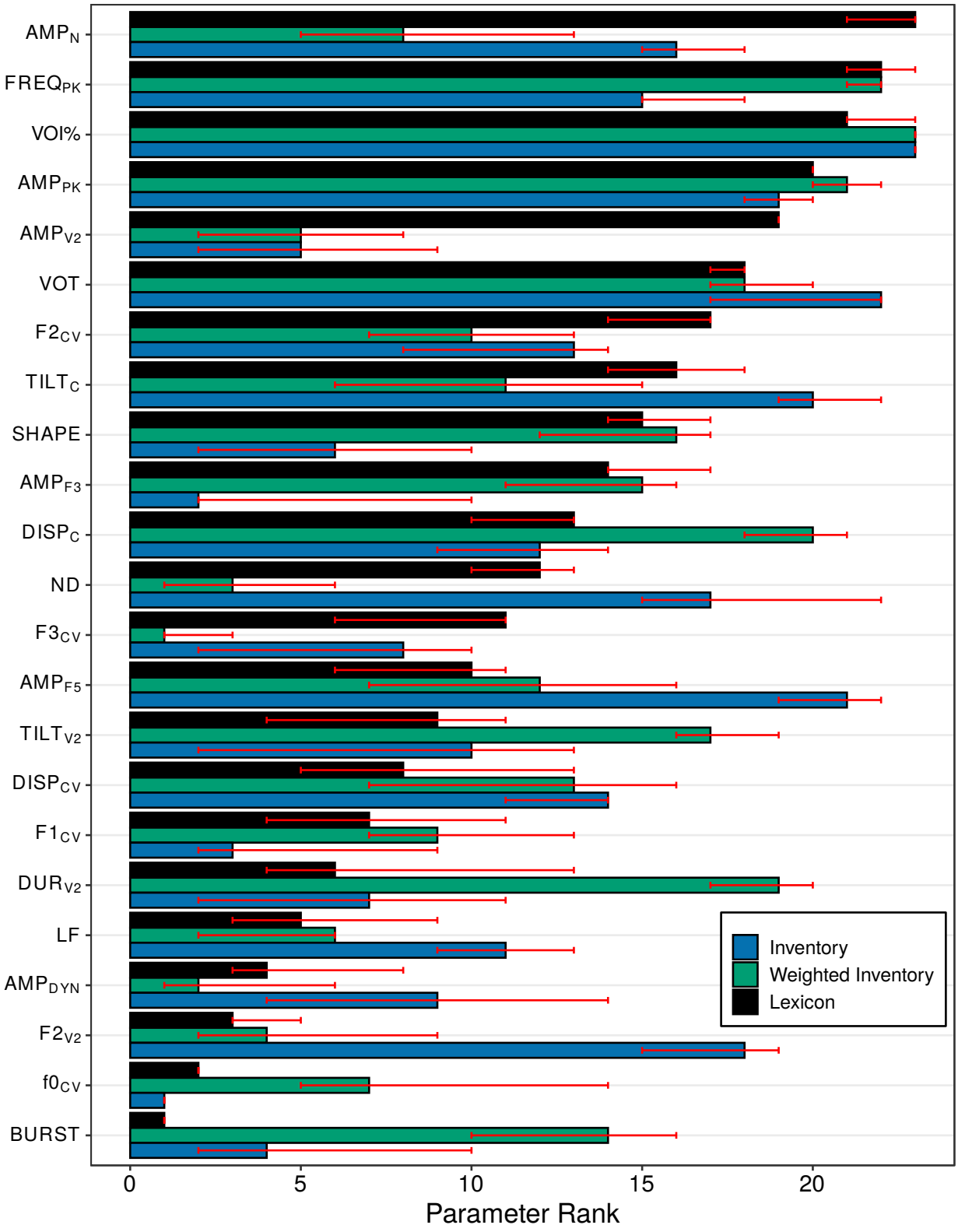

Figure 4.36: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-initial contrasts fit to listener recognition data. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

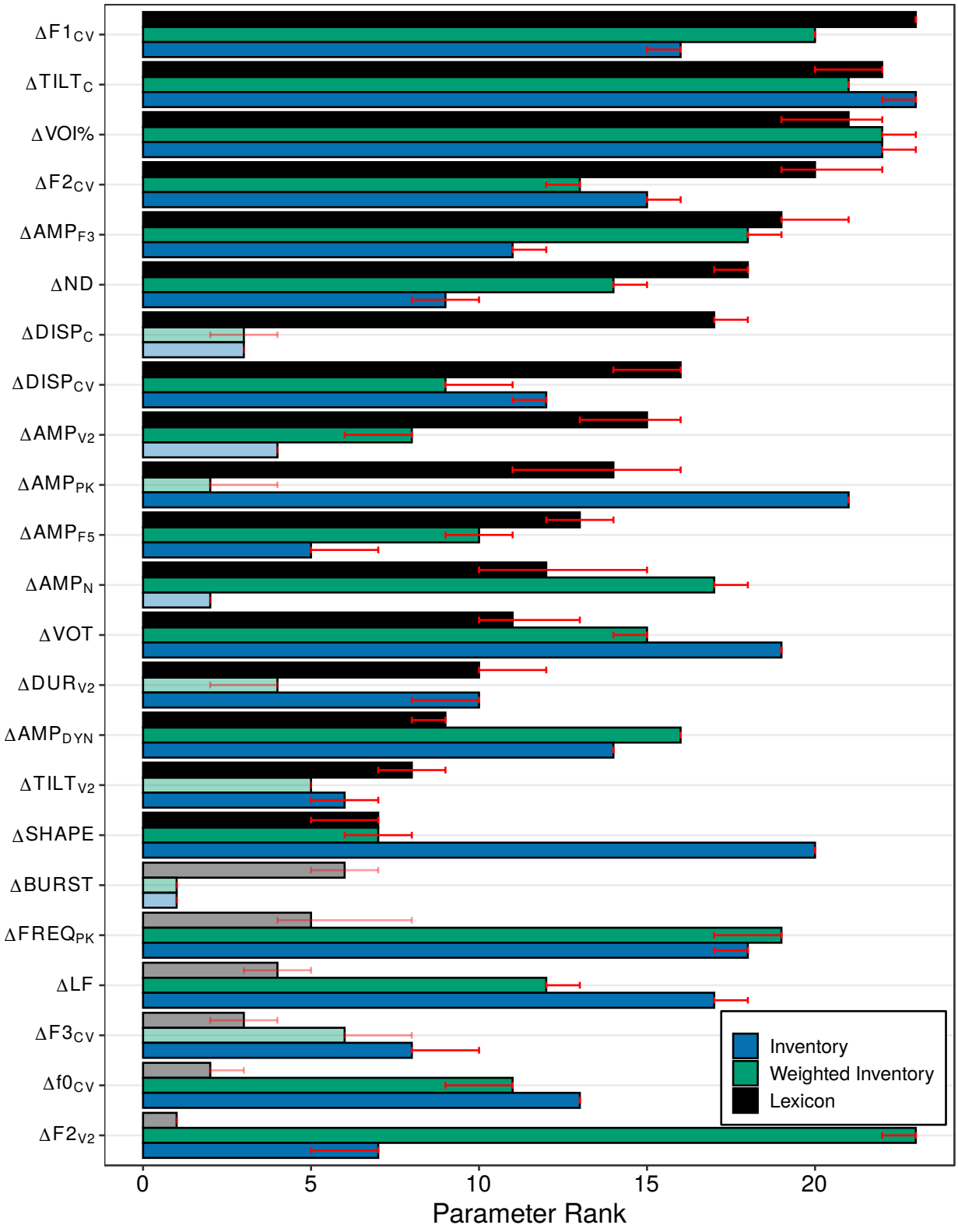

Figure 4.37: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-initial contrasts fit to listener recognition data. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

all frequency of voiced obstruents word-initially, declines in the lexicon. Therefore, this change in contrast distributions has the effect of reversing the role of target noise amplitude in the lexicon relative to the inventory.

Similar results are obtained for spectral peak amplitude, though for AMP_{PK} the threshold primarily distinguishes sibilants and voiceless from voiced plosives and nonsibilant fricatives, which again disadvantages the inventory relative to the lexicon as the former set is relatively larger, while in the lexicon the general sparsity of nonsibilants forces the model to further distinguish between voiced plosives, voiceless plosives, and sibilants (particularly [s]), which are all highly frequent in word-initial minimal pairs. This distributional difference is also why the change in predicted accuracy from increases in AMP_{PK} increases more gradually and with a later inflection point in the lexicon (28 dB) than in the inventory (25 dB). This distributional explanation is also consistent with the greater agreement between the weighted inventory and lexicon models, though we must acknowledge that listeners may also differ in their reliance on amplitudinal information when perceiving real words embedded in multi-talker babble as compared with controlled syllables due to differences in the acoustic and informational similarity between the signal and noise.

The two spectral parameters, $FREQ_{PK}$ and SHAPE, show greater conformity between the three models. Regarding spectral peak frequencies, listeners are predicted to be most accurate on obstruents with spectral peaks in the mid-frequency range ($\sim$1500–3000 Hz), largely corresponding to velar plosives and the voiced alveolar plosive [d], while over the range above 3000 Hz there is a moderate increase in predicted accuracy with increasing frequency, capturing the influence of alveolar sibilants and the relatively low accuracy on postalveolars. However, predicted accuracy over this range never reaches the level of the mid-frequency range. Finally, listener accuracy on obstruents in the low-frequency range (below 1500 Hz), which largely reflects labials and dental fricatives, is predicted to decline in all three models with increasingly lower peak frequencies, though this decline is greater in the weighted inventory and lexicon than in the inventory. Regarding spectral shape, all three models show a sharp decline in predicted accuracy between 0 and 5 dB, a result which is due to the generally poor recognition of postalveolar obstruents word-initially,

which are the only obstrents with spectral shapes above this threshold.

Finally, both the inventory and lexicon models show a consistent relation between VOI% and predicted accuracy wherein obstruents in the range of 10–20% (largely [p, t, k, ʧ, h]) are predicted to be more accurately perceived than the remainder of the set. The weighted inventory model shows a similar predicted advantage for obstruents with lower voicing percentages, but the threshold is shifted lower by approximately 5%. This result is due primarily to the influence of the voiceless postalveolar affricate [ʧ], which is 5% lower VOI% in the lexicon than the inventory, which places it in the range of [f] and [s], two highly frequent and accurately perceived obstruents. Given that [ʧ] is one of the least accurate target obstruents in the lexicon (see Figure 3.2 in Chapter 3), its influence in this set causes a reduction in predicted accuracy over the low VOI% range. This means that a discrepancy in the acoustics of controlled syllables versus real words causes a shift in the aggregate role of VOI% in the weighted inventory when the former is used to predict perception of the latter. Lastly, similar points of agreement and discrepancy are observed among the low-ranked target cues in the lexicon, among them burst presence (BURST), vowel-onset f0 ($f0_{CV}$), F2 at the midpoint of the following vowel ($F2_{V2}$), dynamic amplitude (AMP_{DYN}), and low-frequency energy (LF). However, there is much greater variance among low-ranked cues, and only burst presence is consistently uninformative in the lexicon in both Experiment 1a and 1b (see Figures A.73 and A.74 in the appendix for details).

Turning next to contrast cues, which are the focus of the remainder of this section, Figure 4.37 shows the distribution of cue ranks for the contrast parameters in each model, while Figures 4.38 and 4.39 show rank correlations and rank differences between the lexicon model and the inventory models. Overall, the cue ranks in both the inventory and weighted inventory models are moderately correlated with those in the lexicon model, though notably less so than in the ideal perceiver models, a result which is largely due to the added perceptual layer of differentiation between the two sets (in addition to the aforementioned acoustic and distributional differences discussed in Section 4.3).

Regarding cue-rank discrepancies between the three models, Figure 4.39 shows that there are

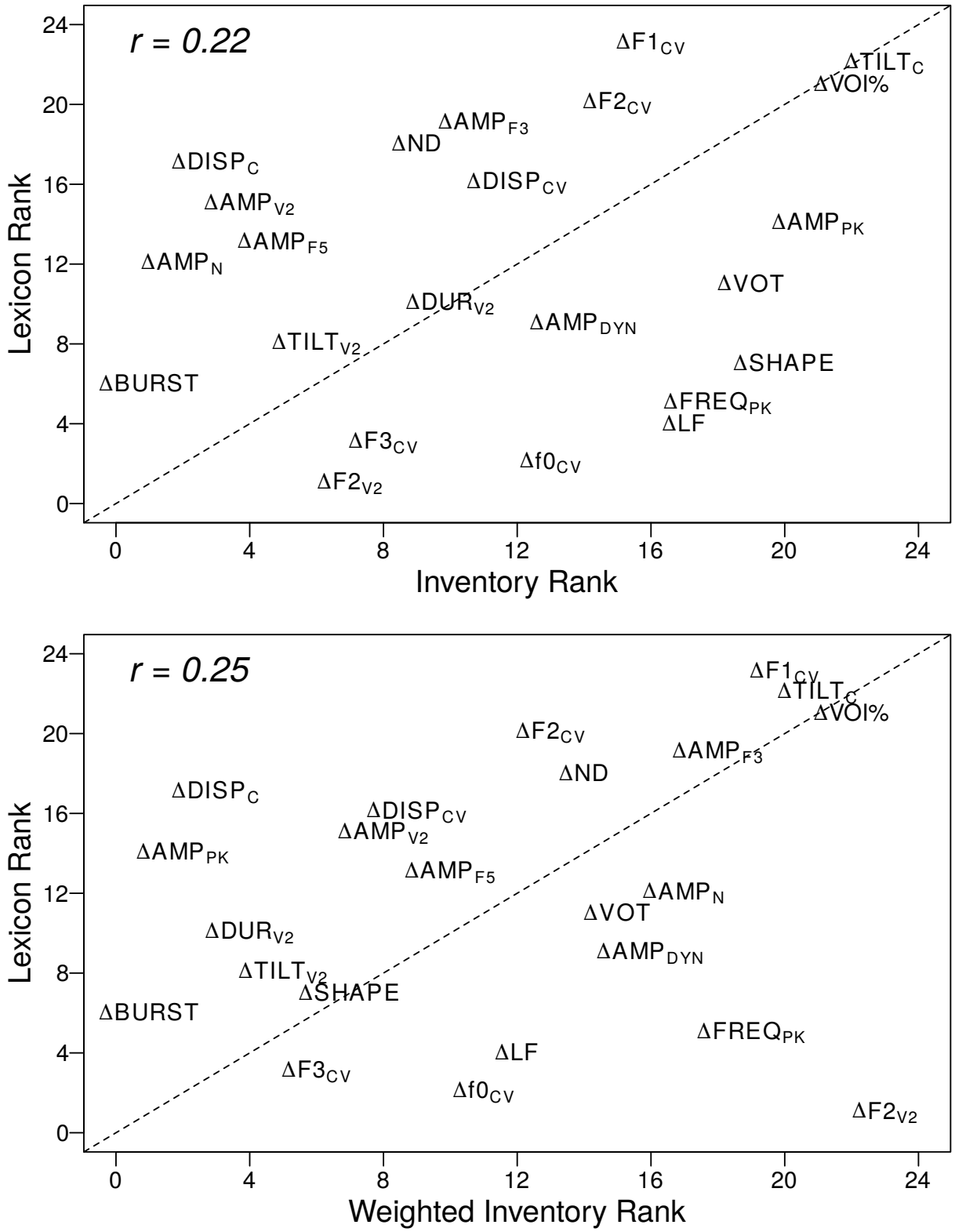

Figure 4.38: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in CV position. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

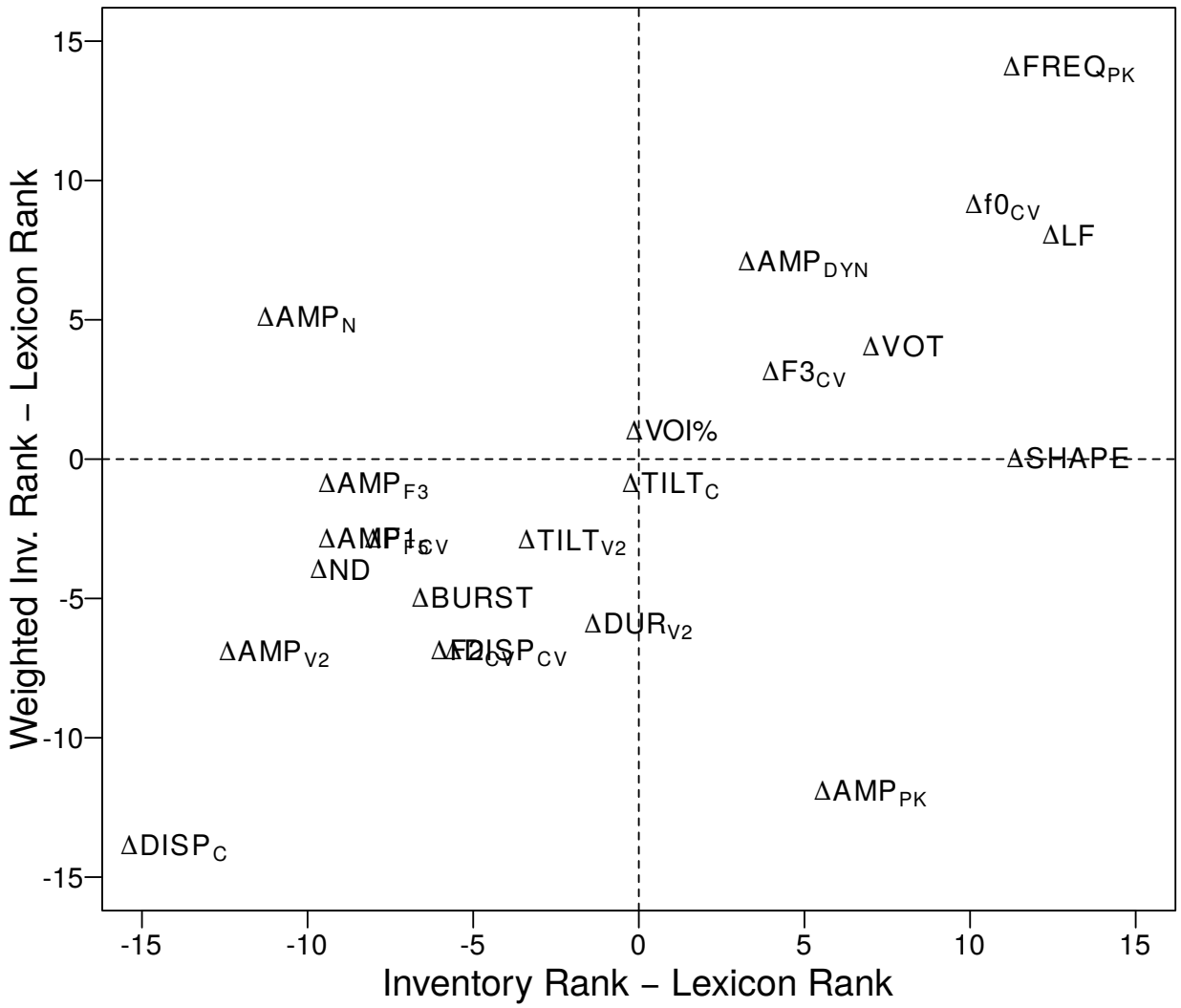

Figure 4.39: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in CV position. Dashed lines indicate equivalence relations between each pair of models. $F2_{V2}$ was excluded for visual clarity; its coordinates are (6, 22).

relatively few cues exhibiting close agreement between the inventory, weighted inventory, and lexicon, among them VOI% and $TILT_C$, and to a somewhat lesser extent $TILT_{V2}$ and $F3_{CV}$, though the cue in this set that is most comparable in agreement across the two sub-experiments in $TILT_C$. $TILT_{V2}$ and $F3_{CV}$ are also behave similarly in Experiments 1a and 1b, though with greater fluctuation in directionality, while VOI% shows close agreement between all three models in Experiment 1b, but is highly overestimated by the inventory models in Experiment 1a, where it is the lowest-ranked cue in the lexicon model. In terms of points of potential distributional agreement, spectral

shape is highly overestimated in the inventory model but is brought into good agreement with the lexicon once contrast distributions in the latter are accounted for in the weighted inventory model, while AMP_{F3} shows the opposite pattern: substantial underestimation by the inventory model but close agreement between the lexicon and the weighted inventory. AMP_{F5} and $F1_{CV}$ are also similarly underestimated by the inventory in terms of their role in the lexicon, but with slightly less agreement from the weighted inventory relative to AMP_{F3}. Among this set, the role of AMP_{F3} in the three models is best replicated across Experiments 1a and 1b, while the consistency of $F1_{CV}$ is somewhat lower, and SHAPE and AMP_{F5} are the least consistent, though the directionality of all four cues is the same in each sub-experiment.

Acoustic discrepancies are much more variable with few good exemplars, but DUR_{V2} and $F2_{V2}$ are two such cues that show good agreement between the inventory and lexicon both overall in Experiment 1, and separately in Experiments 1a and 1b, the former being consistently underestimated in the weighted inventory, and the latter consistently overestimated. Finally, as implied by the relatively low correlations between cue ranks in the three models in Figure 4.38, there are many cues that exhibit potential composite points of disagreement. Among the cues whose role is overestimated by both inventory models, the only cue which is consistent across sub-experiments is LF, though the weighted inventory model in Experiment 1a brings this cue into much closer agreement with the lexicon than in Experiment 1b. All other cues in the upper right quadrant of Figure 4.39 are highly variable across sub-experiments. In terms of cues which are underestimated by both inventory models, the only cues that consistently exhibit this relation are $DISP_C$ and BURST, though in general cues whose role in the lexicon is underestimated (i.e., those in the lower left quadrant) retain this role better than the cues whose role is overestimated in the inventory models.

In the sections below we investigate in greater detail the behavior of four cues which best exemplify points of agreement and disagreement between the inventory and lexicon. These cues are the following: spectral tilt of the consonant ($TILT_C$; cue *agreement*), relative F3 amplitude (AMP_{F3}; *distributional disagreement*), following vowel duration (DUR_{V2}; *acoustic disagreement*), and spectral dispersion of the consonant ($DISP_C$; *composite disagreement*). Finally, we should

emphasize that there is an important perceptual component potentially involved in all comparisons between the inventory and weighted-inventory/lexicon models, but which is unspecified in the taxonomy above. We will note such effects where they arise in the discussion below, but retain the terminology from the ideal perceiver models in Section 4.3 for ease of comparison.

4.4.1.1 Cue Agreement: Spectral Tilt of the Consonant ($\mathbf{TILT_C}$)

Figures 4.37–4.39 show that not only is there close agreement between all three models in terms of the role of the spectral tilt of the consonant noise spectrum in predicting listener recognition, but TILT_C is highly weighted in each model. The partial dependence functions for TILT_C are shown in Figure 4.40 alongside distributions of $\Delta\text{TILT}_\text{C}$ in each model according to different predicted accuracy classes; i.e., predicted accuracies are drawn from the partial depencence functions, and thus represent (to an approximation) variation in listener recognition due to $\Delta\text{TILT}_\text{C}$ while controlling for other parameters in the model.

While all three models show a monotonic increase in predicted listener accuracy over most of the $\Delta\text{TILT}_\text{C}$ range, they differ slightly in the form of this function, with the inventory model increasing more-or-less linearly between $\sim$4 and 20 dB/kHz, while in the lexicon model the relation between spectral tilt and predicted accuracy is shallower but linear over the full range above 5 dB/kHz, below which there is a relatively sharp drop to the next quantile at 3 dB/kHz. The weighted inventory model is somewhat intermediate between the two, showing a sharp increase in predicted accuracy between 5 and 7 dB/kHz, and a relatively flat partial dependence function for the remainder of the $\Delta\text{TILT}_\text{C}$ range. From the $\Delta\text{TILT}_\text{C}$ distributions in Figure 4.40 we see that there is close agreement between the three models in terms of the ranges of spectral tilt distinctions occupied by low-, mid-, and high-accuracy items. The interval with low predicted accuracy in all three models generally falls below 10 dB/kHz, while the mid-accuracy ranges between approximately 5 and 25 dB/kHz, with the cutoff slightly higher in the lexicon and weighted inventory than in the inventory model. And finally, among items predicted to be the most accurate as a function of $\Delta\text{TILT}_\text{C}$, this range is relatively broad and flat in each model, extending from around 10 dB/kHz to

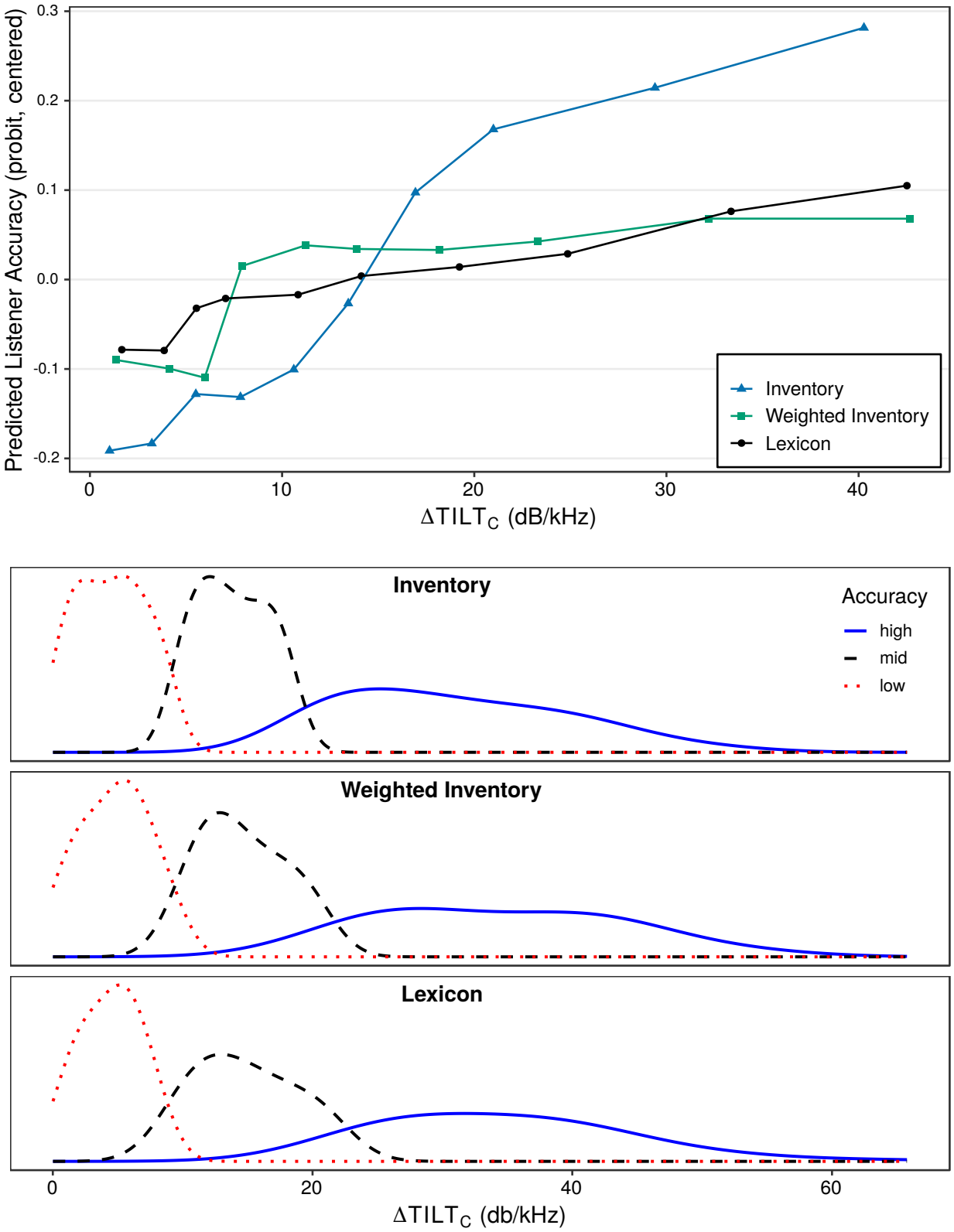

Figure 4.40: Partial dependence functions (top panel) and distributions (bottom panels) of TILT_C in the inventory, weighted inventory, and lexicon models of listener recognition in word-initial position.

over 55 dB/kHz. Thus the three models agree in the broad relation between $\Delta TILT_C$ and predicted accuracy, and only differ in the relative rate of change in accuracy over different $\Delta TILT_C$ intervals.

Figure 4.41 further confirms this close relation between the inventory and lexicon in the high correlation between spectral tilt distinctions by contrast in both controlled syllables and real words, where contrasts are colored by predicted accuracy in the lexicon in the top panel, and in the inventory in the bottom panel. The most frequent contrasts in the lexicon which are highly distinct in their spectral tilts—and correspondingly of high predicted accuracy along the $\Delta TILT_C$ dimension—are those between the voiceless alveolar sibilant [s] and voiceless plosives and the glottal fricative [h], and to a somewhat lesser extent between voiceless fricatives differing in place of articulation, confirming that spectral tilt is a broadly discriminative cue, reflecting characteristics of both the point of constriction and the noise source. Similarly, at the low end of the $\Delta TILT_C$ range are contrasts between plosives that are consistently predicted to be less accurate as a function of spectral tilt in both the inventory and lexicon. Contrasts occupying the intermediate range of predicted accuracy, with spectral tilt distinctions of between 7 and 15 dB/kHz, are manner contrasts between labials and posterior obstruents (i.e., [k] vs. [h]), each of which are distinct either in their primary noise source or in their place of articulation, but not both.

4.4.1.2 Distributional Disagreement: Relative F3 Amplitude (AMP_{F3})

Turning next to a case of disagreement in cue rankings between the inventory and lexicon that appears to be due primarily to differences in the distribution of contrasts in the two systems, Figure 4.42 shows partial dependence functions and cue distributions for relative F3 amplitude in each model. As with consonantal spectral tilt, ΔAMP_{F3} is of moderate-to-high weight in all three models (Figure 4.37), but when the asymmetric distribution of contrasts in the lexicon is accounted for in the weighted inventory model, the relative weight on ΔAMP_{F3} is increased even further beyond its value in the inventory, and thus is brought into closer agreement with the lexicon. We can see in Figure 4.42 that while there is a relatively constant increase in predicted accuracy in the lexicon between 0 and 10 dB, this relation is more varied in the inventory, with positive relations

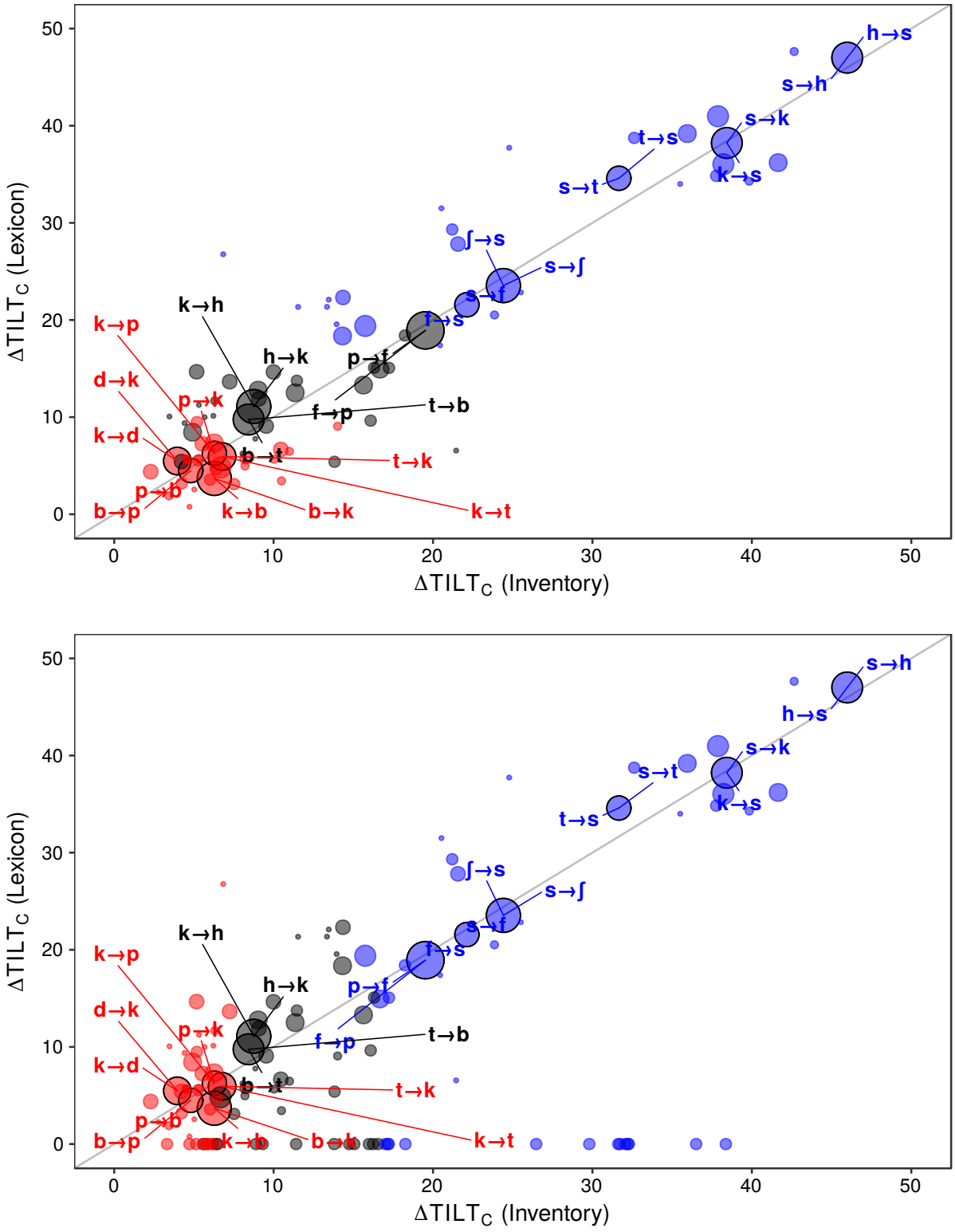

Figure 4.41: Relationship between $\Delta\text{TILT}_\text{C}$ means by phonetic contrast in the inventory and lexicon models in CV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown in at $y = 0$.

occurring over two discontinous intervals, the first between approximately 3 and 7 dB, and the second between 14 and 20 dB. The general form of the partial dependence function in the weighted inventory model is closer to the lexicon in exhibiting an approximately linear increase over a wider range of ΔAMP_{F3} values, but which is shifted to the right by approximately 7 dB, consistent with the generally greater relative F3 amplitudes in the controlled syllable data relative to the real-word productions comprising the model lexicon.

The distributions in Figure 4.42 further reveal that this scaling from inventory to lexicon via the weighted inventory model is accomplished by shifting both the low- and high-accuracy items to lower ΔAMP_{F3} values, while contrasts in the intermediate predicted accuracy range are shifted away from the low-accuracy set, though all such shifts are not complete in matching the distributions in the lexicon. Additionally, the accuracy-based distributions in the inventory are bimodal in all but the highest accuracy range, which is why the partial dependence between ΔAMP_{F3} and listener accuracy in the inventory model does not increase monotonically.

Figure 4.43 shows the breakdown of ΔAMP_{F3} by contrast and predicted accuracy in the inventory and lexicon. As predicted in the introduction of relative F3 amplitude in Chapter 2, ΔAMP_{F3} is highly discriminative of the voiceless sibilant distinction [s, ʃ], while voiceless lingual stops are also relatively distinct from [s] in AMP_{F3}, where [t] and [k] play a particularly important role given their frequent occurrence in contrasts with [s] in the lexicon. At the lower end of the ΔAMP_{F3} range are contrasts between plosives and between plosives and nonsibilant fricatives; however, this set of contrasts is much less accurate in aggregate in the inventory than in the lexicon. Further, there are many contrasts in the low ΔAMP_{F3} range that relatively lower in predicted accuracy in both models, but which either occur less often in the lexicon (shown in the relative size of each point in Figure 4.43) or do not occur at all in minimal-pair contrasts in the lexicon (shown in the points aligned along the bottom of the lower panel of Figure 4.43).

Thus, here we have an example of disagreement in cue behavior in the inventory and lexicon that reflects both distributional differences in the two systems, and differences in the perception of words versus syllables in Experiment 1 of the present study and the Woods et al. (2010) study,

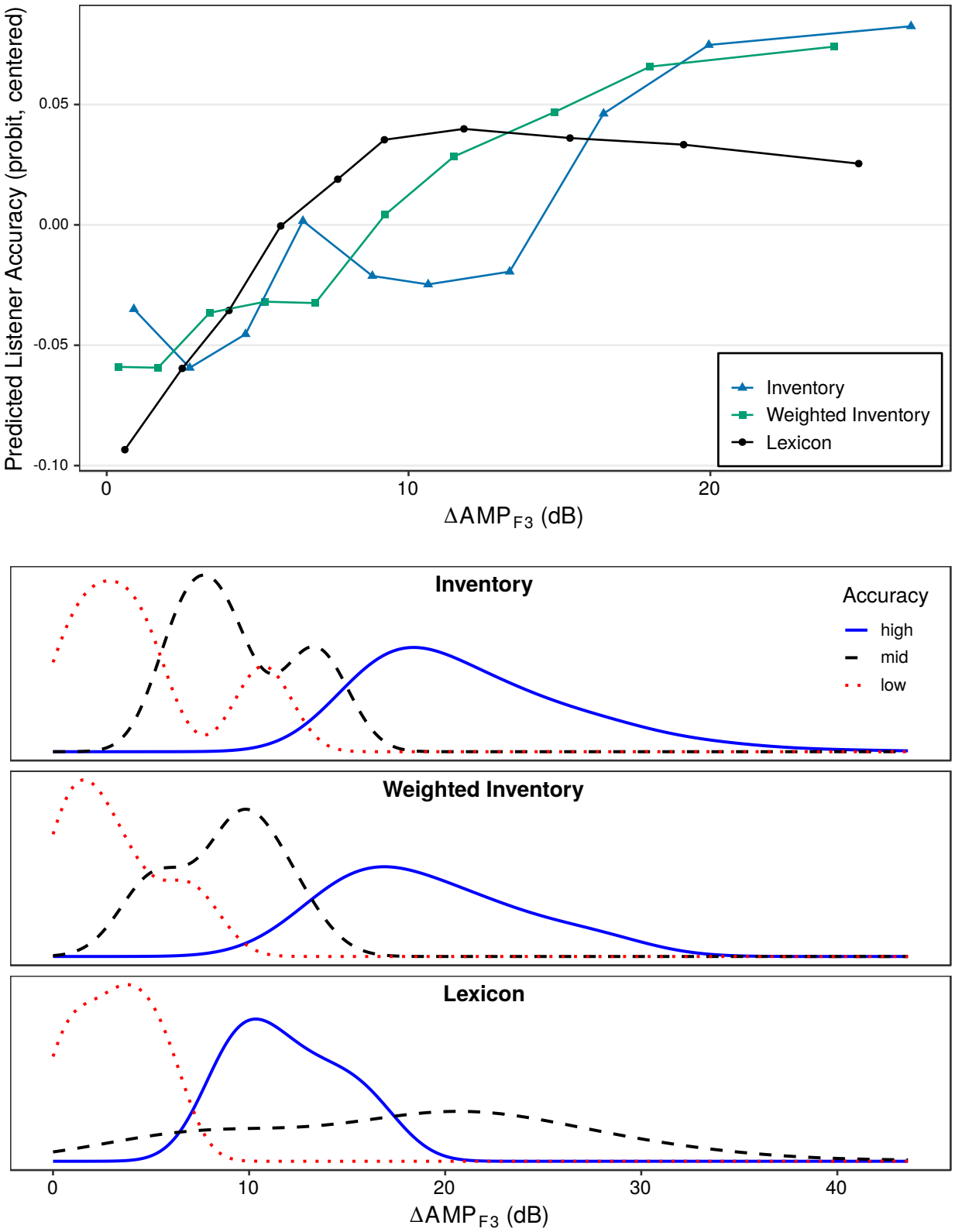

Figure 4.42: Partial dependence functions (top panel) and distributions (bottom panels) of AMP_{F3} in the inventory, weighted inventory, and lexicon models of listener recognition in word-initial position.

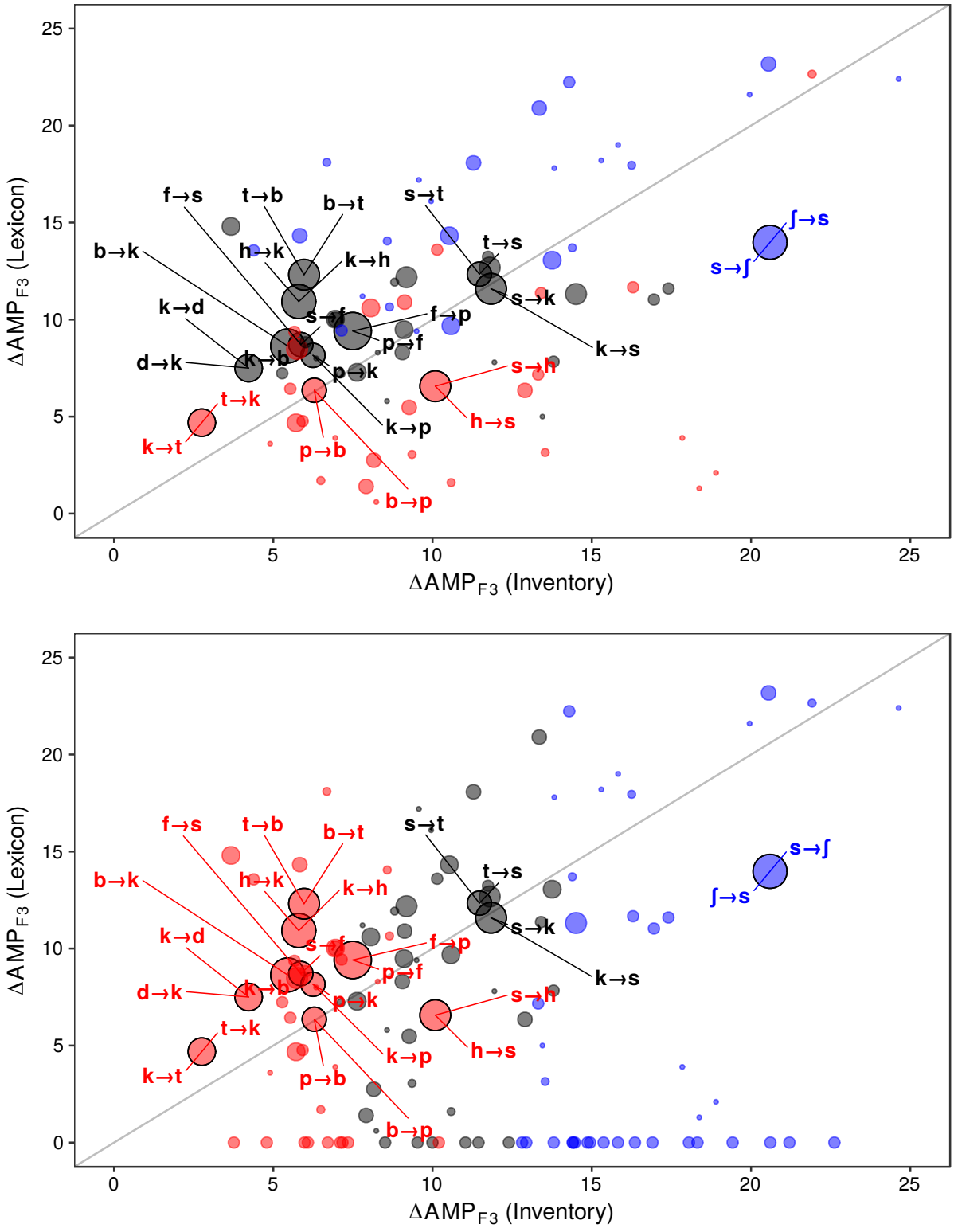

Figure 4.43: Relationship between ΔAMP_{F3} means by phonetic contrast in the inventory and lexicon models in CV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown in at $y = 0$.

respectively. Therefore, while the closer agreement between the weighted inventory and lexicon models relative to the balanced inventory model suggests some promise for the scaling of the relative cue weight on AMP_{F3} derived from controlled syllable recognition to its weight in real-word recognition, perceptual differences between the two tasks mean this scaling cannot be fully accomplished by taking into account asymmetries in the distribution of contrasts in the lexicon. Further study on both task and linguistic unit constraints on listener perception is therefore necessary to fill this gap.

4.4.1.3 Acoustic Disagreement: Following Vowel Duration (DUR_{V2})

In Chapter 2 we found that while preceding vowel duration is highly discriminative of word-final voicing distinctions, following vowel duration is much less informative for the identification of word-initial obstruents, though the directionality of the voicing effect (*voiceless* < *voiced*) is consistent between the two positions. Nevertheless, as the cue rankings in Figure 4.37 and the partial dependence functions in Figure 4.44 show, following vowel duration still plays a moderate role in the prediction of listener recognition of word-initial contrasts. In the lexicon there is a sharp increase in predicted accuracy for contrasts differing in DUR_{V2} by more than 20 ms, and another more gradual increase above 40 ms. In the lexicon model there is a linear increase in predicted accuracy between 15 and 40 ms. Thus both models show similar effects of vowel duration on listener accuracy, but they achieve this relation over different ΔDUR_{V2} intervals. As such, there is poor scaling between the inventory and lexicon via the weighted inventory model, which largely exhibits no direct relationship between vowel duration and listener accuracy. This poor scaling is further illustrated in the ΔDUR_{V2} distributions, which show little overlap between low- and high-accuracy sets in the inventory and lexicon, while the weighted inventory shows considerable overlap between all three sets (note in particular the bimodal distribution in the low-accuracy set).

Figure 4.45 shows the relative correlation between ΔDUR_{V2} in the inventory and lexicon, as well as the relation between vowel duration, phonetic contrasts, and predicted accuracy in each model. Here we have confirmation that DUR_{V2} does indeed index obstruent voicing, as the ma-

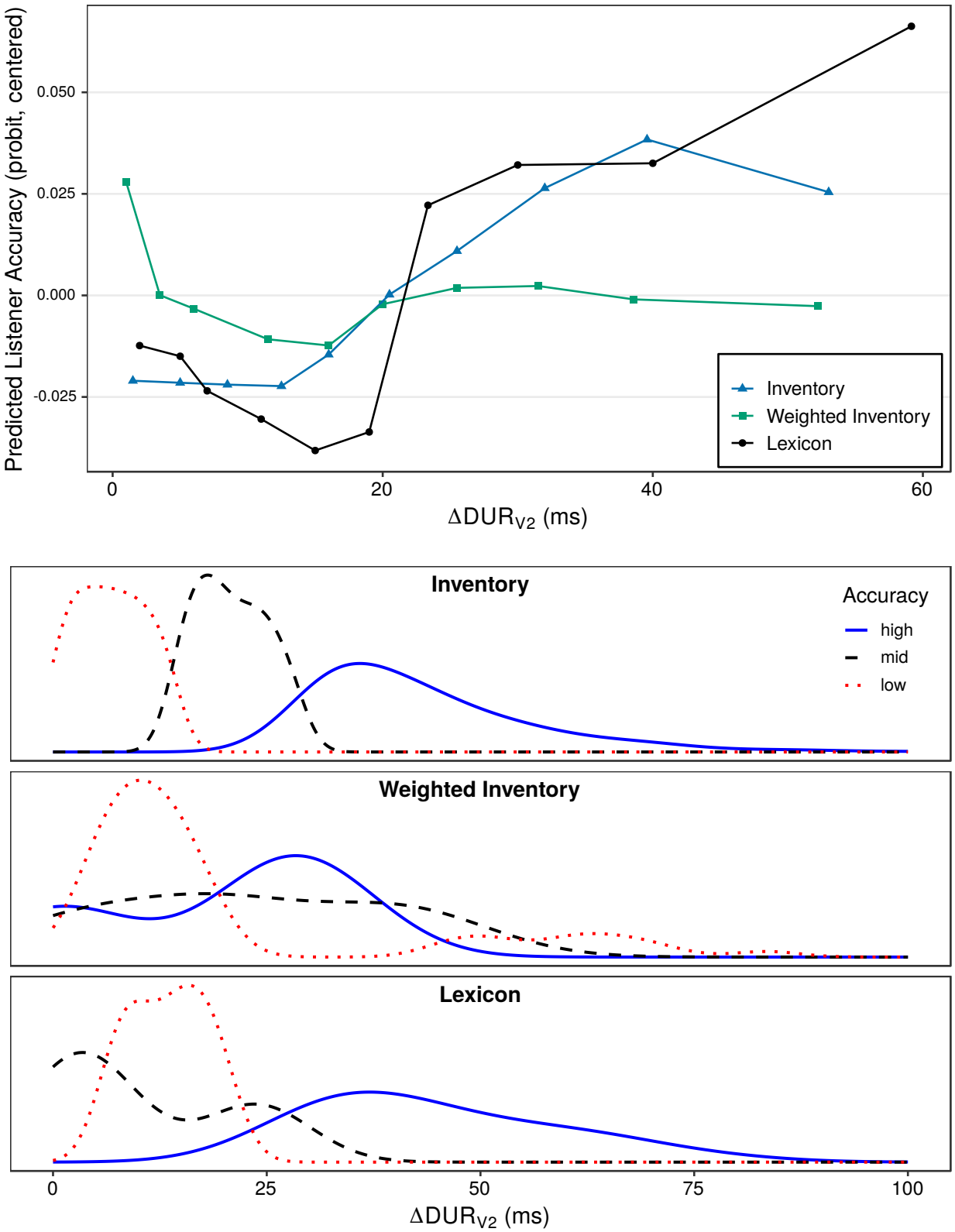

Figure 4.44: Partial dependence functions (top panel) and distributions (bottom panels) of DUR_{V2} in the inventory, weighted inventory, and lexicon models of listener recognition in word-initial position.

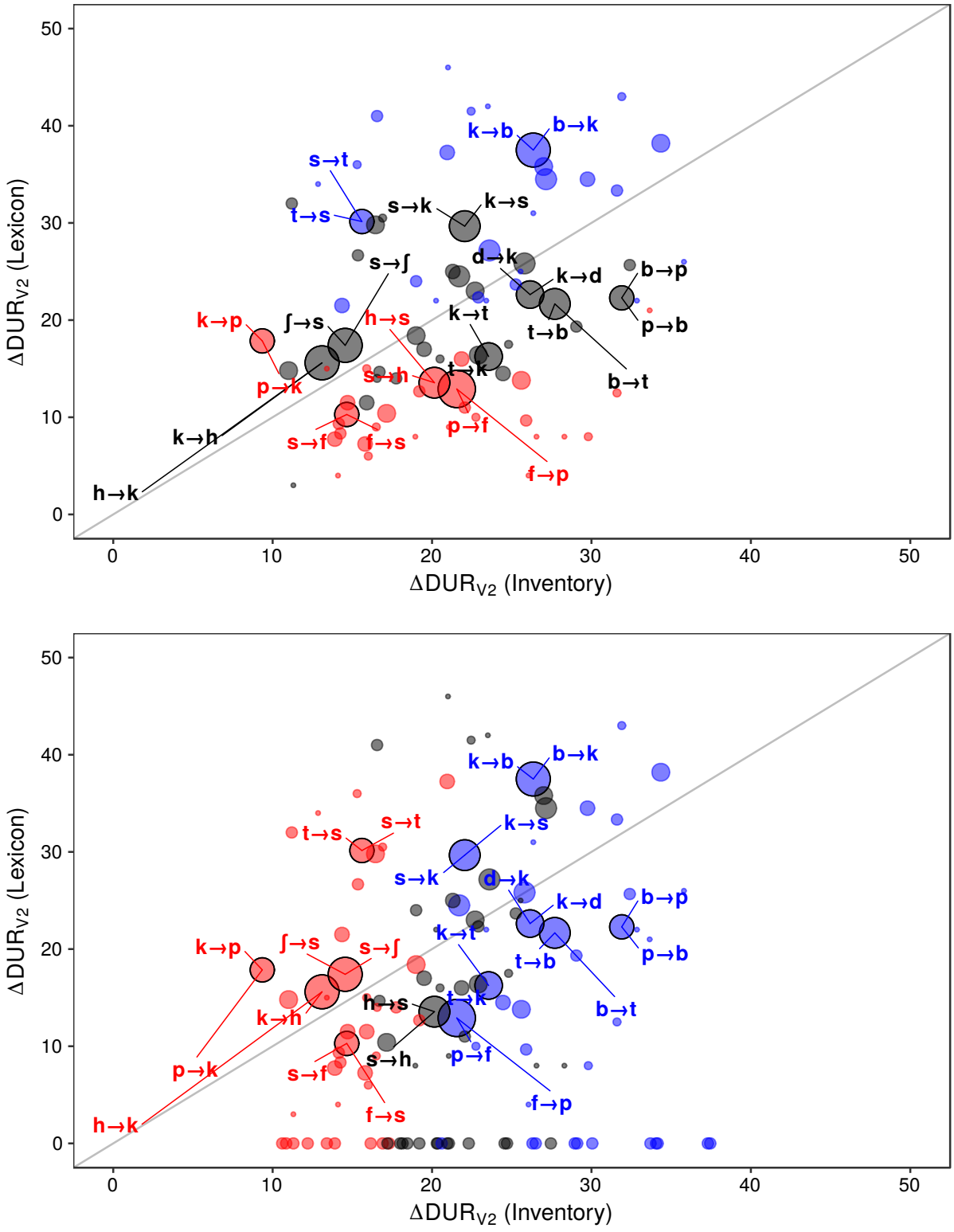

Figure 4.45: Relationship between ΔDUR_{V2} means by phonetic contrast in the inventory and lexicon models in CV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown in at $y = 0$.

jority of contrasts with the highest predicted accuracy in Figure 4.45 (those shown in blue), both in the lexicon and inventory models, differ in voicing, while those at the low accuracy range are predominantly within-voicing distinctions. However, aside from this commonality the correlation between vowel duration distinctions in the controlled syllable data and those in minimal pairs in the lexicon is relatively low, resulting in poor scaling in the weighted inventory model as contrasts such as *p–f*, which are relatively low in predicted accuracy, exhibit considerably larger DUR_{V2} distinctions in the inventory than in the lexicon, while at the same time contrasts like *s–t* which are predicted to be highly accurate exhibit much narrower vowel duration differences in the controlled syllable data. This combination of effects breaks the generally monotonic relationship between $\Delta\text{DUR}_{\text{V2}}$ and listener accuracy found in the lexicon model, leading to the flat partial dependence curve in Figure 4.44 and the negligible role of $\Delta\text{DUR}_{\text{V2}}$ in the weighted inventory model.

4.4.1.4 Composite Disagreement: Spectral Dispersion of the Consonant (DISP_{C})

Finally, we examine the dispersion (Wiener entropy) of the consonant noise spectrum as a case of *composite* disagreement. Figure 4.46 shows the partial dependence functions and distributions of DISP_{C} in the lexicon, inventory, and weighted inventory models, each of which exhibits a distinct relationship between $\Delta\text{DISP}_{\text{C}}$ and predicted listener accuracy. In the inventory model the majority of the change in predicted listener accuracy occurs between $\Delta\text{DISP}_{\text{C}}$ values above and below 5, where dispersion distinctions above 5 are notably lower in predicted accuracy, and thus the aggregate cue weight on $\Delta\text{DISP}_{\text{C}}$ is negative, indicating that listeners do not appear to be using spectral dispersion as a cue in the discrimination of obstruent contrasts. By comparison, in the lexicon model there is a substantial increase in predicted accuracy between dispersion values of approximately 1 and 3. Predicted accuracy then decreases over the 3–7 range and then increases again, albeit more gradually, over the remainder of the $\Delta\text{DISP}_{\text{C}}$ range. Thus spectral dispersion appears to be operative in word recognition as a cue to some distinctions—namely between low- and mid/high-$\Delta\text{DISP}_{\text{C}}$ contrasts, as well as among high-$\Delta\text{DISP}_{\text{C}}$ contrasts—but not between the majority contrasts as in the cases of spectral tilt and relative F3 amplitude described above. Finally,

the partial dependence function for the weighted inventory model is approximately flat over half of its distribution ($DISP_C > 4$) and shows a negative relation over the low-$\Delta DISP_C$ range, and thus is unable to replicate the role of spectral dispersion in the lexicon.

These results are further summarized in Figure 4.46 in the form of $\Delta DISP_C$ distributions among contrasts predicted to be of low, mid, and high accuracy (controlling for other cues) in each model. In the inventory model the three accuracy sets largely coincide, except for $\Delta DISP_C > 5$, which is composed primarily of low- and mid-accuracy contrasts. The weighted inventory model shows a similar pattern as the inventory model, but with the threshold between high- and mid/low-accuracy contrasts lowered to between 2 and 3. Finally, all three accuracy sets in the lexicon model exhibit bimodal distributions leading to a distinction between low- and mid/high-accuracy items around $\Delta DISP_C = 2$, and further distinctions between low- and mid-accuracy distributions at $\Delta DISP_C = 8$, and between mid- and high-accuracy distributions at $\Delta DISP_C = 12$.

Figure 4.47 shows that the ultimate source of these discrepancies is perceptual, though there are small acoustic discrepancies in the form of lower correlations between $\Delta DISP_C$ in the inventory and lexicon relative to spectral tilt (Figure 4.41), and small distributional discrepancies in the form of greater overlap in the inventory data between low- and mid-accuracy contrasts at the lower $\Delta DISP_C$ range, and reduced $DISP_C$ distinctions among high-accuracy contrasts in the $\Delta DISP_C > 5$ range. However, the primary difference between the inventory and lexicon is in the relative predicted accuracy on each contrast, the distinctions between stops/sibilants and nonsibilant fricatives ($\Delta DISP_C > 5$) being relatively accurate in the lexicon but inaccurate in the inventory, with the reverse pattern obtained for $DISP_C$ distinctions below 3. Overall, where spectral dispersion does appear to cue listener recognition (in the lexicon model), the primary distinction is between the nonsibilant fricatives, which exhibit relatively dispersed spectra, and the remainder of the obstruent set, which exhibit greater energy concentration in a restricted frequency range. This distinction accounts for the accuracy difference over the upper $\Delta DISP_C$ range, while the distinction between low- and mid-$\Delta DISP_C$ contrasts is primarily driven by differences between cross-manner/voicing contrasts and within-manner/voicing contrasts, where the former are more spectrally distinct than

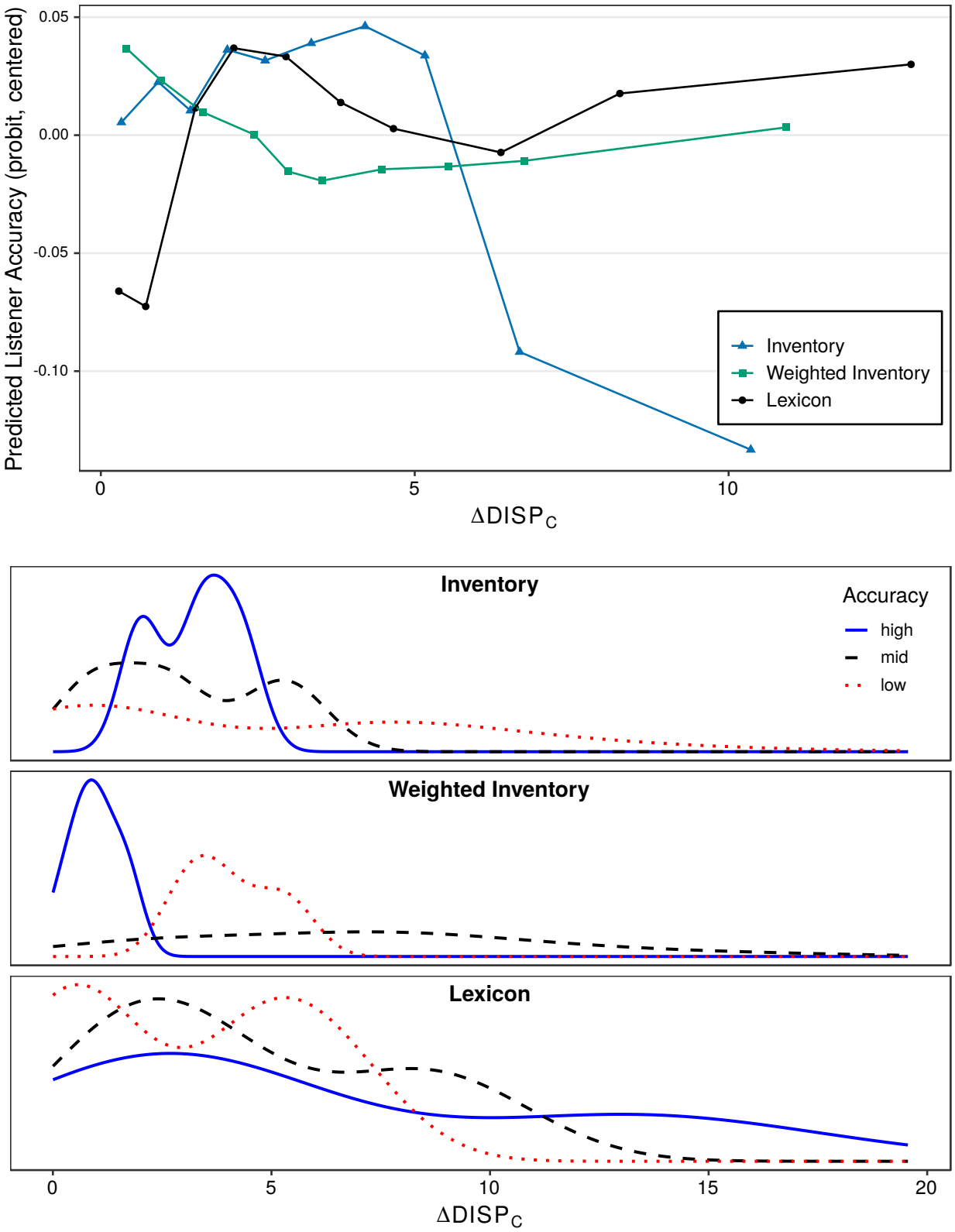

Figure 4.46: Partial dependence functions (top panel) and distributions (bottom panels) of $DISP_C$ in the inventory, weighted inventory, and lexicon models of listener recognition in word-initial position.

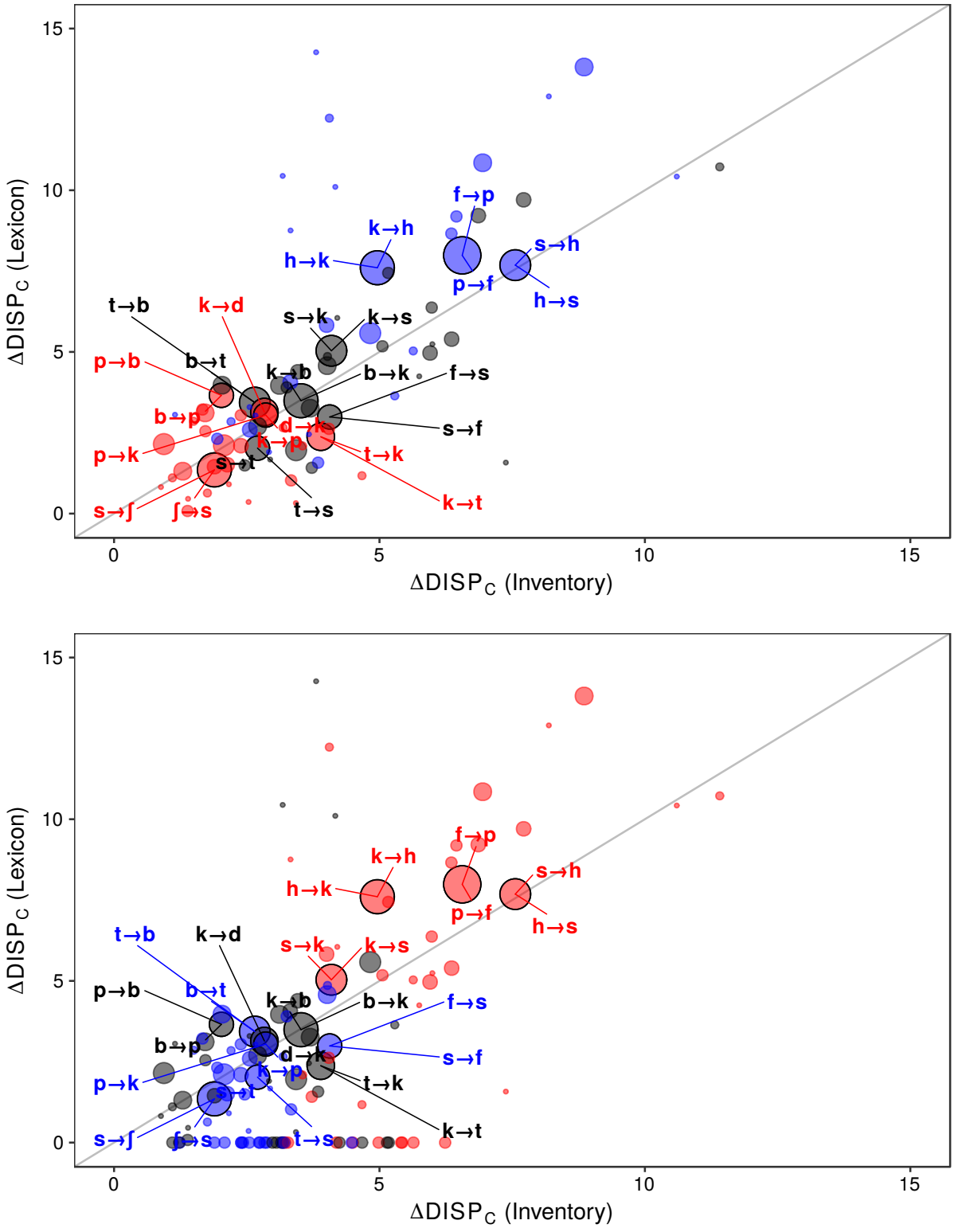

Figure 4.47: Relationship between $\Delta DISP_C$ means by phonetic contrast in the inventory and lexicon models in CV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown in at $y = 0$.

the latter and thus predicted to be of higher accuracy in listener perception.

4.4.2 Word-medial position (VCV)

Models of listener recognition of intervocalic contrasts in Experiment 1 provide similar though slightly improved fits to the data relative to the word-initial models discussed above ($RMSE_{winv}$ = 0.384, $RMSE_{lex}$ = 0.322; as compared with 0.419 and 0.355, respectively, in CV position); however, models of listener recognition of controlled syllable recognition in Cooke & Scharenborg (2008), both from the inventory and reference acoustics, are notably poorer in fit ($RMSE_{inv}$ = 0.330, $RMSE_{ref}$ = 0.324; as compared with 0.258 and 0.309, respectively, in CV positon), though all four models show a monotonic increase in predicted accuracy with increases in listener accuracy (Figure 4.48). The generally poor fit of the inventory and reference models can be attributed to the greater variance in the stimuli presented to listeners in Cooke & Scharenborg (2008), who unlike Woods et al. (2010) neither control the vowel contexts different obstruents are presented in, nor the speakers who contribute the data from each VCV type; i.e., for a given vowel context different obstruents come from different speakers, while for each obstruent, variation in vowel context is also accompanied by variation in the speaker who produced the stimulus. This variation poses challenges for modeling listener perception from the stimulus acoustics, as well as introducing the potential that many errors in Cooke & Scharenborg (2008) may be due to *switch costs* from changes in speaker; i.e., an increase in the processing burden on working memory that weakens the relationship between linguistic structure and listener perception (Lim et al., 2019).

Nevertheless, despite the generally poor fit between the inventory/reference models and the syllable perception data, the relative rankings of contrast parameters derived from each model correlate significantly ($r = 0.347$, $p = 0.041$). The correlation between inventory and reference target parameter ranks was not significant ($r = 0.113$, $p > 0.1$), but given that target parameters are more variable in general than contrast parameters (across models and positions), and that the primary focus of the analysis is on the behavior of contrast parameters in each model as an indicator of cue integration in obstruent discrimination, this discrepancy is not a major concern.

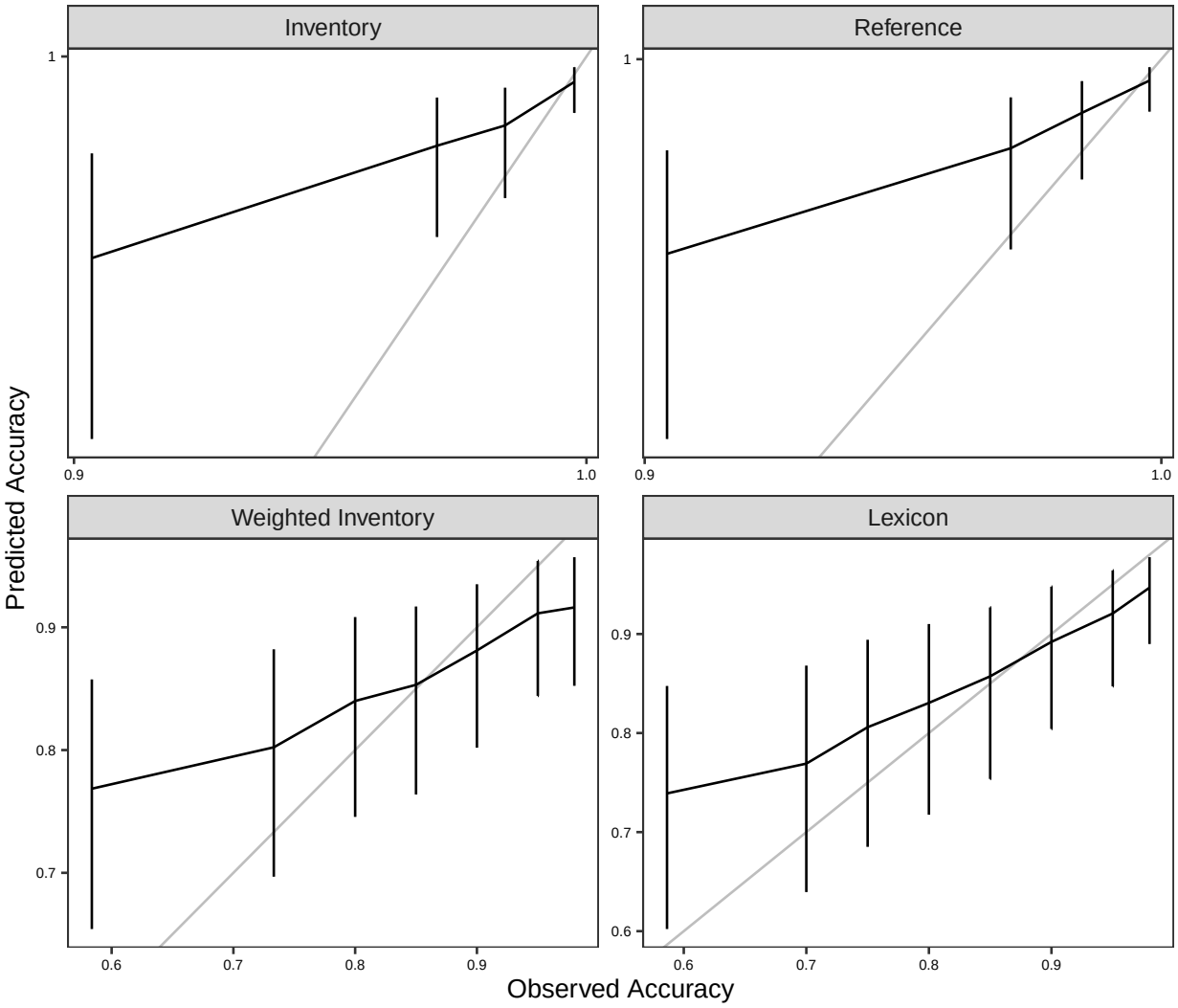

Figure 4.48: Listener model fit in the inventory, reference, weighted inventory, and lexicon models of word-medial contrasts. Lines indicate posterior medians averaged at the observed quantiles (0, 0.1, ..., 1). Error bars indicate the middle 90% of the posterior distribution. Solid grey lines show the identity function.

Further model fits from transformed predictions via the meta-modeling approach described in the previous section are shown in Figure 4.49. Overall, the relative improvement in fit between the initial model predictions (Figure 4.48) and the transformed predictions (Figure 4.49) is slightly higher than that observed in CV position, at a 7% reduction in the RMSE of the inventory model, a 5% reduction in the reference model, and error reductions of 5 and 7%, respectively, in the weighted inventory and lexicon models. Thus, as in the modeling of word-initial contrast recognition, these fits indicate that the three target speaker models—inventory, weighted inventory, and

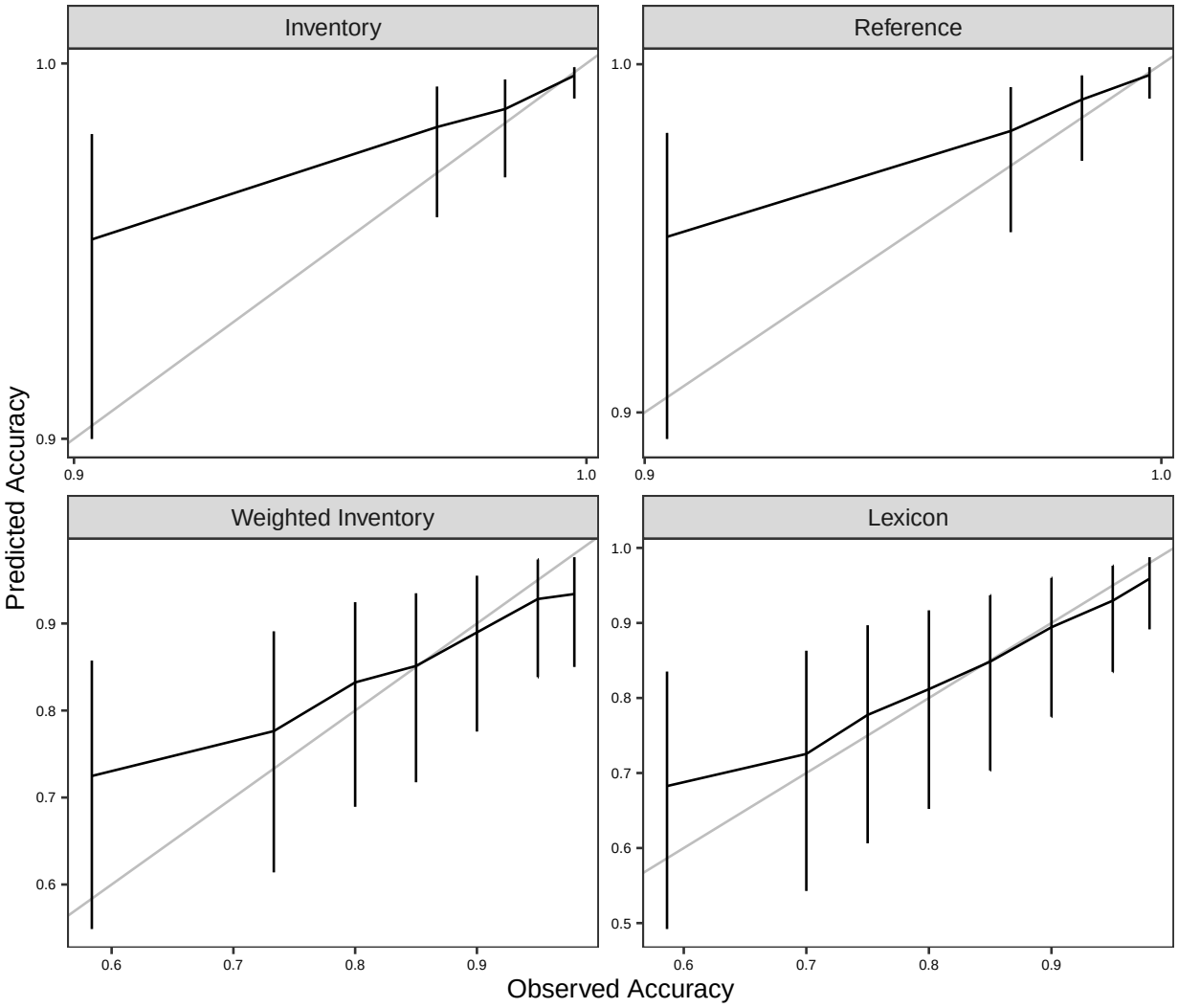

Figure 4.49: Transformed listener model fit based on a meta-model of the inventory, reference, weighted inventory, and lexicon predictions of word-medial contrasts. Lines indicate posterior medians averaged at the observed quantiles (0, 0.1, ..., 1). Error bars indicate the middle 90% of the posterior distribution. Solid grey lines show the identity function.

lexicon—can be treated as a viable approximation to the differential weighting of acoustic cues in the perception of intervocalic obstruent distinctions.

Figures 4.50 and 4.51 show target and contrast cue ranks in the inventory, weighted inventory, and lexicon models of VCV contrasts. Among the target cues, the most highly ranked cues in the lexicon across sub-experiments (see Figures A.81 and A.82 in the appendix for details) are consonant voicing percentage (VOI%), spectral peak frequency (FREQ_{PK}), spectral dispersion of the consonant (DISP_{C}), F3 at following vowel onset (F3_{CV}), noise amplitude (AMP_{N}), and the

spectral tilt of the consonant ($TILT_C$), each of which correlates strongly with listener recognition irrespective of the corresponding cue value in the competitor item.

Most of the above cues delineate place and sibilance classes; however, the consistently high weight on target VOI% illustrates that there are also major differences between voiced and voiceless obstruent perception word-finally. In general, voiced obstruents are more poorly perceived in VCV position than voiceless obstruents, both in Experiment 1 (see the target phone results in Figure 3.3 of Chapter 3 for details) and in the Cooke & Scharenborg (2008) study, a result which is consistent with lexicon/inventory model predictions based on consonant voicing percentage. The threshold after which predicted accuracy drops substantially in all three models is between 15 and 20%, which broadly separates voiceless fricatives (excepting [h]) and affricates from voiceless plosives and voiced obstruents, though the voiceless plosives vary about this threshold and comprise the majority of the transition region from high to low predicted accuracy.

The remaining cues may differ in the primary feature they distinguish, $FREQ_{PK}$, $F3_{CV}$ and $TILT_C$ primarily reflecting differences in place of articulation, while $DISP_C$, and AMP_N vary mostly according to obstruent sibilance; manner effects on both sets can also be seen, though in many cases these effects are conflated with differences in place and sibilance. However, most of these cues index multiple featural distinctions of differing roles in the inventory and lexicon; thus, in the discussion below we treat each cue separately, focusing primarily on how each aggregate cue rank derives from variation in predicted listener accuracy along a given acoustic dimension.

Spectral peak frequency shows notable variation in its behavior in the inventory and lexicon. Among minimal-pair contrasts in Experiment 1, there are three main classes of obstruents which differ in the relation between peak frequency and predicted accuracy. At the lower end of the $FREQ_{PK}$ range are the voiced nonsibilant fricatives [v, ð], the labial plosives [p, b], and the glottal fricative [h], all exhibiting peak frequencies generally below 1000 Hz and substantially lower in predicted accuracy relative to the remainder of the obstruent set. Over the mid-$FREQ_{PK}$ range, predicted accuracy from changes in peak frequency is relatively flat. This set includes the stop consonants and voiceless nonsibilant fricatives which are of intermediate predicted accuracy. Fi-

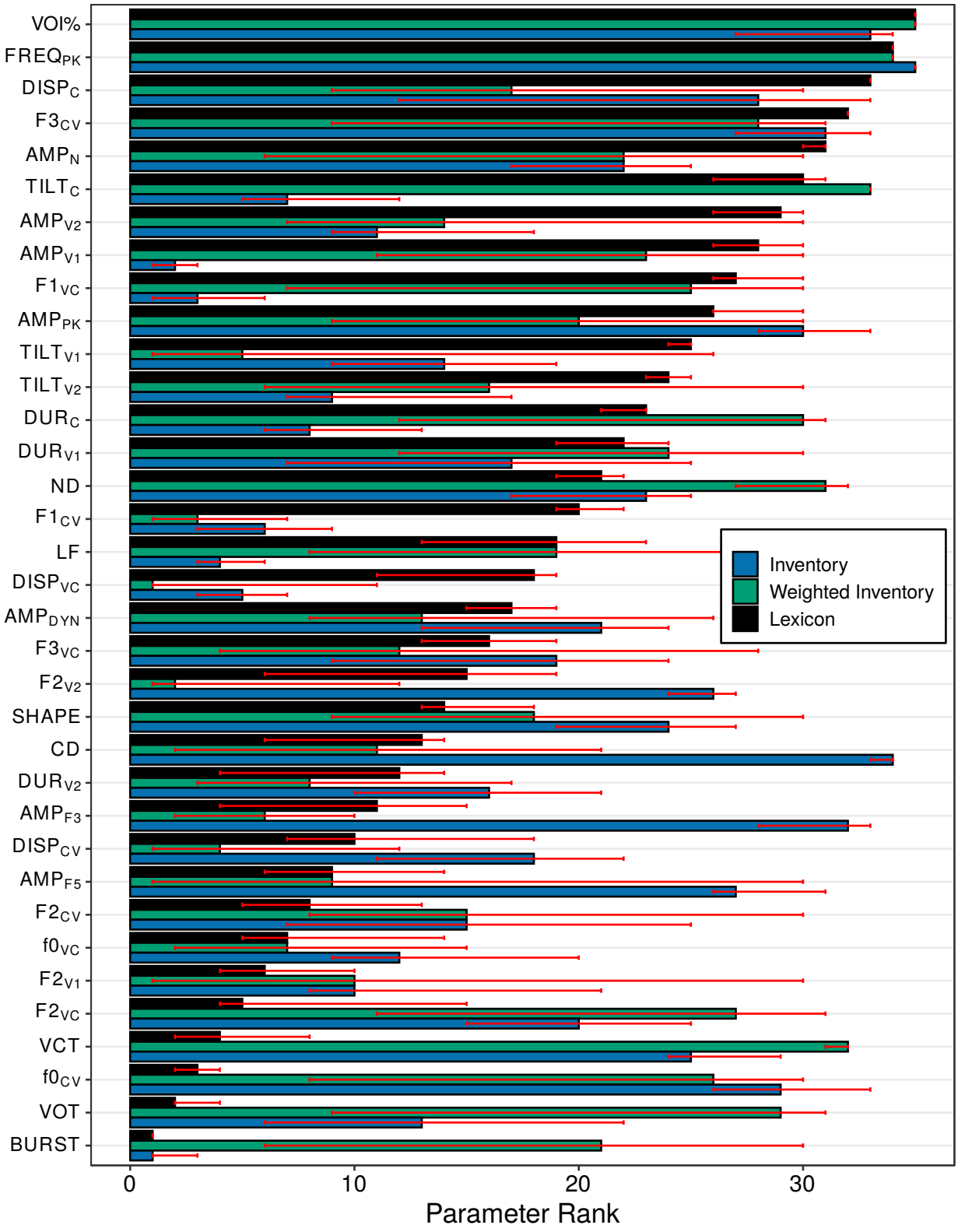

Figure 4.50: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-medial contrasts fit to listener recognition data. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

nally, at the upper end of the peak frequency range are the alveolar sibilants [s, z], which are notably higher in predicted accuracy relative to the remainder of the set. Thus, in the lexicon model FREQ_{PK} primarily indexes the relation [p, b, v, ð, h] < [t, d, k, g, ɾ, ʧ, ʤ, f, θ] < [s, z], where higher peak frequencies generally correspond to greater predicted accuracy, though such changes occur in discrete steps rather than linearly over the FREQ_{PK} range. In the inventory model two classes can generally be distinguished: the plosives (excepting [b]), postalveolars, and the voiceless labiodental fricative [f], which are all of intermediate spectral peak frequencies in the controlled syllable data; and the remainder of the obstruent set, [b, v, ð, h] at the lower margin of the FREQ_{PK} range, and [θ, s, z] at the upper FREQ_{PK} margin. The former is predicted to be exhibit higher accuracies, based on listener recognition data in Cooke & Scharenborg (2008), than the latter, though there is no distinct separation between low- and high-FREQ_{PK} sets in terms of predicted accuracy in the listener model. Finally, in mapping controlled syllable acoustics onto word recognition data in Experiment 1, the weighted inventory model largely reflects the pattern in the lexicon model, though the separation between low- and mid-FREQ_{PK} sets is reduced in the weighted inventory relative to the lexicon, while the mid- versus high-FREQ_{PK} distinction is notably enhanced, though the directionality of all such effects is consistent between the two.

The frequency of the third formant at the onset of the following vowel (F3_{CV}), shows a stark separation between obstruents with F3 onsets below 2500 Hz and those above 2500 Hz, where the former largely corresponds to plosives and labial/glottal fricatives, and the latter is largely comprised of fricatives and affricates. This is a natural partition of the target obstruents intervocalically into generally low- and high-accuracy sets, as the greater duration, amplitude, and spectral resolution of fricatives and affricates intervocalically provides robust cues to listeners in perception (see the category and contrast accuracy results in Figures 3.3 and 3.12 in Chapter 3 for details). No such relation is found in the inventory models, which are both more variable in their alignment of these classes along the F3_{CV} dimension, and in the case of the inventory model, less distinct in the accuracy differences between plosives and fricatives/affricates, perhaps due to the greater presence of within-manner contrasts in the balanced inventory, which reduces the likelihood that a

high-onset F3 can cue the target item as being a fricative/affricate and therefore distinct from the expected value for a competitor plosive. To be clear, this does not mean that in such cases listeners cannot use F3 as a cue to manner of articulation, but that in a balanced inventory these cases will not arise as often as they do in the lexicon.

The spectral tilt of the consonant noise spectrum behaves similarly to $F3_{CV}$, distinguishing lingual fricatives and affricates, which largely exhibit spectral tilts above -5 dB/kHz, from plosives and labial/glottal fricatives, which have much steeper negative-tilting spectra. However, in this case the acoustics more closely align between the inventory and lexicon, resulting in close alignment between the behavior of $TILT_C$ in the weighted inventory model and in the lexicon model. Thus the inventory model primarily differs from the two regarding this cue ranking because again the discontinuity in recognition accuracy between these two manner/place classes does not arise in the balanced syllable recognition experiment of Cooke & Scharenborg (2008). Noise amplitude is similar in distinguishing these two sets at around 57 dB, and is consistent in form in two inventory models, but with a somewhat shallower relation between AMP_N and predicted listener accuracy.

Finally, the dispersion of the consonant noise spectrum behaves similarly to the cues above, but is more distinct in separating obstruents into three general categories: plosives and glottals; postalveolars; and fricatives. These sets increase monotonically in spectral dispersion, with the fricatives the most dispersed, and the plosives/glottals the most concentrated in their spectral energy distributions. And given that these sets largely correspond to a similar increase in listener accuracy, there is a straightforward relationship between $DISP_C$ and predicted accuracy in the lexicon model. A similar aggregate ranking is obtained in the inventory model, but it derives from a much narrower distinction between the set [t, f] and the remainder of the obstruents. This acoustic discrepancy then results in poor scaling of $DISP_C$ in the weighted inventory model, where its relationship with listener accuracy in word recognition is largely flat.

Figure 4.51 shows the relative ranking of each contrast cue in the three models of intervocalic obstruent perception, with the relative correlations and cue-ranking differences further plotted in Figures 4.52 and 4.53, respectively. Among the most discriminative of these cues in the lexicon,

both overall and by sub-experiment (see Figures A.83 and A.84 in the appendix for details) are consonant voicing percentage (ΔVOI%), relative F3 amplitude (ΔAMP_{F3}), consonant spectral tilt ($\Delta TILT_C$), and F2 at preceding vowel offset ($\Delta F2_{VC}$). ΔVOI% and $\Delta TILT_C$ show close agreement between all three models, while ΔAMP_{F3} and $\Delta F2_{VC}$ show greater agreement between the inventory and lexicon. The lower end of the cue ranking in the lexicon model is primarily composed of cues in the adjacent vowels and vowel transitions, though some consonantal cues such as spectral peak amplitude (ΔAMP_{PK}), burst presence (ΔBURST), and voice cessation time (ΔVCT) are also ranked relatively low. Several of these cues are given a much higher weight in the inventory models, however, such as F2 and F3 at vowel onset ($\Delta F2_{CV}$, $\Delta F3_{CV}$), and spectral dispersion at the VC transition ($\Delta DISP_{VC}$). Finally, we should note that the cues in this set that exhibit low rankings in the lexicon model all show negative partial density functions, indicating that it is difficult to derive a valid inference from the model regarding listeners' use of such cues, as they suggest that listeners are more accurate when there is greater similarity between the target and competitor along a particular acoustic dimension. Such a result is likely due to interactions with other parameters in the model, and so we will ignore these cases at present, focusing instead on those cues which are powerful independent predictors of listener recognition behavior.

Figures 4.52 and 4.53 show the complete distribution of cues in terms of the relative agreement between the lexicon and inventory models. Overall, the rank correlations in Figure 4.52 are slightly higher than in CV position, though still well below the rank agreement in the ideal perceiver models. This result is again due to the additional variable of listener perception of real words versus controlled syllables. From Figure 4.53, which shows the rank differences between the lexicon and inventory models, we see that the majority of cues extend along the diagonal, indicating that where a cue's role in the lexicon is underestimated in the inventory model, it tends to also be underestimated in the weighted inventory, and vice versa for overestimated cue ranks. This trend is consistent across sub-experiments, though Experiments 1a and 1b do differ in the overall strength of parameter rank correlations (Exp. 1a showing much higher correlations than Exp. 1b), and in the relative ranking of particular parameters (see Figures A.85–A.88 in the appendix for details).

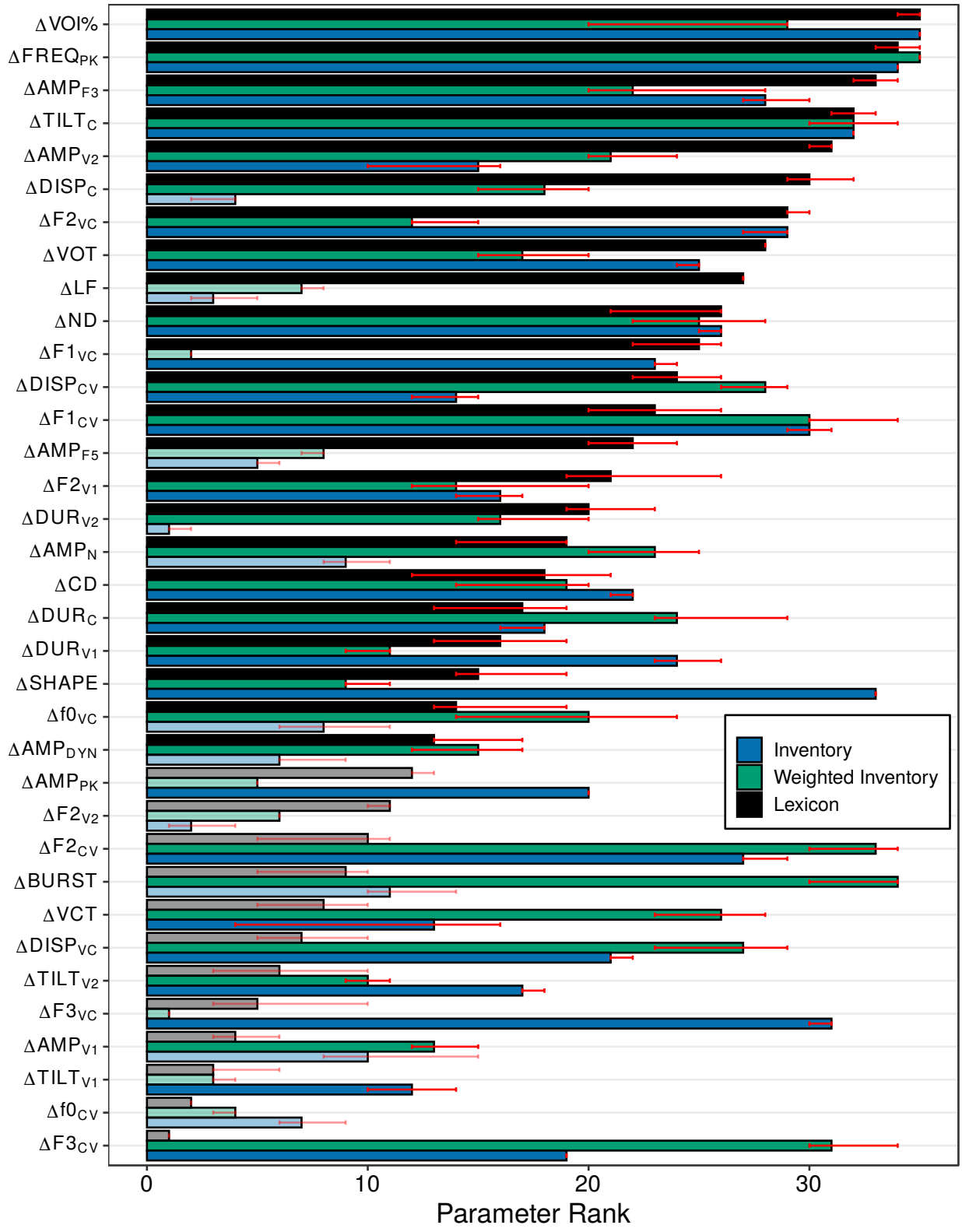

Figure 4.51: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-medial contrasts fit to listener recognition data. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

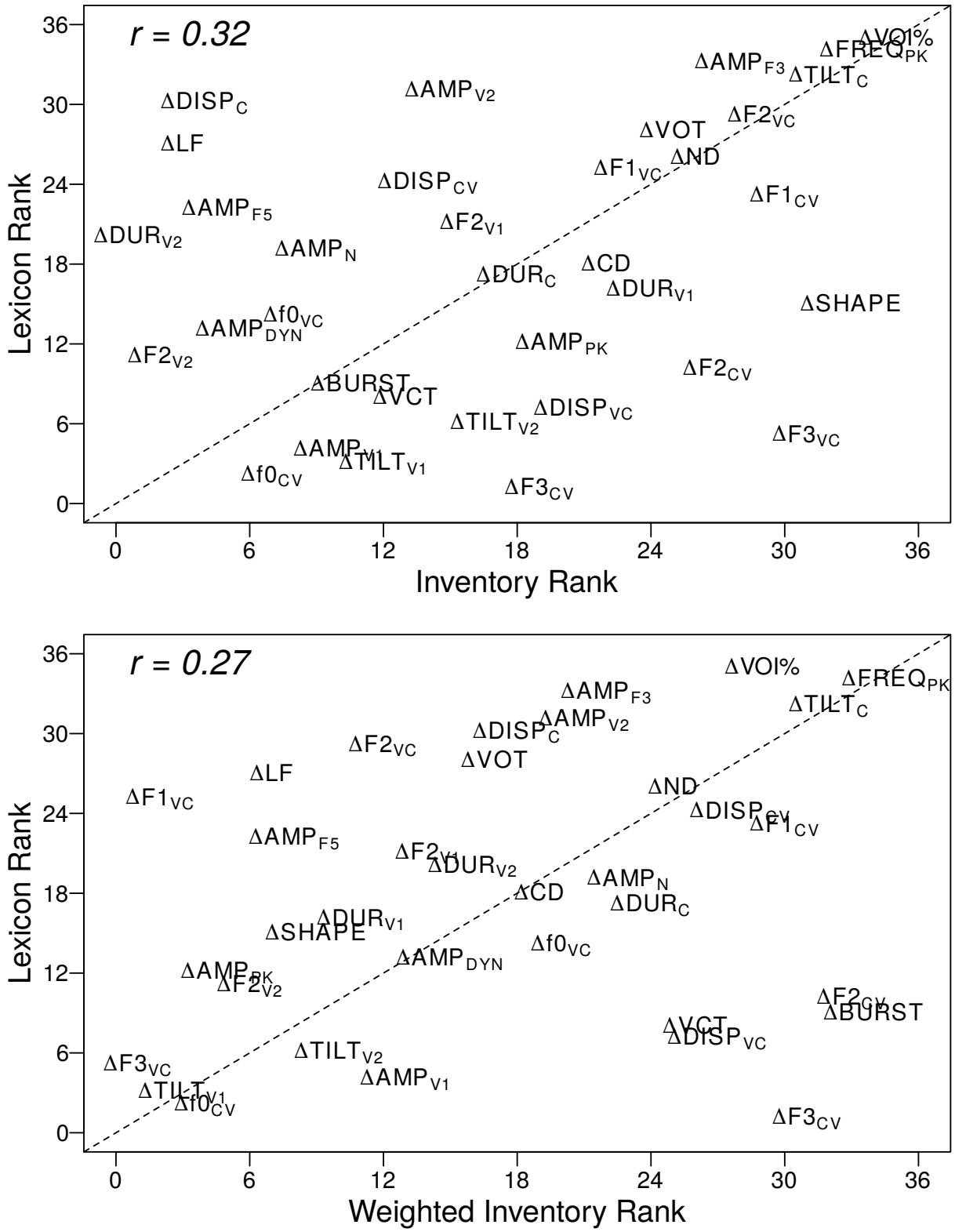

Figure 4.52: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VCV position. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

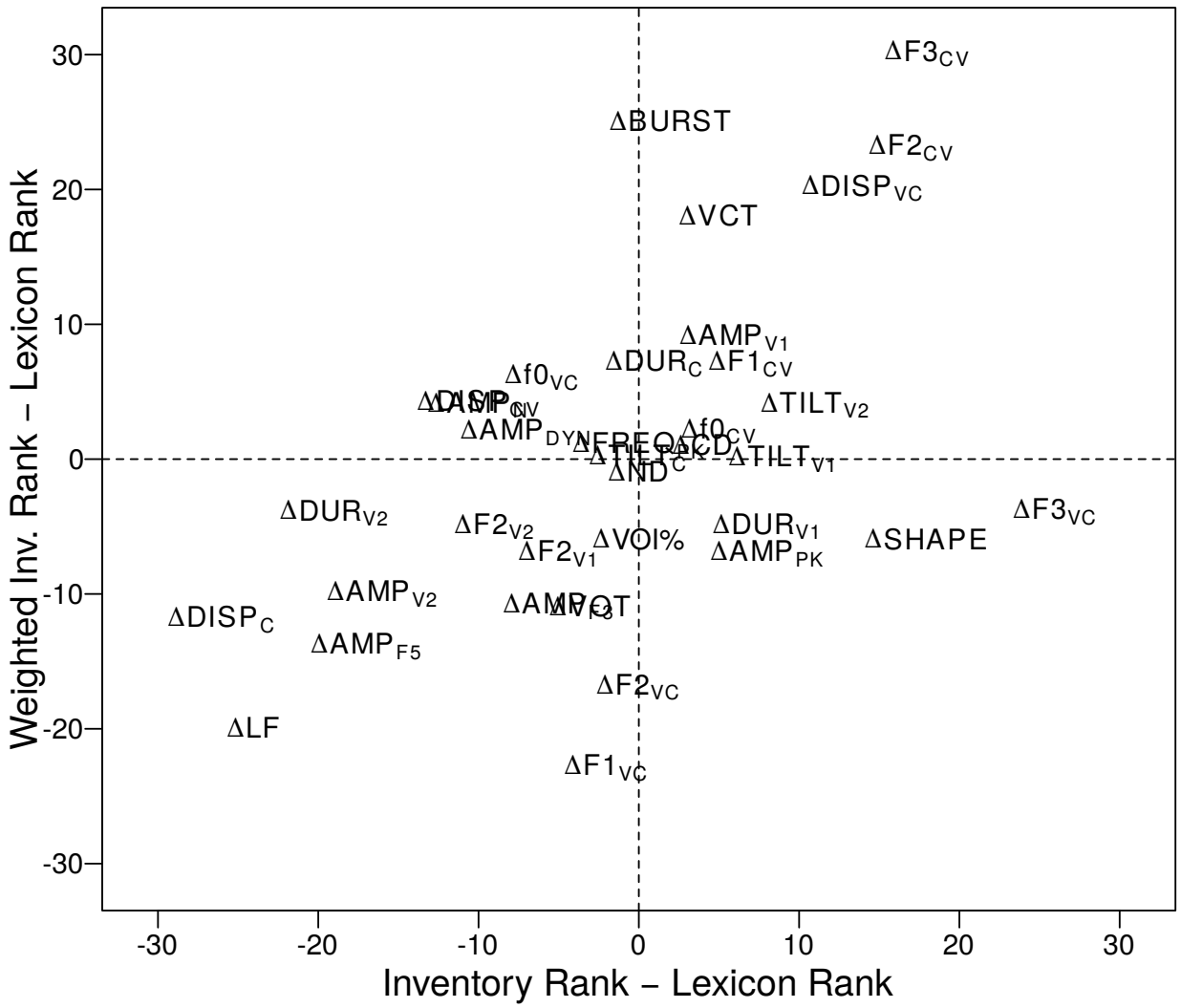

Figure 4.53: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VCV position. Dashed lines indicate equivalence relations between each pair of models.

Among the cues that are consistently overestimated by the inventory models in terms of their role in the lexicon are low-frequency energy (ΔLF) and relative F5 amplitude (ΔAMP_{F5}). Those that are consistently underestimated (across sub-experiments) include F2 and F3 at the onset of the following vowel ($\Delta F2_{CV}$, $\Delta F3_{CV}$). Cues that align along the *x*-axis represent potential points of distributional disagreement, as they indicate predictions from controlled syllable data can be brought into agreement with the lexicon model by sampling the inventory set to match the lexical distribution of obstruent contrast. This set is less consistent across sub-experiments, with spectral shape (ΔSHAPE) the only cue that is consistently overemphasized in the inventory but effectively

downweighted in the weighted inventory. In terms of acoustic disagreement—cues that are brought out of alignment in the lexicon when used to predict listener word recognition in an unbalanced contrast set—several cues closely align across sub-experiments. The best exemplar of this set is F2 at preceding vowel offset ($\Delta F2_{VC}$), which is ranked much lower in the weighted inventory model than in the inventory and lexicon models. Finally, turning to points of cue agreement, we have the set: noise duration (ΔND), consonantal spectral tilt ($\Delta TILT_C$), and spectral peak frequency ($\Delta FREQ_{PK}$) all of which are highly ranked in the three models. In the sections below we investigate in greater detail four exemplars of the above points of agreement/disagreement between the three models: consonantal spectral tilt ($TILT_C$; *cue agreement*), spectral shape (SHAPE; *distributional disagreement*), F2 at preceding vowel offset ($F2_{VC}$; *acoustic disagreement*), and low-frequency energy (LF; *composite disagreement*).

4.4.2.1 Cue agreement: Consonantal Spectral Tilt ($TILT_C$)

Figure 4.54 shows the partial dependence functions and distributions of $TILT_C$ in the inventory, weighted inventory, and lexicon models of intervocalic contrast perception. All three models show a consistent monotonic relationship between increases in $\Delta TILT_C$ and greater contrast discriminability. This agreement is further shown in the distributions of $\Delta TILT_C$ by predicted accuracy, all of which align closely at low, mid, and high accuracies.

The relationship between consonantal spectral tilt and specific contrast perception is shown in greater detail in Figure 4.55, which illustrates that the greatest distinctions in spectral tilt occur between sibilants and nonsibilants, which are generally accurately perceived in both word and syllable recognition. At the lower end of the $\Delta TILT_C$ range are contrasts between stops and sibilant voicing contrasts, all of which are relatively less robust perceptually. Thus, due to the close agreement in both the acoustics of consonantal spectral tilt and the relative perceptability of contrasts aligning along this dimension, $TILT_C$ scales well between the inventory and lexicon. In the next section we examine a cue that shows poor agreement in this respect, spectral shape (SHAPE), but which can scale to the lexicon once lexical contrast distributions are incorporated into the model.

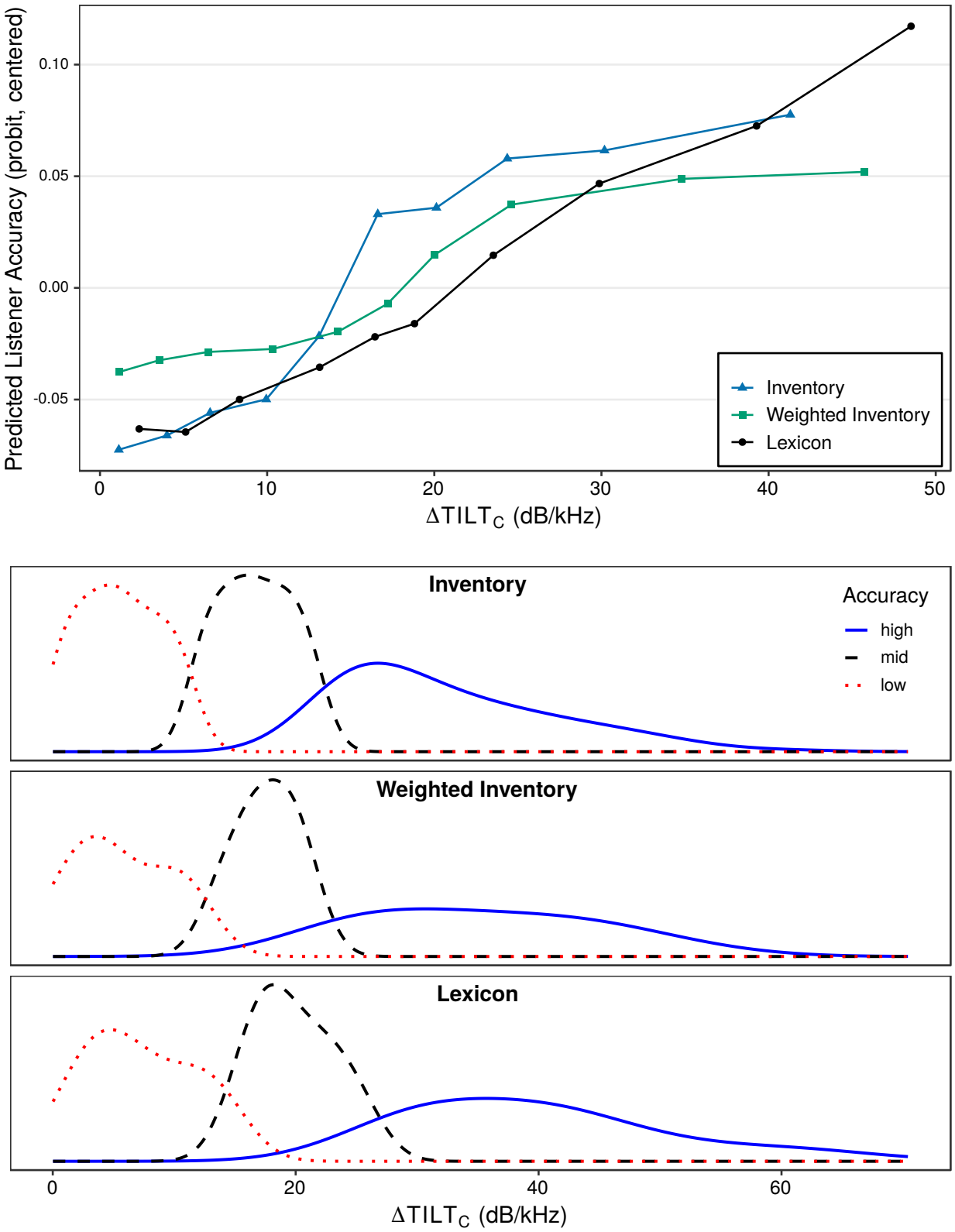

Figure 4.54: Partial dependence functions (top panel) and distributions (bottom panels) of TILT_C in the inventory, weighted inventory, and lexicon models of listener recognition in word-medial position.

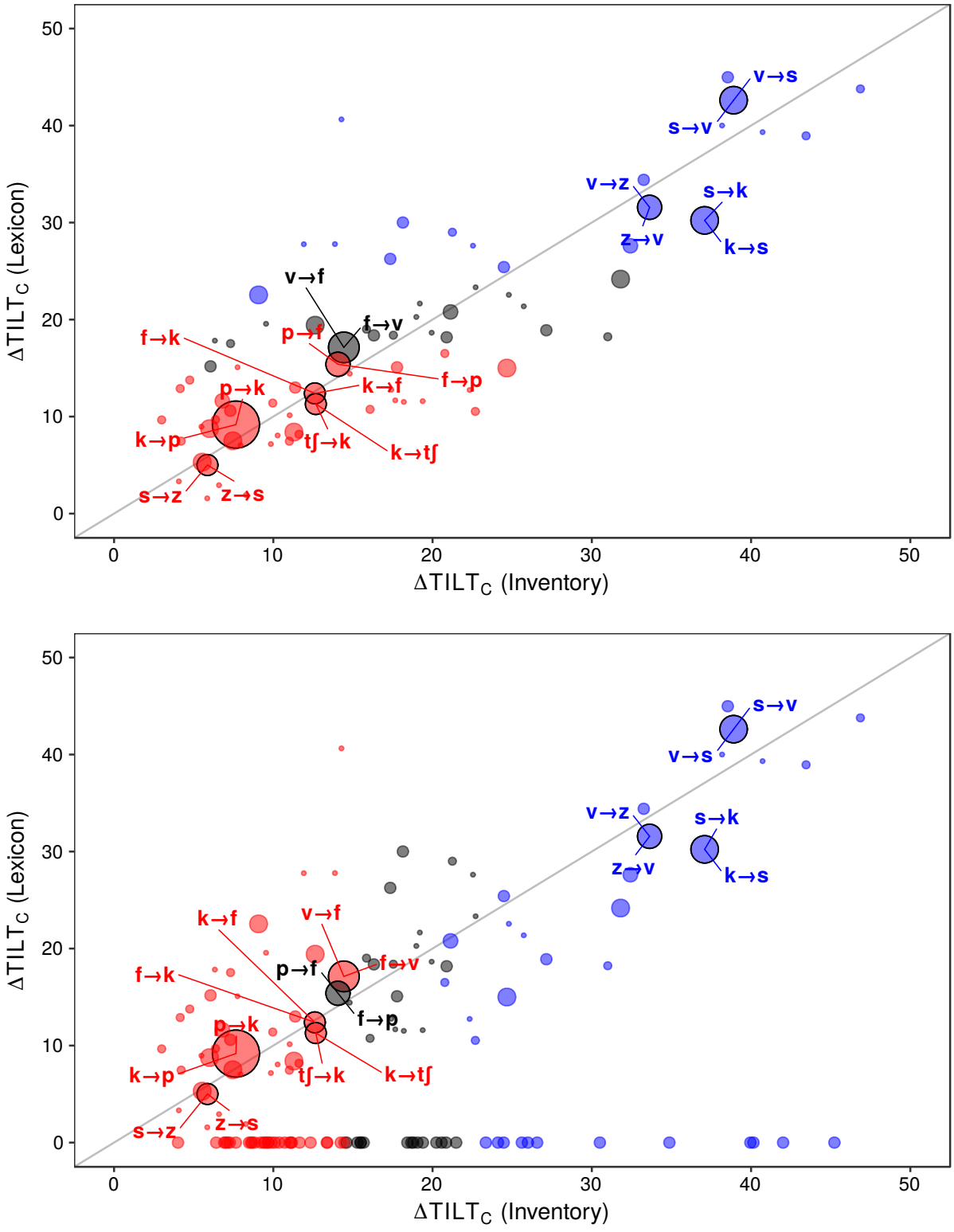

Figure 4.55: Relationship between $\Delta TILT_C$ means by phonetic contrast in the inventory and lexicon models in VCV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

4.4.2.2 Distributional disagreement: Spectral Shape (SHAPE)

As noted in Chapter 2, and at several other points in the analysis of models of cue integration, spectral shape serves primarily to distinguish postalveolars from non-postalveolars, as the spectra of the set [ʃ, ʒ, ʧ, ʤ] exhibit a uniquely prominent mid-frequency peak (even more so than velars) that shows a steep rise in energy over the low-frequency range and a steep fall in energy above the peak. Taking the difference of these two slopes yields high SHAPE values that are well above those of the remainder of the English obstruents. Thus, this cue has high discriminative potential acoustically, but depends on the relative prevalence of distinctions between postalveolar and non-postalveolar obstruents in the system under study. In the inventory model, because all contrasts are given equal weight, there are many such distinctions, which is why spectral shape is highly predictive of balanced syllable recognition in Figure 4.56. However, such contrasts are relatively sparse in the lexicon, and so little weight is attached to spectral shape in the lexicon and weighted inventory models. The distributions in Figure 4.56 further confirm this relationship as there is much greater overlap between the low-, mid-, and high-accuracy sets in these models.

Figure 4.57 further reveals that even among contrasts between postalveolars and non-postalveolars, the spectral shape distinctions in controlled syllables are much greater than in the lexicon, showing a stark discontinuity between 5 and 10 dB/kHz along the *x*-axis that is not present along the *y*-axis corresponding to spectral shape differences among real words. Thus, by downweighting the role of such contrasts in the weighted inventory model, we are able to accurately estimate the minimal utility of spectral shape in predicting listener word recognition behavior over a set of contrasts that are more representative of the English lexicon.

4.4.2.3 Acoustic disagreement: F2 at Vowel Offset ($F2_{VC}$)

The behavior of vowel-offset F2 in models of intervocalic contrast perception illustrates a case where it is possible to arrive at the same estimates of lexical cue utility from a balanced inventory model while capturing differences in the distribution of that cue and its relationship with listener perception. Figure 4.58 shows that in aggregate, both lexicon and inventory models are highly

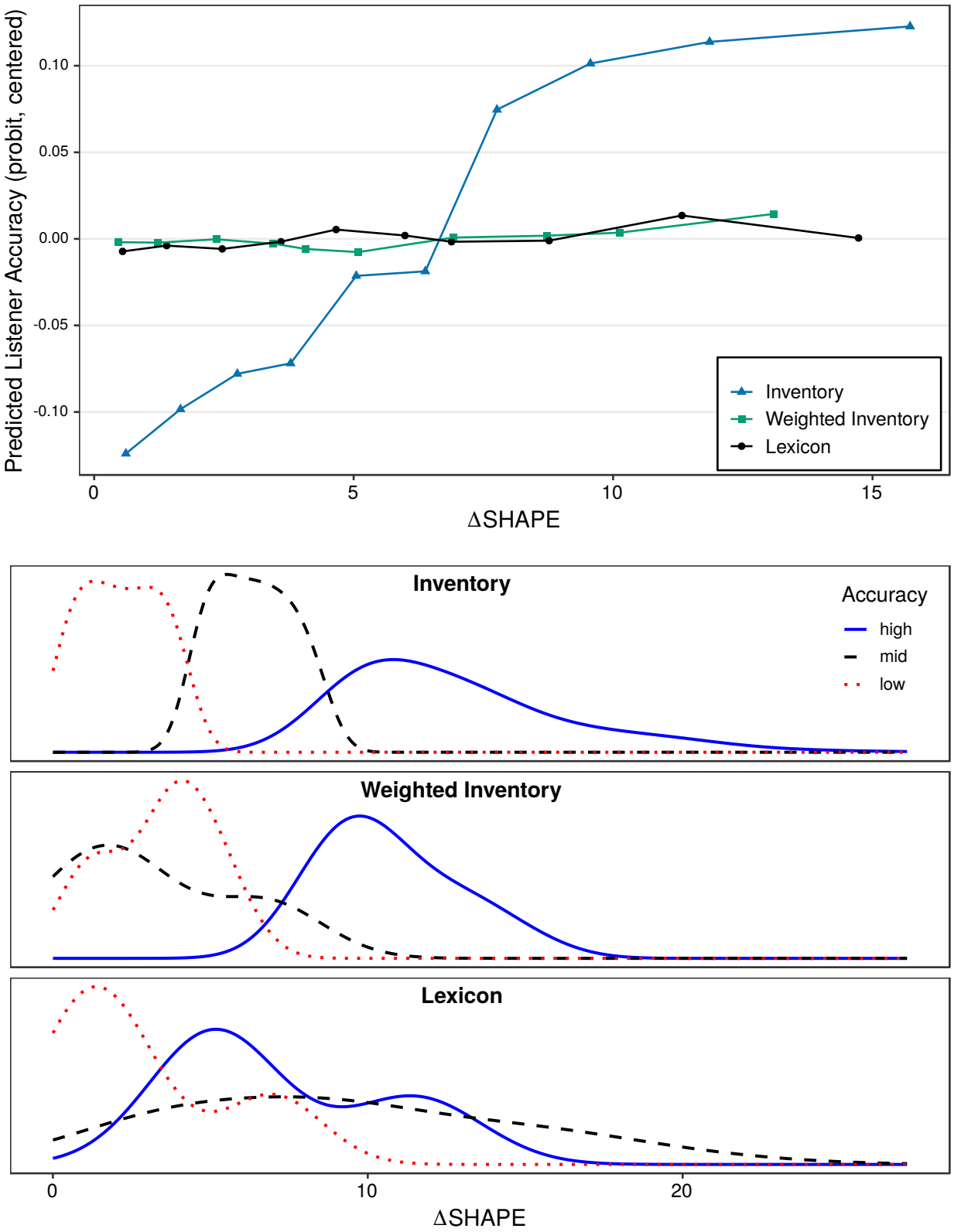

Figure 4.56: Partial dependence functions (top panel) and distributions (bottom panels) of SHAPE in the inventory, weighted inventory, and lexicon models of listener recognition in word-medial position.

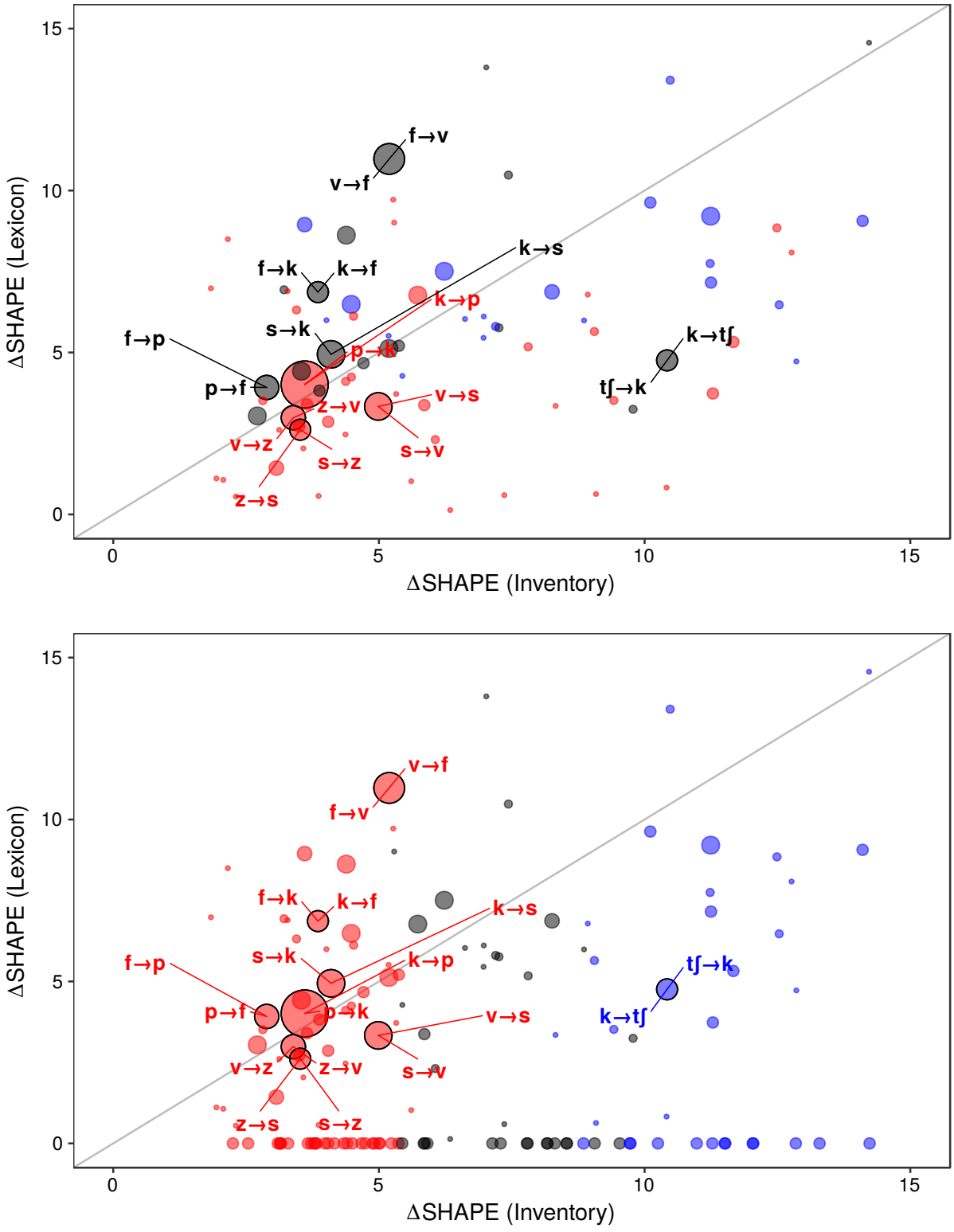

Figure 4.57: Relationship between ΔSHAPE means by phonetic contrast in the inventory and lexicon models in VCV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

predictive of listener recognition over some range of $\Delta F2_{VC}$, but the two ranges differ, as the inventory model shows a steep increase in predicted accuracy between 0 and 75 Hz, while the lexicon model shows an approximately linear increase between 0 and 300 Hz. This discrepancy is further highlighted in the corresponding distributions in Figure 4.58, wherein only high- and low-accuracy items are distinguished by $F2_{VC}$ in the inventory model, while in the lexicon model each group is clearly separated from the other, with higher-accuracy items exhibiting the greatest $F2_{VC}$ distinctions, followed by mid-accuracy items at 100 $\Delta F2_{VC} < 250$, and low-accuracy items below 100 Hz. Now, the weighted inventory model does show much closer agreement with the lexicon in this general pattern, but with greater overlap between the three sets.

Figure 4.59 clarifies that the discrepancy between the two models is both acoustic and perceptual. Acoustically, there is a clear separation of contrasts along the *y*-axis that is less clearly captured in the inventory (*x*-axis). At the upper end of the $\Delta F2_{VC}$ range are sibilance contrasts, followed by place distinctions (velar–labial, velar–postalveolar), with obstruents of the same or similar place at the lowest end. This decline matches the predicted decline based on listener data from Experiment 1, but is less consistent with the syllable recognition patterns from Cooke & Scharenborg (2008), where obstruent place perception is relatively poorer. Thus, the inventory model appears primarily to be capturing the highly salient sibilance contrasts, which both appear more often in the inventory and are relatively lower in $\Delta F2_{VC}$. This combination of incomplete acoustic and perceptual agreement leads to the poor scaling we find in the weighted inventory model, which neither has access to the syllable perception data nor the relative expansion of $F2_{VC}$ distinctions in the lexicon.

4.4.2.4 Composite disagreement: Low-Frequency Energy (LF)

Finally, we examine low-frequency energy (LF) as a case of *composite* disagreement, where both acoustic and distributional discrepancies—and potentially perceptual discrepancies as well—result in both inventory models underestimating the utility of LF in word recognition. Figure 4.60 shows the partial dependence functions of LF in each model, and while the lexicon model shows a con-

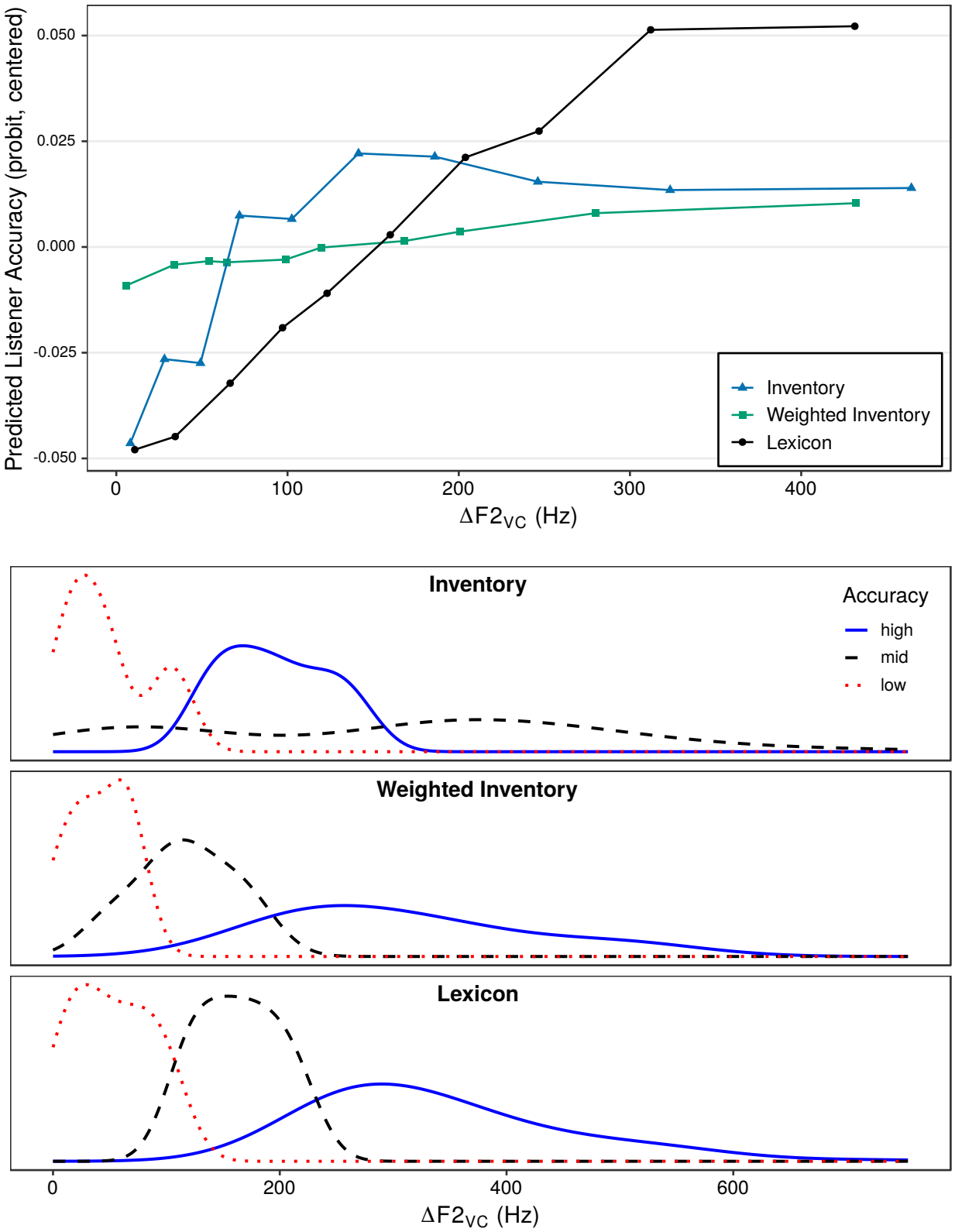

Figure 4.58: Partial dependence functions (top panel) and distributions (bottom panels) of $F2_{VC}$ in the inventory, weighted inventory, and lexicon models of listener recognition in word-medial position.

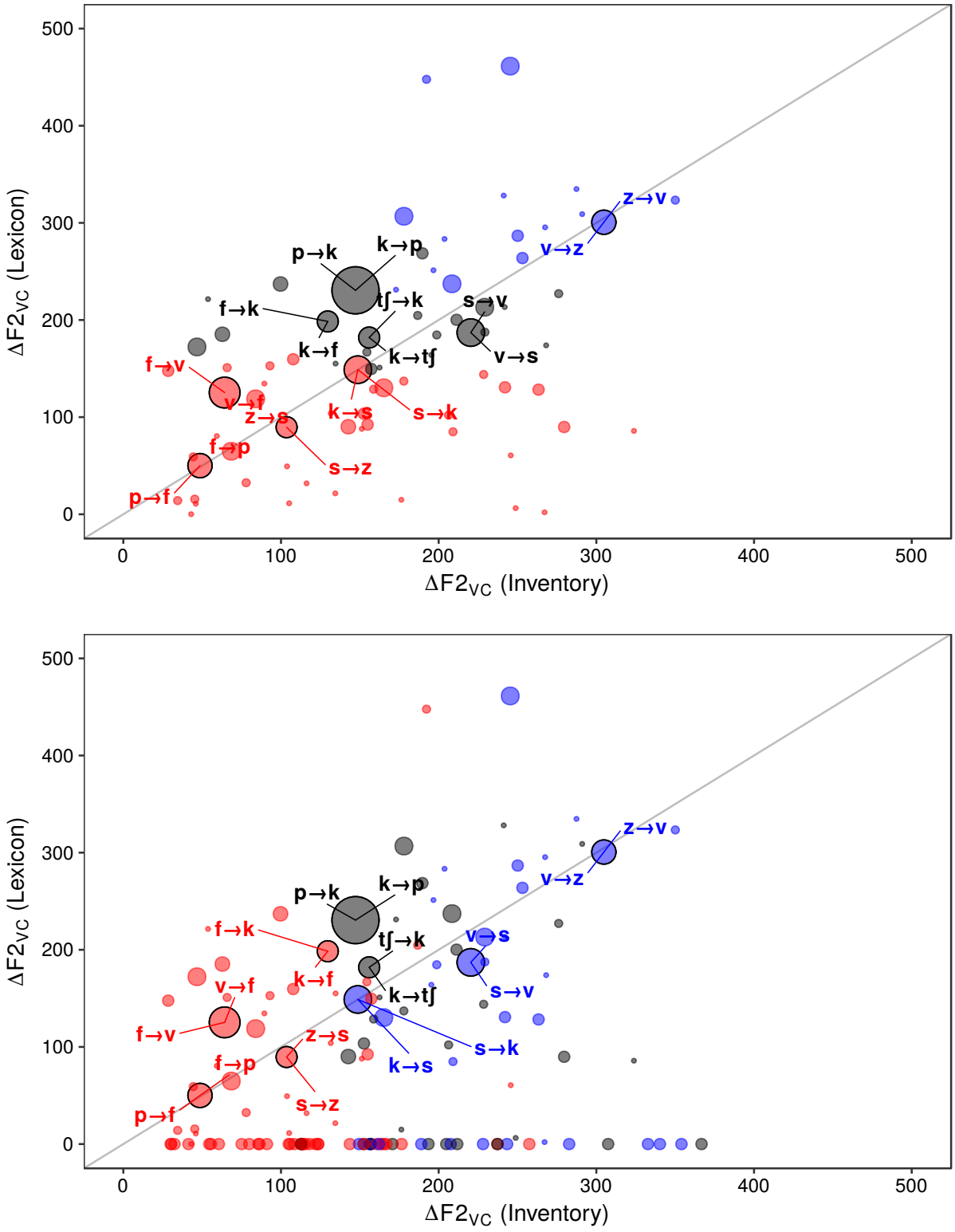

Figure 4.59: Relationship between $\Delta F2_{VC}$ means by phonetic contrast in the inventory and lexicon models in VCV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

sistent linear increase in predicted accuracy with greater distinctions in the amplitude of low-frequency energy in the consonant noise spectrum, the inventory and weighted inventory models exhibit no such relation, being either flat or erroneously predicting poorer recognition with increases in LF contrastiveness. Indeed, the distribution of low-, mid-, and high-accuracy contrasts along the ΔLF dimension in the inventory model is the reverse of what it should be if LF served as a robust cue in listener recognition. The weighted inventory model shows some improvement in aligning the high-accuracy set more closely with that in the lexicon, but the highest LF distinctions are still predicted to be the least accurate. Further, the set predicted to be of intermediate accuracy as a function of low-frequency energy distinctions occupies the lowest range of ΔLF. Such discrepancies are indicative of poor scaling from a host of potential differences between balanced syllable and representative word recognition.

Figure 4.61 shows that while distinctions in low-frequency energy profiles do more than distinguish voicing contrasts—manner distinctions are in fact the best discriminated in the lexicon according to this measure, capturing differences in supralaryngeal versus laryngeal noise sources—these distinctions are less accurately perceived in controlled syllable recognition. Further, listeners appear to be most accurate at perceiving contrasts among voiced obstruents in the inventory model, a result which is directly counter to the generally greater robustness of intervocalic voiceless obstruents in Experiment 1. Finally, even if the inventory data is used to predict listener recognition in Experiment 1, as in the weighted inventory model, the acoustic relationship between LF in the two data sets is too poorly correlated to allow for such scaling. Thus, the role of low-frequency energy in the three models of intervocalic obstruent perception provides a clear illustration of the number of assumptions that need to be met in order for a cue weight in a balanced inventory design operating on controlled syllable data to accurately generalize to the naturally unbalanced and acoustically variable problem of distinguishing words in communication.

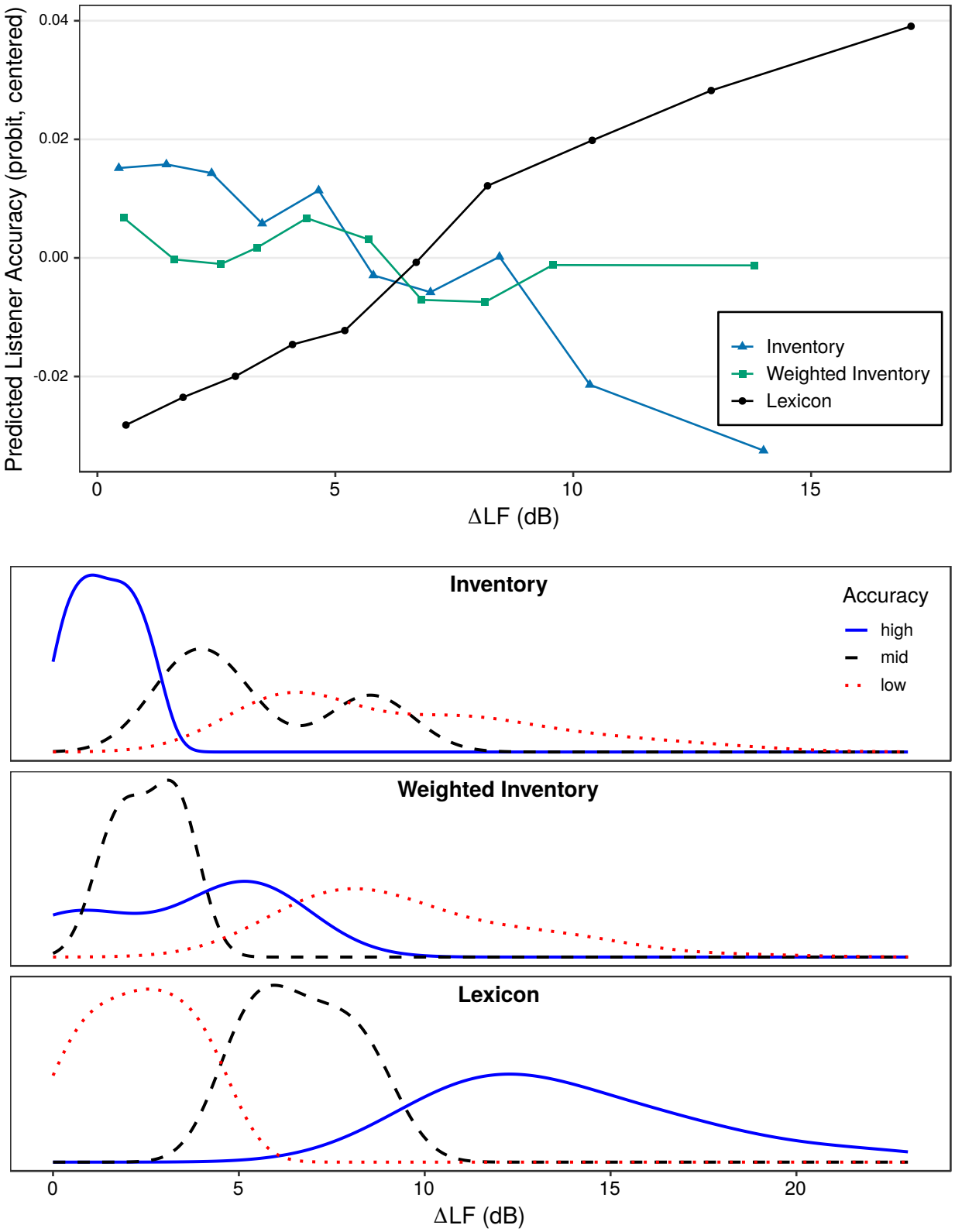

Figure 4.60: Partial dependence functions (top panel) and distributions (bottom panels) of LF in the inventory, weighted inventory, and lexicon models of listener recognition in word-medial position.

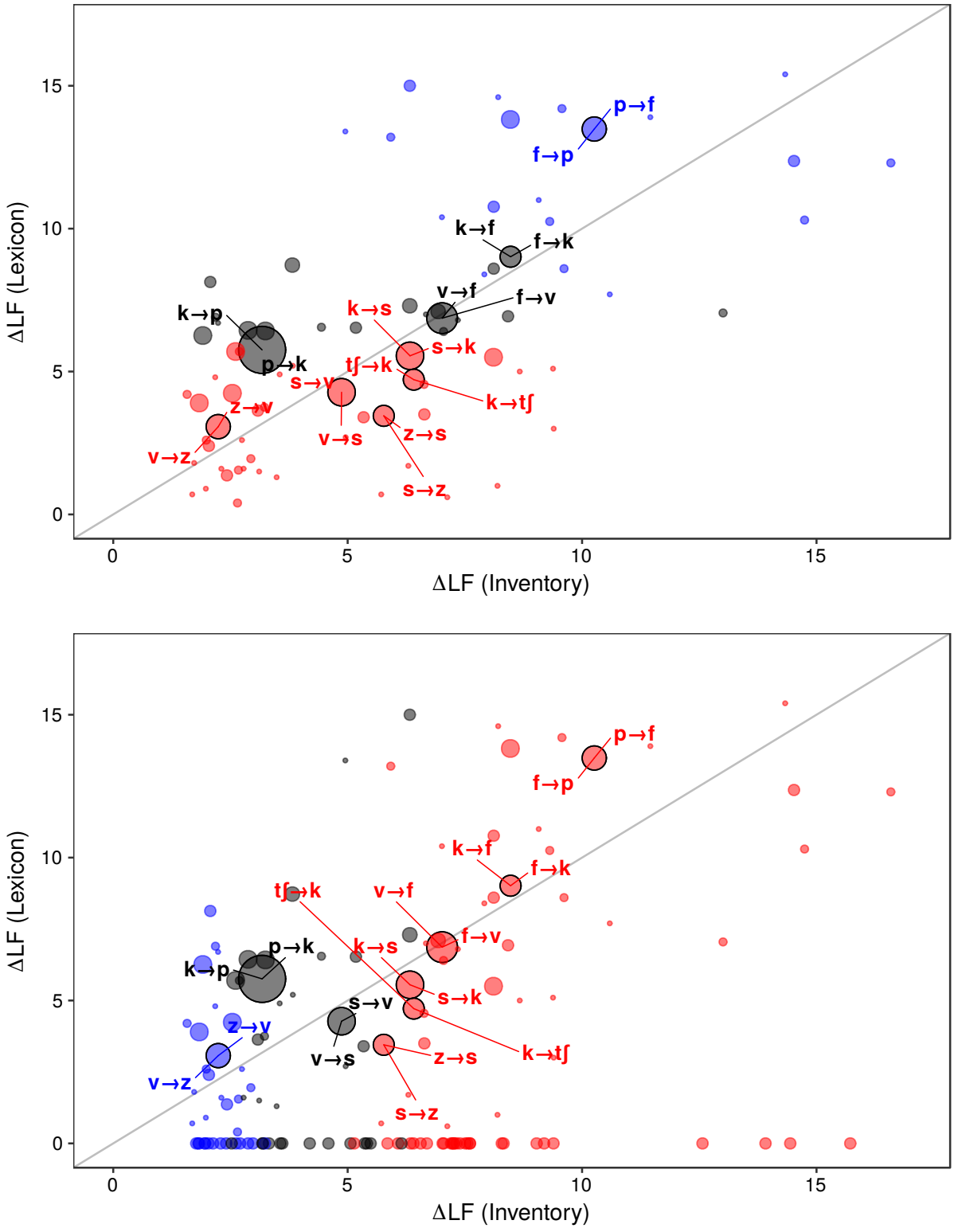

Figure 4.61: Relationship between ΔLF means by phonetic contrast in the inventory and lexicon models in VCV position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 25% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

4.4.3 Word-final position (VC)

We conclude the analysis of cue integration in listener recognition with an examination of models of word-final obstruent perception in the inventory and lexical systems. Figure 4.62 shows the fit of the inventory, reference, weighted inventory, and lexicon models, and as in the other two positions each model shows a consistent monotonic increase in predicted accuracy over the observed accuracy range. The overall fit of the VC models is comparable to that in CV and VCV positions, with the inventory model exhibiting the lowest RMSE (0.231) and the closest fit of any listener model. The reference model fit is somewhat worse at 0.322, which is intermediate between the CV and VCV models; however the contrast cues in the inventory and reference models correlate significantly ($r = 0.611,\ p = 0.001$), though as in VCV position the correlation between target cues in the two models was not significant ($r = 0.191,\ p > 0.1$). Finally, the weighted inventory and lexicon models exhibit the worst fits of the four at 0.422 and 0.382, respectively. We can see in Figure 4.62 that this difference in model fit is due to the word-recognition models systematically underestimating listener accuracy as listeners approach ceiling performance ($> 90\%$), and overestimating listener accuracy in the range approaching chance performance ($< 70\%$).

The final model fits after applying a linear transformation to the BART model predictions are shown in Figure 4.63. All models show a benefit from this meta-modeling approach of between 2 and 4 percent, where the greatest change is observed in the lexicon model (4% reduction, from an RMSE of 0.382 to 0.369), and the smallest change is in the reference model (2% reduction, from 0.322 to 0.318). These improvements are the smallest of the three positions, but are nevertheless critical in generating closer agreement between the model and the data at the accuracy extrema; i.e., at ceiling and chance performance. As noted in the previous sections, this transformation from the meta-model does not affect any of the cue weights, just the model predictions that will be used in the cue perturbation analysis in Chapter 5.

Figures 4.64 and 4.65 show the target and contrast cue ranks, respectively, in the word-final listener models (see Figures A.89–A.92 in the appendix for the corresponding cue ranks in each sub-experiment). The target cues that are most predictive of listener recognition in Experiment

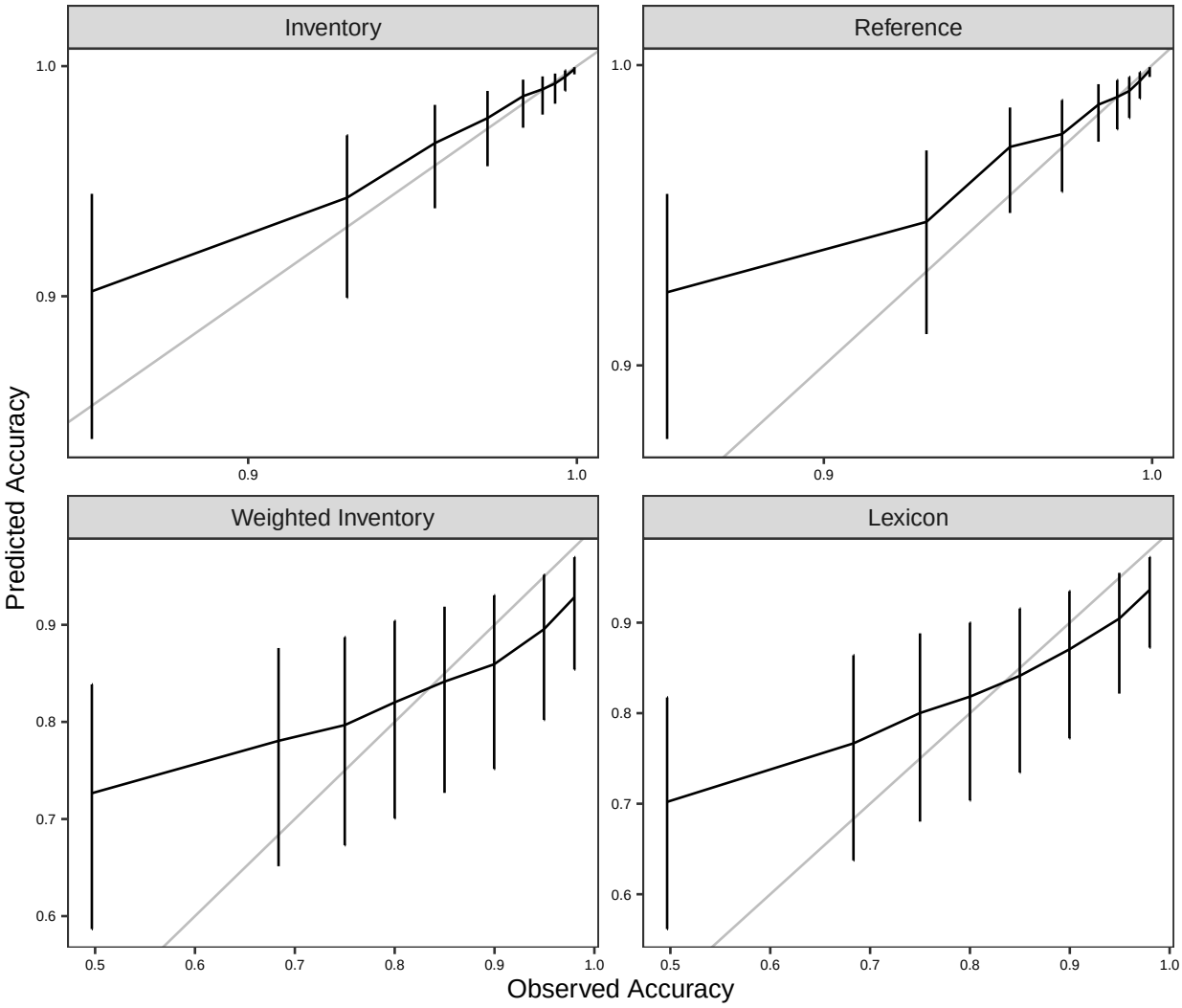

Figure 4.62: Listener model fit in the inventory, reference, weighted inventory, and lexicon models of word-final contrasts. Lines indicate posterior medians averaged at the observed quantiles (0, 0.1, ..., 1). Error bars indicate the middle 90% of the posterior distribution. Solid grey lines show the identity function.

1, and whose role replicates across the disjoint item sets in Experiments 1a and 1b, are noise amplitude (AMP_N), noise duration (ND), F1 and F3 at vowel offset ($F1_{VC}$, $F3_{VC}$), and spectral dispersion of the consonant noise interval ($DISP_C$). The impact of noise amplitude is relatively straightforward: listeners are more accurate in general on louder obstruents. This result extends across obstruents, but also highlights the relative differences between sibilants and nonsibilants, the former being more reliably perceived on average.

Regarding noise duration, obstruents with longer noise intervals tend to be better perceived,

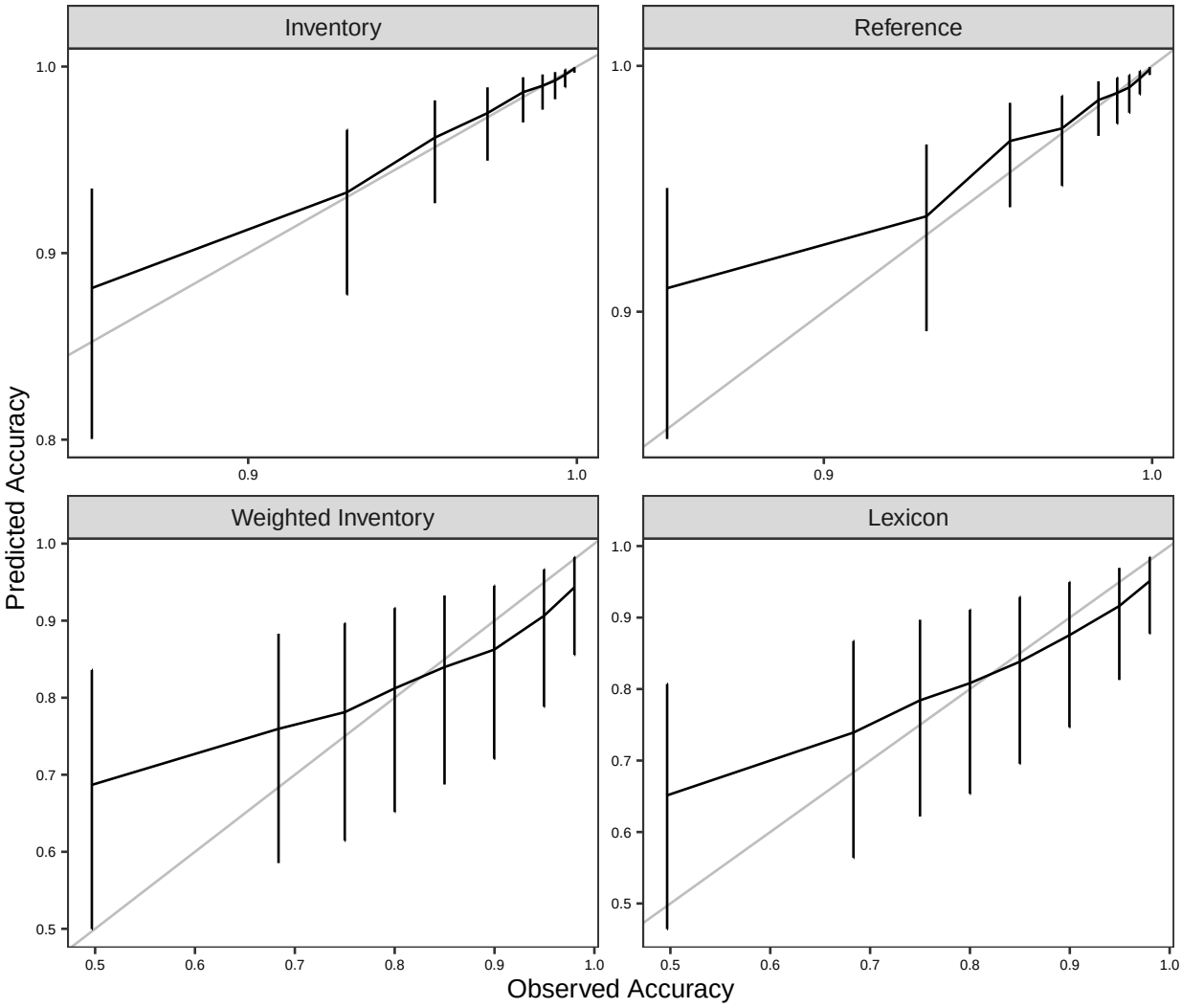

Figure 4.63: Rescaled listener model fit based on a meta-model of the inventory, reference, weighted inventory, and lexicon predictions of word-final contrasts. Lines indicate posterior medians averaged at the observed quantiles (0, 0.1, ..., 1). Error bars indicate the middle 90% of the posterior distribution. Solid grey lines show the identity function.

particularly above 150 ms, which distinguishes the voiceless fricative set from the remainder of the obstruents. This result is consistent with the perceptual patterns in Chapter 3, and reflects the fact that the generally weak word-final environment makes the duration of the noise interval particularly important at governing the likelihood of listeners parsing spectral cues from the consonant. Noise duration is further useful as a voicing cue, but this role will be covered in more detail in the discussion of contrast parameters.

The first and third formants at vowel offset are primarily considered to serve as cues to ob-

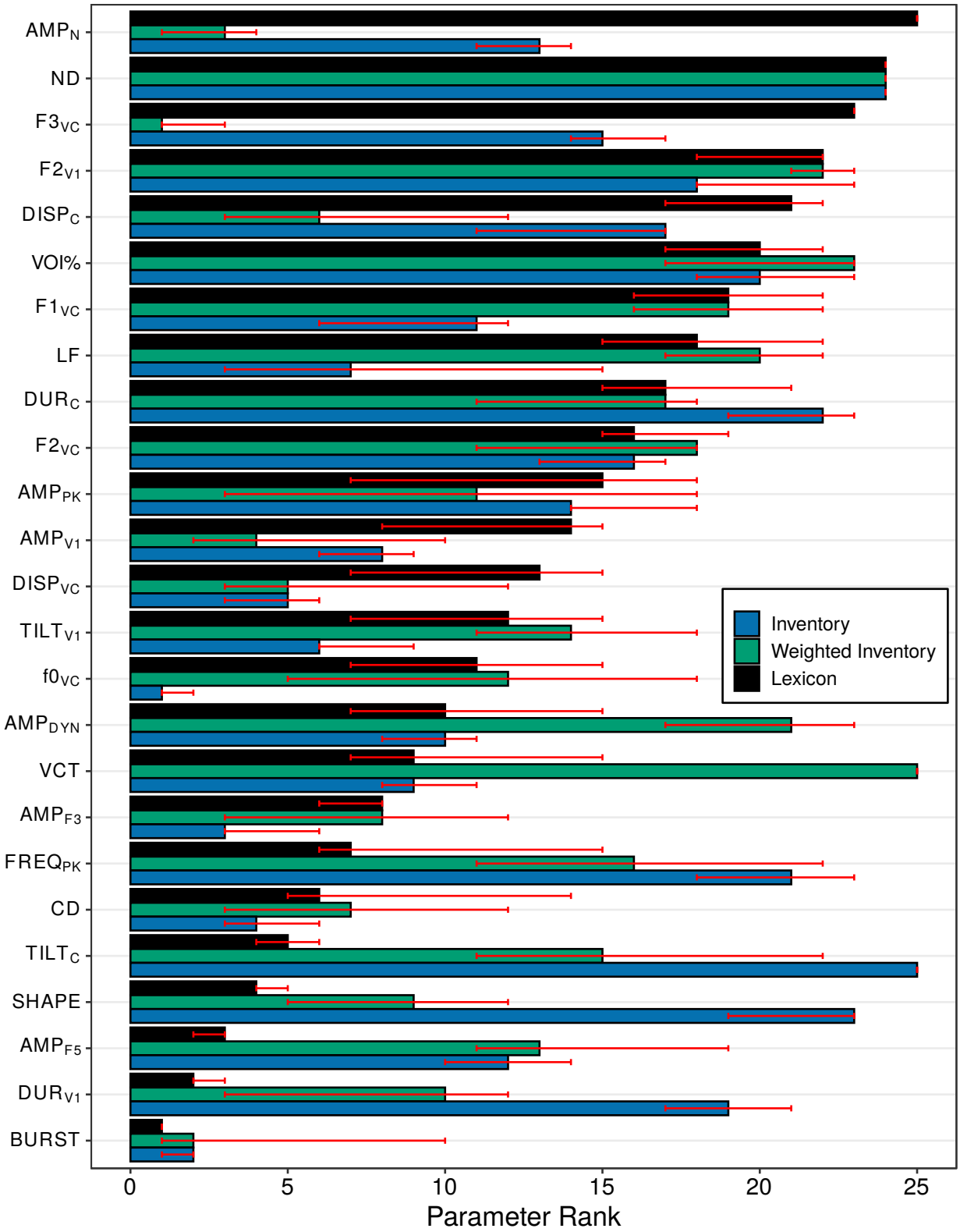

Figure 4.64: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-final contrasts fit to listener recognition data. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

struent voicing and place/manner, respectively; however, their absolute values in the target word also correlate highly with listener recognition. In the case of F1, voiceless obstruents are better perceived in general than voiced obstruents in VC position, which is why listener accuracy is predicted to increase on average with higher F1 values, the lower range corresponding almost exclusively to voiced obstruents. F3 show an even starker jump in predicted accuracy at higher offset F3 values (above approximately 2500 Hz), but serves rather to partition the obstruent set into plosives and non-plosives, the latter being generally more accurately perceived word-finally than the former. Finally, consonantal spectral dispersion provides a clear separation between two obstruent classes differing broadly in manner and place. Among those obstruents with relatively more dispersed spectra are the fricatives [f, θ, v, ð, s, z] and the plosive [t], most of which are accurately perceived word-finally (see Figures 3.4 and 3.13 in Chapter 3 for details). The remainder of the obstruents, particularly the voiced plosives, have spectra with a much less dispersed distribution of energy, and given that this set also tends to be poorly perceived word-finally, DISP_C becomes a useful predictor of listener accuracy, though it remains to be seen whether there is evidence for the use of spectral dispersion as a cue in perception, or if it merely correlates well with listener behavior.

Figure 4.65 shows contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-final contrast perception. The most highly weighted cues in the lexicon that are consistently predictive in both sub-experiments are preceding vowel duration ($\Delta\text{DUR}_\text{V1}$), noise duration (ΔND), and spectral peak amplitude ($\Delta\text{AMP}_\text{PK}$), while those that are consistently poor as word-final cues include consonant voicing percentage (ΔVOI%), relative F3 amplitude ($\Delta\text{AMP}_\text{F3}$), low-frequency energy (ΔLF), and F3 at vowel offset ($\Delta\text{F3}_\text{VC}$). This latter result answers our question regarding the extent to which F3 is used as a cue in contrast discrimination, or simply indexes those obstruents that tend to be better perceived overall, supporting the latter. Other cues that are highly ranked in Figure 4.65, but in the inventory models rather than the lexicon, are consonant voicing percentage (both models), consonant duration (inventory), voice cessation time (weighted inventory), spectral peak frequency (inventory), and dynamic amplitude (weighted inventory), while ND, TILT_C, and AMP_PK are highly ranked in all three models. Figures 4.66 and

4.67 further clarify the relationship between cue ranks in the three models by presenting rank correlations and rank differences, respectively, between the lexicon and inventory models. In general, cue ranks among the word-final models are much more closely correlated than those in CV and VCV positions. However, there are still many points of disagreement evident in both the outliers in Figure 4.66 and the diagonal spread of rank differences in Figure 4.67.

Beginning with points of agreement, Figure 4.67 shows a tight cluster of cues around the origin, indicating a close correspondence in cue ranks across all three models. Some of these cues, such as LF, AMP_{V1}, and CD, agree in being similarly low-weighted in all three models, and so we will focus rather on the cues that are in agreement in being highly weighted in each model. Among the higher-ranked cues, noise duration is the only one that shows consistent ranks in all three models that replicate across Experiments 1a and 1b (see Figures A.91 and A.92 in the appendix for details). This suggests that noise duration is the least dependent on lexical contrast distributions, likely due to its pervasiveness as a cue to multiple featural distinctions—most notably manner of articulation, but also voicing and place to a lesser extent (see Section 2.4.4 of Chapter 2 for details).

Potential points of distributional disagreement—i.e., cues exhibiting much closer rank agreement between the weighted inventory and lexicon models than between the lexicon and (unweighted) inventory—are shown aligned more along the *x*-axis than the *y*-axis in Figure 4.67, and are relatively few in number, $DISP_{VC}$, AMP_{F3}, and $FREQ_{PK}$ being the most prominent in this set in the aggregate Experiment 1 model. However, only spectral peak frequency consistently exhibits this pattern in both sub-experiments wherein the relative rank of spectral peak frequency is downweighted in the weighted inventory model and thereby brought into closer agreement with the role of $FREQ_{PK}$ in the lexicon.

Cues aligning along the *y*-axis in Figure 4.67 indicate potential points of *acoustic disagreement*: i.e., cases where the cue ranks between the inventory and lexicon are similar but reflect different underlying acoustic/perceptual distributions, resulting in poor scaling between the two when data from the former is used to predict recognition of the latter. Cues in this set include preceding vowel duration (DUR_{V1}), f0 at vowel offset ($f0_{VC}$), and voice cessation time (VCT), where only DUR_{V1}

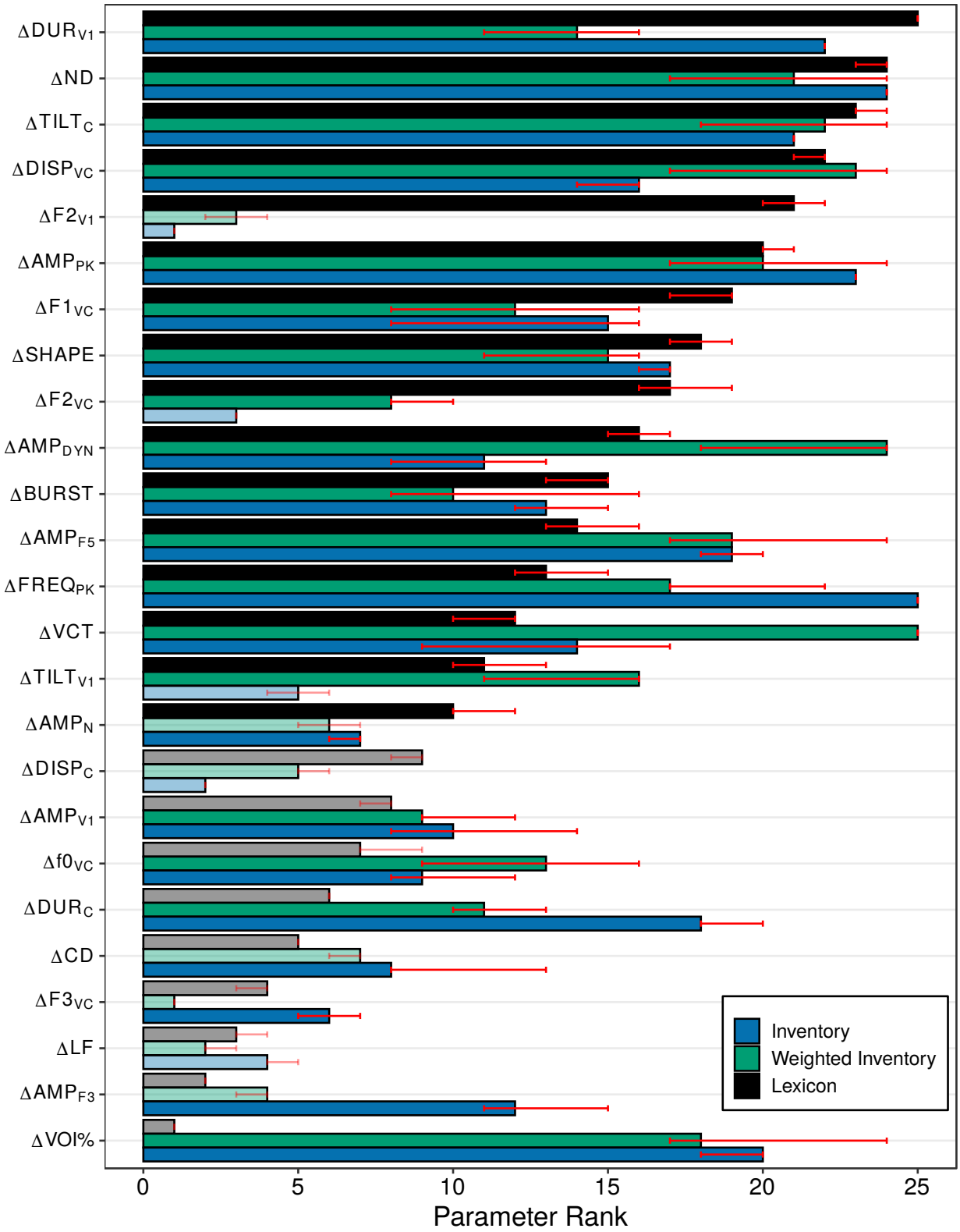

Figure 4.65: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-final contrasts fit to listener recognition data. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

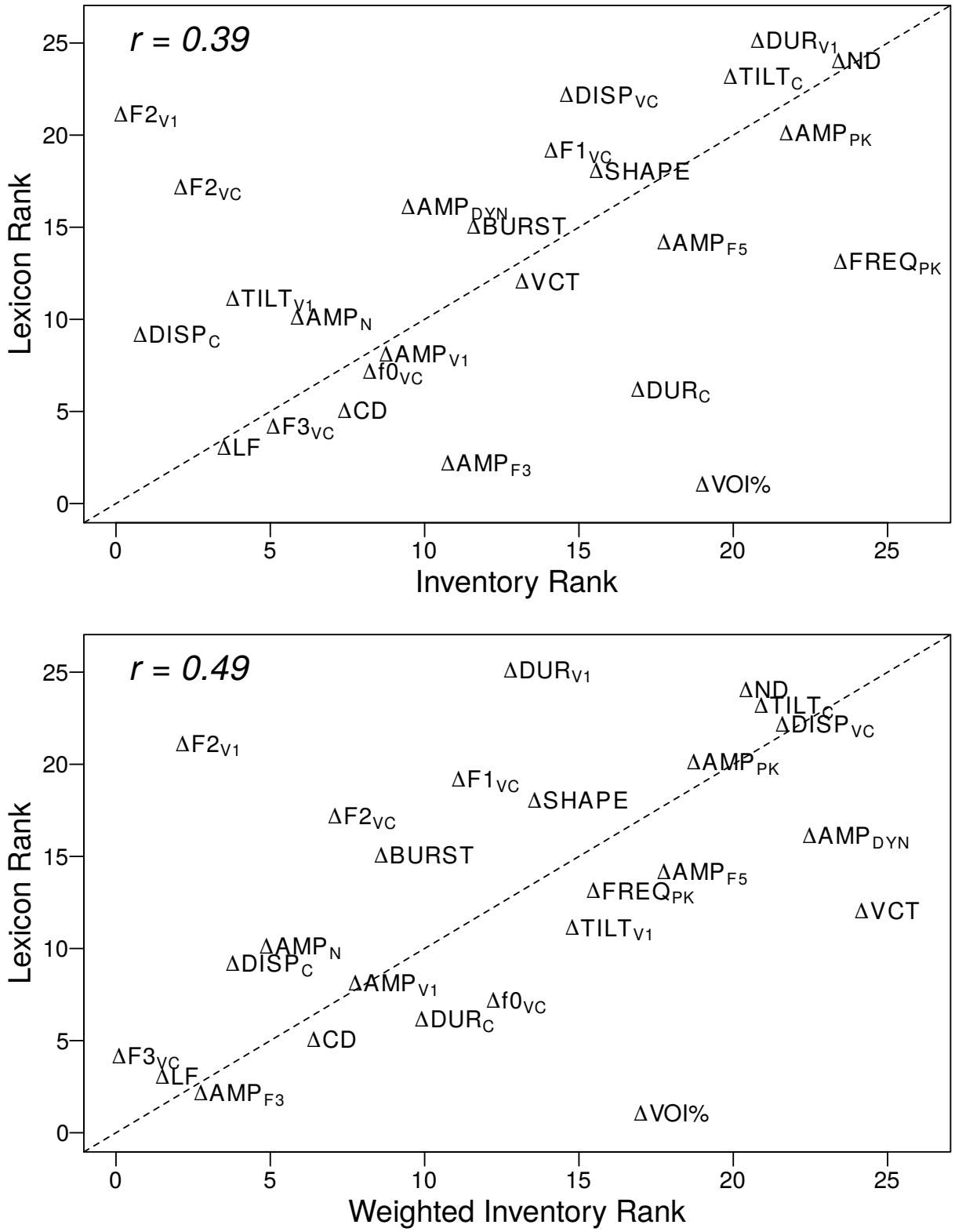

Figure 4.66: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VC position. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

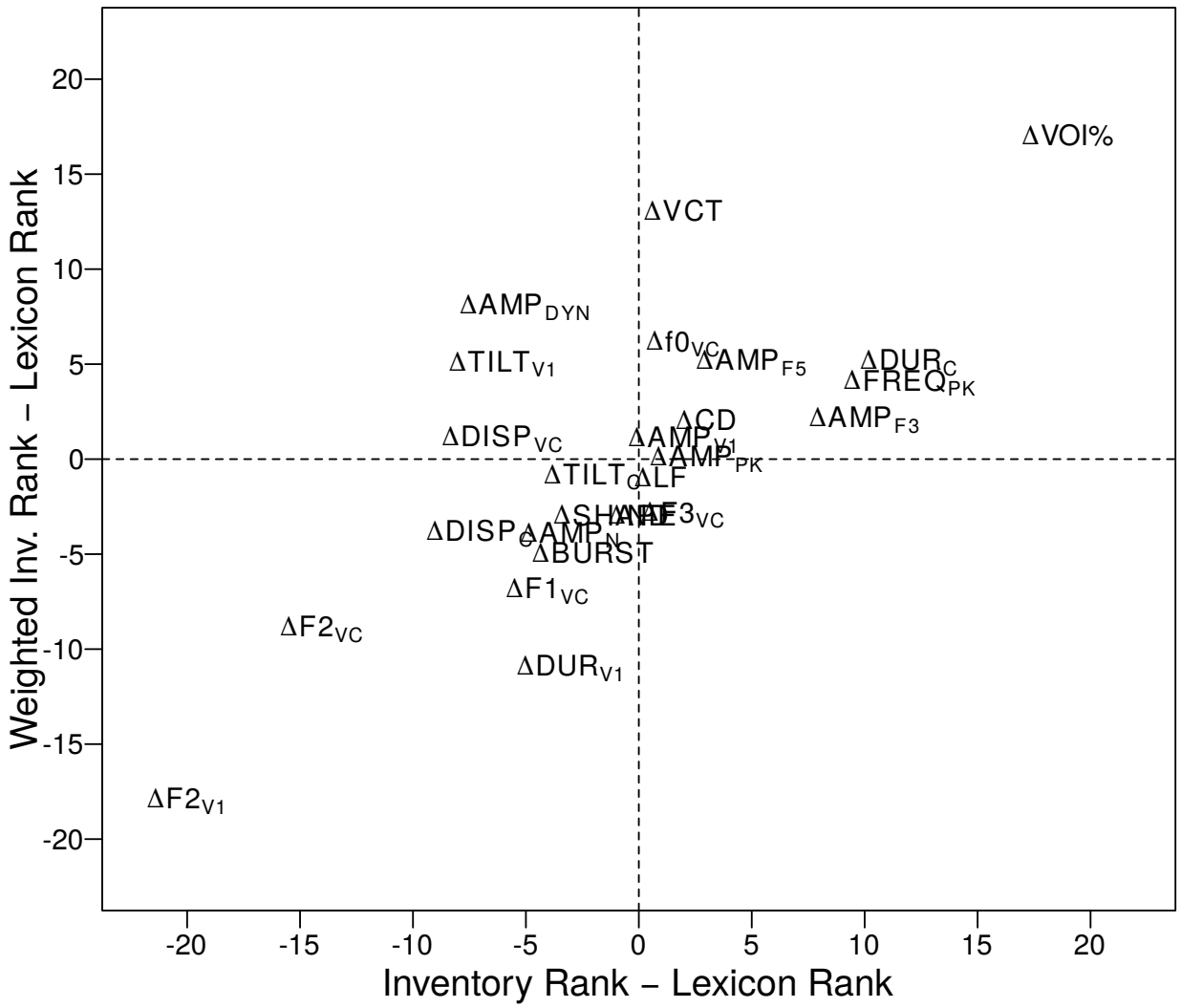

Figure 4.67: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VC position. Dashed lines indicate equivalence relations between each pair of models.

and VCT exhibit this relation consistently across sub-experiments.

Finally, points of *composite disagreement* are classified as those where both distributional and acoustic factors appear to be involved in the poor scaling between the inventory and lexicon. Such cues align along the $y = x$ line in Figure 4.67, with those in the upper right quadrant (positive values for both $rank_{inv} - rank_{lex}$ and $rank_{winv} - rank_{lex}$) representing points of joint overestimation of cue utility in the lexicon by the inventory models, and those in the lower left comprising the cues whose role in the lexicon is similarly underestimated. The most notable cues in this set are consonant voicing percentage (VOI%) as a case of joint overestimation, and F2 at both preceding

vowel offset and midpoint ($F2_{VC}$, $F2_{V1}$) as cases of joint underestimation. VOI% and $F2_{VC}$ are the most consistent in this respect across sub-experiments (see Figures A.95 and A.96 in the appendix for details). Next we review in detail four exemplars of this cue-scaling taxonomy: noise duration (ND; *cue agreement*), spectral peak frequency ($FREQ_{PK}$; *distributional disagreement*), preceding vowel duration (DUR_{V1}; *acoustic disagreement*), and consonant voicing percentage (VOI%; *composite disagreement*).

4.4.3.1 Cue agreement: Noise Duration (ND)

Figure 4.68 shows partial dependence functions and cue distributions for noise duration in the lexicon, inventory, and weighted inventory models of word-final contrast perception. All three models show monotonic increases in predicted listener accuracy over most of the ΔND range, with the inventory model showing the strongest relationship of the three, while the weighted inventory exhibits the shallowest partial dependence function, which is also non-monotonic-increasing over the bottom 45% of the distribution (< 90 ms). The lexicon model is somewhat intermediate between the two, being broadly shallower than the inventory model, but with relative agreement in two points where predicted accuracy increases rapidly with increases in ΔND: between approximately 25 and 45 ms, and after 125–150 ms. Figure 4.68 also confirms the acoustic source of this agreement in the similar ΔND distributions in each model, both in terms of their general range, and in the degree of separation between low-, mid-, and high-predicted-accuracy sets.

Figure 4.69 reveals the phonetic source of this agreement is in the consistent separation of manner and fricative voicing distinctions from within-manner distinctions, particularly among the plosives [p, t, k, d] and the fricatives [v, z]. This partition coincides with predicted listener accuracy as a function of ΔND in both Experiment 1 and in the controlled syllable perception data of Woods et al. (2010); however, the relative separation of each set in the inventory is slightly higher than that in the lexicon. This is also the source of the modest downweighting of noise duration in the weighted inventory model, where there is greater overlap between low- and mid-accuracy contrasts due to the enhancement of *z–v* and *t–d* distinctions and reduction of *p–k* and *t–v*, the former

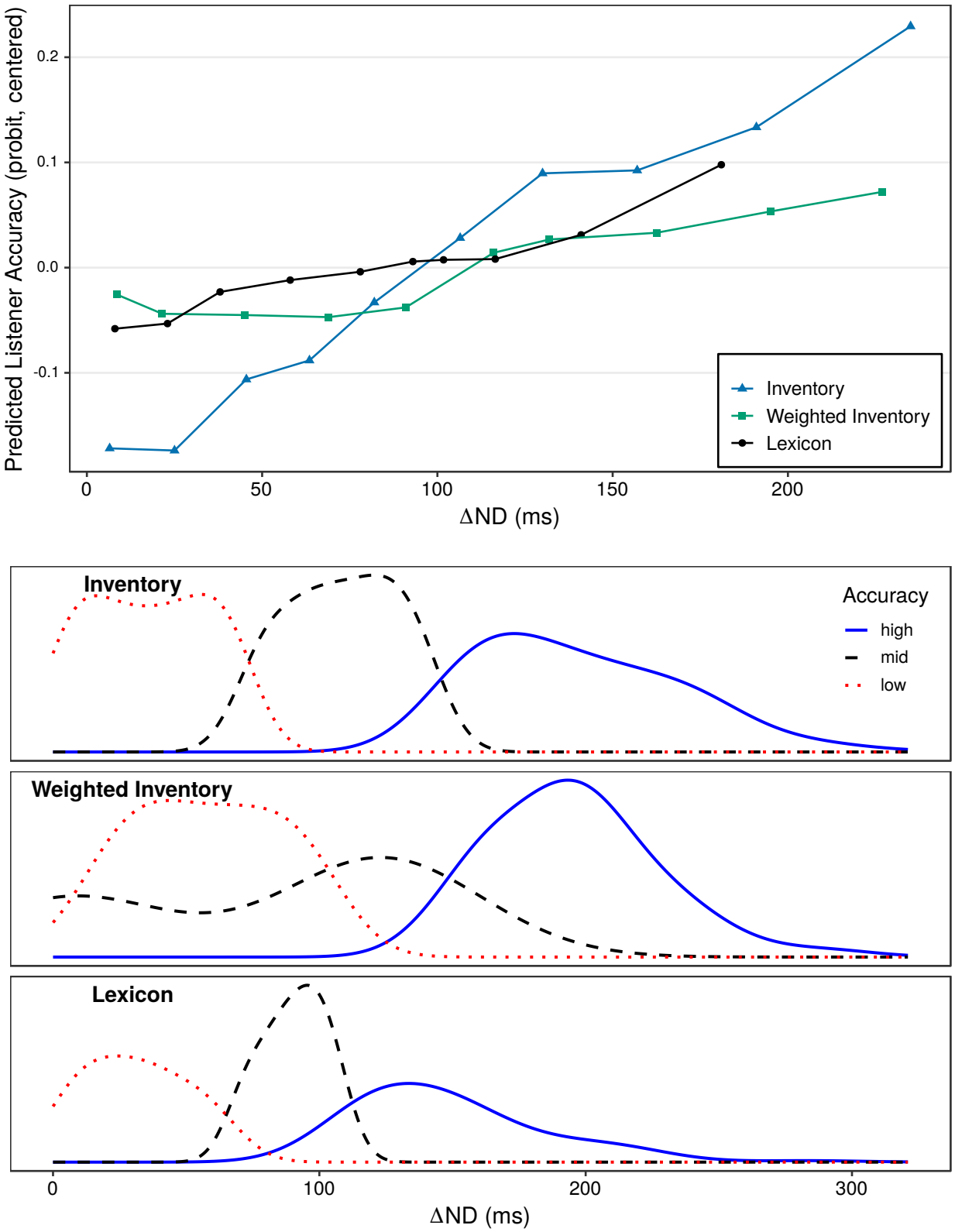

Figure 4.68: Partial dependence functions (top panel) and distributions (bottom panels) of LF in the inventory, weighted inventory, and lexicon models of listener recognition in word-final position.

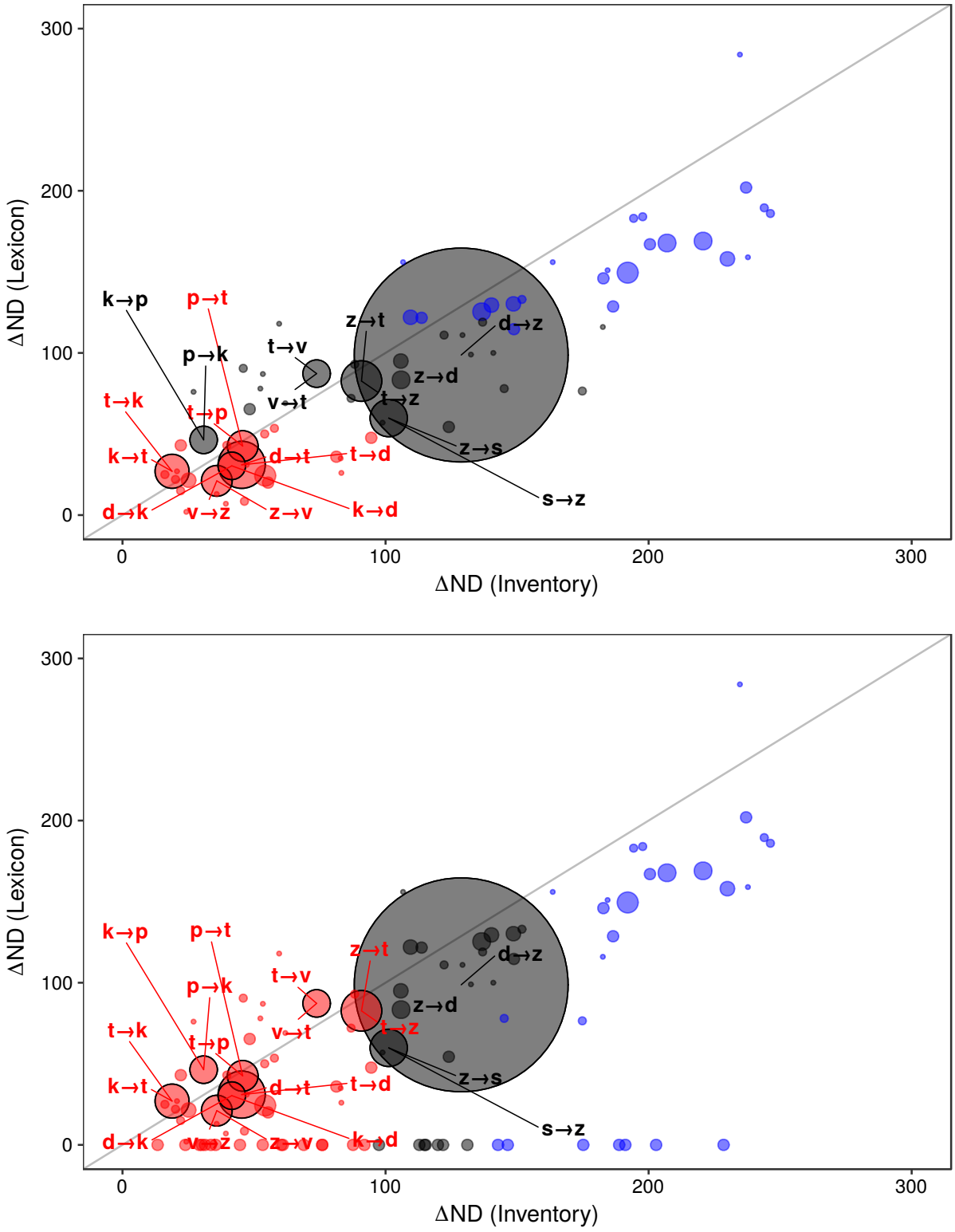

Figure 4.69: Relationship between ΔND means by phonetic contrast in the inventory and lexicon models in VC position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 40% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

exhibiting lower perceptual dependence on ND in word recognition than the latter. However, due to the close acoustic agreement between inventory and lexicon estimates of ΔND for *d–z*, which comprises 20% of VC contrasts in the lexicon, the weighted inventory model comes into closer agreement above 100 ms. Here we have a clear example of how significant asymmetries in lexical contrast distributions impact the general scaling problem between the inventory and lexicon.

4.4.3.2 Distributional disagreement: Spectral Peak Frequency ($\mathbf{FREQ_{PK}}$)

Turning next to spectral peak frequency, a case of distributional disagreement in cue ranks, Figure 4.70 shows stark separation of contrasts in the inventory model below and above $\Delta FREQ_{PK}$ = 1600 Hz. The weighted inventory and lexicon models, however, exhibit shallow partial dependence functions, indicating that there is little-to-no relationship between distinctions in spectral peak frequency and listener recognition of word-final contrasts. The distributions in Figure 4.70 further reveal that all three models do in fact distinguish between low- and high-accuracy items as a function of $\Delta FREQ_{PK}$, the difference being that the accuracy range in the lexicon and weighted inventory models is much narrower than that in the inventory model.

Figure 4.16 illustrates that spectral peak frequency exhibits both greater variation in the inventory, and greater separation of sibilance/place contrasts from within-sibilance/place distinctions (plosive place contrasts are generally less robust than those among fricatives, but not in the case of contrasts involving [t]). This difference directly relates to the predictive power of spectral peak frequency in the inventory and lexicon/weighted-inventory models, and is attributable to the overall greater weakening of obstruents in VC position in real words versus controlled syllables, particularly in di- and tri-syllabic items.

4.4.3.3 Acoustic disagreement: Preceding Vowel Duration ($\mathbf{DUR_{V1}}$)

The opposing configuration of model ranks in the lexicon, inventory, and weighted inventory, is one where the former two agree in cue ranks while the latter does not. That is, in such cases weighting the inventory data to match lexical contrast distributions causes a decrease rather than

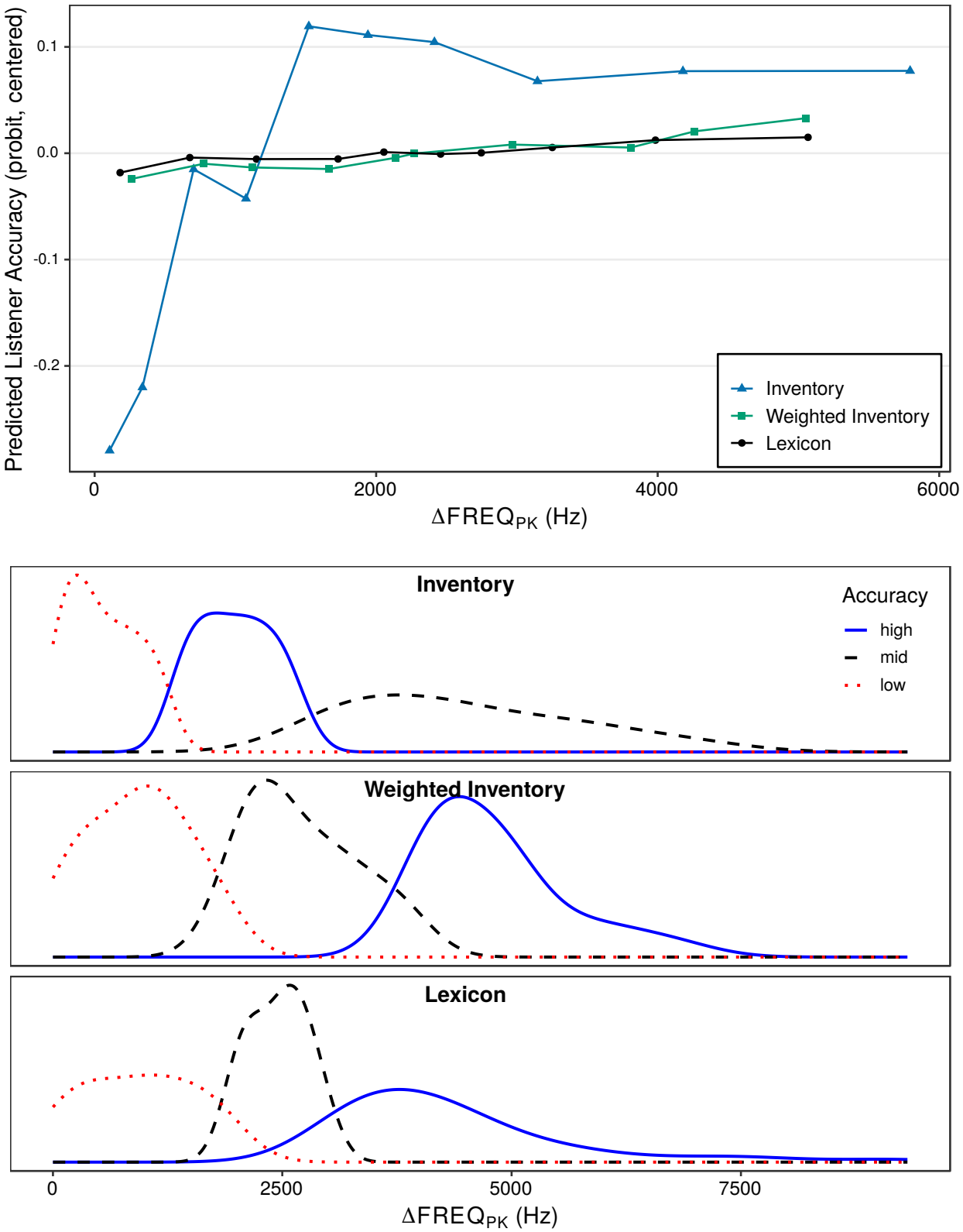

Figure 4.70: Partial dependence functions (top panel) and distributions (bottom panels) of FREQ_{PK} in the inventory, weighted inventory, and lexicon models of listener recognition in word-final position.

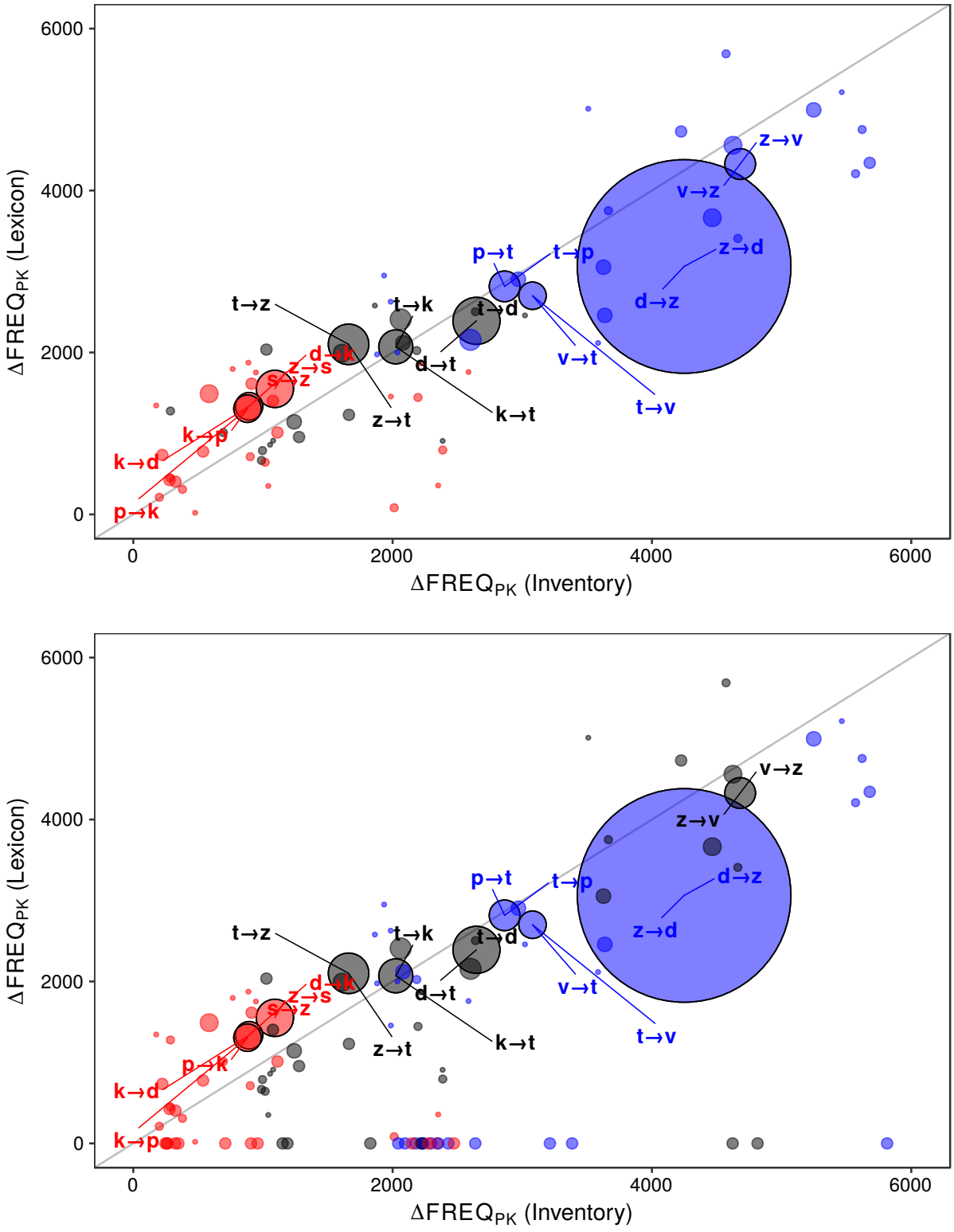

Figure 4.71: Relationship between $\Delta\text{FREQ}_{\text{PK}}$ means by phonetic contrast in the inventory and lexicon models in VC position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 40% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

an increase in cue-rank agreement with the lexicon. Figure 4.72 illustrates this very result for preceding vowel duration (DUR_{V1}), where both inventory and lexicon models show monotonic increases in predicted accuracy for contrasts differing by 50 ms or more in DUR_{V1}. The partial dependence of accuracy on ΔDUR_{V1} in the weighted inventory model, on the other hand, is relatively flat, showing only a modest increase over the final 10% of the distribution (> 90 ms). This acoustic disagrement is further reflected in the distributions in Figure 4.72, which in addition to showing a lack of alignment between the weighted inventory and lexicon models (particularly between mid- and high-accuracy items), are bimodal in all three accuracy terciles in each model. This latter result poses a greater challenge for distributional scaling because it requires greater fidelity to the DUR_{V1} relations between smaller subsets of obstruent contrasts.

Figure 4.73 shows that such contrast relations can primarily be categorized into a single ΔDUR_{V1} continuum: *within-voicing/manner* $<$ *cross-manner* $<$ *cross-voicing*. The discontinuities between these sets, however, are much greater in the controlled syllable data than in real words. Further, there is less of a distinct relationship between vowel duration and perception over the lower ΔDUR_{V1} range in word recognition, meaning that listeners appear to be using vowel duration primarily to distinguish voicing from non-voicing contrasts, a result which makes the acoustic distinction between manner- and non-manner classes in the controlled syllable data less useful at predicting listener word recognition behavior in the weighted inventory model. Finally, there is a distributional component to this distinction that is masked by the general discussion of acoustic misalignment. In the word-final models in particular, there are a great many contrasts which are not present in the lexicon (28% of the inventory total). These contrasts are shown aligned at $y = 0$ in Figure 4.73, and are consistent with the general relationship between ΔDUR_{V1} and listener accuracy in the rest of the data, improving the role of preceding vowel duration in the inventory model at the expense of more accurate scaling to the lexicon where such contrasts do not occur.

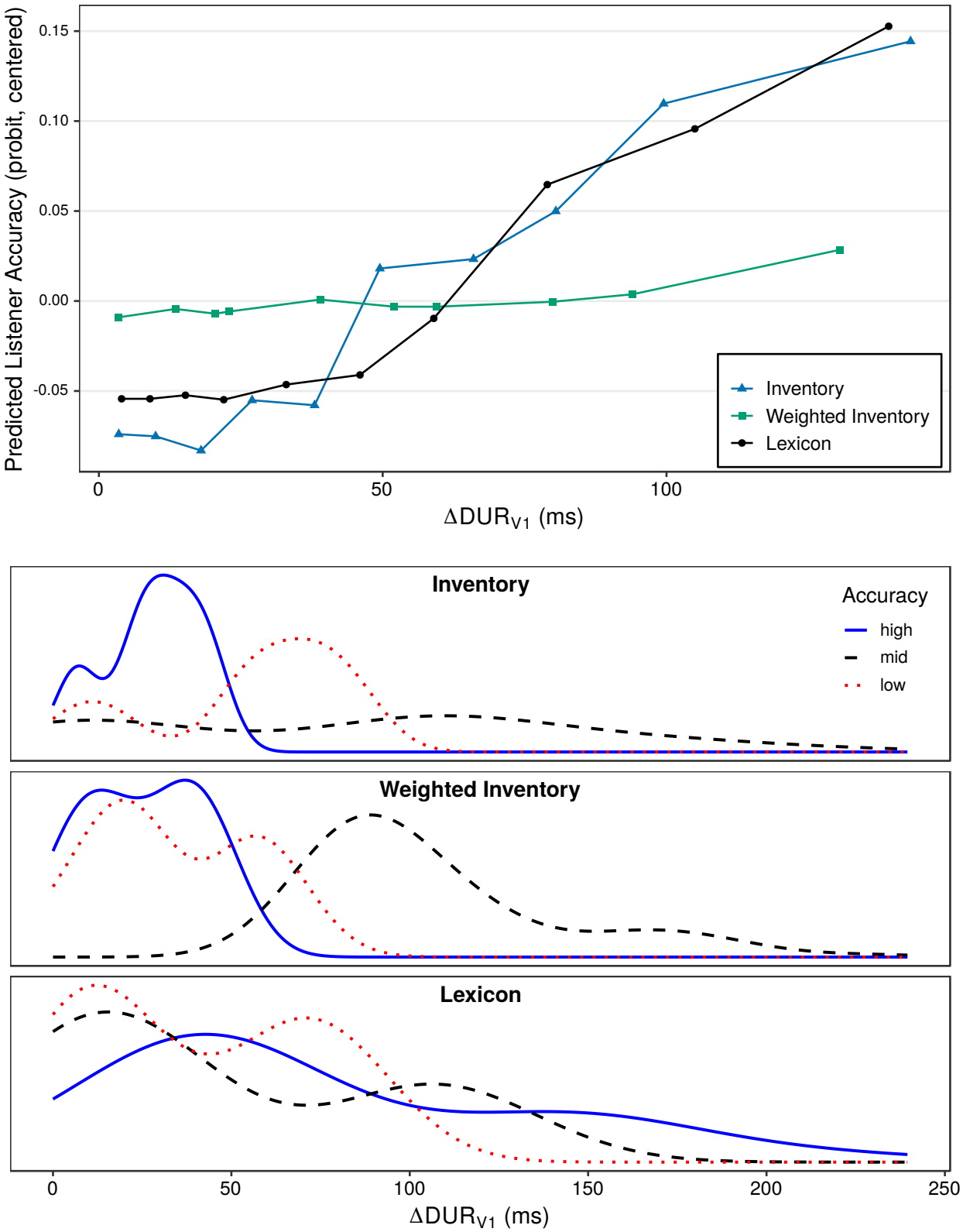

Figure 4.72: Partial dependence functions (top panel) and distributions (bottom panels) of DUR_{V1} in the inventory, weighted inventory, and lexicon models of listener recognition in word-final position.

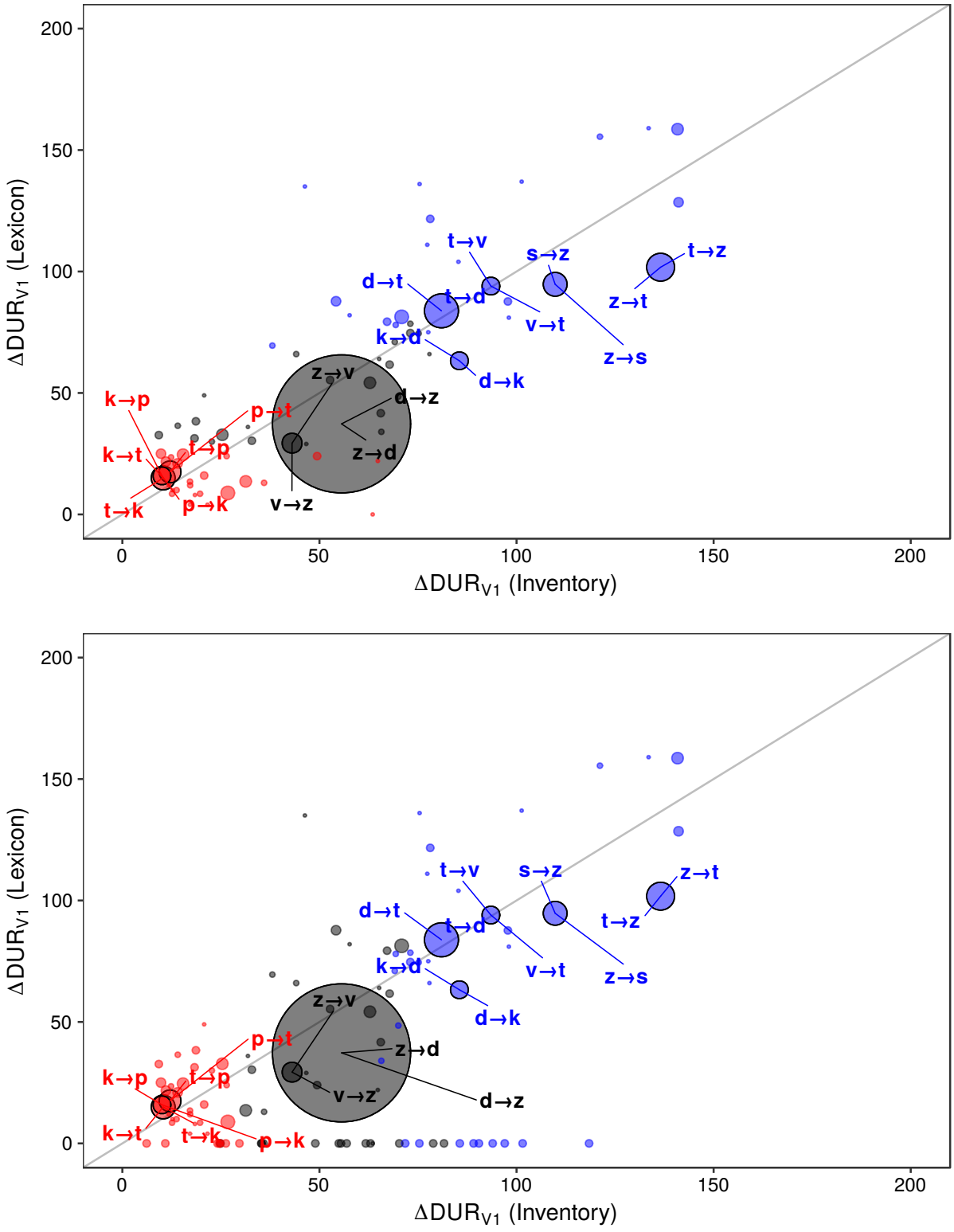

Figure 4.73: Relationship between ΔDUR_{V1} means by phonetic contrast in the inventory and lexicon models in VC position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 40% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

4.4.3.4 Composite disagreement: F2 at Vowel Offset ($F2_{VC}$)

Finally, we examine the behavior of vowel-offset F2 in the lexicon, inventory, and weighted inventory models. $F2_{VC}$ is ranked much higher in the lexicon than in the weighted inventory, and even lower in the balanced inventory model. The low ranking of $F2_{VC}$ is due in part to a negative relationship between increases in F2 separation and predicted accuracy between 0 and 50 ms, while both inventory models fail to capture the steep increase in predicted accuracy for $\Delta F2_{VC} > 200$ Hz. From the distributions in Figure 4.58 we see that the weighted inventory model does indeed approximate the lexical distributions much better than the inventory model, but the relationship between $\Delta F2_{VC}$ values in mid- and high-predicted-accuracy sets shows greater overlap in the weighted inventory than in the lexicon model.

Figure 4.75 reveals in greater detail the source of the discrepancy between the three models. Overall, $\Delta F2_{VC}$ values in obstruent contrasts in controlled syllable are poorly correlated with distinctions observed in real-word minimal pairs. For example, place distinctions are generally more distinct in the former than in the latter (shown in Figure 4.59 as points below the gray identity line), while voicing distinctions are more robust in the lexicon than in the inventory. Further, for the contrasts that are most frequent in the lexicon, the inventory acoustics (x-axis) show much greater overlap between contrasts on the basis of predicted accuracy in word recognition than the lexical acoustics (y-axis), where increases in $\Delta F2_{VC}$ generally correspond to increases in listener accuracy. Thus the scaling problem exemplified by vowel-offset F2 is not regarding whether or not listeners use the cue in recognition—all three models show a positive relation over some interval in the $\Delta F2_{VC}$ range—but which contrasts they use the cue for: voicing, and to a lesser extent place/sibilance, in the lexicon; and primarily place/sibilance in the inventory.

4.4.4 Discussion

The examination of cue integration in the prediction of listener recognition behavior was motivated primarily on the basis of improved validity as a measure of cue utility in communication over an ideal perceiver framework with no mechanism to account for relative differences in obstruent con-

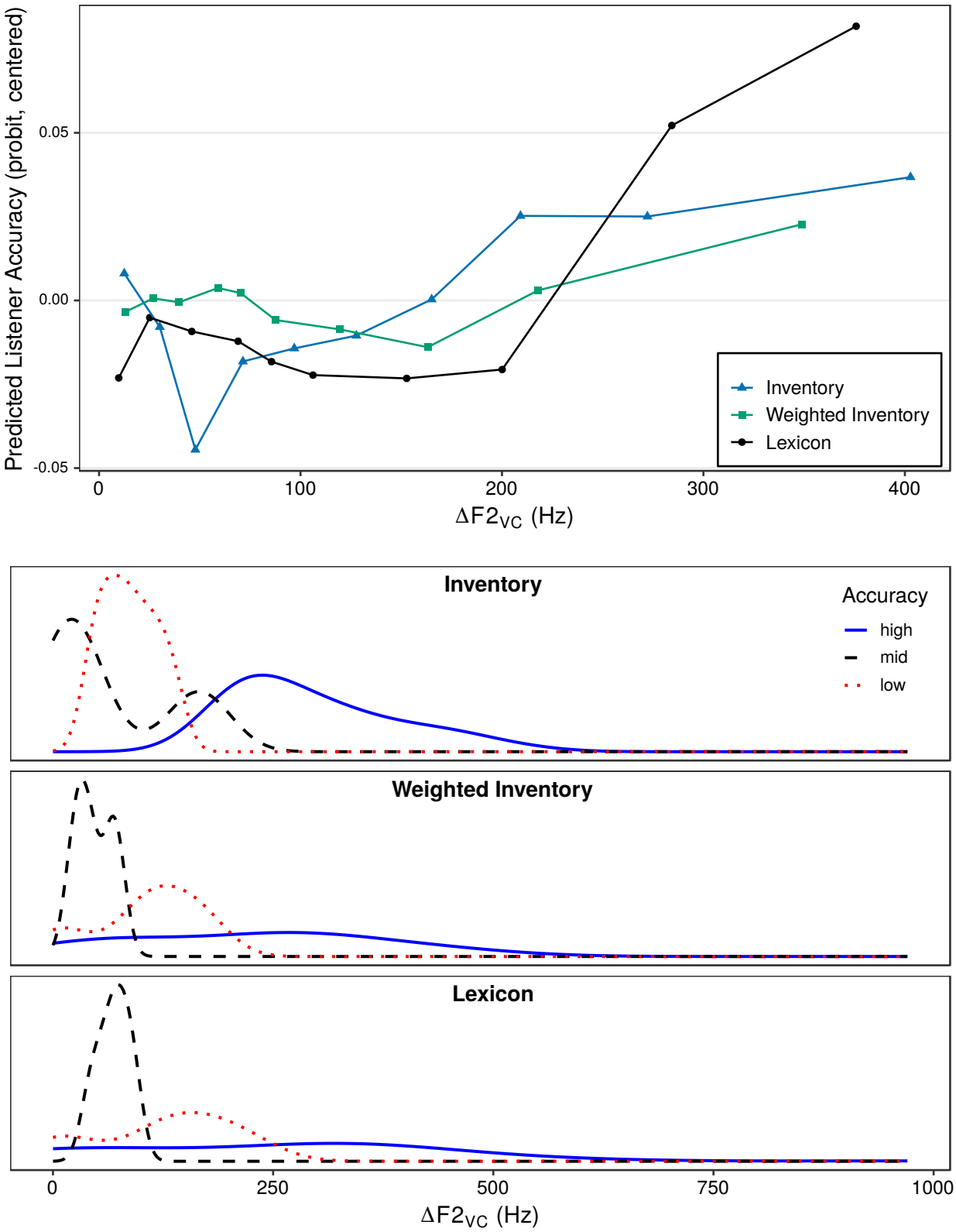

Figure 4.74: Partial dependence functions (top panel) and distributions (bottom panels) of $F2_{VC}$ in the inventory, weighted inventory, and lexicon models of listener recognition in word-final position.

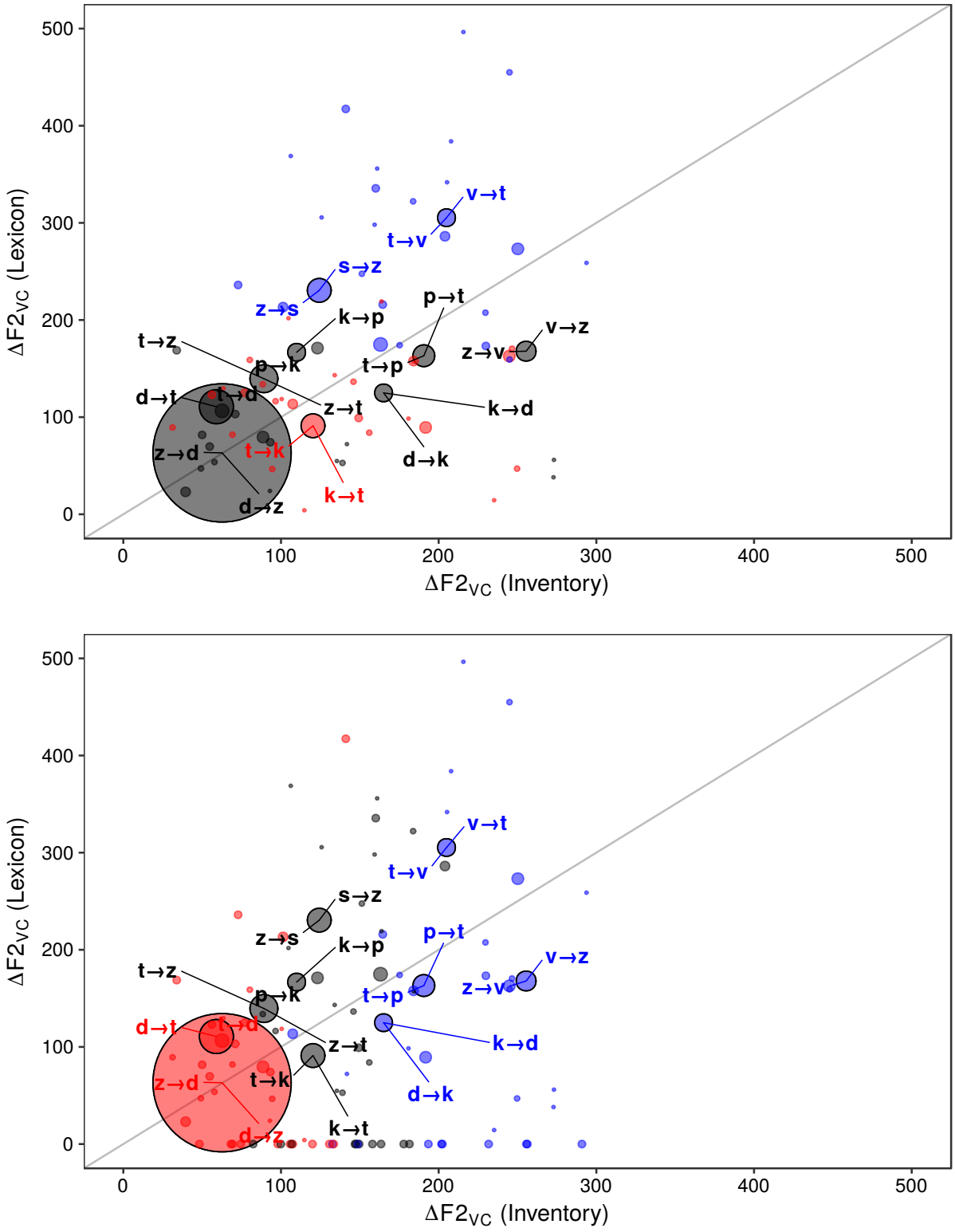

Figure 4.75: Relationship between $\Delta F2_{VC}$ means by phonetic contrast in the inventory and lexicon models in VC position. Points are scaled to match lexical frequency, and colored to reflect accuracy in the lexicon (upper panel) and in the inventory (lower panel). Contrasts comprising the top 40% of items are labeled. Contrasts absent from the lexicon are shown at $y = 0$.

trast perceptibility. However, this section also introduces an important methodological point in the scaling problem between inventory-based estimates of cue weighting and cue weighting in spoken word recognition. The inventory model operates on both controlled syllable acoustics and perception; and indeed, much of the work on speech perception in the phonetic literature relies on perception data from either balanced syllable stimuli or a small sample of real words meant to exemplify a particular subset of the phonetic system. Given that in addition to acoustic and distributional discrepancies, there were many discrepancies in cue weights attributable to perception differences between the two experiments. This raises the question as to whether listeners' cue parsing behavior is at all sensitive to the item and choice constraints of the task.

Among the highest ranked target cues in the listener models of word recognition were noise amplitude, spectral peak frequency, and consonant voicing percentage in word-initial and word-medial positions, and noise amplitude, noise duration, and F1/F3 at vowel offset in word-final position. Among contrast cues, which are the focus of model comparisons and the general question of cue scaling between the inventory and lexicon, F2, consonantal spectral tilt, and consonant voicing percentage are highly ranked in the CV and VCV models, as are F1 and relative F3 amplitude in CV and VCV positions, respectively, while in word-final position preceding vowel duration, noise duration, and spectral peak amplitude are the most informative.

As in the ideal perceiver models in Section 4.3, points of cue disagreement were classified into *distributional*, *acoustic*, and *composite* types, where in addition to these sources *perceptual* discrepancies may also be involved in the poor scaling between the two systems. Examples of distributional disagreement presented above include the relatively higher weighting of AMP_{F3} word-initially in the lexicon, and the the lower weighting of SHAPE and $FREQ_{PK}$ in VCV and VC positions, respectively, in the lexicon relative to the inventory. Acoustic disagreements included poor scaling of cue weights from following vowel duration, F2 at vowel offset, and preceding vowel duration in CV, VCV, and VC positions, respectively. Finally, among the cues in the *composite disagreement* class, where both inventory and weighted inventory models fail to accurately estimate the relative weight of a given cue in the lexicon, this set includes $DISP_C$, LF, and $F2_{VC}$ in

CV, VCV, and VC positions, respectively.

More important than the specific cue results is the demonstration in this section of the means by which cue integration can be modeled under different assumptions about the structure of the system, and where discrepancies between two or more systems arise, what components of the system—the acoustic data, the phonological distribution, the stimulus and task constraints of the perceptual baseline—are primarily responsible. This procedure informs both the direct study of scaling between the inventory and lexicon, and the more general question of modeling assumptions in the application and development of phonetic theory.

4.5 Experiment 2: Cross-splicing validation

The cue integration model developed in Section 4.4 provides estimates of the relative utility of different acoustic parameters in predicting listeners' ability to discriminate obstruent contrasts in English minimal pairs. However, we do not have evidence, based on model output of this sort, that these parameters are even properly *cues*,[6] because the consistent co-occurrence of certain acoustic parameter values with corresponding listener word recognition behavior does not entail that listeners, in their decisions, are responding directly to changes in these parameters. In other words, the acoustic results thus far have largely been *correlative* in nature; the goal of the present chapter is to provide *causal* evidence by validating that model predictions correspond to changes in listener accuracy on items that have been manipulated to reflect predicted increases and decreases in accuracy. In particular, Experiment 2 uses cross-splicing to replace the target phone and its adjacent vowel, henceforth referred to as the *target diphone*, with diphones from other items in the database whose acoustic parameters yield cross-spliced versions of two forms: (1) an *enhanced* item predicted to show an increase in listener accuracy, and (2) a *reduced* item predicted to show a decrease in listener accuracy (see Figure 4.76 for details). To the extent that listeners significantly

[6]We have till now been imprecise in the use of certain terminology regarding properties of the acoustic signal available to listeners in word recognition. The terms *cue* and *parameter* have been used interchangeably, but strictly speaking, *cues* refer to properties of the signal as perceived by listeners and used in a linguistic decision, whether that decision is to distinguish phones in nonword syllables, or as in the present study, to recognize words.

improve on enhanced items relative to reduced,[7] we have initial evidence of the causal validity of the model in Section 4.4 that forms the basis for both our assessment of cue weights in different paradigms of phonetic system structure (inventory vs. lexicon), and our simulation in Chapter 5 of effects of cue perturbation on the distributed system of contrasts in the lexicon.

4.5.1 Methods

4.5.1.1 Participants

Ten native speakers of American English were recruited from the University of Kansas student population for participation in the experiment. The experiment was originally designed for 40 participants, but due to the COVID-19 outbreak and subsequent campus closure, all further recruitment was halted. Participants received either $10 USD or course credit from the Department of Psychology as compensation for their time. All participants were administered a language background questionnaire prior to the experiment, and those reporting speech or hearing impairments, non-native speakers of English, and simultaneous bilinguals were excluded from the study.

4.5.1.2 Materials

Two hundred items were chosen from CV and VC minimal pairs in Experiment 1 based on model predictions of the relative increase/decrease in accuracy when acoustic cues from the target phone were replaced with those of another item in the database.[8] Items were evenly distributed between the two contrast positions (100 CV, 100 VC), and were chosen based on two criteria: (1) the overall predicted accuracy difference between enhanced and reduced versions of that item, (2) the cue which exhibits the greatest independent contribution to stimulated increases/decreases in accuracy on that item. For example, given the target item 'bit', word-initial [bɪ] diphones from other words in the database (e.g., 'bid', 'bitter', 'big') are identified, following which the effect of cross-splicing

[7] Enhanced and reduced accuracies are directly compared, as opposed to each version being compared with the baseline accuracy on the unaltered item, in order to account for potential artifacts of the splicing manipulation itself.

[8] VCV contrasts were excluded from Experiment 2 because of the relative sparsity of items sharing the target obstruent and both preceeding and following vowels, which resulted in too narrow a set of potential candidates for cross-splicing to provide adequate flexibility in targeting specific cues for tests of contrast enhancement/reduction.

these diphones onto the target is simulated by swapping their acoustic parameters with those of the target and retrieving the model's predictions on the new modified parameter set. What remains from the original target parameterization are the two word frequency measures (Absolute Target Frequency and Relative Target Frequency) and the competitor acoustic parameters against which the new target parameters are relativized (e.g., a new relative difference in VOT is computed by taking the absolute value of the difference between the new target VOT and the competitor VOT).

From this procedure we identify potential candidates for cross-splicing by setting a lower threshold of 10% on the predicted accuracy difference between enhanced and reduced items. Any items where the simulated manipulation effect is below 10% are removed from further consideration. Within the candidate set we then sought to identify, as noted in (2) above, 100 items (50 CV, 50 VC) where the primary cue responsible for the change in accuracy was the same for both enhanced and reduced versions. This subset of the data serves as a validation of the causal role different cues play in word recognition. Independent cue contributions to enhancing/reducing accuracy on a given item were measured by swapping a single cue at a time between the candidate item for splicing and the target item, while holding all other cues constant. Model predictions were then applied to each such parameter set, and cues showing the greatest change in predicted accuracy in the direction of the manipulation—positive for enhanced, negative for reduced—were identified as the primary cues responsible for a given enhancement/reduction.[9]

In many cases the cue pairing as described above resulted in cues that were similar though not identical in enhanced/reduced items. For example, the enhancing cue might be the spectral tilt of the consonant, while the reducing cue is the spectral tilt of the vowel. These two cues measure approximately the same thing, and are involved in derived cue of Lahiri et al. (1984) tracking the dynamics of spectral tilt in CV transitions. And so for the purpose of identifying cue-matched items in the present experiment, these cues were treated as equivalent. Two such composite cues were created in this manner: Spectral Tilt {tilt of the consonant, tilt at vowel onset}, and Spectral

[9]In some cases enhancement was recorded as a small negative change in accuracy relative to that in reduced, and vice versa for reductions as small positive changes. These cases arose primarily when the baseline predicted accuracy of the target was near ceiling or floor, respectively.

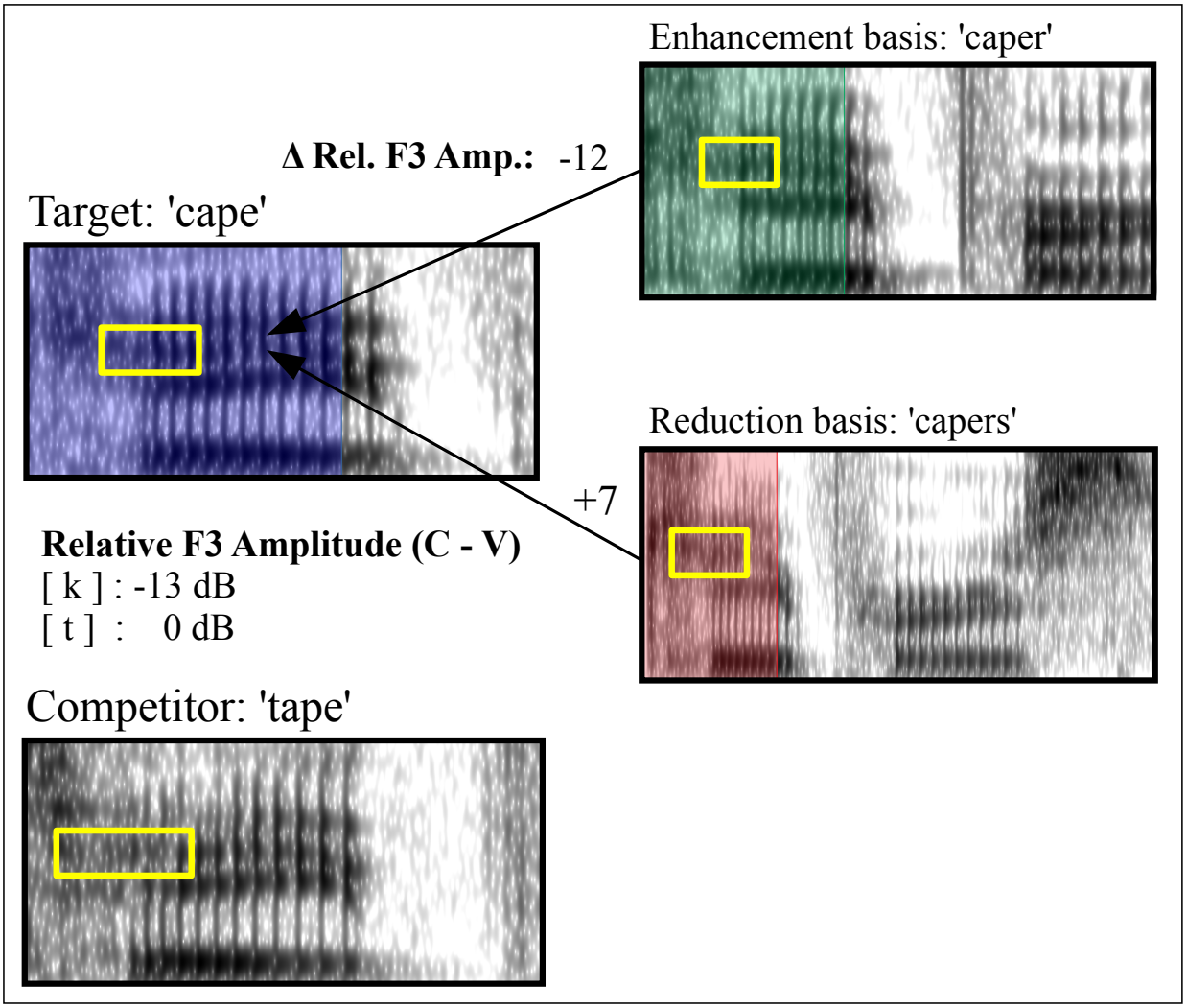

Figure 4.76: Schematic of cross-splicing methodology. On the left are spectrograms of target and competitor items from a minimal pair in Experiment 1. The target word 'cape' is the item that appears as a target in Experiment 2, while the competitor only appears on the screen, though its acoustic features are used to generate model predictions. The two spectrograms on the right represent words that are the basis for diphones that enhance ('caper') or reduce ('capers') the identifiability of the target word relative to the competitor. Here, the primary cue responsible for the model-predicted increase/decrease in accuracy on the target word 'cape' is Relative F3 Amplitude, which is made more distinct from the competitor when the target region (blue box), whose Relative F3 Amplitude is -13 dB, is replaced by the region from 'caper' (green box), which is a further 12 dB lower in relative amplitude, at -25 dB (as compared with 0 dB for the competitor). Conversely, replacing the target diphone with the initial diphone of 'capers' (red box) results in a reduced contrast with 'tape', because the Relative F3 Amplitude of 'capers' is 7 dB higher than the target, at -6 dB, resulting in an amplitude transition in the F3 region of [k] that is much more typical of [t] than either the original or enhanced [k] in 'cape'. On the basis of the difference in Relative F3 Amplitudes as a result of the cross-splicing manipulation, listeners should therefore be more accurate at identifying the enhanced target than the reduced.

Dispersion {dispersion of the consonant noise spectrum, dispersion at vowel onset}.

Maximal diversity in the representation of different cues in the 100-item set was sought, however the constraint that overall predicted accuracy differences must be above the 10% threshold (otherwise they are unlikely to be detected) limited the range and balance in cue representation that was possible. The following cue pairs were included in Experiment 3: Noise Amplitude (39 items), Spectral Peak Amplitude (25), Vowel Duration (9), F3 (8) Spectral Dispersion (7), F1 (5), Spectral Tilt (4), F2 (2), f0 (1).

Finally, an additional 100 items (50 CV, 50 VC) were identified from those candidates showing the greatest predicted effects of the manipulation. The resulting 200-item set had a mean predicted accuracy difference between enhanced and reduced of 16.9% (17.1% in CV, 16.7% in VC). All items were cross-spliced in Praat, and amplitude normalized and embedded in noise following the same procedure adopted for Experiment 1.

4.5.1.3 Procedure

The procedure in Experiment 2 followed that in Experiment 1. On each trial, a noise-masked word was presented binaurally over headphones, after which two words appeared on the screen, one being the target word and the other its minimal pair competitor. Each word was associated with left and right buttons on a button box, corresponding to the placement of the words on the screen, and participants were instructed to push the button corresponding to the word they heard (screen position / button order was counterbalanced across participants). No time pressure was applied to this choice, and after selecting an option the next trial began (ITI = 1 sec). Listeners were instructed to guess in cases where they were unsure or did not perceive a word in the stimulus.

Ten practice trials were given before the main experiment, all using minimal pairs distinct from the experimental items and not exhibiting obstruent consonant contrasts, though they were otherwise similar in exposing listeners to words of mono-, di-, and tri-syllabic lengths. The 200 experimental trials were divided into four blocks of 50 items each. Between blocks, participants were given up to a 1 minute break, though they were able to start the next block whenever they

	+2 dB		−2 dB		
	Enhanced	Reduced	Enhanced	Reduced	Total
CV	250	251	240	259	1000
VC	258	241	252	249	1000
All	508	492	492	508	2000

Table 4.4: Number of responses in Experiment 2 by Position (CV, VC), Manipulation (enhanced, reduced), and SNR (+2 dB, −2 dB).

were ready. In total, the experiment took between 15 and 20 minutes.

4.5.2 Validating overall model performance

Given that participant recruitment was designed for 40 but prematurely halted at 10, and that participants were randomly assigned items with enhanced or reduced manipulations at +2 or −2 dB, we first review the number of responses recorded in each condition in this truncated data set. Table 4.4 shows response counts by Position (CV, VC), Manipulation (enhanced, reduced), and SNR (+2 dB, −2 dB). Overall, the number of items in each combination of conditions was relatively even, at between 240 and 259 items. The only major concern raised by the distribution in Table 4.4 is the slightly greater number of reduced items (relative to enhanced) presented at −2 dB than at +2 dB. This imbalance is potentially problematic as it means that in an overall analysis of the effect of the stimulus manipulation on listener responses, accuracy on reduced stimuli will be driven more by responses at the lower SNR, and conversely for enhanced stimuli at the higher SNR, resulting in a confound between Manipulation and Noise Level given that the SNR biases above are in the same direction as the predicted difference between enhanced and reduced stimuli. For this reason, all analyses of manipulation effects will focus on the interaction with noise level.

Figure 4.77 shows listener accuracies alongside model-predicted accuracies for enhanced and reduced stimuli by Position and Noise Level. Listeners show a marked decline in accuracy from enhanced to reduced stimuli across CV and VC positions and at both +2 and −2 dB SNR. In VC position, the accuracy difference between enhanced and reduced stimuli is relatively constant

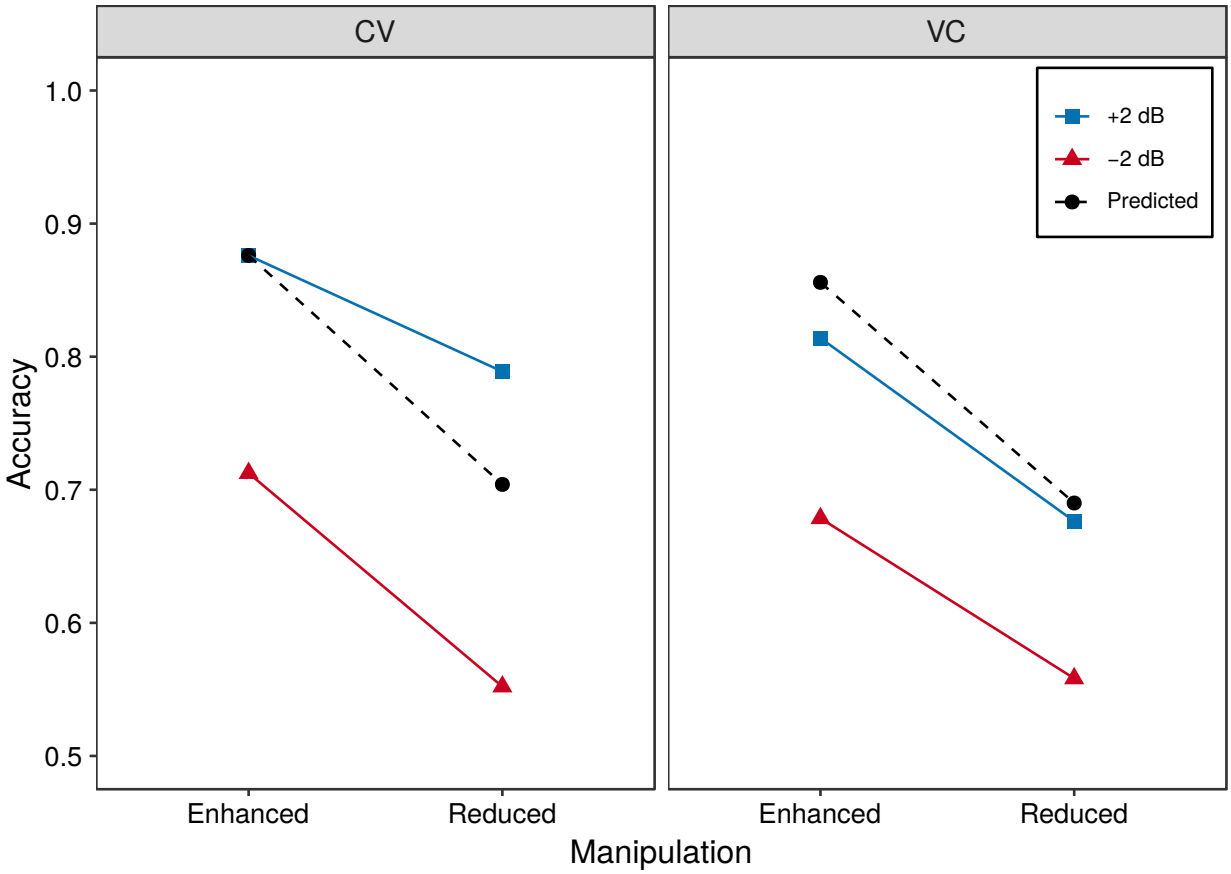

Figure 4.77: Listener (red/blue lines) and model-predicted (black dashed line) accuracies on enhanced and reduced stimuli in Experiment 2 as a function of Position (CV, VC) and SNR (+2 dB, −2 dB).

across SNRs—13.7% at +2 dB, and 12.0% at −2 dB—as well as in comparison to the predicted 16.6% difference from the model. In CV position, we see greater variability in manipulation effects by SNR. At +2 dB, the gain in accuracy on the enhanced stimuli relative to reduced is only 8.7%, half that observed at −2 dB (16.0%), and half the model-predicted difference (17.2%).

These effects were then modeled in a mixed-effects logistic regression predicting Accuracy (correct = 1, incorrect = 0) from fixed effects of Position (CV [ref], VC), Manipulation (enhanced [ref], reduced), and Noise Level (−2 dB [ref], +2 dB) and a Listener random intercept. In this model, the three-way interaction between Position, Manipulation, and Noise Level was not significant ($\chi^2(1) = 0.42,\ p > 0.1$), nor were the two-way interactions: Position × Noise Level ($\chi^2(2) = 5.60,\ p = 0.061$), Position × Manipulation ($\chi^2(2) = 0.519,\ p = 0.771$), and Noise Level × Manipulation ($\chi^2(2) = 0.698,\ p = 0.706$). The two consistent effects were Noise

Level ($0.514 < \beta < 1.067$, $2.715 < z < 5.522$, $p < 0.007$), and Manipulation ($0.519 < \beta < 0.697$, $2.644 < z < 3.633$, $p < 0.008$).

Thus, from these results we have initial confirmation that the cross-splicing manipulation predicted by the cue-weighting model to enhance and reduce the discriminability of obstruent contrasts does in fact yield significant comparable changes in listener accuracy. This effect was present in both CV and VC positions, and at both SNRs, and closely matched model predictions, particularly at the higher SNR, where there is a closer match between the acoustic information available to the listener, and the 'clean' (non-noisy) acoustic data that served as input to the model. Having shown that the statistical model in Section 4.4 is capable of generating predictions that can be causally verified, we turn next to the individual cue weights derived from the model, and the extent to which they can be shown to similarly influence listener behavior.

4.5.3 Validating the role of individual cues

Table 4.4 shows the number of responses recorded in Experiment 2 by Cue, Position, SNR, and Manipulation. This distribution is even more critical to the analysis of listener accuracies than in the previous section, as once responses are broken down by cue, there is considerable sparsity in the data. This sparsity reduces the reliability of any accuracy estimates by both increasing the variance and introducing a greater potential for bias, as the truncated participant recruitment means not all items are evenly represented at a given manipulation and noise level. For this reason, only a descriptive analysis of patterns in listener accuracy is provided below, as the data remains too sparse for reliable statistical inference.

From among the set in Table 4.5, the four most represented cues at each position were analyzed for listener accuracy patterns: Noise Amplitude and Spectral Peak Amplitude in both CV and VC positions, F1 and Spectral Tilt in CV, and Vowel Duration and F3 in VC. Figures 4.78 and 4.79 show listener accuracies on CV and VC stimuli, respectively, where these cues are the primary determinant of predicted increases and decreases in accuracy after cross-splicing.

Robust effects of the stimulus manipulation (i.e., enhanced $>$ reduced) were found for the

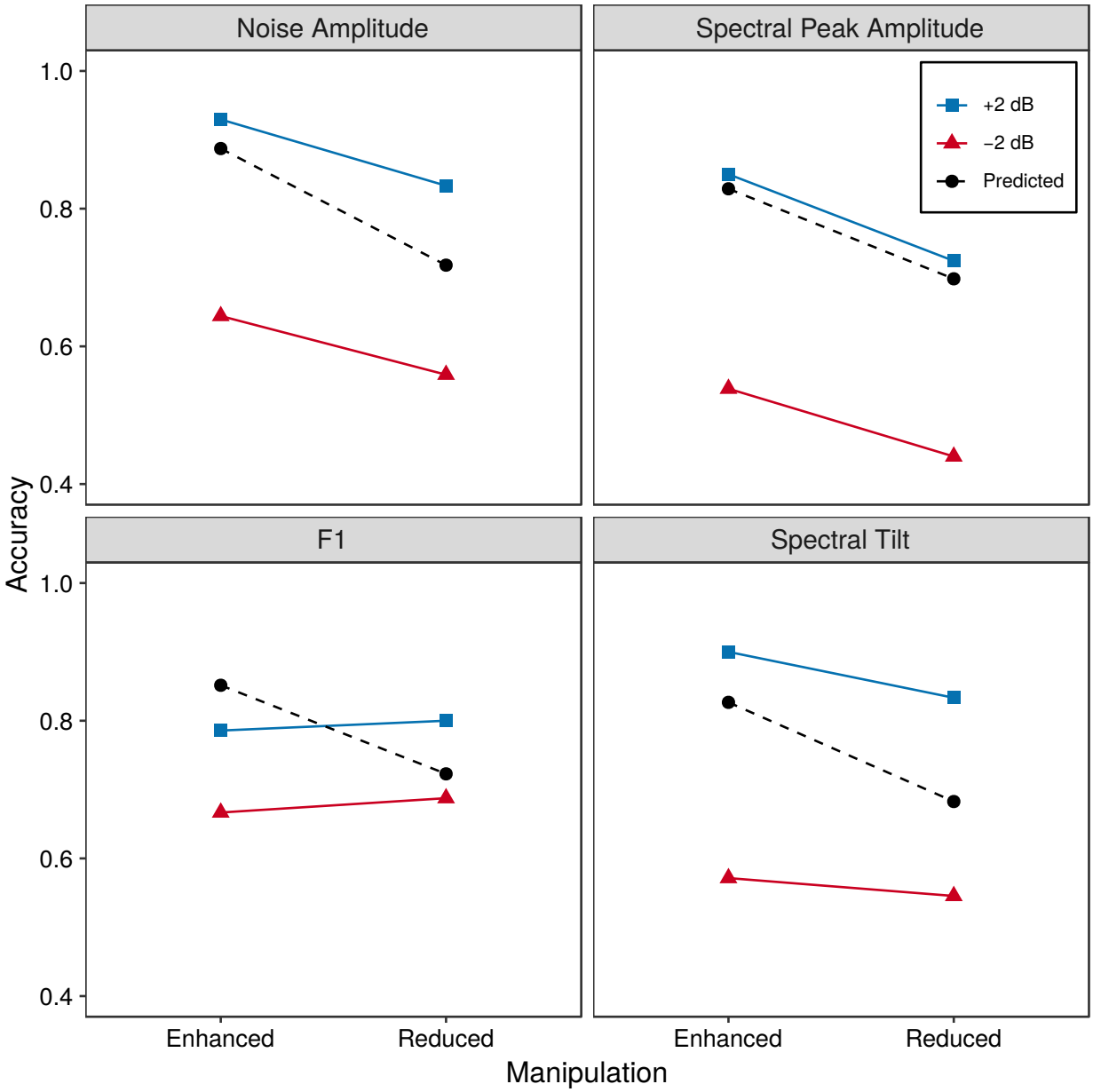

Figure 4.78: Listener (red/blue lines) and model-predicted (black dashed line) accuracies on enhanced and reduced CV stimuli in Experiment 2 as a function of the Primary Cue responsible for the predicted change in accuracy (Noise Amplitude, Spectral Peak Amplitude, F1, Spectral Tilt) and the SNR of stimulus presentation (+2 dB, −2 dB).

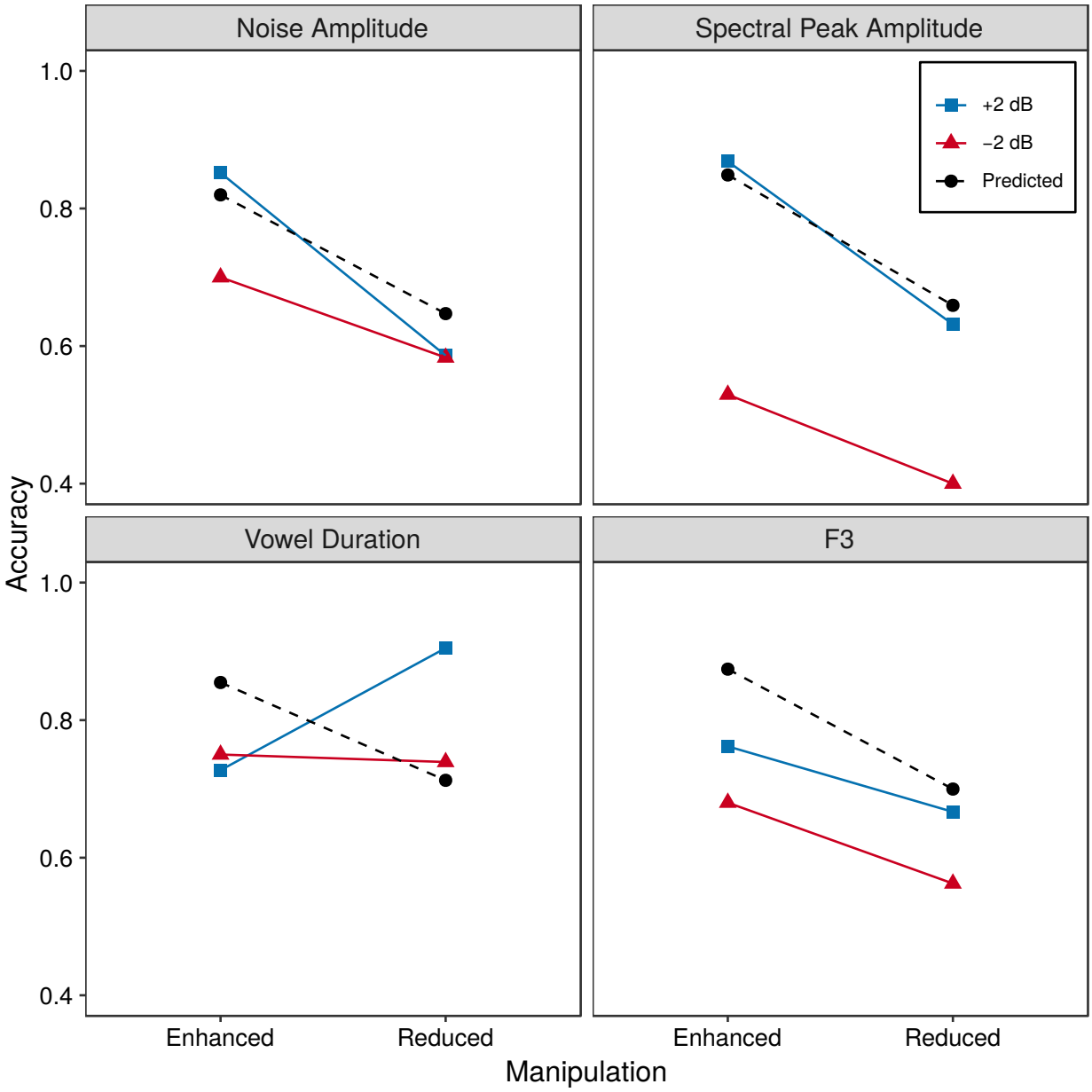

Figure 4.79: Listener (red/blue lines) and model-predicted (black dashed line) accuracies on enhanced and reduced VC stimuli in Experiment 2 as a function of the Primary Cue responsible for the predicted change in accuracy (Noise Amplitude, Spectral Peak Amplitude, Vowel Duration, F3) and the SNR of stimulus presentation (+2 dB, −2 dB).

	CV				VC					
	+2 dB		−2 dB		+2 dB		−2 dB		Totals	
Cue	Enh.	Red.	Enh.	Red.	Enh.	Red.	Enh.	Red.	Rec.	Plan.
Noise Amplitude	71	72	59	68	28	32	33	27	390	1560
Spectral Peak Amp.	20	29	26	25	38	38	34	40	250	1000
Vowel Duration	0	0	0	0	22	21	24	23	90	360
F3	0	0	0	0	21	18	25	16	80	320
Spectral Dispersion	4	4	1	1	14	12	17	17	70	280
F1	14	5	15	16	0	0	0	0	50	200
Spectral Tilt	10	12	7	11	0	0	0	0	40	160
F2	6	7	3	4	0	0	0	0	20	80
f0	0	4	0	6	0	0	0	0	10	40

Table 4.5: Number of responses in Experiment 2 by Cue targeted, Position (CV, VC), Manipulation (enhanced, reduced), and SNR (+2 dB, −2 dB). Total counts for each cue are shown for the present recorded data and the planned data under the 40-participant recording design.

two amplitude parameters—Noise Amplitude and Spectral Peak Amplitude—in both positions, with the Spectral Peak Amplitude manipulation at +2 dB consistently the closest match to model predictions. Effects for the other four parameters—F1 and Spectral Tilt in CV, Vowel Duration and F3 in VC—were more modest, with F1 and Vowel Duration showing either no difference between enhanced and reduced stimuli, or differences inconsistent with expectations. The Spectral Tilt manipulation in CV position showed a slight decline in accuracy from enhanced to reduced at +2 dB, but little difference at −2 dB. Finally, the F3 manipulation showed consistent effects at both +2 and −2 dB that were in line with expectations, providing some evidence that listeners use the F3 transition at the offset of the vowel to discriminate word-final obstruent contrasts. However, the latter four cues are all under-powered in the present study, so caution must be applied in interpreting these patterns before the full set of data can be collected at a future date.

That the most robust effects occurred on the two amplitudinal cues deserves some comment. This result is partly a reflection of their greater representation in the stimuli, and likely also due to the fact that we are measuring word recognition in noise, meaning changes in amplitude not only effect the amplitudinal cue dimension between sounds (e.g., the fact that sibilants are louder than nonsibilants, and therefore noise amplitude can be a cue to sibilance) but the likelihood a listener

can even perceive certain cues against the background noise. And these two explanations are not in fact independent, as the greater number of amplitude-based cues targeted for splicing derives from the fact that such cues were more likely than others to be the main drivers of change in predicted accuracy when cross-splicing effects were simulated in the cue integration model.

4.5.4 Discussion

The results from the cross-splicing manipulation in Experiment 2 provide preliminary support for the validity of the cue integration model as one that can generate causal predictions about listener behavior, and thereby allow inferences about the cues listeners attend to in real word recognition on a diverse set of items and minimal pair contrasts. However, further data is required to provide more robust evidence of these effects, particularly in the investigation of individual cues. Finally, while the targeting of primary cues responsible for relative increases and decreases in accuracy following cross-splicing provides initial evidence of the relevance of those cues in listener perception, we cannot guarantee that listener responses are in fact driven by a given cue and not other cues that are numerically secondary in our model predictions, but may have greater perceptual weight in reality.

Nevertheless, we believe cross-splicing is a promising method for deriving causal evidence for different cues in the face of the complex multivariate cue structure required for adequate modeling of word recognition. Synthetic manipulation, though able in most cases to precisely target a single acoustic parameter while holding all else constant, has the potential to introduce artifacts in the form of cue relationships that are not observed in real speech data and may even be physiologically implausible. A potential middle ground to be explored in future research is to use a large set of speech data to map the space of possible synthetic speech outputs whose cue structure is consistent with real speech data. However, given the size of the parameter space, such a procedure would require a much greater quantity of data than is presently available in this thesis.

4.6 General discussion

The goal at the outset of this chapter was to determine the extent to which relative cue weights depend on the assumptions of the system, either explicit or implicit, both in terms of the fundamental units each assumes (independent phones/contrasts versus real-word distinctions), and the acoustic and phonological distributions presumed to be relevant in both perceptual experimentation and model-building. In both *ideal perceiver* and *listener* frameworks, statistical models of cue integration under balanced inventory, weighted inventory, and lexical contrast assumptions were built and compared in terms of the relative utility of each acoustic cue in each model. Finally, among the discrepancies in cue ranks that were found, several exemplars of *distributional*, *acoustic*, and *composite* sources of disagreement were demonstrated. Thus, the analysis in Sections 4.3 and 4.4 provides both an assessment of the relative impact of the inventory assumption on the predictability of cue weights in word recognition, and a roadmap for how discrepancies between two systems can be decomposed into more fundamental structural differences that offer insight into the theory that informed each system configuration.

However, the cue-integration results in Sections 4.3 and 4.4 remain correlational, and so Experiment 2 (Section 4.5) was run to test the causal validity of the model by measuring listener sensitivity to a cross-splicing-based manipulation of stimuli based on aggregate and cue-specific model predictions. Overall, the results of Experiment 2 provide initial evidence that the models in Section 4.4 can predict both overall shifts in listener accuracy, and shifts in accuracy due to a particular cue, though the latter results are more sparsely represented in the data due to the small sample size caused by the premature halting of the study during the COVID-19 pandemic.

Overall, differences in cue integration between the inventory and lexical systems provide the most compelling evidence that the canonical assumption of a balanced, independent inventory of phones as the basis of the system has many unforeseen consequences: chief among them the tendency to under/over-estimate the predictive power of different acoustic cues in word recognition. Some of these gaps can be overcome by adjusting the relative role of different contrasts in the system, but others require a reevaluation of the acoustic and perceptual data that informs the model.

Chapter 5

System structure

Outline

This chapter investigates the structure of the phonetic system embedded in the distributed complex of minimal-pair contrasts in the lexicon. Two general architectures of the lexicon are considered in this analysis: a *set* architecture and a *network* architecture. Structural characteristics of each system architecture are examined in Sections 5.2–5.4 in three ways. First, a description of the phonological structure of each system is presented, which serves as a baseline for subsequent work on gradient acoustic properties of lexical contrasts. Next, we examine the response of the system of contrasts in the lexicon to perturbation by noise, using the results of Experiment 1 to predict the overall impact of signal masking from background noise. Finally, we use the cue-integration models in Section 4.4 to predict the impact of perturbing individual cues on the global maintenance of contrast in the lexicon.

5.1 Introduction

5.2 Phonological structure
Measures of the number of minimal pairs and functional load on different obstruent categories and contrasts in the lexicon are presented and used as a baseline for the study of gradient contrast reduction under acoustic perturbation.

5.3 Noise perturbation
The impact of signal perturbation by background noise (multitalker babble at $+2$ and -2 dB SNR) is presented based on the results of Experiment 1. Here we focus on the differential auditory robustness of phonetic contrasts to noise masking, and the role of the distribution of such contrasts in determining both points of stability and vulnerabilities in the lexicon.

5.4 Cue perturbation
Model predictions from Section 4.4 are used to study the global role of individual acoustic cues by perturbing each cue (simulating cue ambiguity numerically) and measuring the model-predicted discrimination accuracy on the perturbed cue set. This analysis, though similar to the cue-weighting analysis in Section 4.4, rather than measuring the degree of covariation in listener behavior to changes in a given cue, focuses on the response of the lexicon to a loss in information from that cue. Further, these simulations are run on a comprehensive set of obstruent contrasts in the lexicon, beyond the subset that was presented to listeners in Experiment 1, and thus provide a more accurate picture of the global role of each cue in the lexicon.

5.5 Discussion

5.1 Introduction

Chapters 2 and 3 provided detailed descriptions of the acoustics and perception of obstruent contrasts in the lexicon, while Chapter 4 assessed both the degree to which the latter may be predicted from the former, and whether the cue weights derived from the prediction of lexical contrast perception conform with those based on a balanced inventory of contrasts. In making this comparison we were able to locate both points of agreement and disagreement between the two approaches, and where disagreements arose we identified the primary source of the discrepancy in differences in the distribution of obstruent contrasts in the lexicon and inventory, differences in the acoustics of real words versus controlled syllables, or a combination of the two. This analysis was primarily *listener-centric* in studying both the information available to listeners, and how they appear to be integrating that information in perception based on their word recognition behavior.

The present chapter may be described as *lexicon-centric* in that it examines the impact on the system of contrasts in the lexicon of the relative discriminability of items as a function of the uncertainty introduced by background noise and cue loss. More broadly, we use the simulation of contrast weakening in the lexicon under acoustic perturbation to study how an obstruent system that is embedded in higher-order distinctions is structured. This analysis provides both a new measure of *functional load* that is gradient—i.e., it reflects the relative likelihood that the information contained in obstruent contrasts in the lexicon is preserved in the presence of background noise—and a new measure of the functional load of specific acoustic cues that reflects the degree to which a given cue contributes to the maintenance of lexical form distinctions. Both analyses are fundamental to an approach to acoustic phonetics that is fundamentally linked to the higher-order systems that are encoded in the signal.

In the sections below we begin with a review of the phonological structure of two primary lexical architectures that are employed in the simulations: one that is *set*-based (i.e., the general *lexicon-as-list* approach that is the default assumption in linguistics and psychology) and one that is *network*-based (Vitevitch, 2008). Sections 5.3 and 5.4 then present simulations of the impact of noise and cue perturbation, respectively, on each architecture, while Section 5.5 concludes.

5.2 Phonological structure

Before examining the effect of noise and cue perturbation on the maintenance of contrast in the lexicon, it is first necessary to establish the formal structure of the lexicon that forms the baseline for each simulation. Below we review two architectures of the lexicon that feature most widely in the literature: a *set*-based architecture that treats the lexicon as an unordered list of phonological forms, and a *network* architecture as in Vitevitch (2008) that treats the phonological lexicon as a network of minimal-pair oppositions. To be clear, minimal pairs are studied in both frameworks, but whereas in the former, minimal pairs are associations that are derived by the linguist but otherwise do not contribute structurally to the system, in the latter the minimal-pair relation is fundamental to the structure of the system. Formally, this distinction amounts to the mathematical difference between a *graph* and a set of subsets, where in the latter each minimal-pair subset is independent of the others, while in the former all such pairs that exhibit overlap in their membership (e.g., minimal pairs $\langle cat, bat \rangle$ and $\langle bat, back \rangle$) are dependent.

The most critical consequence of this difference for the study of system structure under perturbation is the degree to which reductions in contrast discriminability under noise masking or cue loss impact the system as a whole. In the set architecture, perturbations are largely locally contained, impacting only those contrasts that they directly apply to. However, the network architecture allows for 'ripple' effects where the increased merger potential between words *A* and *B* also serves to bring all of *A*'s minimal-pair neighbors into closer phonological proximity to *B*'s neighbors, which means that in such a system it not only matters which words a given phonetic contrast or cue serves to distinguish, but how those words relate phonologically to other words in the lexicon. To be clear, this analytical distinction is not meant to evaluate which architecture is *correct*, but rather to present the simulation results in multiple forms so that they are informative for researchers working in different frameworks. However, we hope that the simulations below will provide for more explicit predictions about cue adaptation in language change and language learning that will aid in the evaluation of these formal architectures in future research.

5.2.1 Set architecture

The simplest and most fundamental organization of items in the lexicon is in the form of a *set*. That is, at a phonological level,[1] the lexicon can be represented as $L = \{w_1, w_2, ..., w_n\}$, where w_i is the phonological form of word i over a vocabulary of n words. This representation does not explicitly commit to the abstract form of w_i, be it a string of symbols, a matrix of feature values, a gestural score of articulator activation sequences, or some other encoding framework. For simplicity, in the analysis below we will use strings of phones to describe such items and the minimal pair relations between them. Further, while many homophones exist in English, we adopt the simplifying assumption that all items in L are distinct in their phonological form; i.e., $w_i \neq w_j \; \forall \; 1 \leq i, j \leq n$. Below we review two key measurements that have been used to describe the distribution of phonological information in the lexicon: the count of minimal pairs by obstruent category and contrast (Labov, 1994; Wedel et al., 2013) and the functional load of each category/contrast (Martinet, 1952; Hockett, 1967; Surendran & Niyogi, 2003; Wedel et al., 2013).

5.2.1.1 Minimal pairs

The analysis of minimal pair counts in English, as well as all subsequent analyes in this chapter, is performed on the Lex95 database, which as described in Section 4.2 is composed of all words in the MALD database that comprise 95% of tokens in each of several corpora, with the database further processed for direct comparison with the network architecture by excluding all items that either have no minimal pairs, or which are not contained within the largest component of the lexicon (i.e., the largest set of items which are all reachable via phoneme substitution, deletion, or addition from another item in the set). This set contains 3,406 words participating in 11,972 minimal-pair contrasts, of which 2,501 are contrasts between obstruent consonants. In total, 1,649 unique words comprise the set of obstruent-contrastive minimal pairs used in this model lexicon.

[1] Note that here we assume, along with the majority of the field, that a much more complex array of information is stored in the lexicon, including morphological, syntactic, and semantic information, among other systems of linguistic and paralinguistic knowledge, and thus our discussion of the *set* of lexical items and any relations therein is limited to phonological structure independent of the other systems, though we acknowledge that this is a simplification, and that these systems are all interdependent.

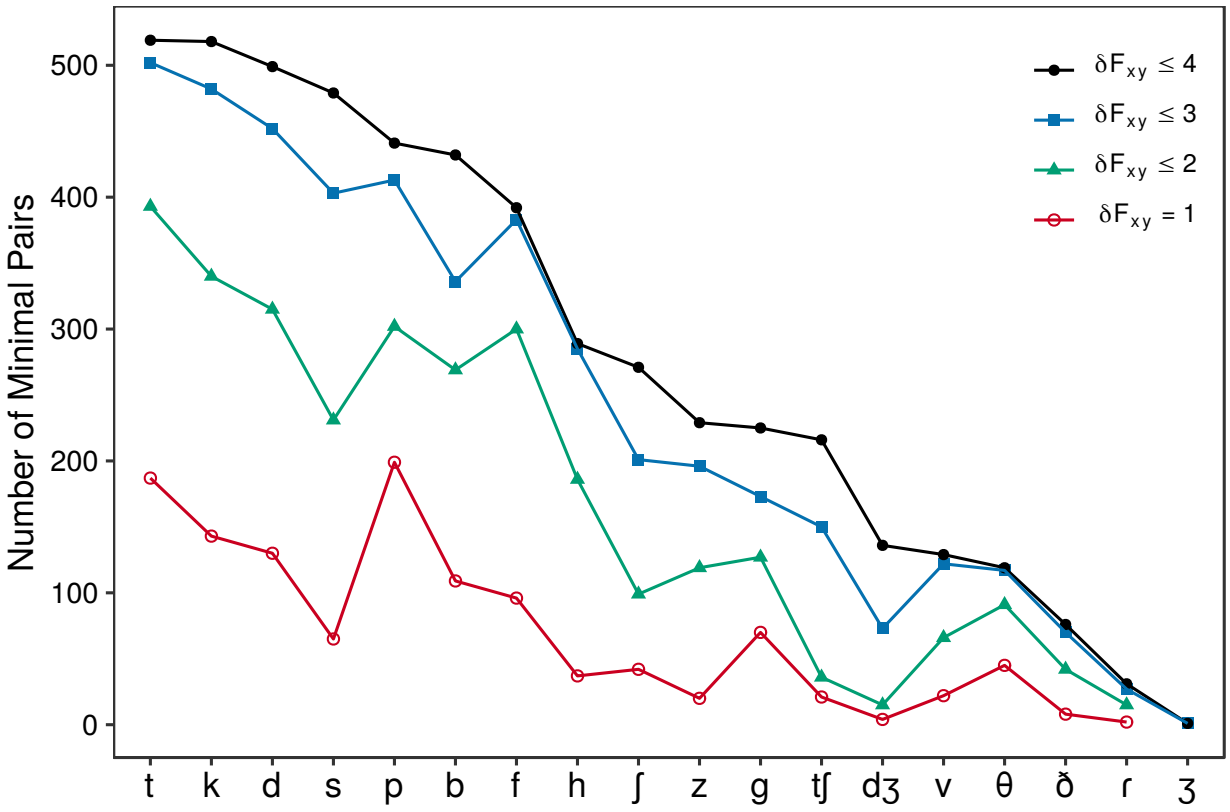

Figure 5.1: Number of obstruent-contrastive minimal pairs in the lexicon by constituent phone.

Figure 5.1 shows the number of obstruent-contrastive minimal pairs in the lexicon that each phone participates in as a function of the number of features distinguishing the contrast, δF_{xy} (x and y being the contrasting phones). In Figure 5.1, taking all contrasts together ($\delta F_{xy} \leq 4$) we see a fairly linear decline between the phone that occurs most frequently in contrasts with other obstruents, [t], and the least frequent constituent of such contrasts, [ʒ]. There are two significant jumps in the distribution, however, that allow us to divide the set into three groups of high-, mid-, and low-frequency phones. The high-frequency phones, [t, k, d, s, p, b, f], include all of the plosives except [g], as well as the voiceless fricatives [s] and [f]. By comparison, the mid-frequency set, [h, ʃ, z, g, ʧ] is dominated more by fricatives and affricates, including both voiceless postalveolars, [ʃ, ʧ], the glottal fricative [h], the voiced alveolar sibilant [z], and the voiced velar plosive [g]. Finally, the low-frequency set, [v, ʤ, θ, ð, ɾ, ʒ], includes both voiced postalveolars, [ʒ, ʤ], both dental fricatives, [θ, ð], the voiced labiodental fricative [v], and the alveolar flap [ɾ].

The minimal pair distributions among more constrained featural contrasts, such as those differing by 1 or 2 features, are of particular interest because these are the contrasts that are most likely to be perceptually confusable, and therefore play a role in the weakening of lexical contrast when the acoustic signal is perturbed. Contrasts within this set are also the most likely to undergo mergers as the language changes over time, a fact which will be revisited in the next section when the impact of different mergers is directly studied in the measurement of *functional load*. Regarding single-feature contrasts, for instance, the most notable changes in the relative ranking of phones in Figure 5.1 are the emergence of [p] as the most frequent contrastive phone, the relative drop in minimal pairs with [s], and the relative enhancement of [g] and [θ]. In contrast sets differing by up to two features, the general distribution is more comparable to the full (unconstrained) minimal pair distribution, though the sibilants remain relatively infrequent in comparison to nonsibilant obstruents, a result that is partly a consequence of asymmetries in the association between features and phones, but which nevertheless means that contrasts involving sibilants are formally more robust, consistent with their acoustic/perceptual robustness that was established in Chapters 2–3.

These patterns are further summarized by feature class in Table 5.1, which largely repeats the distributional results discussed in Chapter 3, though it is worth emphasizing that these minimal pair counts are drawn from a different data set from that used in Experiment 1. The most notable asymmetries in featural class contributions to lexical contrast in Table 5.1 are the disproportionate role of voiceless obstruents, plosives, fricatives, and [LOW] coronals. Other asymmetries such as the difference between sibilants and nonsibilants are largely consistent with the number of phones that comprise each class. In terms of the distribution of minimal pairs among feature classes for different δF_{xy} thresholds, among contrasts with fewer featural distinctions, the above patterns generally remain, though there is greater parity between labials and [LOW] coronals, as well as greater enhancement of the sibilant–nonsibilant distinction as discussed above.

Turning next to minimal pair counts by *contrast*, Figure 5.2 shows that this distribution declines exponentially (a power-law fit is shown for consistency with the functional load analysis, but is a poor fit overall), with less than 25% of contrasts accounting for over 50% of the minimal pairs in

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
$\delta F_{xy} = 1$	835	365	838	335	25	2	426	457	67	213	37	152	1048
$\delta F_{xy} \leq 2$	1978	968	1746	1134	51	15	937	1206	150	467	186	500	2446
$\delta F_{xy} \leq 3$	2936	1450	2358	1778	223	27	1254	1767	425	655	285	1024	3362
$\delta F_{xy} \leq 4$	3244	1758	2634	1985	352	31	1394	1952	624	743	289	1332	3670

Table 5.1: Minimal pair counts by feature class and feature difference (δF_{xy}).

the database. From the lower two panels of Figure 5.2 we see that the majority of contrasts occur between plosives and between the plosive set and the fricatives [f, s, z], with contrasts involving [h] and [ʃ] also appearing in this set but at the lower end of the range. This result reflects primarily the distribution of obstruents among word-initial and word-final contrasts, as word-medial contrasts are much less prevalent in the lexicon, perticularly among the higher-frequency core of words used in the Lex95 set, where they comprise only 4% of contrasts, as compared with 65% for CV position and 32% for VC.

When broken down by featural contrast (Table 5.2), overall, place and manner of articulation are by far the most prevalent, with over 75% of contrasts differing in place of articulation, and 62% differing in manner. Voicing and sibilance, by comparison, are contrastive in 48% and 44% of minimal pairs, respectively. This result is not surprising given the greater number of classes comprising place and manner, but it is also a useful reminder that in general the vocal tract allows greater variability in the location and degree of supralaryngeal constrictions than in the number of distinct laryngeal states, or the number of vocal tract configurations resulting in a secondary turbulent noise source at the teeth. However, when reduced to single-feature distinctions—i.e. 'pure' *voicing*, *manner*, *place*, and *sibilance* contrasts—voicing contrasts are twice as prevalent as those due solely to manner of articulation, while pure place contrasts remain widespread at over half of all such contrasts.

In analyzing the role of different acoustic cues in the system of obstruent contrasts in the lexicon, we expect such asymmetries to determine in part whether perturbing a given cue results in widespread reductions in lexical discriminability, or if it has more localized effects that pose

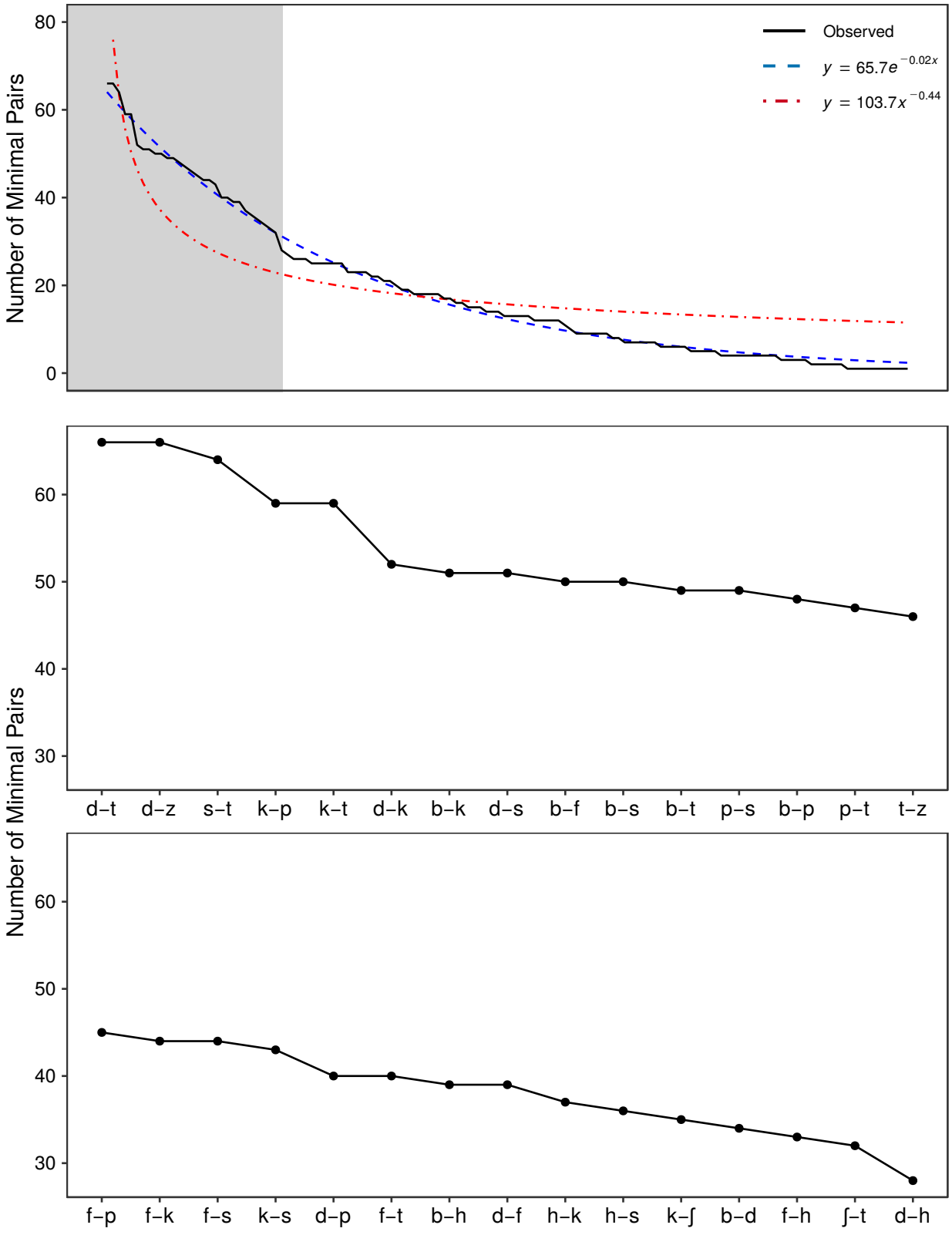

Figure 5.2: Number of minimal pairs in the lexicon by obstruent contrast. The top panel shows the full distribution. The bottom panels show the top 30 contrasts (the shaded region in the top panel), which comprise approximately 55% of obstruent-contrastive minimal pairs in the database.

	Voicing	Manner	Place	Sibilance
$\delta F_{xy} = 1$	177 (30%)	90 (15%)	311 (52%)	22 (4%)
$\delta F_{xy} \leq 2$	528 (36%)	554 (38%)	956 (65%)	308 (21%)
$\delta F_{xy} \leq 3$	900 (41%)	1238 (56%)	1576 (72%)	792 (36%)
$\delta F_{xy} \leq 4$	1208 (48%)	1546 (62%)	1884 (75%)	1100 (44%)

Table 5.2: Minimal pair counts (and percentages) by featural contrast and feature difference (δF_{xy}).

minimal risk to the system. In Section 5.2.2 we examine whether the same patterns hold in a lexical architecture that assumes dependencies between minimal pairs, where contrasts that are relatively less frequent may still play a critical role depending on where they occur in the lexicon.

5.2.1.2 Functional load

The concept of the *functional load* of a phone, contrast, feature, or other phonological class has been pursued by several authors, including early work by Martinet (1952), Hockett (1967), King (1967), and Carter (1987), and broadly attempts to capture the information loss associated with the collapse in a phonological distinction. In other words, the measurement of functional load answers the question: *how much ambiguity in the encoding of linguistic messages is introduced by some perturbation of that code?*, or alternatively, *to what degree does some element of the speech code serve in preventing ambiguity and thereby aiding in message transmission?* The most comprehensive review of the various definitions proposed in these and other work, as well as new generalizations, is provided in Surendran & Niyogi (2003), who provide the following general definition:

$$FL_{\mathrm{T}}(\theta) = \frac{H(L_{\mathrm{T}}) - H(L_{\mathrm{T}_\theta})}{H(L_{\mathrm{T}})} \ ,$$

where L_{T} is a sequence of what Surendran and Niyogi refer to as T-objects, linguistic units of type T, where T can be a phone, syllable, word, etc. Here we use the *word* unit as it provides the closest match with the definition of lexical contrast, and is also the most natural means of interpreting how a loss in some phonological distinction might affect message transmission through the multiplication of homophones in the lexicon. $H(L_{\mathrm{T}})$ is then the entropy of the sequence of

words—or equivalently, the set of words and their corresponding frequencies—in the language, where entropy is defined as $-\sum p_i \log pi$ following Shannon (1948), in this case i being a word and p_i its probability of occurrence based on the composite corpus frequency measure discussed earlier. The new set of (form-unique) words resulting from some change in the code, θ, such as the merger of two phones, is then a new language L_{T_θ} from which we can compute a new entropy $H(L_{T_\theta})$. Finally, by taking the difference in entropies between the two languages (with and without the θ-modification), and normalizing by the original entropy $H(L_T)$, we arrive at a measure of the relative change in entropy, or the relative loss in information, induced by θ.

Beginning with the analysis of obstruent phones, while Surendran and Niyogi do formulate a definition for the functional load of a single phone, this definition is tentative and subject to several assumptions regarding which contrasts are likely to merge, and how a given merger relates to other potential mergers. Surendran and Niyogi define the functional load of phone x as: $FL(x) = \sum_y P(x,y)FL(x,y)$, where $FL(x,y)$ is the functional load of the contrast between x and y, $P(x,y)$ is the probability of a merger between x and y, and y ranges over some subset of phones that are similar to x and thus exhibit some merger potential. Because we do not have direct access to merger probabilities at present, we will set $P(x,y)$ to be a function of the number of features differentiating x and y, where each featural difference reduces the likelihood of merger by half. That is, $P(x,y) = 1/2^{\delta F_{xy}}$, where δF_{xy} is the feature difference between x and y. Finally, we can further control the measure of $FL(x)$ by varying the similarity sets that y ranges over. Here we will again use distinctive features as a guide, and measure $FL(x)$ for different subsets $S(y)$ where $\delta F_{xy} = 1$, $\delta F_{xy} \leq 2$, and $\delta F_{xy} \leq 3$, the fourth in the sequence then reducing to the aggregate measure described above. In Sections 5.3 and 5.4, where simulations of perturbation from noise and cue ambiguity are presented, we will be able to model more directly the likelihood of a given merger based on acoustic/perceptual similarity.

Figure 5.3 shows the functional load of each obstruent phone according to the number of features (δF_{xy}) distinguishing each contrast included in the measurement, where for simplicity we have restricted the minimal pair set in the analysis to obstruent contrasts, as the influence of the re-

mainder will hold constant across analyses of subsets of the obstruent system. This non-obstruent-contrastive set will be included in the network analysis in the next section, where the broader configuration of phonological contrasts in the lexicon is considered.

Overall, many of the key patterns in Figure 5.3 agree with the minimal pair results in Figure 5.1, such as the high weight on the plosives [t, d, k] and the voiceless sibilant [s], and the relatively lower weight on voiced obstruents, dental fricatives, the alveolar flap [ɾ], and the voiced postalveolar fricative [ʒ]. However, some notable discrepancies between the two distributions emerge as well. First and foremost, while functional load declines linearly between [k] and [ʒ], there is a stark jump in functional load between [k] and [d], and further still between [d] and [t], meaning that the alveolar plosives, when accounting for the frequencies of the words they occur in, play a substantially greater role in the lexicon than is evident in minimal pair counts, well-above that of the remainder of the obstruent set. Second, counter to the elevated role of [t, d], the labial plosives [p, b] exhibit much lower functional loads relative to their contribution to minimal-pair contrasts in the lexicon. This result is important given the high confusability of [p, b] in a wide range of obstruent contrasts in Experiment 1, because it means the choice of measurement—minimal pair count versus functional load—will yield different predictions for the impact of noise/cue perturbation on the maintenance of form distinctions in the lexicon.

Finally, when $FL(x, y)$ is restricted to contrasts of a single feature difference (the most likely candidates for merger) the weight on the plosives [p, b, g], as well as the voiceless postalveolar fricative [ʃ], is raised relative to that on the fricatives [s, f, h, z]; these results are largely, though not entirely, consistent with the minimal pair distribution in Figure 5.1: [ʃ], for instance, participates in relatively few single-feature minimal pairs but exhibits a relatively high functional load within this set, meaning that the contrasts [ʃ] does participate in are of high joint token frequency and thus more likely to impact the information potential of the lexicon.

Table 5.3 shows a further breakdown of obstruent phones according to the aggregate functional load of *voicing*, *manner*, *place*, and *sibilance* feature classes. As in Figure 5.3, the distribution of FL by feature class partially agrees with the number of minimal-pair contrasts each class partici-

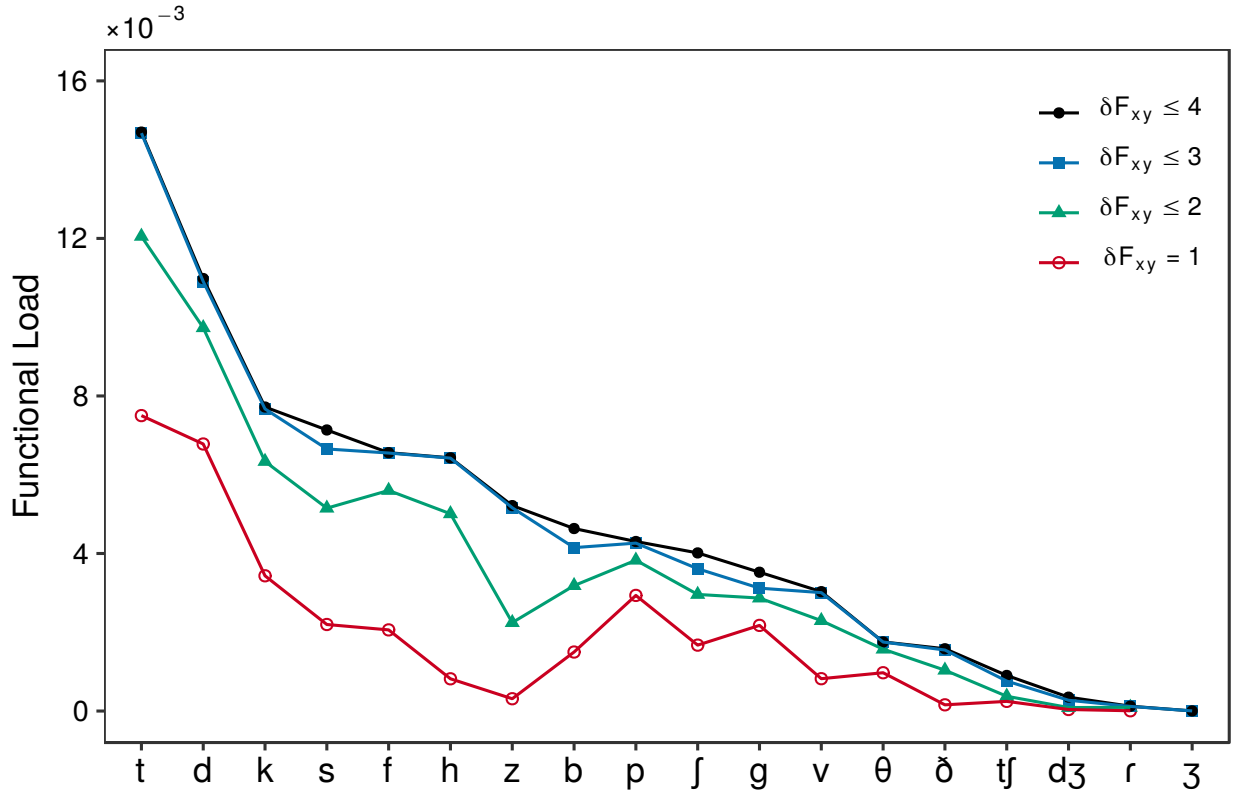

Figure 5.3: Functional load of obstruent phones by feature difference (δF_{xy}).

pates in (Table 5.1), while also exhibiting several key discrepancies. First, the distinction between [LOW] coronals and labials widens in the analysis of functional load, as does the distinction between sibilants and nonsibilants, both results which are consistent with the effect of the logarithmic term in the equation for functional load, as well as the outsized role of the alveolar plosives [t, d] (members of the nonsibilant set) in the highest-weight contrasts (Figure 5.2). Second, the *dorsal* > *glottal* relation, which is consistent in the minimal pair analysis in all but the $\delta F_{xy} \leq 2$ set, largely reverses in the analysis of functional load, a result which is due to the many high-frequency items [h] occurs in ($\mu_h = 406$, $\mu_{k,g} = 225$). Other key patterns from the minimal pair results in Table 5.1, such as the consistent voicing distinction across feature differences, and the minimal role of affricates and the alveolar flap, remain present in the functional load analysis in Table 5.3.

Turning next to the more fundamental measurement of the functional load of particular *contrasts*, Figure 5.4 shows both the full distribution of FL over all obstruent contrasts in the lexicon,

	Voicing		Manner				Place					Sibilance	
δF_{xy}	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
≤ 1	21.8	11.8	24.3	9.0	0.3	0.0	7.3	17.9	2.0	5.6	0.8	4.5	29.2
≤ 2	42.9	21.5	38.0	25.9	0.5	0.1	14.9	31.9	3.4	9.2	5.0	10.8	53.6
≤ 3	52.4	28.3	44.8	34.7	1.0	0.1	18.0	40.8	4.7	10.8	6.4	16.5	64.2
≤ 4	53.5	29.4	45.9	35.7	1.3	0.1	18.5	41.5	5.3	11.2	6.4	17.6	65.3

Table 5.3: Aggregate functional load of obstruent feature classes and feature difference (δF_{xy}).

and a detailed view of the functional load distribution among the 30 most highly weighted contrasts. Unlike the minimal pair distribution in Figure 5.2, the functional load of individual contrasts follows a power-law distribution wherein the top 25% of contrasts account for over 70% of the cumulative functional load among obstruent contrasts in the lexicon. Further, just eight contrasts account for one-third of the cumulative distribution. These contrasts include two highly weighted contrasts with the alveolar plosives (*t–z* and *t–d*), several contrasts with the glottal fricative (*h–b*, *h–t*, and *h–s*), two contrasts with the voiceless labiodental fricative (*f–t* and *f–z*), and the contrast between [d] and [z] that features prominently in morphological distinctions in English. Overall, however, there is good agreement between the two highly weighted contrast sets. Nearly two-thirds of contrasts in the top-30 set based on minimal pair counts remain in the top 30 based on functional load. Most notable among the set of highly weighted contrasts in the latter that are not present in the former are the aforementioned contrasts *h–t* and *f–z*, and several contrasts involving the voiceless sibilants [s, ʃ].

Table 5.4 further summarizes these results by featural contrast, where as in Surendran & Niyogi (2003) only single-feature partitions of the obstruent set have been considered; e.g., for voicing we analyze the set {*p–b*, *t–d*, *k–g*, ..., *s–z*}, for place, {*p–t–k*, *b–d–g*, ..., *s–ʃ*}, and so on for manner and sibilance contrasts. As in the minimal pair counts in Table 5.2, among singular featural contrasts ($\delta F_{xy} = 1$), place and voicing far outweigh manner and sibilance, with manner of articulation even further reduced in functional load relative to voicing and place. However, unlike in the minimal pair distribution, where place contrasts outnumber voicing by 5:3, voicing is of higher functional load, reflecting primarily the dominant role of contrasts between [t] and [d], though with the excep-

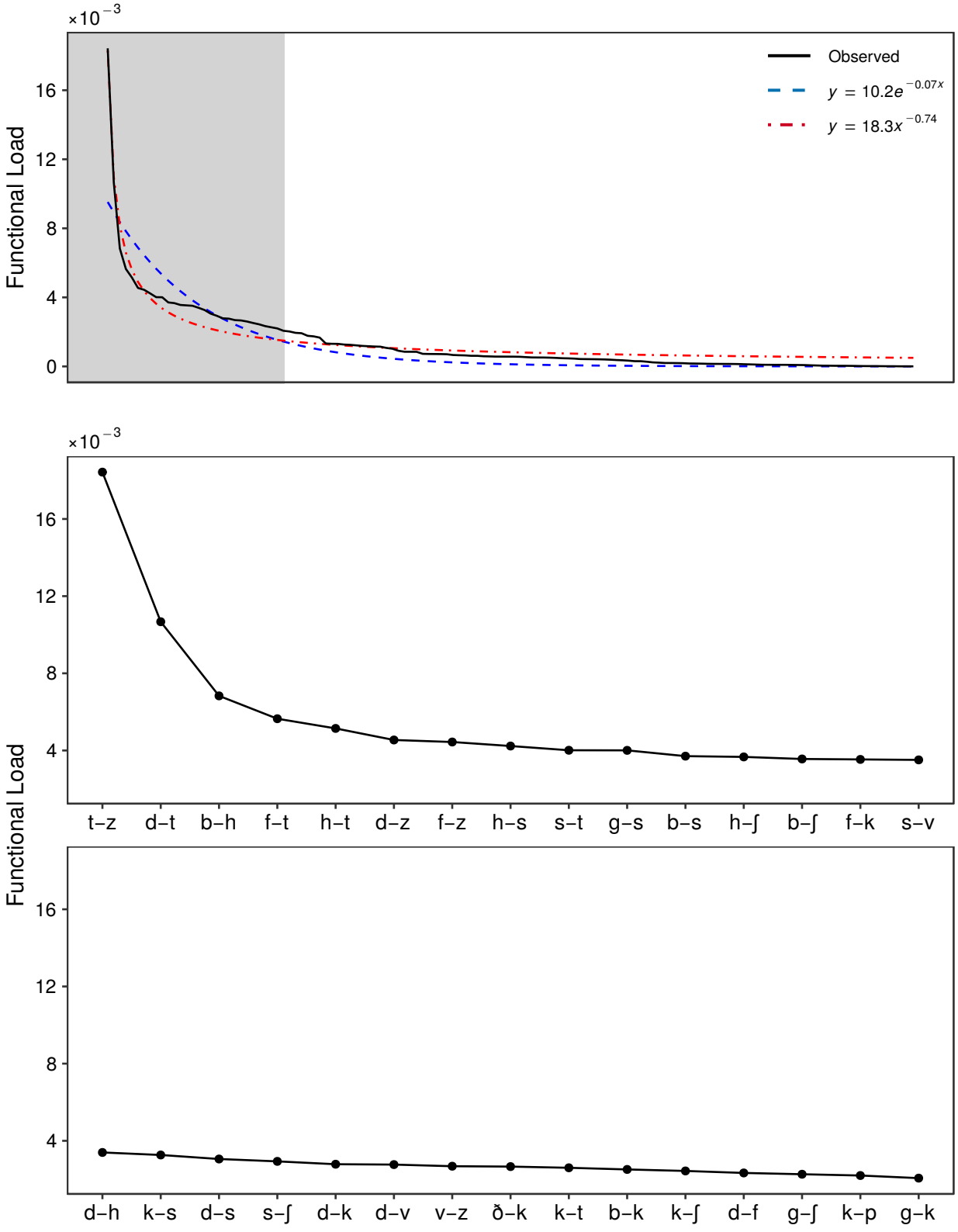

Figure 5.4: Functional load of obstruent contrasts in the lexicon. The top panel shows the full distribution. The bottom panels show the top 30 contrasts (the shaded region in the top panel), which comprise approximately 70% of the cumulative distribution of functional load among obstruent contrasts in the database.

Voicing	Manner	Place	Sibilance
16.3	2.7	14.8	0.8

Table 5.4: Functional load of singular featural contrasts (i.e., $\delta F_{xy} = 1$).

tion of *g*–*k*, all other high-FL voicing contrasts involve multiple featural distinctions ($\delta F_{xy} > 1$). Pure place contrasts are similarly sparse, confirming that as in the discussion of cue integration in Chapter 4, much of the information encoded by obstruents in the lexicon depends on more complex multi-feature distinctions that till now have received little attention in the phonetic literature.

Finally, we must emphasize that where the measurement of functional load and minimal pair counts agrees, much of this compatibility can be attributed to the use of the word as the fundamental unit of functional load—i.e., the T-object is a word, as opposed to a phone, syllable, or some other unit. However, this is a natural assumption given that historical sound mergers typically occur in lexically constrained contexts, diffusing throughout the language as a function of usage characteristics of the words containing such contexts (Wang, 1969; Labov, 1981, 1994), as well as patterns of speech transmission successes/errors that directly involve word-level contrast relations in the form of lexical competition in speech perception (Martinet, 1952; Ohala, 1993). Next we assess the phonological structure of the lexicon under a network architecture, a structure similarly dependent on word-level distinctions and thus compatible with the theoretical motivations above.

5.2.2 Network architecture

The baseline network used in the present study follows the phonological network design of Vitevitch Vitevitch (2008)—vertices represent word types that are connected by an edge if the Levenshtein distance between their phonological transcriptions is equal to one (i.e., if they form a *minimal pair* relation)—but based on the 3,406 words words in the Lex95 database constituting 95% of tokens in several English corpora. General characteristics of the phonological network built on this database follow those in Vitevitch (2008). As in Vitevitch (2008), all network statistics are computed on the giant component of the network: the largest component of the network, where a

component is defined as a subgraph where any two vertices (i, j) are connected by paths—i.e., in the case of a phonological network, w_i and w_j can be derived from each other by traversing a sequence of dependent minimal pairs (within the same network) formed via phone addition, deletion, or substitution—and where no such vertex is connected to (is a minimal pair of) any other vertex in the network (lexicon) outside of that set. In the phonological network studied in Vitevitch (2008), the giant component contained 6,508 words, while the network analyzed in the present study is just under half that at 3,406 words.[2] This difference is expected given the frequency constraint on the development of the Lex95 database, as no such constraints were applied in Vitevitch's analysis.

Despite this discrepancy in network size, the average path length $\bar{\ell}$ in each network—the mean of the shortest-path distance (the number of edges) between every pair of vertices—is quite consistent, at 6.05 in Vitevitch (2008), and 6.25 in the present study. Given that $\bar{\ell}_{V08} \approx \bar{\ell}_{Lex95}$, while $|V(G_{V08})| \approx 2 \cdot |V(G_{Lex95})|$, the result that the Lex95 network is denser, with a global clustering coefficient of 0.304 as compared with 0.126 in Vitevitch (2008), is expected (Strang et al., 2018) and means that information flow in the present network is predicted to be even higher than the values predicted for the phonological network in Vitevitch (2008) and in subsequent studies (Arbesman et al., 2010; Vitevitch et al., 2011).

Finally, as in Vitevitch (2008), the degree distribution of the Lex95 network closely fits an exponential function, where the probability a given vertex has degree k, $P(k) = 0.149e^{-0.134x}$ (RMSE = 0.009), and when end vertices ($k = 1$) are excluded, the distribution follows the exponential even more closely: $P(k|k > 1) = 0.183e^{-0.163x}$ (RMSE = 0.004). Both values agree with Vitevitch's results, both in the general form of the function and in the closeness of the fit. In general, the exponential form of the degree distribution in the phonological network conforms with network models of many biological and social systems (Amaral et al., 2000; Keller, 2005; Clauset et al., 2009), and is a much more plausible form than the power-law distributions that have received much greater attention in the network science literature due to their desirable *scale-free* properties (Price, 1965; Barabási & Albert, 1999). Many such systems exhibit the so-called 'small-world' property

[2] As noted earlier, the Lex95 database is already constrained to include only those words in the giant component of the network.

of having a small number of integral elements that participate in many relations in the system, with a much larger number of elements that are more peripheral in the system but nevertheless connected to the core elements via a relatively small number of relations (Watts & Strogatz, 1998). This description is consistent with with many well-known characteristics of linguistic unit distributions and relations such as Zipf's law, asymmetries in phonological markedness, and phonotactic constraints on syllable structure and sound-sequence combination.

Two general characteristics of *local* and *global* phonological network structure are explored in this chapter as a means of describing the role of different obstruent contrasts in the lexicon, and the acoustic structure that underlies their detection in the signal. The first measure, *edge disjoint degree*, $k_{\ominus}(xy)$, is an extension of *vertex degree* (alternatively referred to in the literature as *neighborhood density*), and is defined as the number of neighbors of vertices x and y (comprising the edge xy) that are not jointly neighbors of both x and y. More formally, $k_{\ominus}(xy) = |N(x) \ominus N(y)|$, where $N(x)$ is the set of neighbors of x, $N(y)$ is the set of neighbors of y, and $\ominus$ represents the symmetric difference between the two—i.e., the size of the disjoint sets in the union of neighborhoods between x and y, or $[N(x) \cap N(y)]^{\mathrm{C}}$. In other words, edge disjoint degree captures the number of new minimal-pair relations resulting from a given contrast merger, and thus represents a kind of second-order measure of functional load.

Figure 5.5 shows sample measurements of edge disjoint degree in a subset of the lexicon, and illustrates how contrasts may differ in the larger ensemble of words they help to differentiate, in addition to the minimal-pair constituents they directly distinguish. The contrasts represented by edges a and b, for instance, each distinguish words occurring in largely disjoint neighborhoods (color-coded in agreement with that of the contrast in question), where the neighborhoods of the former are notably denser than those of the latter. Contrast c, by comparison, exhibits much greater overlap in the neighborhoods of its constituents. This configuration translates to an edge degree of 12 for edge a, 5 for edge b, and 8 for edge c. Contrasts a and b differ primarily in their neighborhood sizes, and given that these neighborhoods are largely disjoint, this difference translates into a substantial $k_{\ominus}(xy)$ distinction. Contrasts b and c, on the other hand, differ in both neighborhood

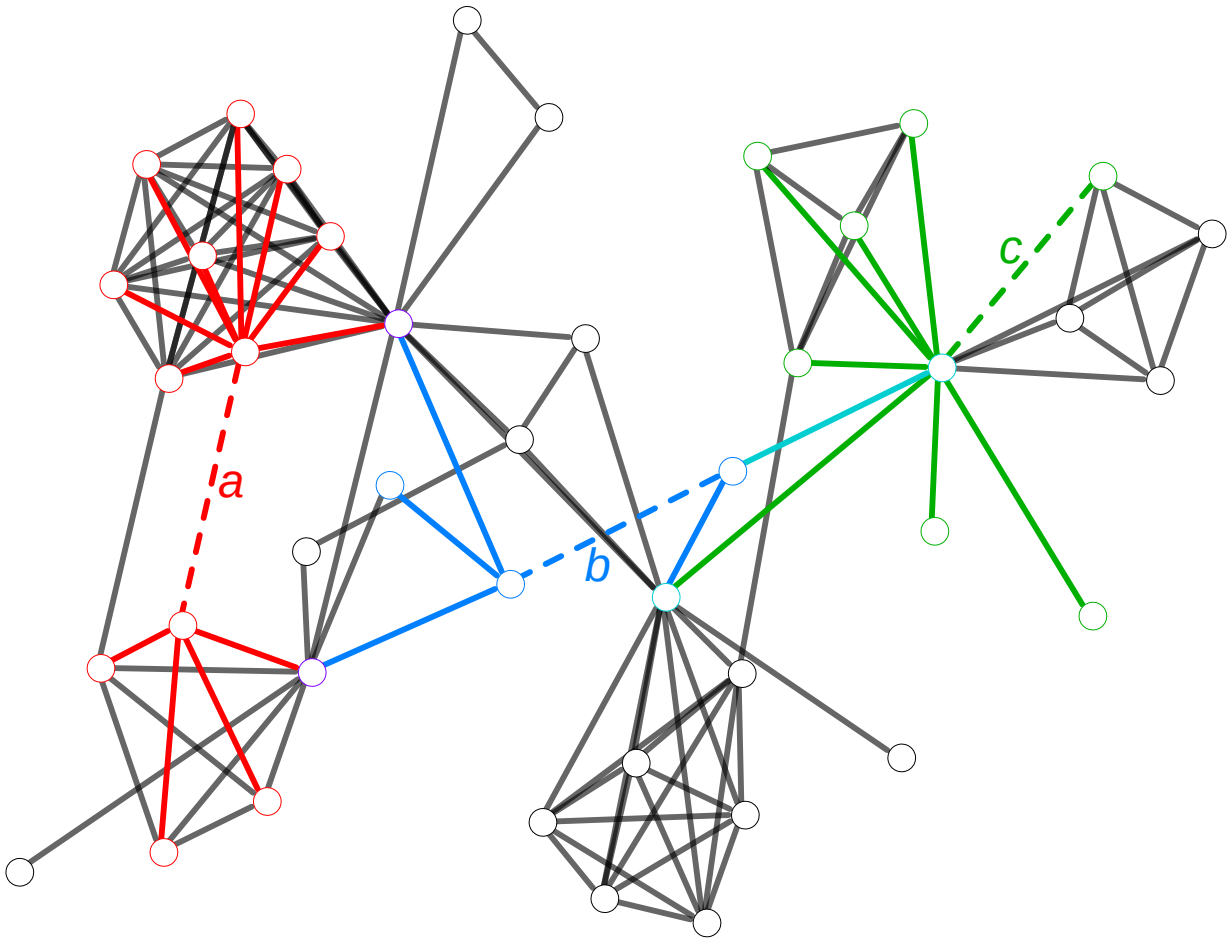

Figure 5.5: Sample measurement of edge disjoint degree. Edges a, b, and c have edge disjoint degrees of 12, 5, and 8, respectively. Disjoint neighborhoods contributing to the quantification of $k_{\ominus}(xy)$ are highlighted.

sizes and their relative degree of overlap, resulting in more similar edge disjoint degrees. Thus, if we were to simulate contrast mergers for this set, we would predict that the merger of a would result in a much larger increase in phonological similarity among items in this subset of the lexicon than would a merger of contrast b, with c somewhat intermediate between the two.

In the analysis of gradient contrastiveness in the simulations of acoustic perturbation in Sections 5.3 and 5.4, we use a weighted network where edges are not simply present or absent, but they are given a weight based on the relative vulnerability of the contrast in perception (measured as the error rate on that minimal pair), and on the predicted vulnerability of the contrast from the cue-integration models in Chapter 4. In a weighted graph the equivalent measure of edge disjoint degree is *edge disjoint strength* (again borrowed from the concept of *vertex strength* in weighted

networks). Here, instead of simply counting the size of each disjoint neighborhood, we take the sum of the weights on those contrasts, and then multiply that sum by the weight on the minimal pair in question. That is, for a given edge xy, where s_x is the vertex strength of word x (the sum of weights on contrasts with x), s_y is the vertex strength of word y, and w_{xy} is the weight of edge xy, $s_{\ominus}(xy) = w_{xy}(s_x + s_y - w_{xy} - [w_{xi} + w_{yi} : i \in N(x) \cap N(y)])$. The final term in this equation is a correction for the weights on contrasts with words that are present in both neighborhoods.

One way of thinking about $s_{\ominus}(xy)$ is that it measures the probability of an increase in local wordform similarity due to a given contrast xy, where for instance, if xy is robust (e.g., contrasts with the voiceless sibilant fricative [s]) and $w_{xy} \approx 0$, the relative similarity of words in the remainder of the set does not matter because xy is unlikely to merge or lose discriminability in perception. However, as w_{xy} approaches 1, the potential increase in similarity relations between neighbors of x and y approaches the result where x and y merge into a single node, and the total strength of the neighborhood of that node equals the sum of the individual vertex strengths s_x and s_y, minus the sum of the weights on edges incident on x and y. Considering Figure 5.5 again in the context of this new weighted measure, if all other edge weights are held constant—e.g., at 0.1, or an error rate of 10%—and edge weights on a, b, and c range between 0.1 and 0.9, the edge strength of a would range between 0.12 at its minimum (12×0.1) and 1.08 (12×0.9) at its maximum, while edge disjoint strengths of b and c are much lower on average and with a narrower range of variation: 0.05–0.45 for contrast b, 0.08–0.72 for contrast c. This weighted measure has a natural interpretation, in that it implies that if words x and y are acoustically/perceptually similar, but y and z are comparatively distinct in this regard, then any merger between x and y will have less of an effect on the resulting minimal-pair relation between x and z than if the yz contrast is also acoustically/perceptually similar. For example, we might expect the set {*bat*, *pat*, *pad*} to behave differently than {*bat*, *pat*, *pass*}, where a *bat–pat* merger is likely to impact the former more than the latter, given the general robustness of [s] in obstruent contrasts.

Thus, any perturbation of contrast a is likely to have a much greater impact on the discriminability of its local neighborhood than if b or c were to be perturbed. Of course, such contrasts are

likely to differ in both their baseline error rates (and thus their edge weight) and in the acoustic and lexical characteristics that make them more or less prone to variation in error rates. This means that while configurations such as those in Figure 5.5 favor contrasts of types *b* and *c* (as points of relative robustness), if *b* and *c* were to be highly confusable and *a* highly discriminable, the impact of noise masking on the former can have greater consequences for the surrounding network. This relationship between acoustic robustness and lexical configuration has been explored elsewhere (Scarborough, 2004; Wright, 2004; Munson & Solomon, 2004; Baese-Berk & Goldrick, 2009), and is one of the motivations for directly studying the acoustics of the lexicon.

The second general measure investigated here, which provides estimates of the global structure of contrast in the phonological network, is *average path length*, $\bar{\ell}$, which as described in the introduction to this section is the mean distance between any two words in the network, or the mean number of contrast mergers necessary to eliminate the phonological distinction between any pair of words in the lexicon. The equivalent measure of $\bar{\ell}$ for weighted graphs is determined by defining the distance between nodes i and j as the sum of edge weights along the path that minimizes this sum. Given that weights are defined in our network as *similarities*, rather than *costs*, as is used in the original definition, for the weighted $\bar{\ell}$ measure we first take the complement of the weights (i.e., by taking $1 - w_{xy} \; \forall \; xy \in E$)—effectively changing the interpretation of edge weight from relative vulnerability (based on error rates) to relative robustness (based on accuracy)—and then measure the mean distance between each pair of nodes according to the definition above.

The role of individual contrasts in the average path length of the network can then be measured via simulation by changing the weight on a given contrast and observing the resulting change in $\bar{\ell}$. Figure 5.5 illustrates that according to this more global measure the roles of contrasts *a* and *b* reverse, while the similarity of *b* and *c* changes as *c* is less central in the network than *b* is. That is, while *b* has relatively few direct neighbors, it serves as a bridge between two largely distinct clusters in the network, and therefore a change in the weight assigned to edge *b* has a greater impact on the overall separation of items in the network than changes in the relative discriminability of *a* or *c*. For example, holding all else constant (at $w = 1$) and varying in turn the weights on edges *a*–*c*

(between 0.1 and 0.9), a change in the discriminability of *b* from 0.9 to 0.1 causes a 4.2% drop in average path length, from 2.71 to 2.60, while contrast *a* shows a 1.7% drop (2.72–2.67) and contrast *c* shows a 1.3% drop (2.72–2.68). This measurement of the change in $\bar{\ell}$ will be particularly useful when noise and cue perturbations are applied to the lexicon as a whole, as it serves as an index of the global reduction in lexical distinctiveness from an increase in uncertainty in the acoustic signal. Such a measure is helpful in identifying the general *spread* of contrasts in the lexicon; i.e., is a given phonetic distinction relatively localized to a particular subset of words of a given phonological structure (e.g., the highly clustered {*bat*, *pat*, *sat*, ..., *hat*} set), or is it more broadly distributed among words of different lengths and phonotactic characteristics. Next we review the role of obstruent categories and contrasts as measured by *edge disjoint degree* in the phonological network, followed by an analysis of *average path length* in Section 5.2.2.2.

5.2.2.1 Edge disjoint degree

Figure 5.6 shows the aggregate disjoint degree/strength of edges involving each obstruent phone, where again *edge disjoint degree*, $k_{\ominus}(xy)$, is defined as the combined neighborhood densities of words x and y, excluding the intersection of the two sets, and *edge disjoint strength*, $s_{\ominus}(xy)$, is the weighted equivalent, summing the edge weights over the same set. Here as an illustration of the operation of $s_{\ominus}(xy)$ on the basis of phonological information alone, we have adopted as edge weights the contrast merger probabilities from the analysis of functional load in the previous section, which defines the probability of a merger between x and y as $P(x,y) = 1/2^{\delta F_{xy}}$, where δF_{xy} is the number of features differentiating the obstruent contrast xy. In the perturbation analysis in Sections 5.3 and 5.4, $P(x,y)$ will be estimated directly from predicted listener error rates.

Overall, the ranking of obstruents according to edge disjoint degree is similar to the results for minimal pair count and functional load. The most critical set is composed primarily of plosives and the anterior voiceless fricatives [s] and [f], while the aggregate role of voiced obstruents according to $k_{\ominus}(xy)$ is relatively low. When the weighted measure, $s_{\ominus}(xy)$, is used, these trends largely remain, with the biggest discrepancies being a sizeable drop in the role of [s], a modest drop in

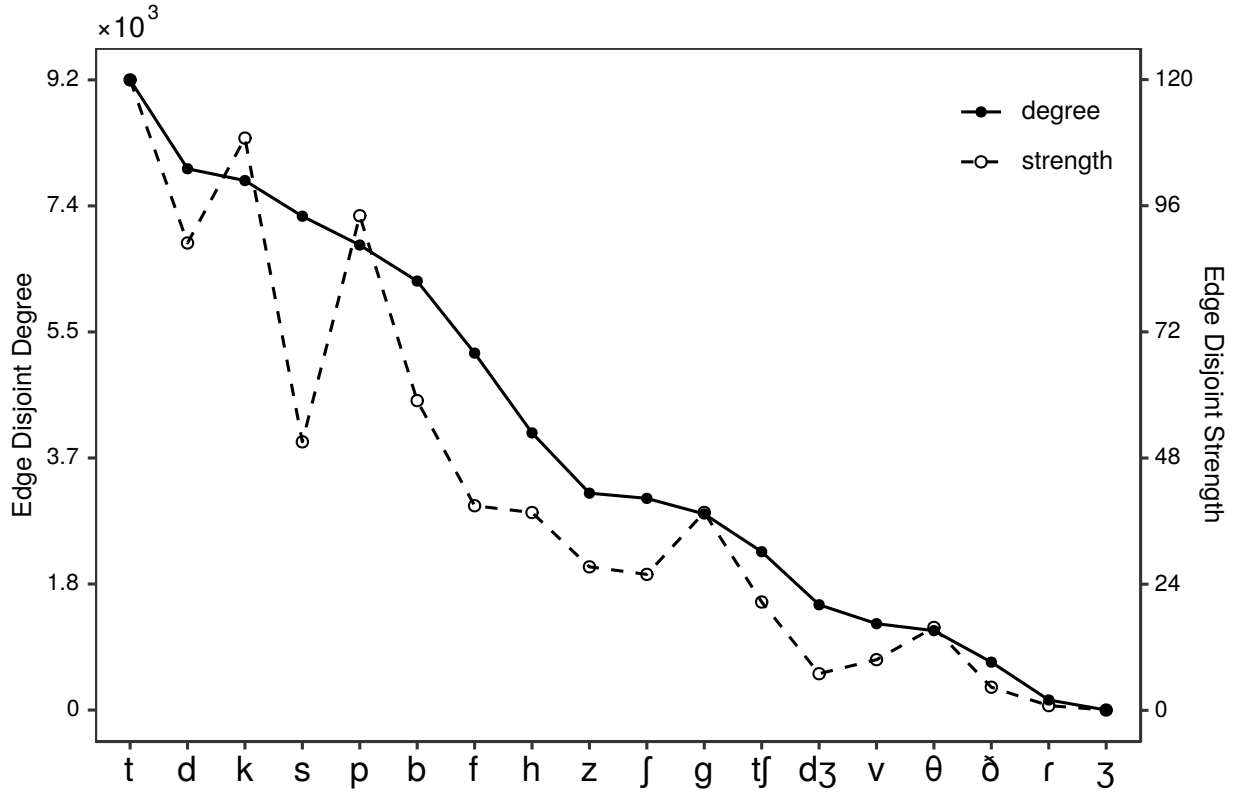

Figure 5.6: Edge disjoint degree/strength aggregated across participating obstruent phones.

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
$k_{\ominus}(xy)$	46.8	23.9	40.8	25.9	3.9	0.2	19.6	29.5	7.0	10.6	4.1	17.4	53.4
$s_{\ominus}(xy)$	0.52	0.24	0.51	0.21	0.03	0	0.20	0.31	0.05	0.15	0.04	0.13	0.62

Table 5.5: Aggregate edge disjoint degree/strength ($k_{\ominus}(xy)$, $s_{\ominus}(xy)$). Degrees are divided by 1000 for clarity.

[d], and an increase in the roles of [k] and [p]. That is, when the featural similarity of obstruent contrasts is taken into account, the voiceless plosives [p, t, k] are the most likely to be the source of a local increase in wordform similarity in the lexicon.

When broken down by feature class, in aggregate (i.e., summing across all constituent phones) voiceless obstruents and plosives distinguish disjoint neighborhoods of nearly twice the next largest class—voiced obstruents and fricatives, respectively. Affricates and flaps, on the other hand, play a substantially more marginal role in this regard. Among places of articulation, coronals and labials are again the most dominant sets, comprising well over half of the cumulative $k_{\ominus}(xy)$ distribu-

tion. Among the remaining places, dorsals still occupy a notable role given that there are only two phones, [k, g], in the set. Finally, sibilants occur in contrasts distinguishing relatively smaller disjoint neighborhoods in aggregate than nonsibilants, a result which is partly due to formal differences in the two feature sets (sibilants representing 6 phones, as compared with 12 for nonsibilants), but given that the two classes differ by greater than a factor of 2 this result is not entirely a consequence of formal definitions. The prominent role of plosives in the lexicon drives much of the aggregate edge disjoint degree among sibilants, as does the role of the nonsibilant fricatives [f] and [h]. These patterns are largely retained in the weighted measure, $s_{\ominus}(xy)$, though due to both their prevalence in the lexicon and their featural similarity, plosives, dorsals, and nonsibilants show a widening of asymmetries in relation to fricatives, [HIGH] coronals, and sibilants, respectively.

Figures 5.7 and 5.8 show $k_{\ominus}(xy)$ and $s_{\ominus}(xy)$ distributions by obstruent contrast. In the unweighted phonological network, where all contrasts are treated as equivalent in acoustic/perceptual similarity, the relative size of disjoint neighborhoods of words distinguished by a given contrast generally follows an exponential distribution, with only the top three contrasts, *d–t*, *k–t*, and *s–t*, increasing above the remainder according to a power-law function (see the curves for exponential and power-law fits in Figure 5.7). Other contrasts exhibiting a high aggregate edge disjoint degree—i.e., summing across all $k_{\ominus}(xy)$ values on minimal pairs a given phonetic contrast distinguishes—include much of the remainder of the plosive contrasts (all pairwise relations between plosives, excluding all contrasts with [g] except *t–g*) as well as many contrasts between voiceless fricatives and between obstruents in each of these sets. This result is similar to the results for minimal pair count and functional load, though with some variability in the relative ranking of contrasts at the upper end of the $k_{\ominus}(xy)$ range. Most notably, there is greater similarity between the minimal pair counts and $k_{\ominus}(xy)$ in this set, as several contrasts of high functional load (*b–h*, *f–t*, *h–t*) exhibit much lower edge disjoint degrees, likely due to the influence of word frequency on the latter.

Turning next to the weighted measure, *edge disjoint strength*, we see in Figure 5.8 that the distribution of $s_{\ominus}(xy)$ more closely follows a power-law than an exponential function, where the top five contrasts—[k, t], [d, t], [k, p], [p, t], and [b, p]—comprise one-third of the cumulative $s_{\ominus}(xy)$

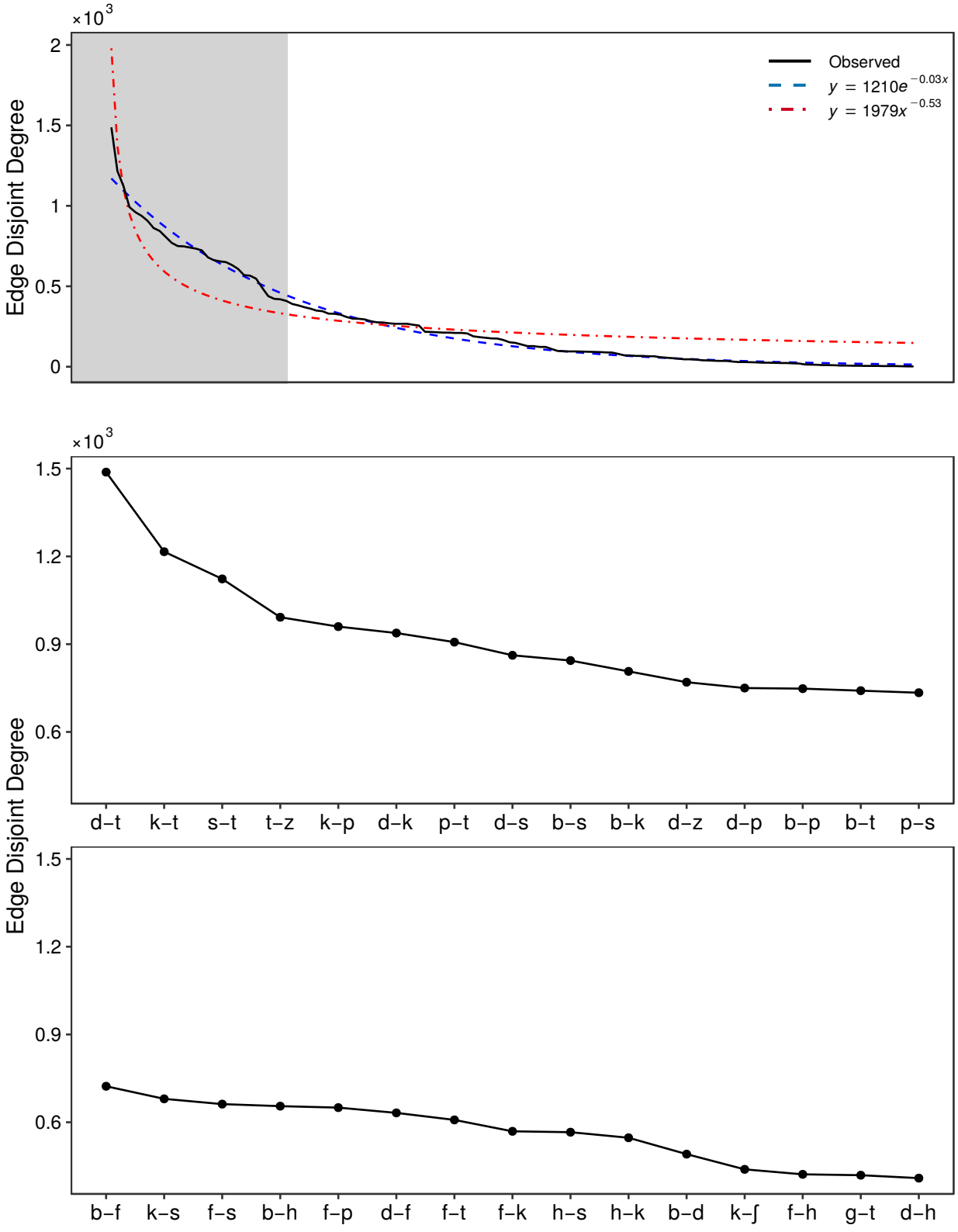

Figure 5.7: Aggregate edge disjoint degree of obstruent contrasts in the lexicon. The top panel shows the full distribution. The bottom panels show the top 30 contrasts (the shaded region in the top panel), which comprise approximately 63% of the cumulative distribution of $k_{\ominus}(xy)$ among obstruent contrasts.

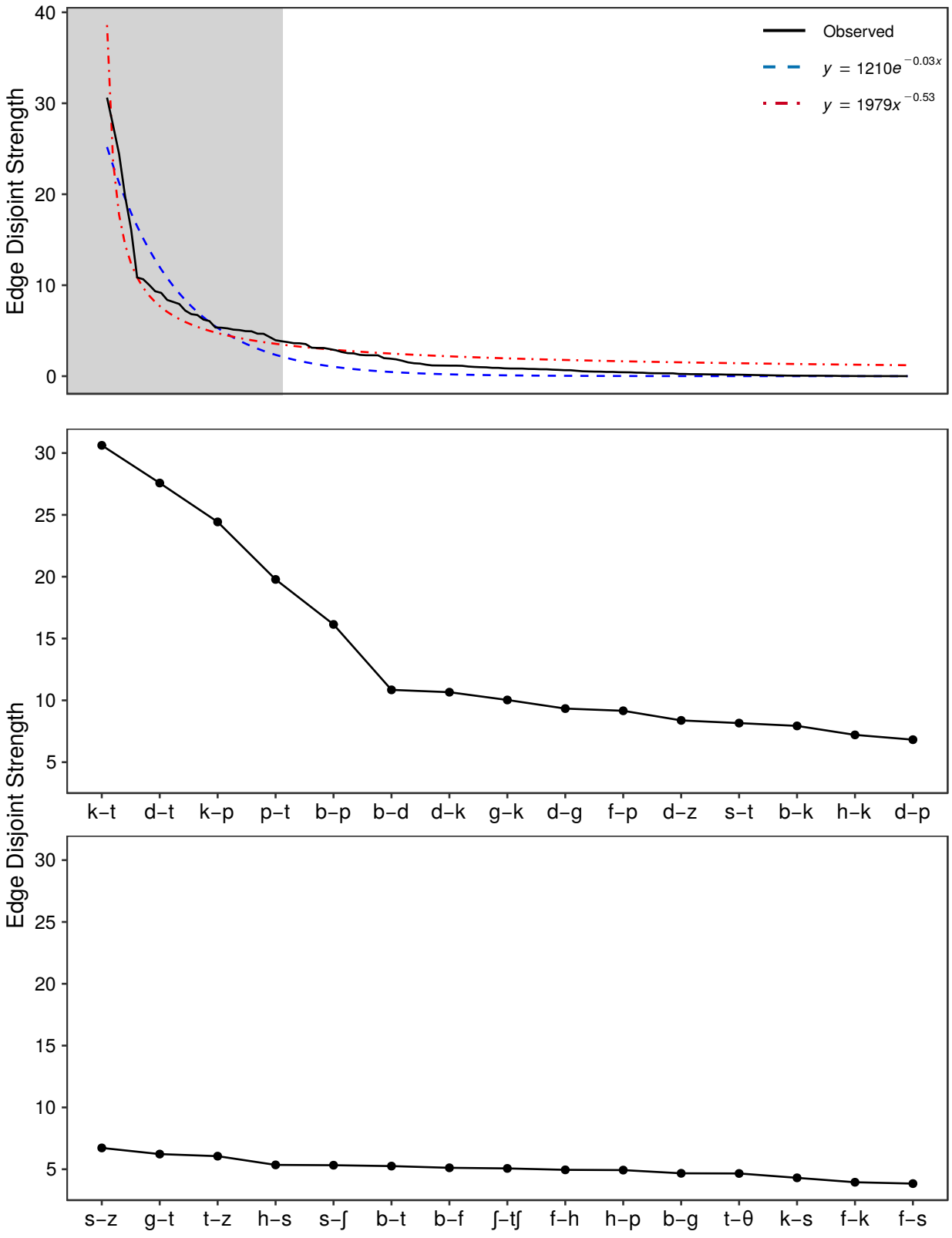

Figure 5.8: Aggregate edge disjoint strength of obstruent contrasts in the lexicon. The top panel shows the full distribution. The bottom panels show the top 30 contrasts (the shaded region in the top panel), which comprise approximately 75% of the cumulative distribution of $s_{\ominus}(xy)$ among obstruent contrasts.

	Voicing	Manner	Place	Sibilance
$k_{\ominus}(xy)$	17.5	20.7	26.1	14.6
$s_{\ominus}(xy)$	151	139	254	83

Table 5.6: Edge disjoint degree/strength ($k_{\ominus}(xy)$, $s_{\ominus}(xy)$) aggregated across featural contrasts. Degrees are multiplied by 1000 for clarity.

distribution. Most of these contrasts (*k–t*, *d–t*, *k–p*, *p–t*) exhibit high edge disjoint degrees as well, indicating that they occur in highly dense sub-networks independent of featural similarity structure. The labial contrast, *b–p*, however, is much higher in $s_{\ominus}(xy)$ than $k_{\ominus}(xy)$, indicating that the featural similarity of obstruent contrasts adjacent to *b–p* minimal pairs in the lexicon is generally higher than that in the remainder of the high-$s_{\ominus}(xy)$ set. This result is particularly important for the analysis of noise- and cue-perturbation, as the labial plosives are less robust—both acoustically and perceptually—than the coronals or velars, though all contrasts in this set are notably more similar than the median obstruent contrast in Experiment 1 (see the contrast accuracy results in Figures 3.11–3.13 for details).

Table 5.6 further summarizes the edge disjoint degree/strength distributions by featural contrast. According to both measures, place distinctions play the largest role, as suggested by the predominance of plosive contrasts in the upper $k_{\ominus}(xy)$ and $s_{\ominus}(xy)$ range. However, the relative roles of the other three features change notably between the unweighted and weighted measures. For instance, manner contrasts show a higher aggregate edge disjoint degree than voicing contrasts, but when the featural similarities of contrasts are accounted for in $s_{\ominus}(xy)$ the relation reverses. Voicing contrasts distinguish relatively larger disjoint neighborhoods of high featural similarity relative to manner contrasts. Finally, the distinction between the role of sibilance and other features is magnified in the weighted network, a result which is consistent with the contrast results in Figures 5.7 and 5.8, where the greater featural robustness of sibilance contrasts makes them a less likely source of local wordform similarity than voicing, manner, or place.

As noted in the introduction to this section, edge disjoint degree/strength provides an estimate of the *local* neighborhood structure distinguished by a given contrast, and as such captures the

most likely immediate impact of a reduction in lexical discriminability due to uncertainty in the signal. In the next section we analyze simulated changes in *average path length* as an estimate of the more *global* role each contrast plays in maintaining phonological distinctions in the lexicon.

5.2.2.2 Average path length

Figure 5.9 shows the relative contribution of each obstruent phone to the average path length in the network, where in the unweighted case (not accounting for featural similarities among obstruents), $\Delta\bar{\ell}$ is measured by setting all edge weights involving a given phone to 0.01, then taking the original average path length and the new value. In other words, this measurement simulates the relative reduction in phonological separation among items in the lexicon under a near-merger of all items in a given contrast set. In the network where contrasts are weighted according to their featural similarity, phone contributions to average path length are measured by setting all weights in a given set to 1 and taking the difference between this value and the original $\bar{\ell}$, thereby answering the question: *to what degree do featural similarities among obstruent contrasts contribute to changes in the mean phonological separability of items in the lexicon?*

Among the obstruents exhibiting the greatest contribution to the average path length in the network, irrespective of contrast feature structure, are many of the phones with high minimal pair counts in general; namely, the plosives [p, t, k, b, d] and the voiceless anterior fricatives [s, f]. What is different between Figures 5.9 and 5.1 is the relative ranking of phones within this set, as [t] plays a notably less central role in the obstruent contrasts which define the structure of the broader phonological network, while contrasts [k, d, s, p, f] are all more prominent in their contribution to the average path length in the lexicon relative to independent minimal pair counts. This means that if contrasts involving such sounds were to be perturbed by background noise or cue loss, for example, such perturbations would be expected to have a greater effect on the global separation of items in the network than would perturbations of other phones such the alveolar flap, or the voiced fricative and affricate series.

However, when featural similarity is taken into account in the weighted network, [t] plays the

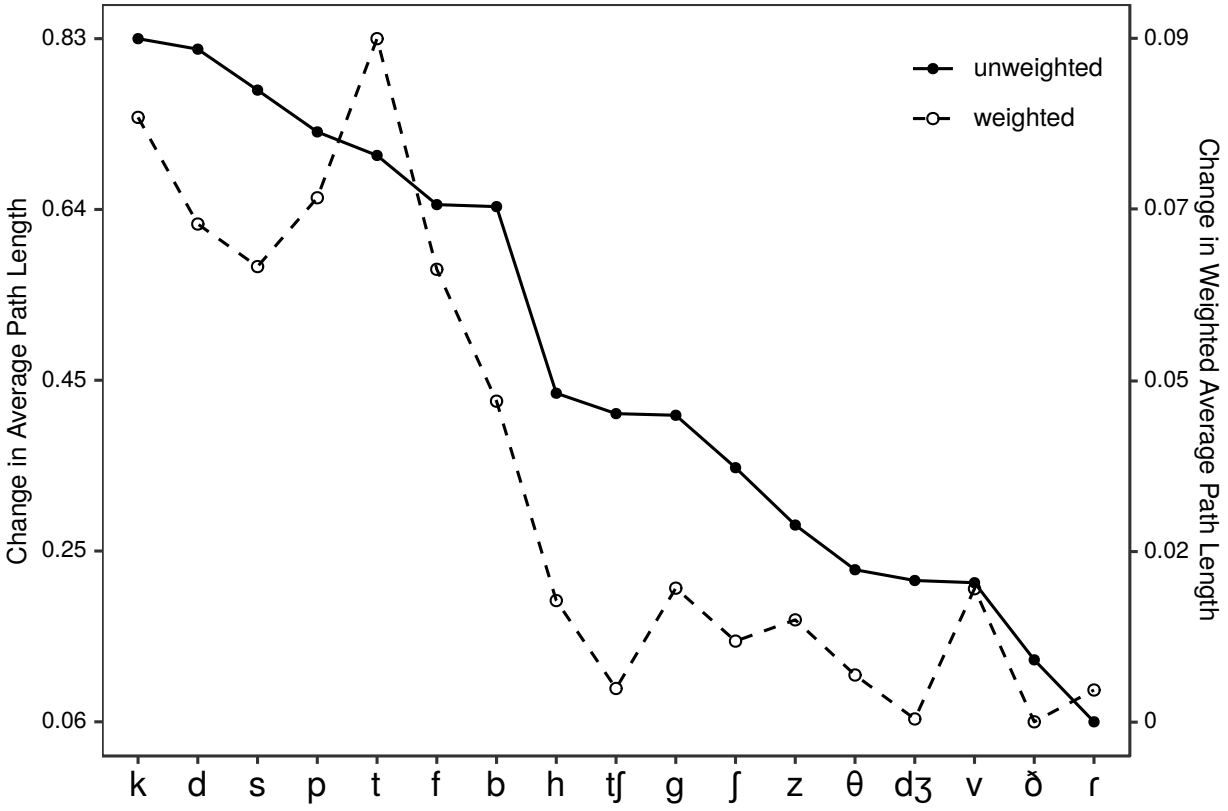

Figure 5.9: Changes in unweighted and weighted average path lengths aggregated across contrasts involving each obstruent phone. The former is measured by calculating the absolute change in $\bar{\ell}$ when the weight on each contrast involving a given phone is changed from 1 to 0.01 (simulating a near-merger), while the latter adopts the opposite procedure, changing contrasts from their initial weight to 1.

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
unweighted	1.38	1.28	1.37	1.29	0.55	0.06	1.21	1.25	0.74	0.98	0.43	1.17	1.42
weighted	0.30	0.15	0.27	0.16	0.01	0.01	0.17	0.21	0.02	0.10	0.02	0.09	0.33

Table 5.7: Changes in unweighted and weighted average path lengths aggregated across contrasts involving each feature class.

greatest role in reducing $\bar{\ell}$, meaning that [t] occurs in many contrasts that are separated by few distinctive features, as well as occurring along low-weighted paths, which in the context of the present study's focus on obstruent contrasts means that many words that are removed from a given [t]-contrastive item by 2 or more segmental changes differ along a chain of obstruent contrasts that consequently are more likely to be perturbed by uncertainty along a given acoustic dimension. By comparison, contrasts involving [s], [b], and [d] are more robust (featurally), while the roles of [k], [p], and [f] are relatively similar between weighted and unweighted measures. This drop in the contribution of [s] to reducing the average path length in the network is expected given its sibilance increases the featural separation between [s] and other obstruents; however, [b] and [d] exhibit many more featural affinities with other obstruents, meaning that their occurrence in minimal-pair contrasts and neighborhoods of obstruent-contrastive items in the lexicon tends to favor more featurally distinct phones relative to the distribution of the voiceless set [p, k, f].

Table 5.7 shows the contribution of obstruent phones to average path length when broken down by feature classes. Overall, these results are comparable to the results for minimal pair count and functional load. Voiceless obstruents play a greater role than voiced obstruents, particularly when the featural structure of obstruent distinctions is taken into account. Similar relations are obtained for *sibilance* (nonsibilants $>$ sibilants), and for the distinctions between plosives and fricatives, and between [LOW] coronals and labials.

Turning next to the relative role of individual contrasts in reducing the average path length in the phonological network, Figures 5.10 and 5.11 show the change in average path length in the unweighted and weighted networks, respectively, when each is either simulated as undergoing a near merger (Figure 5.10) or its relative vulnerability is removed by setting its edge weight

to 1 (Figure 5.11). As with *edge disjoint degree/strength*, the distribution of $\Delta\bar{\ell}$ by contrast in the unweighted network closely follows an exponential distribution, while the distribution of the weighted measure is better characterized by a power-law function. The most critical contrasts in the unweighted network according to this measure are largely sibilance distinctions and distinctions among plosives, while the role of sibilance contrasts in the weighted network is notably reduced, with the upper end of the $\Delta\bar{\ell}$ range primarily composed of plosive distinctions. This result reflects the fact that the greater featural robustness of sibilance distinctions (which also matches listener perception) means that rather than contributing to global wordform similarity in the lexicon, these contrasts are critical in maintaining item separation in the network.

These results are further confirmed in the aggregate role of featural contrasts in Table 5.8, where place of articulation distinctions are both more widely distributed in the lexicon and responsible for the greatest increases in global wordform similarity when featural affinities among obstruent contrasts are taken into account. Sibilance contrasts, on the other hand, though less widespread in the lexicon due to asymmetries in the relative sizes of the sibilant and nonsibilant sets, are still close to voicing distinctions in their distribution in the lexicon while being notably more robust. Finally, voicing and manner of articulation play an intermediate role in both the unweighted and weighted networks, and are generally similar in their contribution to the global phonological structure of the lexicon. This result is interesting because the two features differ in the number of distinct classes they represent, voicing being binary, and manner quaternary. Thus, formally there should be a much greater number of manner contrasts than voicing contrasts, but as we have seen in the minimal pair results, among others, this formal expectation from the inventory does not match the distribution of contrasts in the lexicon. The predominance of plosive contrasts in English significantly skews the roles of both place and voicing, a result which will also have implications for the cue perturbation results in Section 5.4.

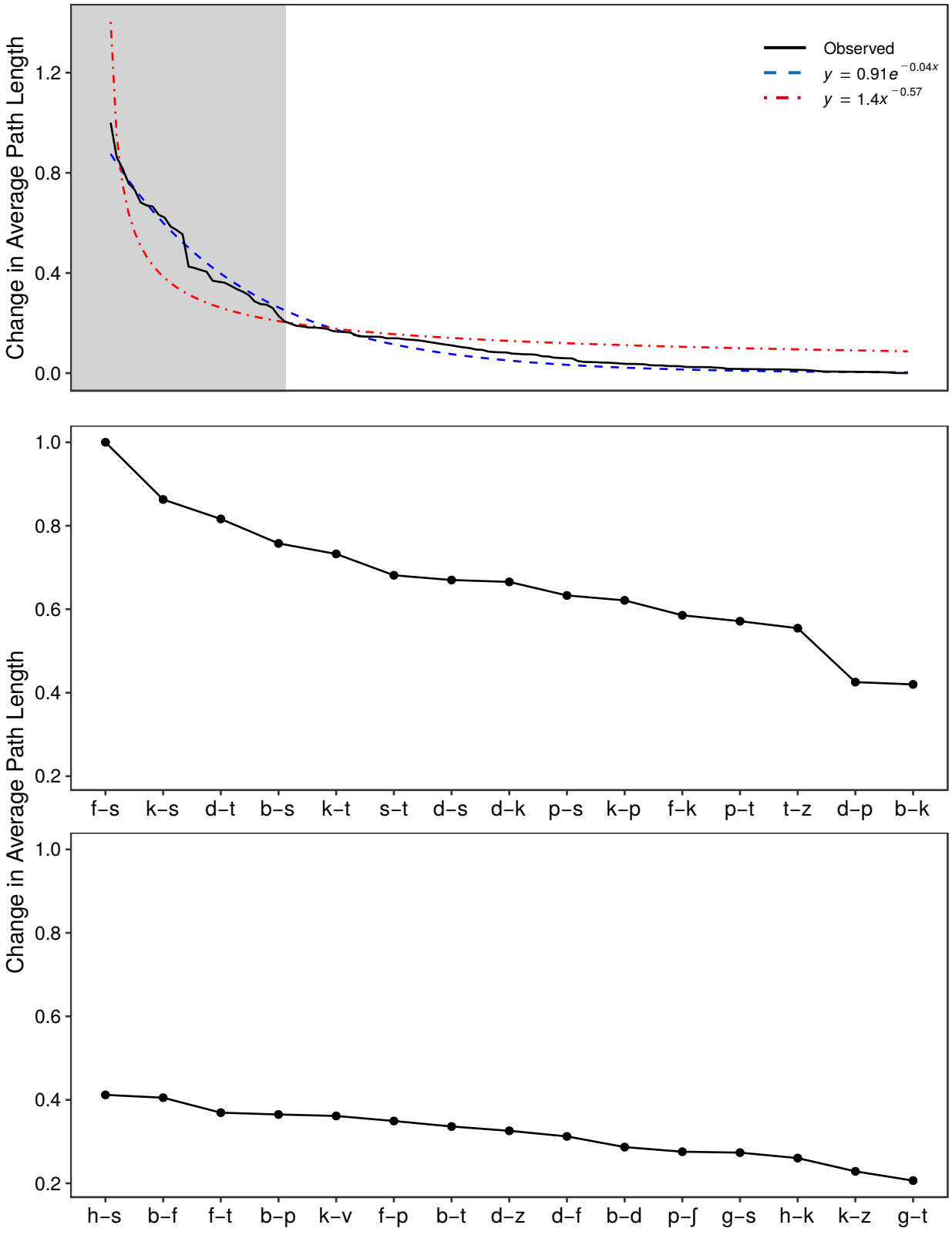

Figure 5.10: Changes in average path length in the unweighted network by obstruent contrast. The top panel shows the full distribution. The bottom panels show the top 30 contrasts (the shaded region in the top panel), which comprise approximately 68% of the cumulative distribution of $\Delta\bar{\ell}$ among obstruent contrasts.

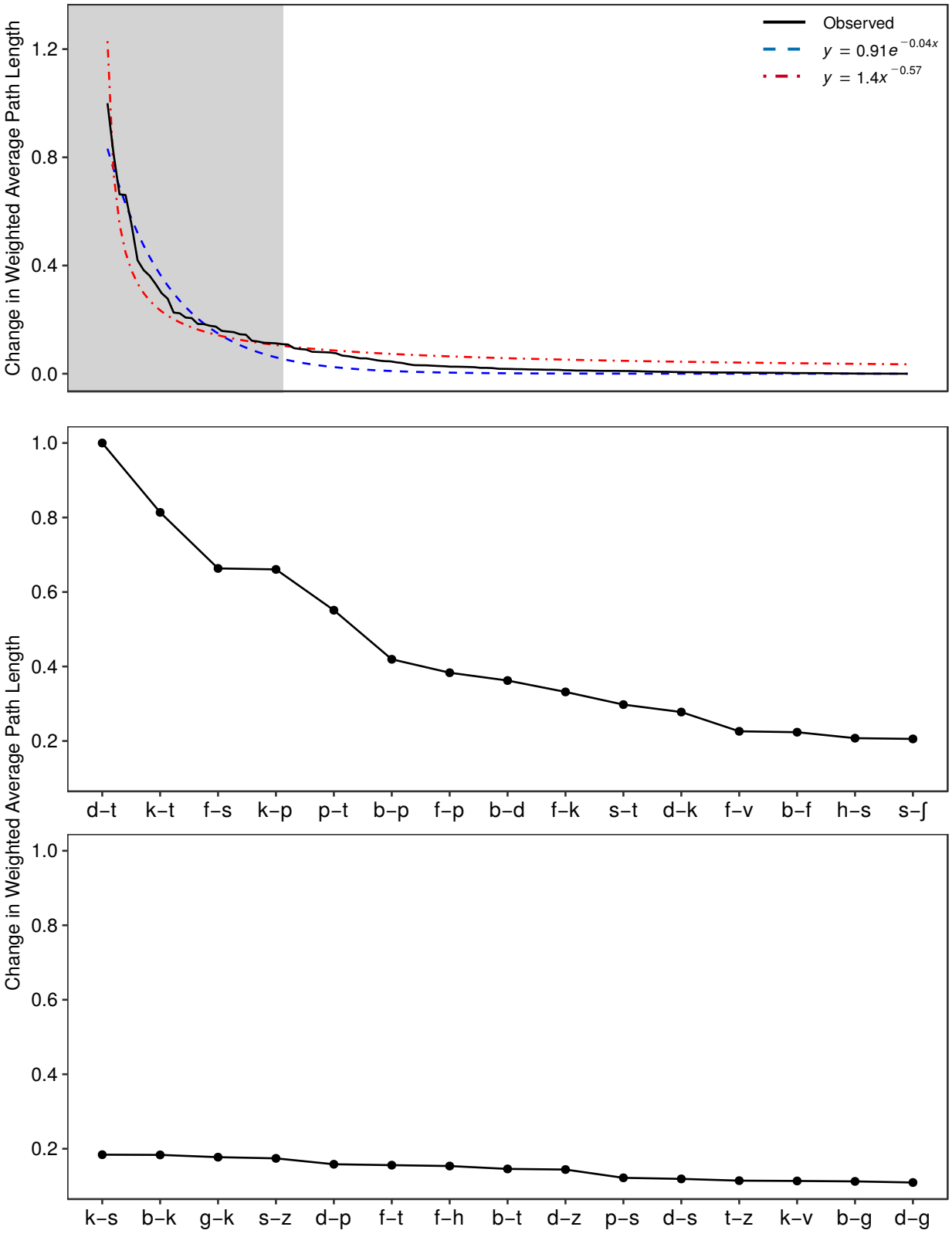

Figure 5.11: Changes in average path length in the weighted network by obstruent contrast. The top panel shows the full distribution. The bottom panels show the top 30 contrasts (the shaded region in the top panel), which comprise approximately 80% of the cumulative distribution of $\Delta\bar{\ell}$ among obstruent contrasts.

	Voicing	Manner	Place	Sibilance
unweighted	1.228	1.311	1.352	1.162
weighted	0.114	0.110	0.211	0.074

Table 5.8: Changes in average path length in the unweighted and weighted networks by featural contrast.

5.2.3 Discussion

The phonological analysis in this section has described several measures of the configuration of obstruent contrasts in the lexicon, where a particular emphasis has been placed on the relationships between such contrasts in addition to raw counts of their type frequency in English. Across all four measurements several consistent patterns were obtained. First, the plosives—particularly the voiceless series—and the voiceless anterior fricatives [f, s], play a prominent role in the English lexicon regardless of the measurement used. Well over half of the cumulative distributions of minimal pairs, functional load, edge disjoint degree/strength, and average path length, involve these phones, and when broken down by contrast the vast majority of critical contrasts occur within this set. The affricates, voiced fricatives, and dental/glottal fricatives, by comparison, are relatively marginal in the lexicon. Further, this relation becomes even more stark when item frequencies are taken into account, as they are in the measurement of functional load. This amplification of asymmetries in contrast distributions occurs not just for frequency-based measures like functional load, but appears in any weighted measurement that accounts for gradient featural similarity in sets beyond the minimal pair, such as in the weighted network measures of *edge disjoint strength* and *weighted average path length*. Each of these contrast distributions, along with functional load, follows a power-law decline rather than an exponential one, suggesting that the phonological system in the lexicon is highly reliant on a narrow set of contrasts. Therefore, in the analysis of noise and cue perturbation in the sections below, we expect this phonological baseline to define the general impact of signal uncertainty on the lexical system as a whole, both in terms of which contrasts are more or less susceptible to perturbation by noise, and which cues are most and least critical to distinguishing contrasts in this core set.

5.3 Noise perturbation

In this section we examine the impact of background noise in the auditory stimulus on the separation of obstruent contrasts in the lexicon. Since we do not have direct access to listener responses to all obstruent-contrastive minimal pairs in the Lex95 database (due to constraints on item repetition in word-recognition experiments) the relative discriminability of such items in the model lexicon was estimated from the listener data in Experiment 1. More precisely, a linear regression predicting probit-transformed accuracy from Target and Competitor obstruents (C_T, C_C), contrast Position (CV, VCV, VC), Noise Level (NL; $+2$ dB SNR, -2 dB SNR), Word Length (WL; mono-, di-, tri-syllabic), Absolute Target Frequency (TF), and Relative Target Frequency (RF) was run on the Experiment 1 data. The specific form of the model (in R syntax), including interaction terms, was the following:

$$probit(y) \sim C_T * C_C * \text{Position} * \text{NL} + \text{WL} + \text{TF} * \text{RF} ,$$

where predicted accuracy was then computed from the Gaussian cumulative distribution function as in a probit regression where $\Pr(y = 1|\mathbf{X}) = \Phi(\mathbf{X}^T\beta)$.[3] The specific form of the predictor set in the model, both the variables included and their interactions, was chosen based on the factors found to be most influential in predicting listener recognition in Experiment 1, though more complex models with higher-order interactions were explored and confirmed to yield no significant improvements in model fit. Overall, this model provides a close fit to the data ($R^2 = 0.49$), which crucially provides contrast accuracy predictions whose rank distribution correlates highly with the observed values, both overall ($\rho = 0.991,\ p < 0.001$) and by Position ($\rho = 0.993,\ p < 0.001$), Noise Level ($\rho = 0.996$, $p < 0.001$), and Word Length ($\rho = 0.885$, $p < 0.001$). Model predictions were then projected onto the Lex95 database and used in the measurement of *minimal pair count*, *functional load*, *edge disjoint strength*, and *average path length* in the sections below.

[3] A linear regression on probit-transformed responses was chosen over a probit regression both for estimation and stability purposes, given the large number of observations and parameters in the model.

5.3.1 Minimal pair count

In order to extend the measurement of minimal pair counts to a gradient measure that incorporates recognition probabilities from the model described above, we define minimal pair count (MP) as the sum of minimal pair recognition probabilities over the lexicon, where non-obstruent-contrastive minimal pairs are given a constant recognition probability of 1 given that such contrasts cannot be estimated from the Experiment 1 data.[4] That is, $MP = \frac{1}{2}\sum_{xy \in E} p_{x \to y} + p_{y \to x}$, where $p_{x \to y}$ is the probability contrast xy is distinguished in listener perception when x serves as the stimulus (and vice versa for $p_{y \to x}$), and E is the set of all minimal pairs in the lexicon. Here the initial sum is divided by two to obtain an average, non-directional, minimal pair count.

Overall, the predicted impact of signal perturbation by background noise on the number of minimal pairs comprising the obstruent system in the lexicon was an 18% reduction in expected minimal pair count: from 2501 to 2055.5, consistent with the designed error rates in Experiment 1. Further, the effect of a 4 dB reduction in SNR was a reduction in predicted minimal pair counts of approximately 11% (from 2176.0 to 1935.1 minimal pairs), again consistent with the general effect of SNR on listener recognition in Experiment 1 (see 3.3.2 for details). These results are not trivial, because the Lex95 data is largely distinct from the stimulus set in Experiment 1, and also includes words of a much higher baseline frequency (again, based on its design as a lexical *core* representing only those words which comprise most of everyday English usage), which could have biased the predictions toward higher accuracies, though this outcome was not obtained.

When broken down by feature class, we see in Table 5.9 that overall, the reduction in minimal pair counts due to noise perturbation is fairly even: ranging between 20 and 26%. The most robust class are the sibilants, at a 20% reduction from the baseline phonological counts to the expected counts at -2 dB, while flaps and affricates are the least robust at 26 and 25% reductions, respectively. Within each feature the relation between classes is consistent with the results of Experiment 1; namely, voiceless obstruents are more robust than voiced obstruents (as well as

[4]This assignment of 1 to non-estimated contrasts is necessary for the formalization of the measure, though it has no practical impact on the analysis below, as we will only be summarizing the impact of noise perturbation on the obstruent system.

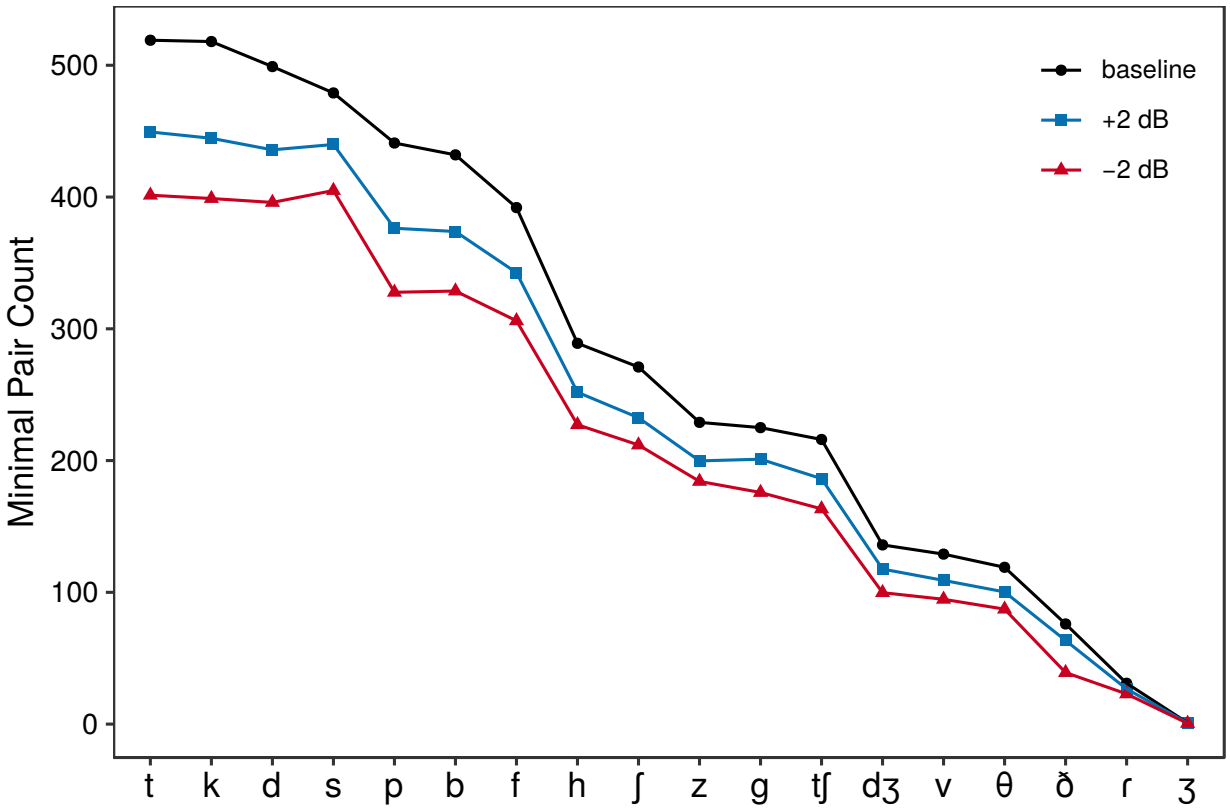

Figure 5.12: Noise perturbation effects on minimal pair counts among obstruent contrasts, aggregated across each constituent phone. The *baseline* counts refer to the minimal pair count in the absence of noise, whereas the +2 and −2 dB counts refer to the minimal pair count which has been weighted by predicted recognition probabilities at each SNR in Experiment 1.

being more frequent members of minimal pairs in general); fricatives are more robust than plosives, but plosives comprise a larger number of contrasts (by a 4:3 ratio); [LOW] coronals are the most frequent place of articulation and are also the most resistant to perturbation by background noise (21%), while labials are the next most frequent class but by comparison are relatively vulnerable to noise perturbation (24%); and finally, sibilants are less frequent overall than nonsibilants but are notably more robust (a 20% reduction from baseline to −2 dB, as compared with 24% for nonsibilants). Thus, the effect of noise perturbation does not re-rank feature classes in terms of their overall role in minimal-pair contrasts in the lexicon; in fact, most such relations are enhanced at −2 dB. That is, the greatest reductions in minimal pair counts tend to occur in the less frequent classes, the most notable exception being the *sibilant–nonsibilant* relation.

Figure 5.12 shows the distribution of minimal pair counts by contrast under simulated noise

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
Baseline	3244	1758	2634	1985	352	31	1394	1952	624	743	289	1332	3670
+2 dB	2824	1528	2281	1740	304	27	1202	1715	537	646	252	1177	3175
−2 dB	2529	1342	2028	1556	263	23	1057	1536	475	575	227	1065	2806
Reduction	22%	24%	23%	22%	25%	26%	24%	21%	24%	23%	21%	20%	24%

Table 5.9: Noise perturbation effects on minimal pair counts by feature class.

perturbation. In general, the contrast distributions at +2 and −2 dB are similar, though a closer examination of the most frequent contrasts reveals several key distinctions between contrasts in terms of both their susceptibility to noise and their relation in perception to the baseline phonological distribution of minimal pairs in the lexicon. The sibilance contrasts, for example, exhibit notably smaller reductions in expected minimal pair counts at −2 dB relative to +2 dB, while within-voicing distinctions among plosives, as well as contrasts involving labial plosives more broadly, show a much greater impact of both decreases in SNR, and the comparison between perception in noise and the baseline phonological distribution in the absence of perceptual information.

	Voicing	Manner	Place	Sibilance
Baseline	604	773	942	550
+2 dB	554	676	821	490
−2 dB	473	600	724	444
Reduction	15%	11%	12%	9%

Table 5.10: Noise perturbation effects on minimal pair counts by featural contrast.

Table 5.10 further summarizes these results by featural contrast, confirming that sibilance contrasts are indeed the most robust to noise perturbation, followed by manner, place, and voicing. Further, this robustness means that sibilance contrasts play a relatively greater role in the obstruent system at lower SNRs, while voicing contrasts are the most likely to be the source of misperceptions when communicating in the presence of background noise.

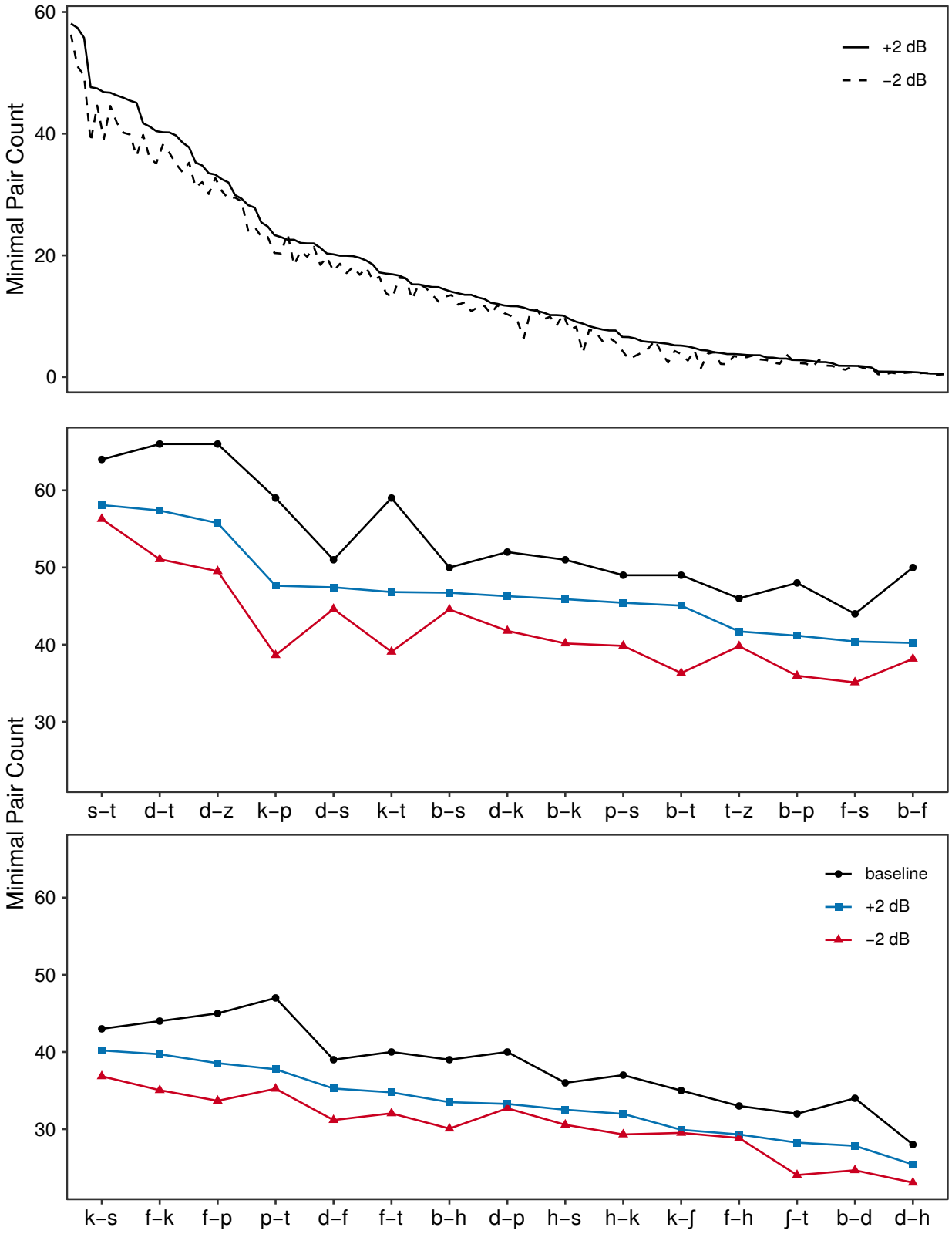

Figure 5.13: Noise perturbation effects on minimal pair counts among obstruent contrasts. The upper panel shows the full distribution. The lower panels show the top 30 contrasts at +2 dB (55% of the cumulative distribution), with minimal pair counts at −2 dB and in the phonological baseline matched for comparison.

5.3.2 Functional load

In analyzing the impact of noise perturbation on the functional load of obstruent phones and contrasts in the lexicon, we measure the gradient impact of contrast mergers by multiplying the functional load of a given contrast by the error probability from the listener recognition model. That is, borrowing the single-phone equation from Surendran & Niyogi (2003), $FL(x) = \sum_y P(x,y)FL(x,y)$, instead of setting $P(x,y)$ to vary as a function of featural similarity, $P(x,y)$ equals the predicted error rate of listeners on a given contrast, with the functional load on a single phone then measured by taking the sum of the resulting weighted FL measure over the set of contrasts that phone occurs in. Figure 5.14 shows the functional load of obstruent phones, both in a baseline model where all contrasts are equally likely to merge, and in a weighted model where the functional load of each contrast is multiplied by its predicted error rate at $+2$ and -2 dB.

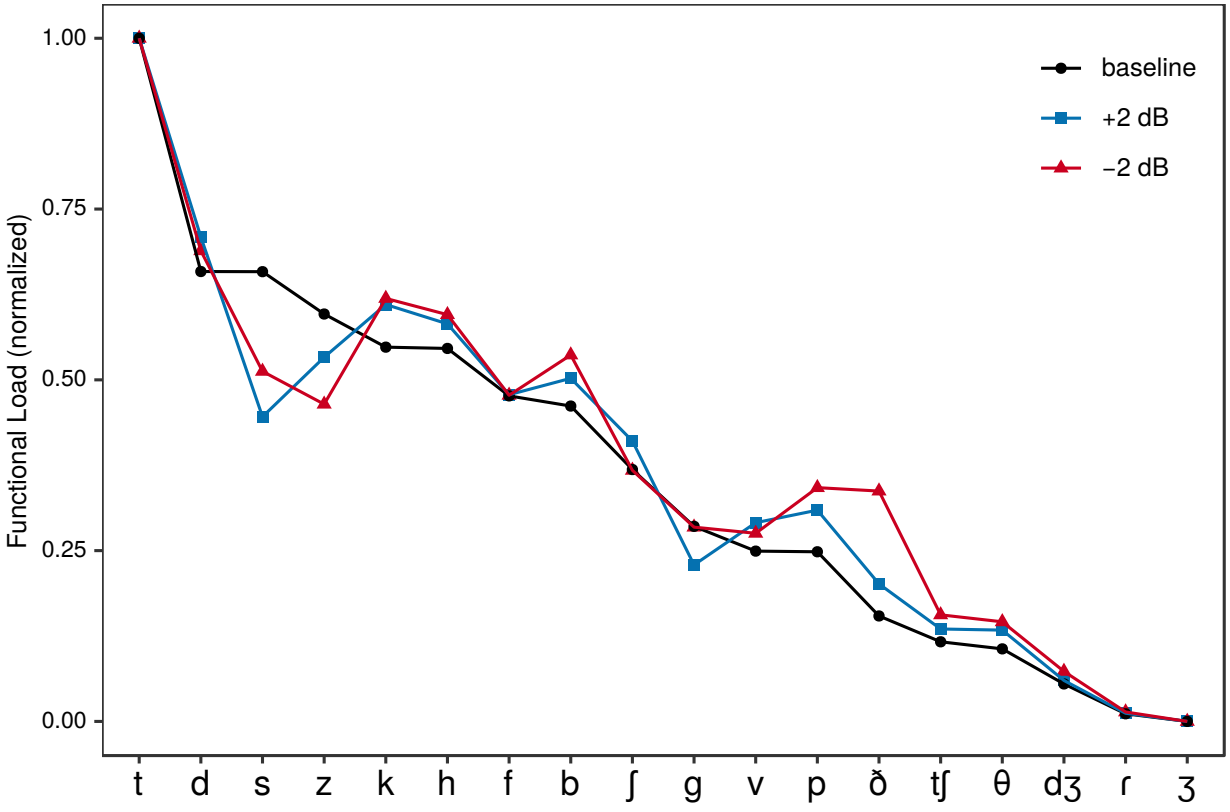

Figure 5.14: Noise perturbation effects on the functional load of obstruent phones. The *baseline FL* refers to the measurement of functional load where all contrasts are assumed to be equally likely to merge. The other two lines show functional load as measured from predicted error probabilities in Experiment 1.

One way to interpret this figure is consider higher weighted functional loads to represent the relative importance of the acoustic information underlying a given contrast, because higher ranked phones according to this measure occur in contrasts that are both highly frequent and easily confusable. However, one could also interpret the difference between the baseline and weighted functional loads as representing the relative reliability of a given phone in perception, where [s] and [z], for example, exhibit high phonological functional loads—i.e., they occur in many contrasts in the lexicon whose merger would result in a considerable loss in information—but when weighted by the error rate on such contrasts their functional load drops notably. This means that the sibilants [s, z] play a greater role in the obstruent system in English relative to [t, d, k, h, b], all of which show a higher potential for information loss when perceptual constraints are considered.

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
Baseline	445	270	350	345	19	1	157	348	59	91	60	196	519
$+2$ dB	27	17	22	20	1	0	10	20	4	6	4	10	33
-2 dB	46	29	37	34	3	0	18	34	7	10	6	17	57

Table 5.11: Aggregate functional load of obstruent feature classes in the baseline model and in models weighted by listener error probabilities at $+2$ and -2 dB SNR.

When broken down by feature classes, voiceless obstruents exhibit consistently higher functional loads than voiced obstruents both in the baseline model, and at $+2$ and -2 dB, with no apparent expansion in this difference as the relative likelihood of listener error is incorporated into the weighted models. Manner of articulation is similar in this regard, as the *flap* $<$ *affricate* $<$ *fricative* $<$ *plosive* relation is relatively stable across weighted and unweighted measures. Regarding place of articulation, [LOW] coronals remain the highest-ranked in functional load, followed by labials, dorsals, and glottals/[HIGH] coronals. As with the other features, there are no sizeable changes in these relations when relative perceptibility in noise is accounted for. Finally, the *sibilant* $<$ *nonsibilant* relation is consistent across models. Thus, while the effect of noise perturbation on the functional load of obstruent phones is informative to some extent, when aggregated across larger classes these differences tend to wash out and largely reflect the baseline estimates.

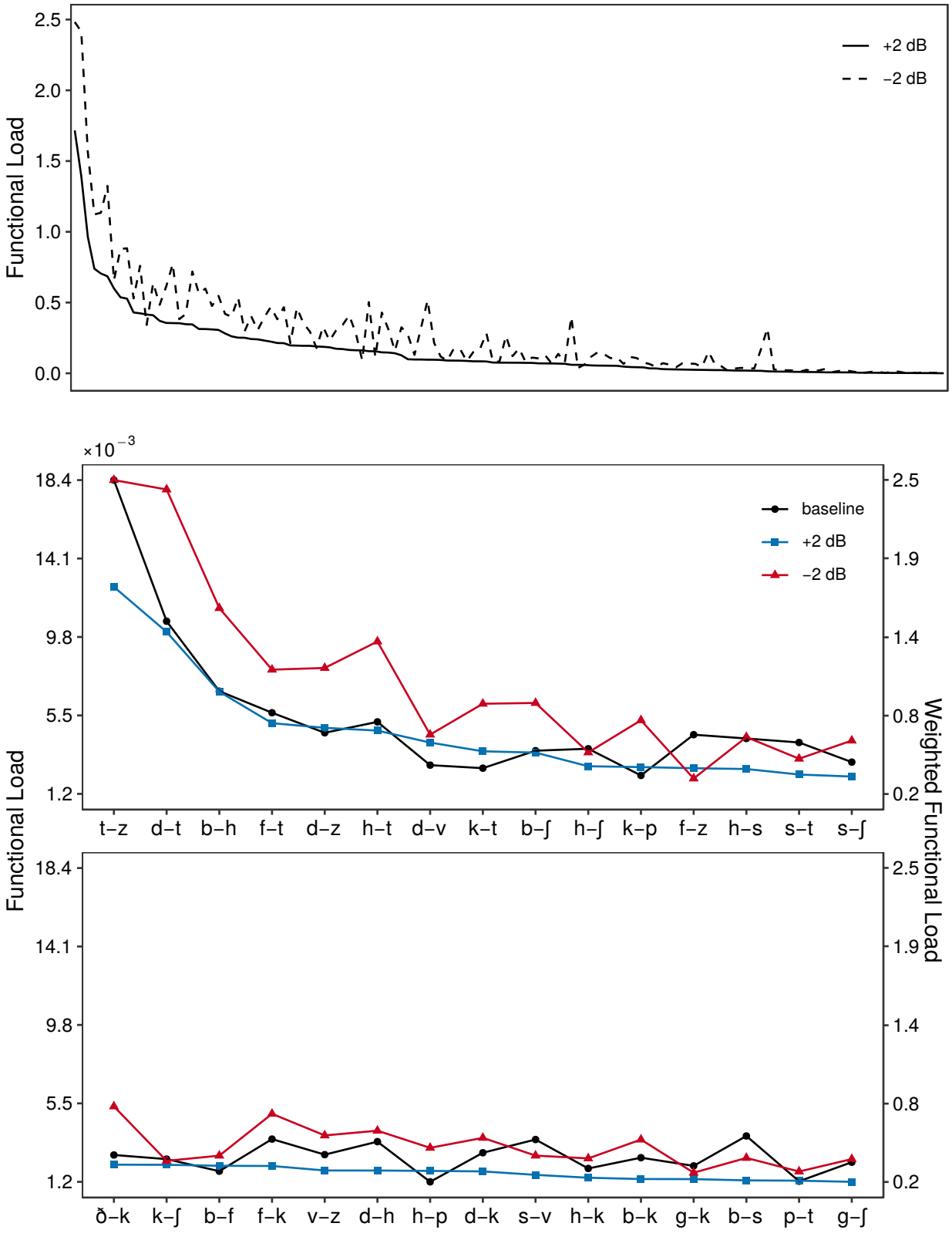

Figure 5.15: Noise perturbation effects on the functional load of obstruent contrasts. The upper panel shows the full distribution. The lower panels show the top 30 contrasts at +2 dB (68% of the cumulative distribution), with functional load at −2 dB and in the phonological baseline matched for comparison.

Figure 5.15 shows the distribution of functional load by contrast and noise level. Overall, the distributions at $+2$ and -2 dB show similar forms, both declining at a power-law rate, though there is variability along the distribution where FL is relatively higher or lower for a given contrast at one SNR versus the other. The lower panels of Figure 5.15 clarify the upper-end of the range, which accounts for 68% of the cumulative distribution. Focusing in on the contrasts near the inflection point (*t–z*, *d–t*, *b–h*, *f–t*, *d–z*, *h–t*), in the phonological baseline the first two contrasts, *t–z* and *d–t*, dominate the distribution, and this relationship remains in general at $+2$ dB; however, at -2 dB the perceptual similarity of *d–t* causes a substantial rise in functional load, indicating that the voicing contrast among coronal plosives not only has a high phonological functional load in the lexicon, but when weighted by listener error rates this contrast is the most likely to cause a significant loss in information when communicating in the presence of background noise. Excepting *h-t*, which is notably more vulnerable at -2 dB than the other contrasts in this set, the relative role of the remainder of the set at -2 dB is fairly consistent with the other two measures. Table 5.12 further summarizes these results by featural contrast, and indicates that overall, just as in the aggregation of functional load across feature classes in Table 5.11, the *sibilance* $<$ *voicing* $<$ *manner* $<$ *place* relation is stable across weighted and unweighted measures. It is only in specific subsets of these larger contrast sets where the impact of noise perturbation on changes in functional load emerges.

	Voicing	Manner	Place	Sibilance
Baseline	202.6	224.3	250.4	174.3
$+2$ dB	11.1	13.3	15.4	8.9
-2 dB	20.1	23.1	27.0	14.5

Table 5.12: Noise perturbation effects on functional load by featural contrast.

5.3.3 Edge disjoint strength

The analysis of changes in edge disjoint strength, $s_{\ominus}(xy)$, under noise perturbation is a straightforward extension of the feature-based edge weighting to one that is based on predicted error rates in listener perception. Figure 5.16 shows the aggregate edge disjoint strength by obstruent phone and

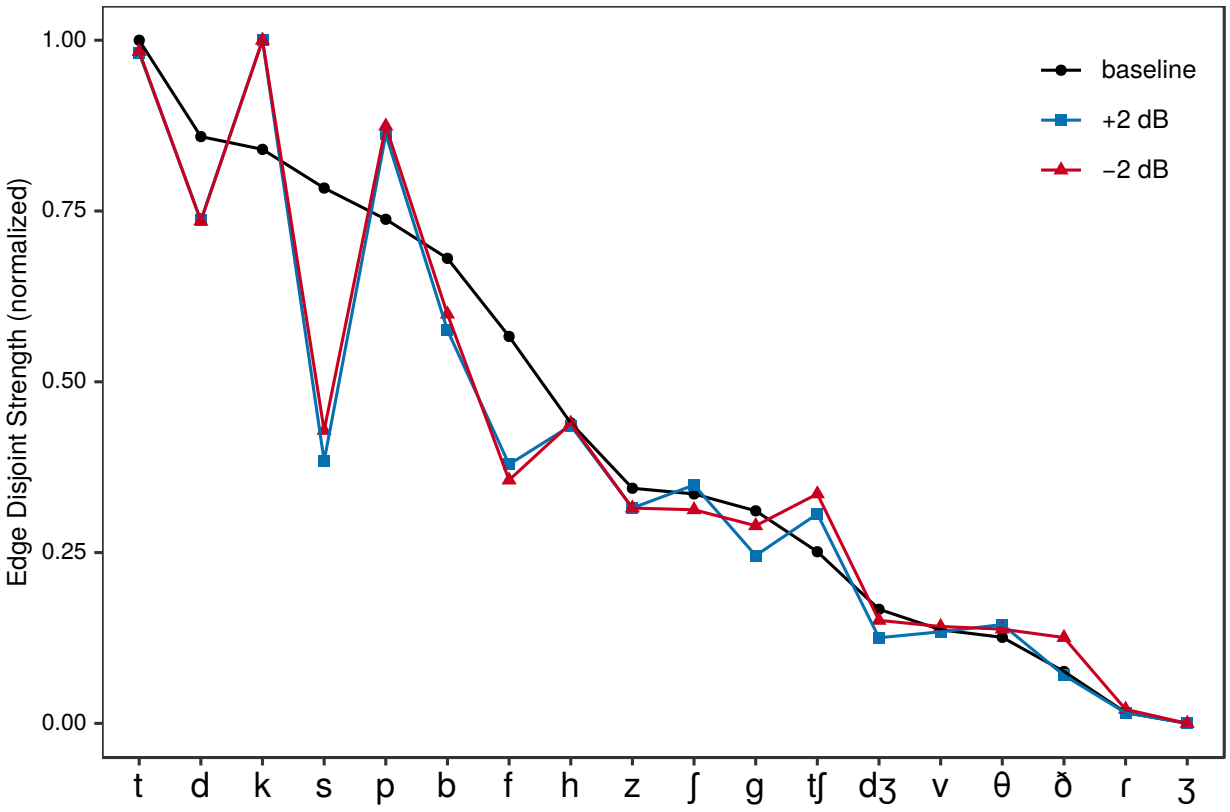

Figure 5.16: Noise perturbation effects on total edge disjoint strength of contrasts involving each obstruent phone.

$+2$ and -2 dB. Here we see that much of the potential change in lexical neighborhood similarities involves the plosives [p, t, k, b, d], while the voiceless anterior fricatives [s, f] drop in $s_{\ominus}(xy)$ relative to their phonological role in the edge disjoint degree results in Figure 5.6. This distinction is present at both SNRs, but is considerably greater at -2 dB (not shown in the figure as values were normalized), where the decreased salience of plosives is predicted to result in large increases in local wordform similarity in the lexicon.

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
$+2$ dB	158	72	143	72	14	1	64	86	25	41	14	48	182
-2 dB	374	183	344	174	37	2	151	211	61	99	34	118	438

Table 5.13: Aggregate edge disjoint strengths of obstruent feature classes when the contrasts they occur in are perturbed by background noise at $+2$ and -2 dB SNR.

When broken down by feature class, the effect of noise perturbation of contrasts involving voiceless obstruents is much greater than that observed for voiced obstruents, though the ratio between the two remains relatively constant across SNRs. The relations among manner, place, and sibilance contrasts are similar in this regard, suggesting that while increases in noise expand the impact of different feature classes on local similarity in the phonological network, this effect is relatively constant, with asymmetries in the impact of noise perturbation primarily confined to specific phones such as the ones mentioned above.

Figure 5.17 shows the impact of noise perturbation on the edge disjoint strength of specific obstruent contrasts. As in the phonological analysis, both distributions exhibit a power-law decline, with the top 30 contrasts comprising over 65% of the cumulative $s_{\ominus}(xy)$ distribution. Within this set, the majority of high-$s_{\ominus}(xy)$ contrasts occur between plosives, particularly voiceless plosives, which comprise the top 3 contrasts at $+2$ dB, and the top 4 at -2 dB (the alveolar voicing contrast, *d–t*, is relatively higher in edge disjoint strength than *p–t* at the lower SNR).

	Voicing	Manner	Place	Sibilance
$+2$ dB	50	63	87	38
-2 dB	126	155	217	94

Table 5.14: Noise perturbation effects on edge disjoint strength by featural contrast.

These results are further summarized in Table 5.14, which confirms that place contrasts—primarily under the influence of voiceless plosive distinctions—bear the greatest responsibility for local increases in wordform similarity due to perturbation of the acoustic signal by background noise. Following place is manner, and then voicing, with sibilance the most robust given that such contrasts are frequent in the lexicon but exhibit below-average error rates in perception.

5.3.4 Average path length

Finally, we examine the global impact of noise perturbation on the separation of items in the lexicon by measuring changes in average path length, $\bar{\ell}$, at $+2$ and -2 dB SNR. Figure 5.18 shows the aggregate change in average path length across all contrasts involving a given obstruent phone,

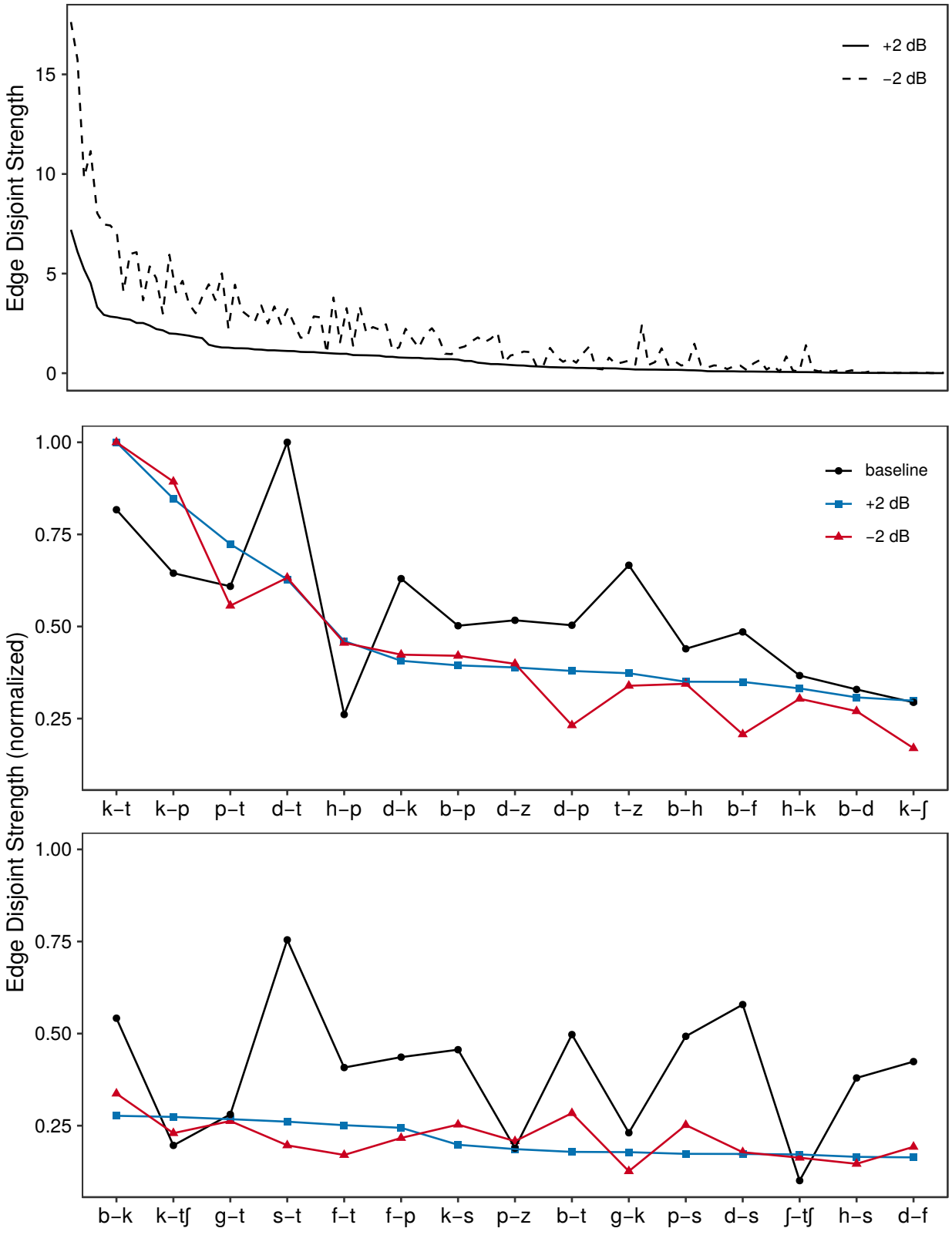

Figure 5.17: Noise perturbation effects on the edge disjoint strength of obstruent contrasts. The upper panel shows the full distribution. The lower panels show the top 30 contrasts at +2 dB (65% of the cumulative distribution), with baseline $k_{\ominus}(xy)$ and $s_{\ominus}(xy)$ at −2 dB matched for comparison.

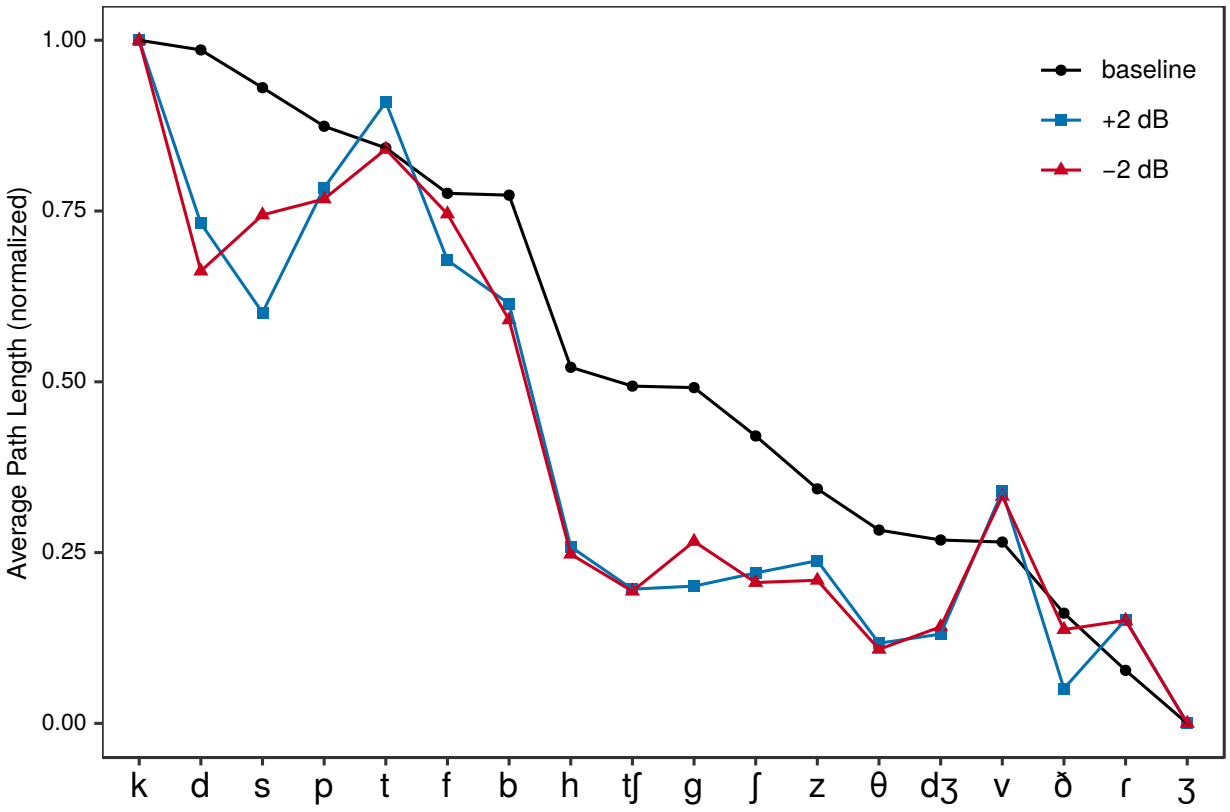

Figure 5.18: Noise perturbation effects on average path length among obstruent contrasts, aggregated across each constituent phone. The *baseline* counts refer to $\bar{\ell}$ in the absence of noise, whereas the $+2$ and -2 dB lines refer to $\bar{\ell}$ which has been weighted by predicted recognition probabilities at each SNR in Experiment 1. All values have been normalized within a given model to range between 0 and 1.

where the greatest distinction that emerges in the noise-perturbed model is consistent with the baseline phonological distinction between the set [p, t, k, b, d, s, f] and the remainder—largely voiced fricatives and postalveolars. And while there is some variation in the relative degree of change in this latter set in terms of the relative role of different phones in contributing to increases or decreases in average path length (e.g., [v] is relatively more vulnerable to perturbation than the other obstruents in this set), the most notable effects occur in the upper $\Delta\bar{\ell}$ range. Here we see that at $+2$ dB, the voiceless plosives [p, t, k] play the greatest role in weakening the global system of contrast in the lexicon, with [s] much lower in this regard; however, at -2 dB, the role of [s] elevates above that of [b] and [d].

This result was unexpected given the perceptual salience of [s], but it is understandable when we consider the kinds of neighborhoods plosives predominantly occur in. The edge disjoint

	Voicing		Manner				Place					Sibilance	
	vl.	vd.	plos.	fric.	affr.	flap	lab.	cor. (L)	cor. (H)	dor.	glot.	sib.	nsib.
+2 dB	0.13	0.08	0.13	0.08	0.01	0.01	0.08	0.09	0.02	0.04	0.01	0.05	0.15
−2 dB	0.24	0.14	0.22	0.15	0.02	0.01	0.14	0.17	0.04	0.08	0.02	0.09	0.27

Table 5.15: Aggregate changes in average path length under noise perturbation by obstruent feature class.

strength results in the previous section revealed that many of the most frequent (and least salient) distinctions in the lexicon are between plosives (consider, for instance, the dense monosyllabic set {*cap*, *tap*, *tack*, *pack*, *back*, *bag*, ...}). What then happens in cases of high noise perturbation (at −2 dB) is that there are many alternative routes through the network that need not include a given contrast. While [s] and [f] also occur in such neighborhoods, they tend to be more widespread in the lexicon (relative to their own distributions), meaning reductions in their perceptual salience will have a greater impact on the global system of lexical distinctions despite being relaively less impactful in local phonological neighborhoods.

When broken down by feature class the same general patterns of *voiceless* > *voiced*, *plosive* > *fricative* > *affricate* > *flap*, *[LOW] coronal* > *labial* > *dorsal* > *[HIGH] coronal* > *glottal*, and *nonsibilant* > *sibilant* are obtained. The effect of an increase in noise level from +2 to −2 dB is also relatively constant across each feature, suggesting that while changes in the distribution of global lexical similarity under noise perturbation may have variable effects on specific phones and contrasts, in aggregate the relative role of different featural distinctions remains constant. This suggests that there are likely other factors such as auditory and articulatory stability that govern the productivity of different features in English word formation.

Figure 5.19 shows the effect of noise perturbation on specific obstruent contrasts, revealing which distinctions are most responsible for changes in the global separability of items in the lexicon, both overall, and at different noise levels. Here again the set [p, t, k, b, d, s, f] plays a prominent role, though unlike in the more local measures of minimal pair count, functional load, and edge disjoint strength, manner distinctions within this set play a relatively greater role. For instance, many contrasts between the plosive series and [f, s] are weighted highly in the degree

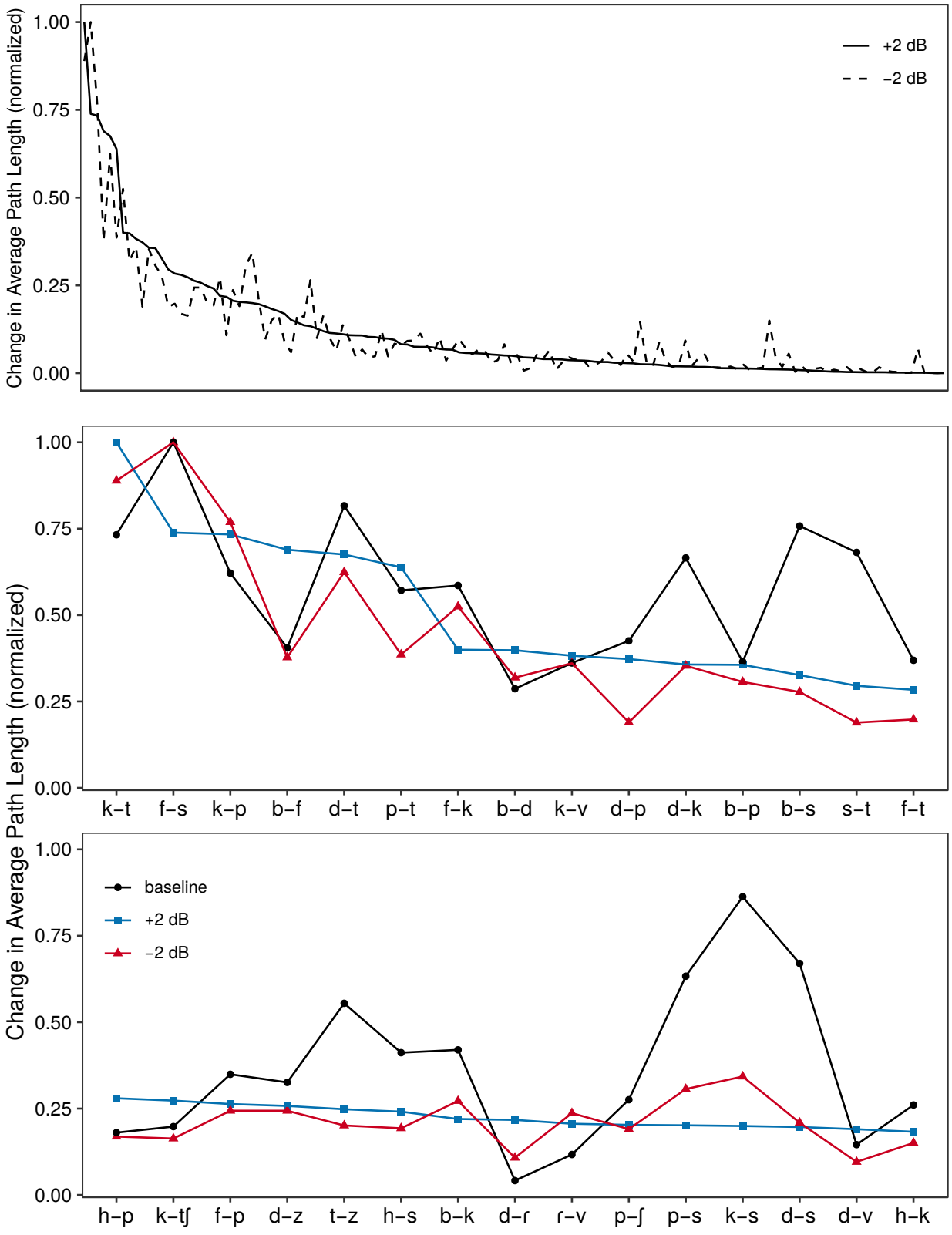

Figure 5.19: Noise perturbation effects average path length by obstruent contrast. The upper panel shows the full distribution. The lower panels show the top 30 contrasts at $+2$ dB (65% of the cumulative distribution), with baseline path lengths and $\bar{\ell}$ at -2 dB matched for comparison.

to which their perturbation by noise impacts the average path length in the network. Contrasts between voiceless plosives are also highly weighted, and behave similarly at +2 and −2 dB. However, it is notable that unlike many of the measures investigated thus far, the impact of noise on the system is less consistent with the overall role of such contrasts at a phonological level. Most notably in Figure 5.19 is the relative robustness of contrasts such as *k–s*, *b–s*, *t–z*, and *d–k*, each of which has a much lower impact on the average path length in the lexicon under noise perturbation than their baseline distribution would suggest. The perceptual robustness of these contrasts is consistent with their greater featural complexity, comprising between 2 and 4 feature differences, but it also reveals something more about the configuration of contrasts in the lexicon. Providing a more central role for such contrasts in distinguishing large subsets of the lexicon makes the system more resilient to perturbation, and motivates further study on how the lexicon might evolve toward more optimal states over time.

	Voicing	Manner	Place	Sibilance
+2 dB	0.05	0.07	0.11	0.04
−2 dB	0.10	0.13	0.20	0.09

Table 5.16: Noise perturbation effects on average path length by featural contrast.

Table 5.16 summarizes the aggregate role of each featural contrast at both noise levels, and retains the general prominence of place distinctions, and the relatively low impact of noise perturbation on sibilance contrasts in the lexicon. The one notable change from the typical patterns observed thus far is the elevated role of manner contrasts. Finally, the effect of increasing the level of noise perturbation does not appear to differentially affect one feature over another, but rather has an additive effect across featural contrasts.

5.3.5 Discussion

In analyzing the relative impact of noise on the system of obstruent distinctions in the lexicon, and the specific phones and contrasts that drive such effects, we arrive at a more general framework for understanding the structure of the system and the phonetic characteristics which underly the

problem of transmitting information over a noisy channel. According to most measures the plosives and the fricatives [f] and [s] play an outsized role, both at a phonological level, and in terms of how decreases in their probability of detection due to background noise affect the broader discriminative potential of the lexicon. The relative ranking of components of this set does vary depending on the measurement used, with more local measures such as minimal pair count, functional load, and edge disjoint strength prioritizing the plosive series and contrasts internal to that set, while more global measures such as average path length prioritize manner distinctions between plosives and fricatives; however, no measure investigated in this section differed in partitioning these core obstruents from the more marginal set composed of voiced fricatives, affricates, dentals/glottals, and the alveolar flap. Thus, in the next section, where perturbation of specific acoustic cues is simulated, we expect the results to reflect this distinction, prioritizing place and voicing cues among plosives, and manner and sibilance cues between [f, s] and [p, t, k, b, d].

5.4 Cue perturbation

Finally, we link the acoustic and perceptual characteristics of obstruent contrasts by studying the impact of cue perturbation on the system of wordform distinctions in the lexicon. The following procedure was adopted for the simulation of acoustic cue perturbation: for each cue in the test set (the Lex95 database), the contrast parameter was set to 0 (e.g., $\Delta\text{FREQ}_{\text{PK}} = 0$), thus simulating a complete loss of information along that acoustic dimension. Mergers in target cue values, however, were not simulated, as such a procedure less directly reflects the use of a given parameter as a cue to obstruent discrimination, and has a tendency rather to capture broad correlations in listener accuracy in Experiment 1, such as the better recognition, on average, of voiceless obstruents relative to voiced obstruents, or fricatives relative to plosives. The lexicon model from Section 4.4 of Chapter 4 was then used to predict listener accuracy on the cue-perturbed data, with changes in minimal pair count, functional load, edge disjoint strength, and average path length after perturbation then used to evaluate the relative role of each cue in the lexicon. Further, this procedure is used as an initial demonstration of how the impact of gradient cue loss/weakening on the structure of

the wider lexical system can be simulated, which we hope will be a starting point for future work predicting the trajectory of historical sound changes that are in progress, as well as reconstructing the mechanism behind sound changes that have already taken place.

5.4.1 Minimal pair count

Beginning with cue perturbation as reflected in changes in expected minimal pair count, in Figure 5.20 we show the relative impact of perturbing each cue on the expected number of obstruent-contrastive minimal pairs in the lexicon. That is, Figure 5.20 plots the difference between weighted minimal pair counts from the initial model predictions of listener accuracy and the recomputed predictions after each cue has been perturbed. Here we see that the most critical cues according to this measure are noise duration (ND), F1 at vowel onset ($F1_{CV}$), relative F3 amplitude (AMP_{F3}), consonantal spectral tilt ($TILT_{C}$), and F2 at vowel onset ($F2_{CV}$), while the least impactful cues are F2 at following vowel midpoint ($F2_{V2}$), low-frequency energy (LF), closure duration (CD), F3 at vowel onset ($F3_{CV}$), and f0 at vowel onset ($f0_{CV}$). These results conform with our initial expectations based on the critical role of plosive contrasts and contrasts with [s] and [f] in the lexicon, as noise duration and F1 provide voicing cues in this set, while consonantal spectral tilt and F2 are robust cues to place of articulation, and relative F3 amplitude cues sibilance distinctions.

Note also that the less informative cues are not necessarily cues that were poor predictors of listener recognition in the cue-integration models in Chapter 4. Some cues in this set could indeed be characterized as such, but other cues are predictive of listener accuracy, but in restricted positions/contrasts, such as low-frequency energy (LF), which is primarily useful intervocalically, making it less prominant in the model lexicon, which represents a frequency-constrained core set of words, than in Experiment 1. This leaves open the question of how the cue weights might have changed in a perception experiment utilizing the Lex95 items as a stimulus set; however, as noted previously, constraints on item repetition make such an experiment infeasible without a much larger participant pool. Finally, there is the question of whether in addition to tracking type distributions in the lexicon, cue weights are attuned to contrasts that are higher in token frequency.

This outcome would certainly be consistent with the paradigm developed in this dissertation, but deserves further research.

Figure 5.21 provides further detail on the relative change in minimal pairs among the 15 most frequent obstruent contrasts (representing one-third of those in the lexicon) due to perturbation of the three most influential cues in Figure 5.20: noise duration, F1 at vowel onset, and relative F3 amplitude. As expected, perturbation of noise duration has the greatest impact on manner distinctions, particularly those which are homorganic in place of articulation: e.g., *b–f*, *s–t*, *d–z*, and *d–s*. Place distinctions among voiceless plosives are the least perturbed by the loss of the ΔND cue. Vowel-onset F1, on the other hand, primarily cues voicing, most notably in the single-feature distinctions of *b–p* and *d–t*, but also in cross-place and cross-manner voicing contrasts. Finally, perturbation of relative F3 amplitude causes a reduction in the discriminability of items contrasting in manner and sibilance, particularly contrasts involving the voiced labial plosive [b], which tends to be higher in AMP_{F3} than most fricatives and plosives.

5.4.2 Functional load

Figure 5.22 shows the change in functional load, aggregated across obstruent contrasts, resulting from a perturbation of each cue. As in the measurement of change in minimal pair counts, the most influential cues are noise duration, consonantal spectral tilt, F1 and F2 at vowel onset, and relative F3 amplitude, while F2 at following vowel midpoint, low-frequency energy, and closure duration do not result in an aggregate increase in functional load when perturbed. In fact, this set shows negative changes in functional load from the baseline model predictions, which as in the measure of minimal pair counts means that the distribution of these cues in the test set runs counter to the model expectations, and thus their presence only confounds predictions of relative contrast discriminability.

Examining the complete distribution of cues, we find the there is remarkable stability in the cue perturbation results between the two measures. This is partly due to their formal similarity, given that we have chosen to measure functional load with respect to words rather than phones (i.e., *T*

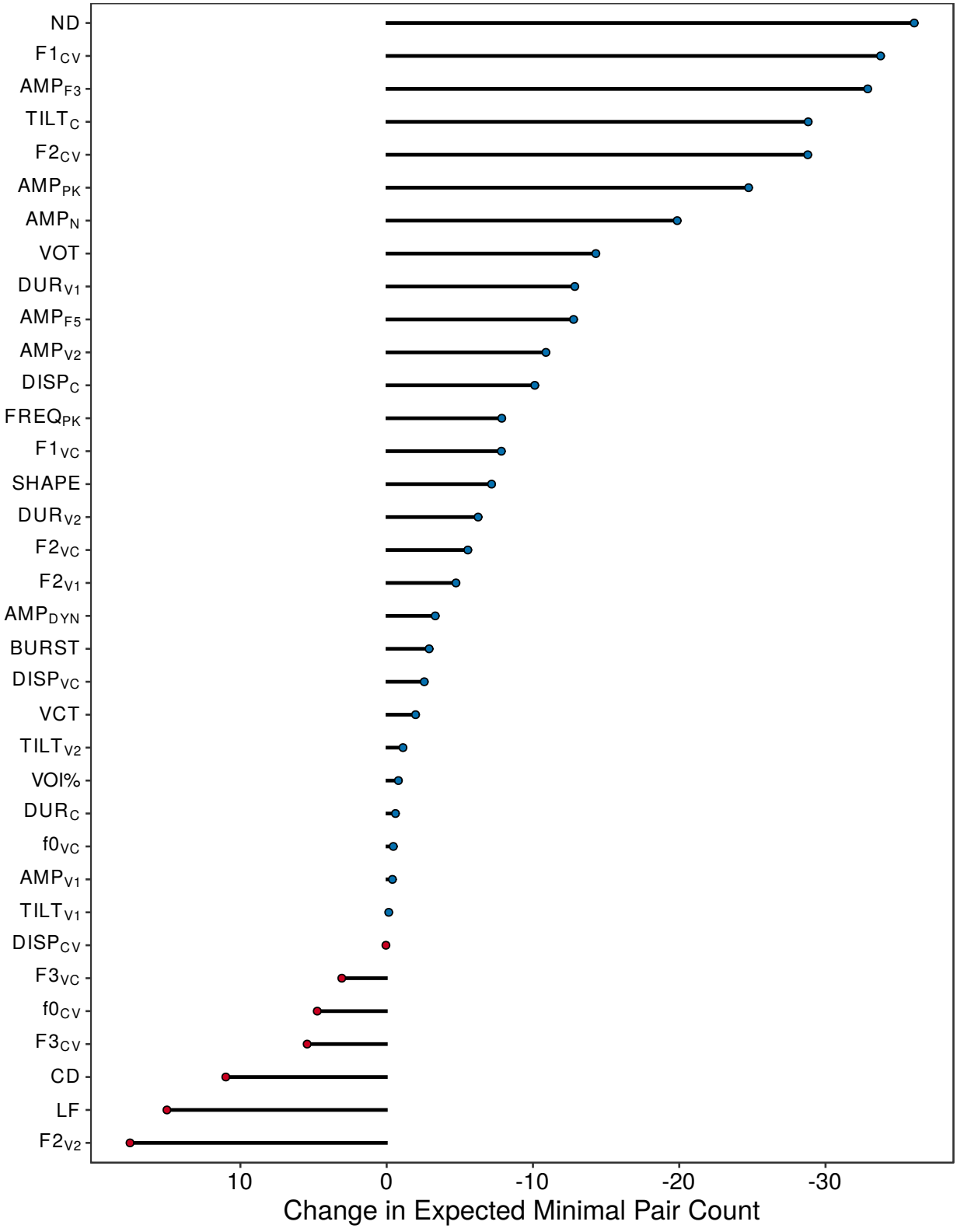

Figure 5.20: Effects of cue perturbation on the expected number of minimal pairs in the lexicon.

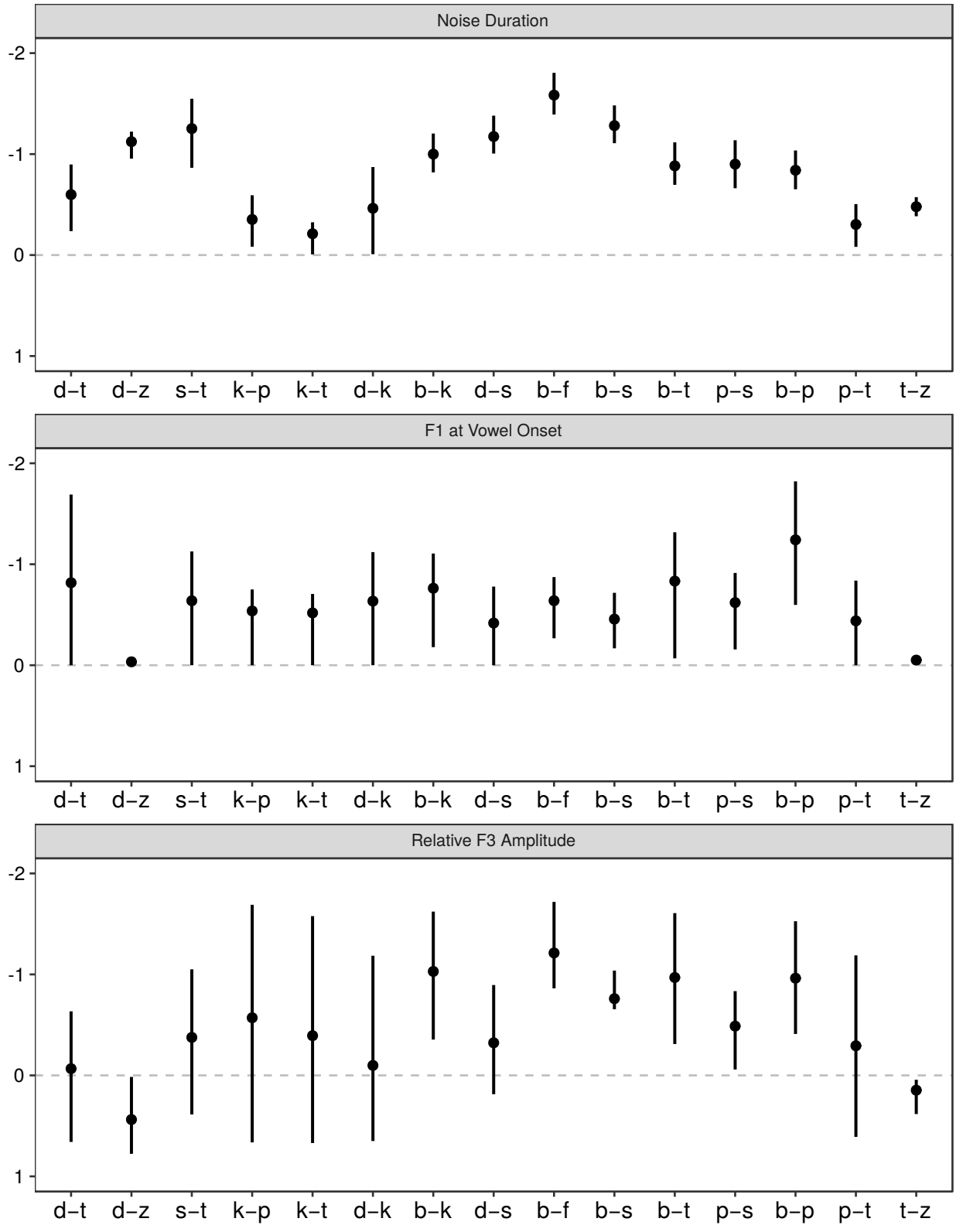

Figure 5.21: Effects of ND, $F1_{CV}$, and AMP_{F3} perturbation on the expected minimal pair count of the 15 most frequent obstruent contrasts in the lexicon. Contrasts are ordered by baseline minimal pair count.

= word in the general FL equation of Surendran & Niyogi, 2003). However, it also reflects the fact that when aggregated across all items in the lexicon, the impact of perturbing a given cue will primarily reflect the distribution of that cue in the full range of contrasts in which it occurs, with effects on particular contrasts and subsets of the lexicon averaged out in the aggregate.

Figure 5.23 provides one window on the differences in specific contrast results that are masked by the aggregate measure. Recall first that the functional load of obstruent contrasts in the lexicon follows a power-law distribution, with a much greater asymmetry in contrast weight than in the approximately exponential minimal pair distribution. Thus the impact of ND perturbation on the manner contrasts *t–z*, *b–h*, and *f–t*, and the voicing contrast *d–t* far outweighs its role in the remainder of the system. This result is featurally consistent with the minimal pair results, but differs in the relative impact of ND perturbation on different contrasts. Consonantal spectral tilt also plays an important role in the system as a sibilance cue, distinguishing *t–z*, *d–z*, and *h–s* in particular, all contrasts with a high functional load in the lexicon. Finally, the voicing (and to a lesser extent manner) cue provided by $F1_{CV}$ makes perturbation of this cue critical for the lexical system because many of the highest-FL contrasts (e.g., *d–t*, *b–h*, *h–t*) differ along these dimensions. This result is similar to that based on minimal pair counts, but with an elevated role of *d–t* as the obstruent contrast with the second-highest functional load, while the impact of $F1_{CV}$ perturbation on *b–p* is less influential in terms of relative functional load.

5.4.3 Edge disjoint strength

Turning now to the first of the network measures of cue perturbation on the system of obstruent contrasts in the lexicon, Figure 5.24 shows that similar to the other *local* measures, the cues whose perturbation is the most influential on the separation of neighborhoods of phonologically similar items in the lexicon include noise duration (ND), F1 at vowel onset ($F1_{CV}$), consonantal spectral tilt ($TILT_{C}$), spectral peak amplitude (AMP_{PK}), and F2 at vowel onset ($F2_{CV}$). Those with little-to-no impact on the system are LF, CD, $F2_{V2}$, VOI%, $F3_{VC/CV}$, and $f0_{CV}$. As noted earlier, edge disjoint strength behaves more similarly to functional load than minimal pair count, which is consistent

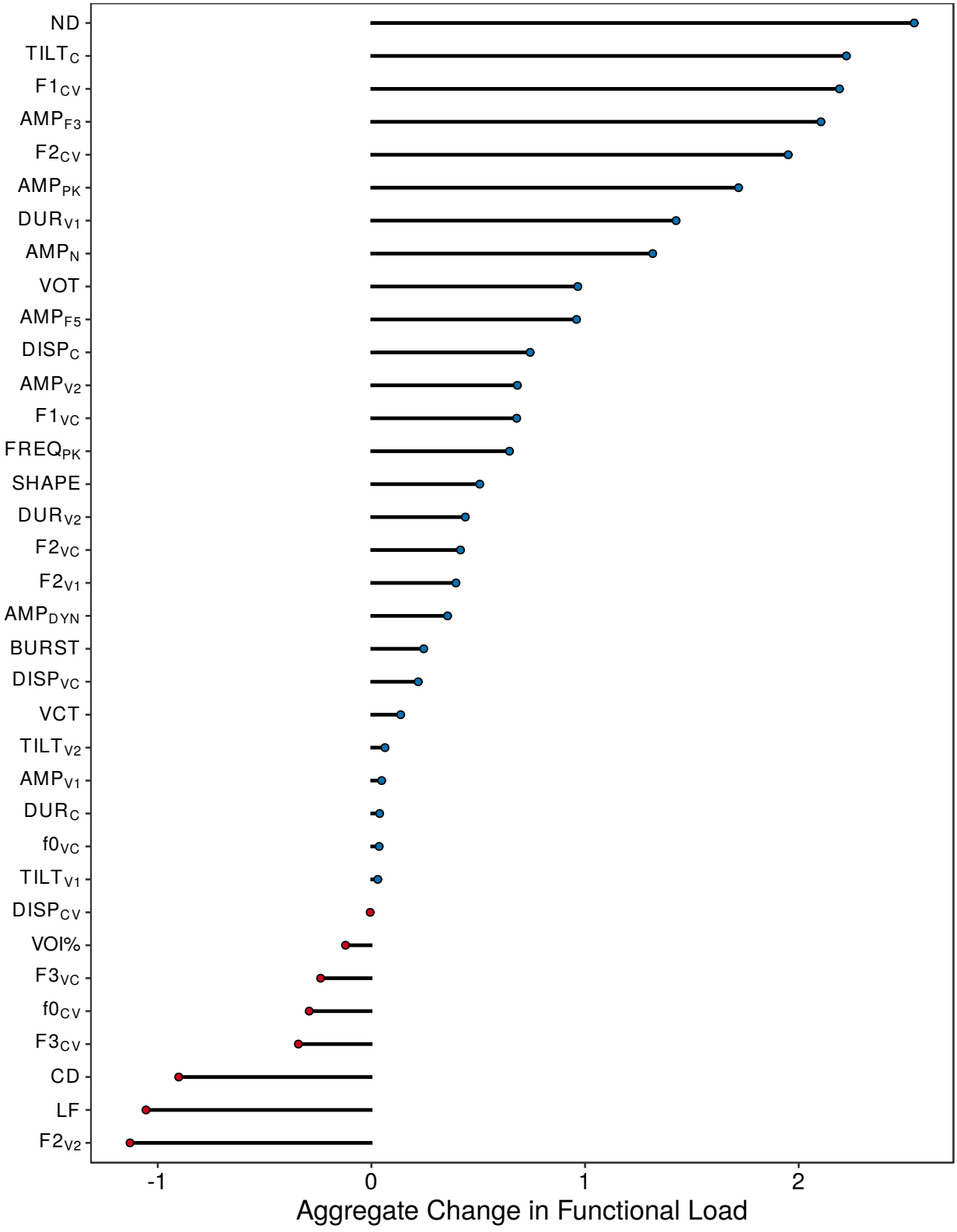

Figure 5.22: Effects of cue perturbation on the aggregate functional load of obstruent contrasts in the lexicon.

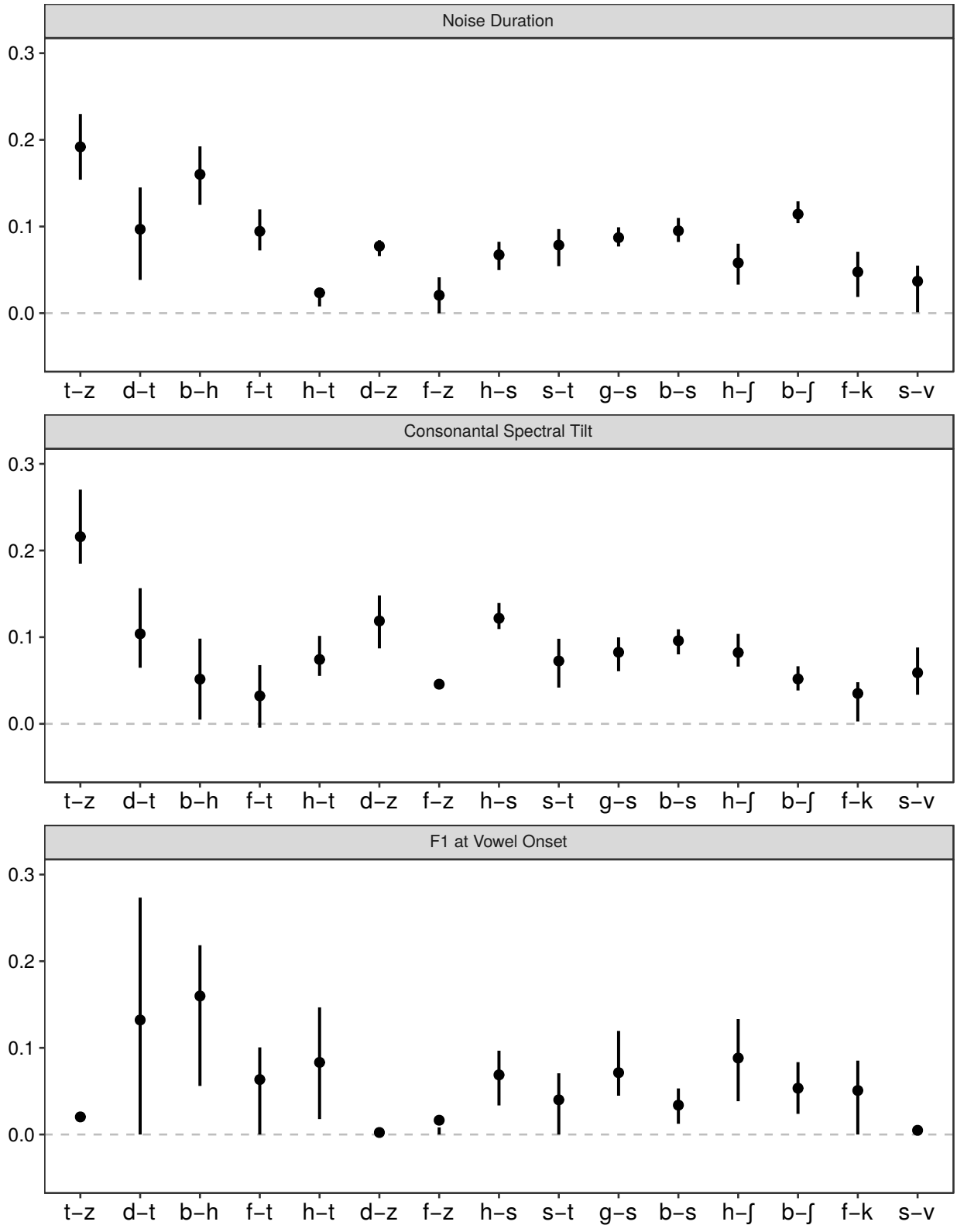

Figure 5.23: Effects of ND, TILT_{C}, and F1_{CV} perturbation on the functional load of the 15 obstruent contrasts with the highest functional load in the lexicon.

with the treatment of $s_{\ominus}(xy)$ as a network measure of potential information loss due to the merger of a given minimal pair, *xy*. Consequently, Figures 5.22 and 5.24 share the three highest-role cues—ND, F1_{CV}, and TILT_{C}—though the distribution of perturbation effects on specific contrasts in Figure 5.25 differs somewhat from that in 5.23.

Beginning with noise duration, Figure 5.25 shows that when contrasts in the wider neighborhood of a given contrast are accounted for, the effect of perturbing the noise duration cue does not align as neatly with expected differences in manner, at least for the focal contrast. For example, while ND-perturbation has the greatest impact on *t–z*, there are several place and voicing contrasts among plosives that are similarly affected. Further, some contrasts such as *s–t* and *p–s* are less impacted than expected. This is because changes in $s_{\ominus}(xy)$ are only partially driven by changes in the relative confusability of *xy* due to the loss of a given cue. A more critical factor in the effect of cue perturbation on $s_{\ominus}(xy)$ is the configuration of contrasts in the neighborhood of *xy*, and the degree to which they depend on a given cue for discrimination. Thus, given that we know from the phonological analysis of edge disjoint degree that the voiceless plosives [p, t, k] occur in dense neighborhoods with many adjacent obstruent contrasts, the result that noise perturbation significantly impacts the role of *k–t*, *p–t*, and *k–p* is not surprising. These contrasts dominate the F1_{CV} results as well, though moving beyond these contrasts F1 remains a clear cue to voicing, as *b–p*, *d–p*, *d–k*, and *d–t* are all notably impacted by F1_{CV} perturbation. Finally, consonantal spectral tilt plays a notable role in several sibilance contrasts, particularly *t–z*, *d–z*, *d–s*, and *p–s*, in addition to impacting the high-$k_{\ominus}(xy)$ voiceless plosive contrasts.

Thus, while edge disjoint strength and functional load respond similarly to various cue perturbations, the manner in which their effects are distributed over different contrasts in the lexicon differs. Functional load focuses directly on the contrasts affected by a given perturbation, irrespective of the number of neighboring obstruent-contrastive minimal pairs. Edge disjoint strength, on the other hand, is sensitive to these neighborhood characteristics, which is particularly beneficial in the context of studying the role of different cues in the lexicon, because acoustic cues are not contrast-constrained; that is, they occur for all contrasts (excepting, of course, some positional

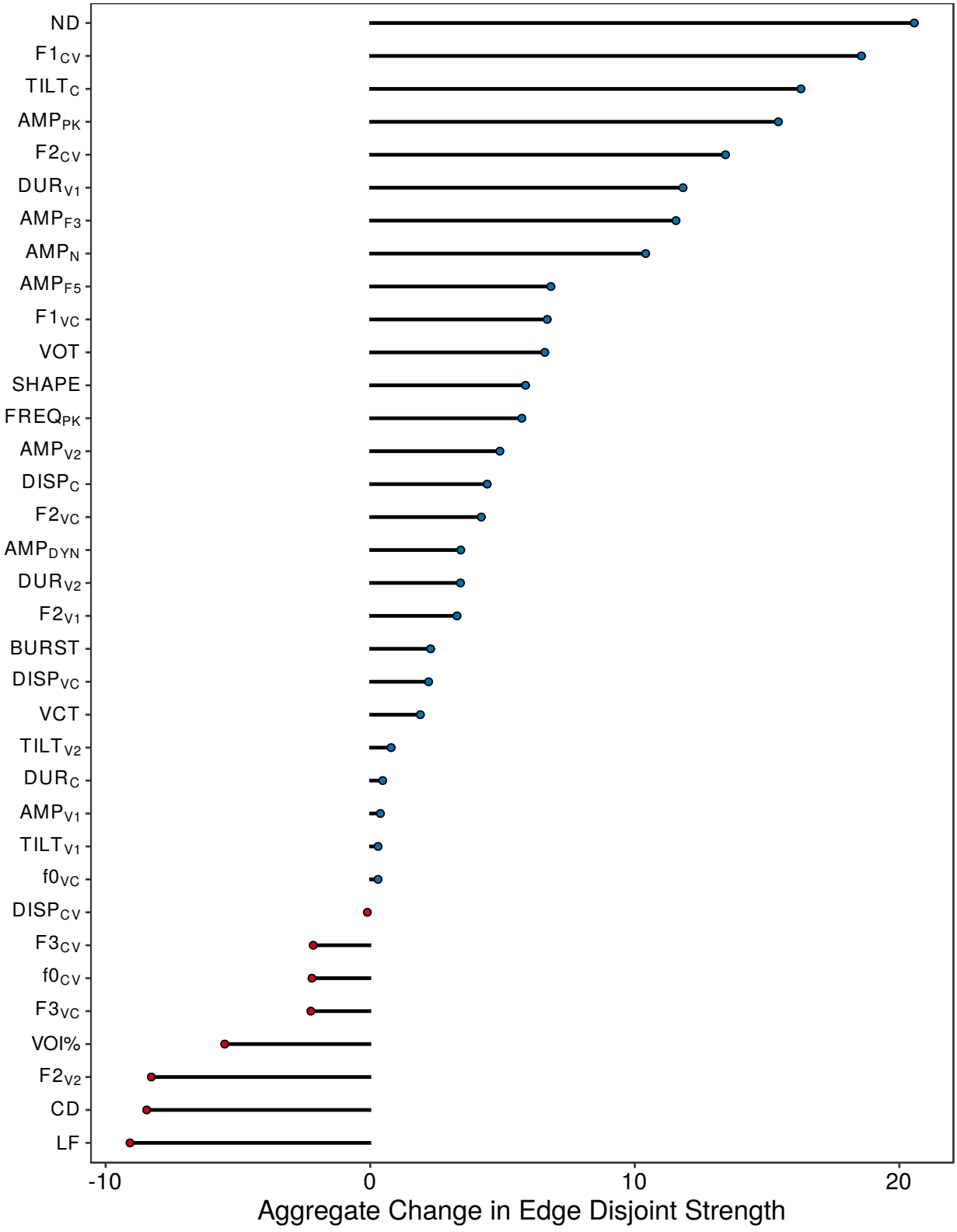

Figure 5.24: Effects of cue perturbation on the aggregate edge disjoint strength of contrasts in the lexicon.

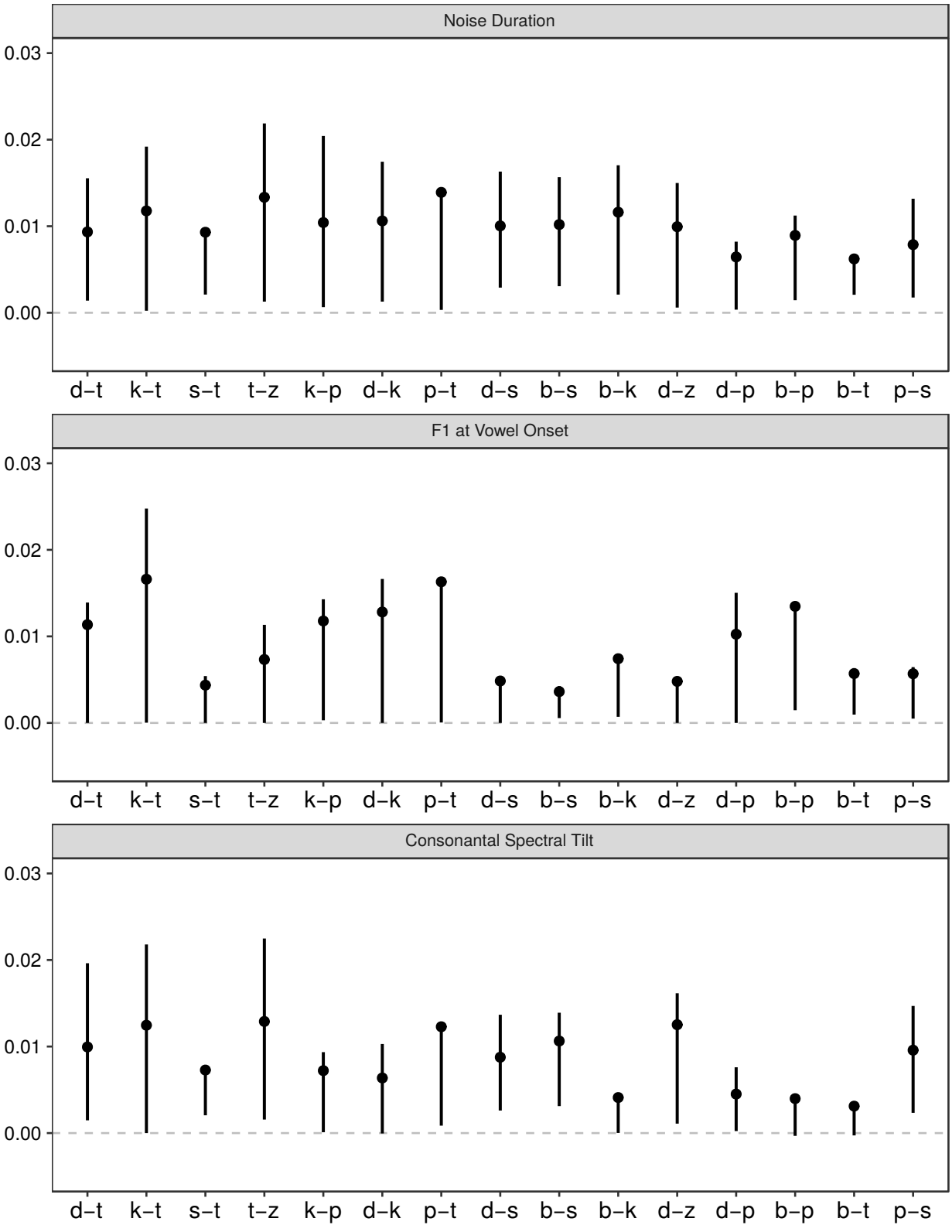

Figure 5.25: Effects of ND, $F1_{CV}$, and $TILT_{C}$ perturbation on the edge disjoint strength of the 15 obstruent contrasts with the highest $s_{\ominus}(xy)$ values in the lexicon.

restrictions like the absence of preceding vowel duration in CV contexts), and only vary in their relative weight. Therefore, it makes sense to consider the impact of a given perturbation on all contrasts in a given neighborhood, not just the contrasts where that cue is known to play a role.

5.4.4 Average path length

Finally, in Figure 5.26 we show the results of perturbing each cue on the average path length in the lexicon, where more negative values indicate larger reductions in global item separability, and more positive values indicate an increase in separation, or greater wordform distinctiveness in the lexicon. As with the other measures, there is general agreement in the role of different acoustic cues when aggregated across the lexicon. The most critical cues are noise duration, consonantal spectral tilt, relative F3 amplitude, and the first and second formants at vowel onset. At the other end of the $\Delta\bar{\ell}$ range are F2 at V2 midpoint, low-frequency energy, closure duration, f0 at vowel onset, and F3 at both vowel offset and onset.

The distribution of percentage change in $\bar{\ell}$ attributable to each contrast under ND, TILT_{C}, and AMP_{F3} perturbation is shown in Figure 5.27, where large decreases in path length due to a given contrast are plotted upwards, and increases in $\bar{\ell}$ plotted downwards. Before examining the effect of each cue on the role of different contrasts in the lexicon, it is worth emphasizing that the most critical contrasts, as far as the global separability of items in the lexicon is concerned, involve manner and sibilance distinctions between [f, s] and the plosive series, which broadly runs counter to the plosive dominance of more *local* measures of phonological structure via minimal pair count, functional load, and edge disjoint strength.

Beginning with noise duration, Figure 5.27 illustrates that ND perturbation most impacts manner contrasts, which occupy a greater proportion of the critical contrasts in the measurement of average path length than most other features (see Table 5.8, for instance). Within this set, the relative discriminability of *b–s*, and the phonological relationships dependent on it, is most impacted by a loss in the noise duration cue. Other manner contrasts involving [f] and [s] are relatively constant in their response to ND perturbation.

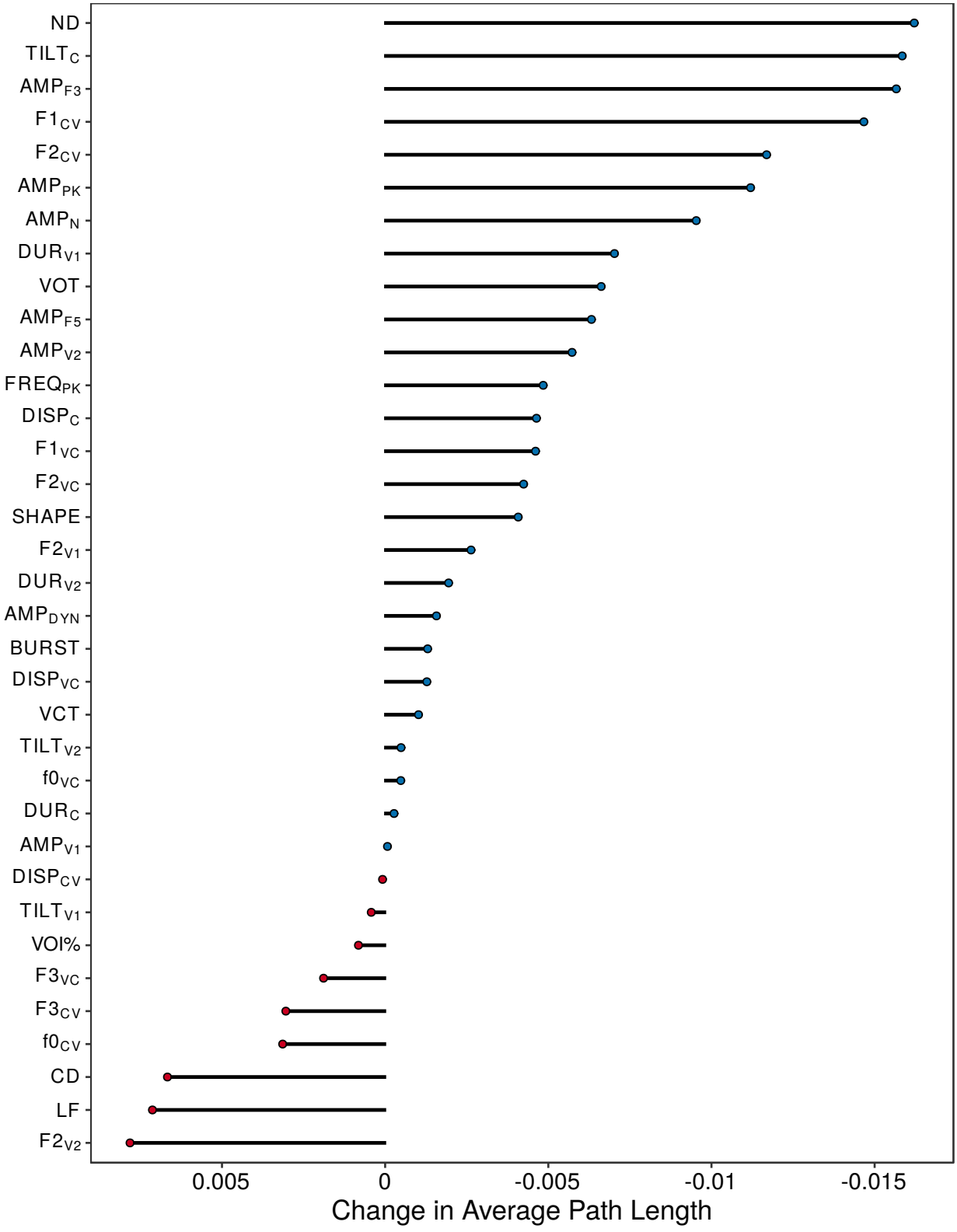

Figure 5.26: Effects of cue perturbation on the average path length in the phonological lexicon.

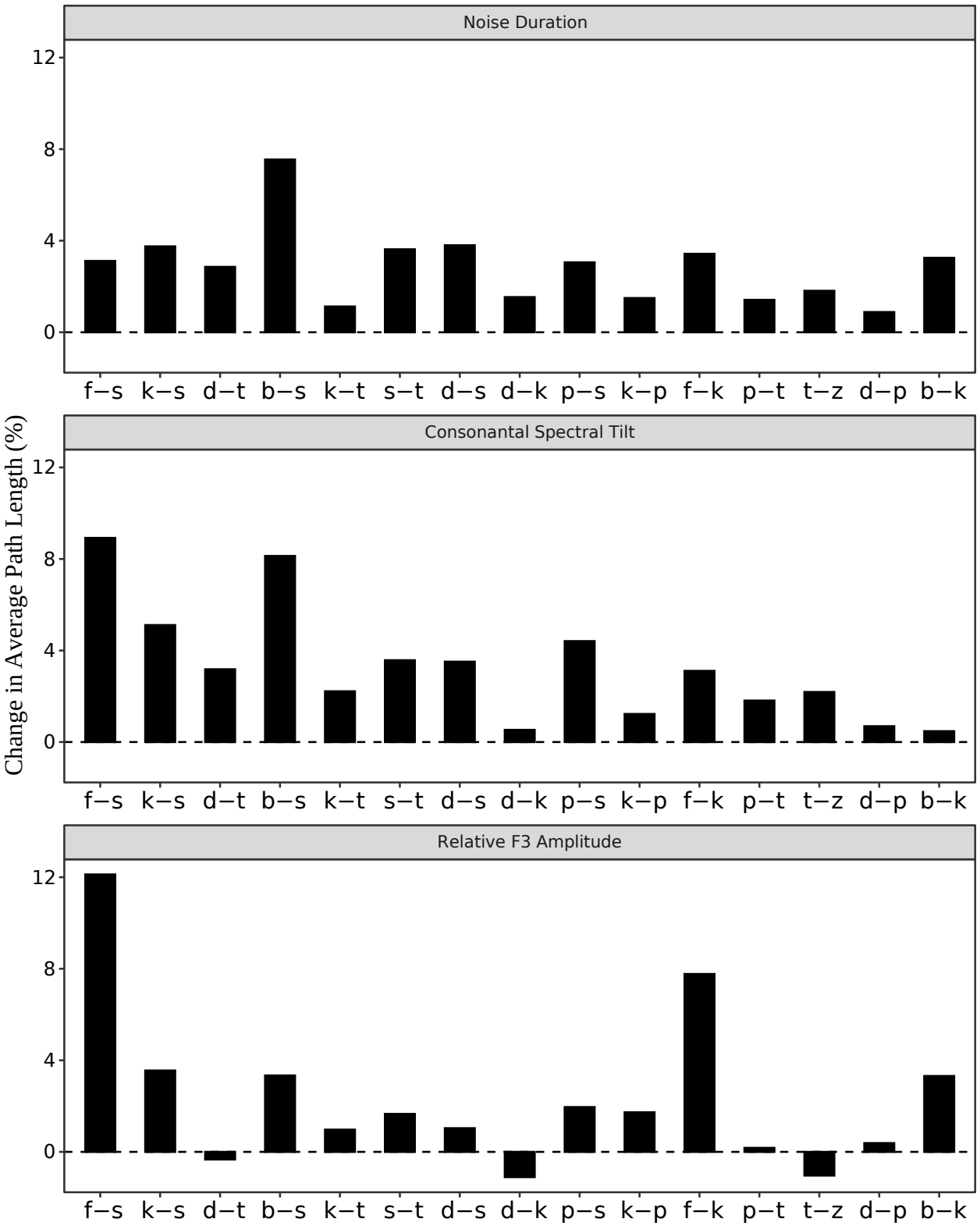

Figure 5.27: Effects of ND, TILT_{C}, and AMP_{F3} perturbation on the average path length in the 15 obstruent contrasts with the highest unweighted path length in the lexicon.

Consonantal spectral tilt, on the other hand, plays a critical role in delineating sibilance distinctions, particularly *f–s*, which is the most phonologically critical contrast of any of the obstruent contrasts in terms of regulating the global separability of items in the lexicon. Finally, the perturbation of relative F3 amplitude does not impact many contrasts, but it has a substantial effect on contrasts with the nonsibilant fricative [f]—i.e., *f–s* and *f–k*. This is because [f] exhibits some of the lowest AMP_{F3} values of any obstruent in English, particularly among real-word distinctions in the lexicon. Thus, we see from the cue-perturbation effects on average path length that the role of a given cue in the lexical system as a whole is not simply a direct function of its type or token frequency. The wider configuration of items in the lexicon also matters in determining the extent of the impact of a loss in information along a particular acoustic dimension.

5.4.5 Discussion

The results of the cue-perturbation analysis in this section provide direct evidence of the impact of asymmetries in lexical contrast distributions on the relative dependence of the system on different properties of the acoustic signal. As in the noise-perturbation analysis, there is a core set of contrasts in the lexicon—within-plosive distinctions and cross-manner/sibilance distinctions between plosives and the anterior fricatives [f, s]—whose relative discriminability far outweighs the remainder of the obstruent set in terms of their impact on the broader system of wordform distinctions in the lexicon. Because of this asymmetry, the acoustic cues which this core set most depends on, such as noise duration, F1, F2, spectral tilt, and relative F3 amplitude, then emerge as the most influential in the lexicon as a whole. And while there were some differences in the effect of perturbing different cues on different measures of system structure, most notably between the more *local* measures of minimal pair count, functional load, and edge disjoint strength, and the more *global* measure of average path length, the four measures were remarkably consistent.

One reason for this result is a distinction between the present study and previous literature that has received little attention in the thesis thus far; namely, the fact that cues are not just defined for restricted subsets of phones, contrasts, and features where they are assumed to be the most relevant.

Cues are defined for all contrasts where they are measurable; for example, voice onset time is defined for both stops and fricatives because the physiological relation it is meant to capture—relative timing differences in laryngeal gestures between the consonant and vowel—applies to fricatives as much as to stops, despite the literature focusing almost exclusively on the latter. Therefore, when spread across nearly all contrasts in the lexicon, the relative impact of perturbing a given cue is much more stable and less dependent on the specific mathematical form of a given structural measurement. This result also has important implications for the organization of control systems in speech production and perception, as the constraints of communication require that a given message be detectable against a range of possible alternatives, not a constrained set varying along only a particular dimension. Therefore, the role of a given acoustic parameter should become more stable as the scope of the system is expanded to include a larger and more representative sample of the higher-order units that must be encoded in the speech signal.

5.5 General discussion

In simulating the response of the system of lexical contrasts to perturbation by background noise and cue loss, we were able to identify the critical components of the system from a lexical perspective. That is, given that we know from Chapter 3 that the distribution of contrasts in the lexicon is highly asymmetric, with the majority of the system dependent on approximately one-third of the obstruent inventory, we asked how such asymmetries affect the impact of acoustic uncertainty on the maintenance of wordform distinctions in English. The result of both the noise- and cue-perturbation analyses is that the perception of the set [p, t, k, b, d, f, s] by-and-large determines the overall performance of the system in terms of successful message transmission. This means that in measuring the encoding potential of the obstruent system in English, both at a theoretical level and a practical level (such as in the assessment of the impact of hearing impairment on communication), much greater attention must be given to plosive place/voicing contrasts, sibilance distinctions, and manner contrasts within this set. This outcome is not only relevant for synchronic research, but implies similar asymmetries in the diachronic evolution of the obstruent system.

Chapter 6

Conclusion

This project represents the first stage in a much wider effort to develop a new paradigm for the analysis of phonetic systems that is fundamentally linked to the higher-order units the system serves to distinguish in communication. In Chapter 2, we examined in detail the distribution of a wide range of acoustic parameters, focusing in particular on the agreement in parameter distributions between the inventory and lexicon data, and on the discriminative potential of each parameter among obstruent contrasts in the lexicon. Chapter 3 provided a similarly comprehensive analysis of the perception of obstruent contrasts in a large set of minimal pairs, with the aim being to show both general patterns in listener accuracy and error distributions, and the degree to which different components of the obstruent system serve as points of relative robustness in the English lexicon, or points of vulnerability; i.e., contrasts which appear frequently in the lexicon and are the source, respectively, of fewer or greater errors than expected in listener word recognition. The acoustic and perceptual distributions were then directly linked in Chapter 4 via a series of statistical cue-integration models designed both to assess the relative agreement in cue weights between the inventory and lexicon, and where points of disagreement arose, to identify the source of the disagreement in distributional, acoustic, or composite differences between the two systems. Further, such analyses were carried out for two main classes of model: an *ideal perceiver* model optimizing discriminative accuracy as a means of uncovering the structure of the acoustic information available in the signal (irrespective of listener parsing behavior), and a *listener* model designed to perform similarly to English perceivers in the recognition task, both in terms of general accuracy and in error distributions. Finally, Chapter 5 analyzed the structure of the system of obstruent contrasts in the lexicon by observing the response of a number of measures of encoding potential to both general perturbation by noise, and perturbation of specific acoustic cues.

The two main research questions identified in Chapter 1 addressed the scalability of canonical, inventory estimates of cue integration to word recognition, and the structure of the lexicon with respect to these estimates—i.e., is there any evidence of lexical optimization in terms of the relationship between critical contrasts and the cues they depend on in perception. Regarding the first question, the results of modeling cue integration under multiple system assumptions in Chapter 4—i.e., in (1) a balanced inventory versus (2) an inventory weighted by phonological distributions in the lexicon versus (3) direct contrasts between lexical items—provided thorough evidence of numerous scaling problems between the inventory and lexicon, both cues that are weighted much lower in a balanced inventory than they should be in the lexicon, and cues whose utility in the lexicon is highly overestimated. As anticipated in the analysis of acoustic and perceptual patterns in the two systems in Chapters 2 and 3, there are notable asymmetries in the lexicon in terms of the relative prevalence of different contrasts in different positions, the acoustic cues that underly this distribution, and the relative role of each contrast in contributing to the overall success of listeners in word recognition. Thus, when a model of contrast discrimination is trained on this data, the resulting cue weights reflect these asymmetries.

For example, in the ideal perceiver models, which reflect differences in acoustics and contrast distributions, the relative utility of consonant voicing percentage (VOI%) in CV and VC contrasts in the lexicon was overestimated by the balanced inventory model because obstruents at real-word boundaries tend to exhibit greater devoicing than those at the margins of controlled syllables, which tend to be hyperarticulated. By comparison, even in cases where the acoustics of the two data sets are closely correlated, such as with preceding vowel duration (DUR_{V1}) in word-final contrasts, lexical cue utility can be underestimated if there are a much greater number of contrasts in the lexicon utilizing this cue than in the balanced inventory. Given that the majority of word-final contrasts in the lexicon involve voicing or manner distinctions, both cued by DUR_{V1}, the effect of this asymmetry is an increase in cue weight relative to that in the inventory, which is more balanced between within- and cross-voicing/manner contrasts.

Such discrepancies are further complicated in the listener models in Chapter 4, where dif-

ferences in controlled-syllable perception versus real-word perception can also be the source of misalignments in cue weights between the two systems. For example, while word-initial consonantal spectral dispersion ($DISP_C$) is acoustically similar in the two databases, listeners were much less accurate at perceiving $DISP_C$ distinctions in the syllable recognition experiment of Woods et al. (2010) than they were in Experiment 1 of the present study. Given that the acoustic and distributional characteristics of the two systems were remarkably similar in this regard, such perceptual discrepancies raise the additional methodological question of whether balanced syllable recognition data is a good proxy for the recognition of such phonological structures in real word recognition. For instance, it could be the case that when the task is constrained to a certain form of linguistic data, such as controlled syllable stimuli, with a balanced representation of different phonological and acoustic features, listeners behave differently from how they would behave if presented with a set of real-word distinctions that are unbalanced in this regard. The question then arises as to which is the more appropriate experimental design, and here we must defer to the particularities of the question being asked. If the question is about listeners' ultimate sensitivity to a particular kind of acoustic or featural distinction, then balanced, controlled syllables may provide an accurate estimate of this boundary value. But if the question is regarding the function of a given cue in the linguistic system, then word- and higher-order unit recognition is paramount, in which case the acoustic and distributional characteristics of such units must be taken into account for the experiment to be ecologically valid.

We should note, however, that in addition to the model discrepancies between inventory and lexical systems, there were also many points of agreement, such as the role of consonantal spectral tilt and noise duration in both *ideal perceiver* and *listener* models. And in all such cases, just as in the analysis of points of cue-weight disagreement, the source of the agreement could be identified as deriving from a combination of distributional and acoustic similarities between the two systems. That is, if a cue is discriminative of a broad range of contrasts, and is acoustically stable in both controlled-syllable and real-word data, then the balanced inventory model will provide a good estimate of that cue's role in the lexicon. Similarly, if a cue is more sparsely relevant for the

discrimination of obstruent contrasts, and the contrasts it does distinguish do not play an outsized role in the lexicon, then estimates of the marginal role of that cue should also scale well between the two systems. However, regarding this latter case there is an important caveat: in either controlled-syllable or word recognition experiments, the size of the contrast set under study is critical. For example, estimates of the utility of closure duration in the discrimination of word-medial stop voicing contrasts, such as in Lisker (1957) or Port (1976), may be overestimated from experiments constrained to only vary CD in the contrasts in which it is most discriminative. In the wider context of obstruent distinctions in English, closure duration is much less reliable as a voicing cue than F1 at vowel onset/offset or the amplitude of low-frequency energy in the noise spectrum.

Therefore, there is ample evidence from Chapter 4, and its descriptive basis in Chapters 2 and 3, that not only are there key discrepancies in the structure of cue integration in models operating under balanced inventory assumptions versus lexical contrast assumptions, but these discrepancies largely have clear phonological and acoustic explanations. This latter result is important because it means that the distinction between the two approaches does not simply force a discrete choice regarding whether or not the canonical inventory approach should be abandoned entirely; it provides a clear roadmap of where inventory assumptions are more and less likely to scale to the lexicon. What is required, rather, is for researchers to motivate both how phonologically representative a given set of phonetic contrasts is of the system of higher-order distinctions in the language as a whole, and how stable acoustic estimates are between (1) controlled syllables or a small number of real words meant to exemplify the contrast, and (2) a larger, more representative sample of real-word distinctions in the language.

The second research question hinges more on the simulations of noise and cue perturbation in Chapter 5, and has a less clear answer than that to Question 1. On the one hand, there are a number of acoustically and perceptually robust contrasts such as those involving the alveolar sibilant [s], that also occupy a central role in the lexicon according to a variety of measures, the most significant among them being *average path length*. This pattern can be seen both in the noise- and cue-perturbation results in Chapter 5, and in the error distributions in Chapter 3. However, for each of

these *robust* contrasts there is an equal or greater number of contrasts that play a prominent role in the lexicon and are poorly discriminated, both in the acoustics and in perception. This set is largely defined by the plosive series, particularly contrasts among voiceless plosives. Thus we have mixed evidence that the system of contrasts in the lexicon represents some kind of optimal distribution of robust contrasts. Further, this question requires not only evidence from analysis of the acoustic and perceptual properties of a model lexicon, but from a larger sample of items produced by multiple speakers and perceived under different task constraints. What is clear from the present study, however, is that the inventory model may provide a false sense of stability in the system as plosive contrasts represent a much smaller subset of the inventory than the lexicon. Therefore, a definition of the system that is independent of the distribution of contrasts among higher-order distinctions in the language will not be able to capture just how critical such distinctions are for the maintenance of contrast in the lexicon, nor the extent to which an external perturbation of those contrasts, such as the frequency and amplitude distortions caused by hearing aids, will impact a speaker/hearer's overall success in communication.

In the Introduction, we outlined the following argument structure motivating the development of a lexical framework for the analysis of the acoustic and perceptual structure of phonetic systems:

P1 The primary basis of phonetic analysis is phonemic.

P2 Phonemes by definition perform a lexically discriminative function.

P3 The distribution of phonemic contrasts in the lexicon is non-uniform, with contrasts differing in their functional load.

$\Rightarrow$ If the speech system is optimized for transmission of phonologically encoded messages, then perceptual weighting of acoustic information in the signal must reflect this distribution.

The three premises are theoretical deductions from the literature. Their implication, however, is conditional on the assumption of cue optimization for message transmission. While we cannot provide direct evidence for optimization of this kind (to do so would require the comparison of cue integration in multiple linguistic systems), Chapter 4 does provide ample evidence of cue

weighting patterns in word recognition that are poorly estimated when distributional asymmetries in the lexicon are not accounted for. More broadly, the acoustic and perceptual data in Chapters 2 and 3, and the noise/cue-perturbation analyses in Chapter 5, provide a picture of the lexicon that is not at all phonetically balanced, and which makes frequent use of a small set of phones, contrasts, and acoustic properties.

This phonetic heterogeneity of the lexicon was foreseen in earlier phonological and computational work, and so the question was not: will similar results be found acoustically and perceptually. The question was what this fundamental fact about the utilization of speech sounds in English word formation would imply for phonetic research programs operating independent of such higher-order distributions. What we find is that there is a wide array of information that is either missing from or distorted in independent, balanced inventory models of the phonetic system. The outsized role of the set [p, t, k, b, d, s, f], and the acoustic properties underlying their discrimination, is masked in a set of 18 phones given equal weight in acoustic and perceptual experimentation. The predominance of multi-feature contrasts ($>$75%) in the lexicon, which promotes the utility of cues such as F1 over VOT in cross-manner voicing perception, is unanticipated by a model where the balance in phone/contrast distributions makes attending to single-feature distinctions more efficient. And biases toward fricatives, sibilants, and voiceless stimuli are only evident when both the distributional and acoustic characteristics of such sounds in a broad set of real-word contrasts are analyzed.

The present study has motivated the need for a lexicon-centric approach to the study of phonetic systems, both on theoretical grounds and from empirical data on the source of scaling errors between cue integration models of syllable and word recognition. In future research we plan to expand the scope of the data—measuring online word recognition in the absence of noise (via eye-tracking) and expanding the choice set beyond minimal pairs and ultimately to models of open-class recognition—as well as the systems under study. This includes cross-linguistic research examining the role of the lexicon in regulating differential attention to cue weights in shared inventories, and clinical research on the lexically distributed consequences of hearing impairment, and developmental research on the role of the growing lexicon in evolving perceptual acuity.

References

Abramson, A. S. & Whalen, D. H. (2017). Voice onset time (vot) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, 63, 75–86.

Agus, N., Anderson, H., Chen, J.-M., Lui, S., & Herremans, D. (2018). Perceptual evaluation of measures of spectral variance. *The Journal of the Acoustical Society of America*, 143(6), 3300–3311.

Amaral, L. A. N., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21), 11149–11152.

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03), 679–685.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database. *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA*.

Baese-Berk, M. & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes*, 24(4), 527–554.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509–512.

Başkent, D. (2012). Effect of speech degradation on top-down repair: Phonemic restoration with

simulations of cochlear implants and combined electric-acoustic stimulation. *Journal of the Association for Research in Otolaryngology*, 13(5), 683–692.

Baum, S. R. & Blumstein, S. E. (1987). Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *Journal of the Acoustical Society of America*, 82(3), 1073–1077.

Behrens, S. & Blumstein, S. E. (1988). On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *J. Acoust. Soc. Am.*, 84(3), 861–867.

Bell-Berti, F. (1975). Control of pharyngeal cavity size for English voiced and voiceless stops. *The Journal of the Acoustical Society of America*, 57(2), 456–461.

Blevins, J. (2006). A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics*, 32(2), 117–166.

Bloch, B. (1948). A set of postulates for phonemic analysis. *Language*, 24(1), 3–46.

Boardman, I., Cohen, M., & Grossberg, S. (1993). *Variable rate working memories for phonetic categorization and invariant speech perception*. Technical report, Boston University Center for Adaptive Systems.

Boberg, C. (2008a). English in Canada: phonology. In E. W. Schneider (Ed.), *Varieties of English: The Americas and the Caribbean*, volume 2 (pp. 144–160).

Boberg, C. (2008b). Regional phonetic differentiation in standard Canadian English. *Journal of English Linguistics*, 36(2), 129–154.

Boersma, P. & Weenink, D. (2016). *Praat: Doing phonetics by computer [Computer software]*. http://www.praat.org/.

Broad, D. J. & Clermont, F. (1987). A methodology for modeling vowel formant contours in CVC context. *Journal of the Acoustical Society of America*, 81(1), 155–165.

Broad, D. J. & Clermont, F. (2002). Linear scaling of vowel-formant ensembles (VFEs) in consonantal contexts. *Speech Communication*, 37(3-4), 175–195.

Broad, D. J. & Clermont, F. (2010). Target–locus scaling methods for modeling families of formant transitions. *Journal of Phonetics*, 38(3), 337–359.

Broad, D. J. & Clermont, F. (2014). A method for analyzing the coarticulated CV and VC components of vowel-formant trajectories in CVC syllables. *Journal of Phonetics*, 47, 47–80.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, 44(4), 991–997.

Byrd, D. (1993). 54,000 american stops. *UCLA Working Papers in Phonetics*, 83, 97–116.

Carter, D. M. (1987). An information-theoretic analysis of phonetic dictionary access. *Computer Speech & Language*, 2(1), 1–11.

Castleman, W. A. & Diehl, R. L. (1996). Effects of fundamental frequency on medial and final [voice] judgments. *Journal of Phonetics*, 24(4), 383–398.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Midland Books.

Chelba, C. & Jelinek, F. (2000). Structured language modeling. *Computer Speech & Language*, 14(4), 283–332.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4), 661–703.

Cohen, L. B., Diehl, R. L., Oakes, L. M., & Loehlin, J. C. (1992). Infant perception of /aba/ versus /apa/: Building a quantitative model of infant categorical discrimination. *Developmental Psychology*, 28(2), 261–272.

Colantoni, L. & Steele, J. (2007). Voicing-dependent cluster simplification asymmetries in Spanish and French. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 109–130). Amsterdam; Philadelphia; J. Benjamins Pub. Co.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, 35(2), 180–209.

Cole, R. A. & Cooper, W. E. (1975). Perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America*, 58(6), 1280–1287.

Cooke, M. & Scharenborg, O. (2008). The interspeech 2008 consonant challenge. *Interspeech, 2008*.

Crane, H. (2018). *Probabilistic Foundations of Statistical Network Analysis*. Chapman and Hall/CRC.

Crystal, T. H. & House, A. S. (1988). The duration of American-English stop consonants: An overview. *Journal of Phonetics*, 16(3), 285–294.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163, 128–145.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159–190.

Davis, M. H. & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research*, 229(1-2), 132–147.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27(4), 769–773.

Denes, P. (1955). Effect of duration on the perception of voicing. *The Journal of the Acoustical Society of America*, 27(4), 761–764.

Dewey, G. (1923). *Relative frequency of English speech sounds*. Cambridge: Harvard University Press.

Docherty, G. J. (1992). *The timing of voicing in British English obstruents*, volume 9 of *Netherlands Phonetic Archives*. Walter de Gruyter.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, 22(2), 109–122.

Evers, V., Reetz, H., & Lahiri, A. (1998). Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of Phonetics*, 26(4), 345–370.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fant, G. (1962). Formant bandwidth data. *STL-QPSR*, 3, 1–2.

Fant, G. (1973). Stops in CV-syllables. *STL-QPSR*, 10(4), 110–139.

Fischer-Jørgensen, E. (1954). Acoustic analysis of stop consonants. *Le Maître Phonétique*, 32, 42–59.

Fischer-Jørgensen, E. (1975). *Trends in phonological theory*. Akademisk Forlag.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, 27(4), 343–350.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am.*, 84(1), 115–123.

Forster, K. I. (1976). Accessing the mental lexicon. In R. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257–287). Amsterdam: North-Holland.

Gerstman, L. J. (1957). *Perceptual dimensions for the friction portions of certain speech sounds*. PhD thesis, New York University, Graduate School of Arts and Science.

Gracco, V. L. (1994). Some organizational characteristics of speech movement control. *Journal of Speech, Language, and Hearing Research*, 37(1), 4–27.

Gray, A. & Markel, J. (1974). A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(3), 207–217.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, 31(3-4), 423–445.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 23(2), 481.

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America*, 47(2B), 613–617.

Harris, K. S. (1974). Mechanisms of duration change. In *Proceedings of the Speech Communication Seminar, Stockholm*.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Effect of third-formant transitions on the perception of the voiced stop consonants. *The Journal of the Acoustical Society of America*, 30(2), 122–126.

Hedrick, M. S. & Ohde, R. N. (1993). Effect of relative amplitude of frication on perception of place of articulation. *J. Acoust. Soc. Am.*, 94(4), 2005–2026.

Heffner, R. S. (1937). Notes on the length of vowels. *American Speech*, (pp. 128–134).

Heinz, J. M. & Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, 33(5), 589–596.

Hockett, C. F. (1967). The quantification of functional load. *Word*, 23(1-3), 300–320.

Hoffman, H. S. (1958). Study of some cues in the perception of the voiced stop consonants. *The Journal of the Acoustical Society of America*, 30(11), 1035–1041.

House, A. S. & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, 25(1), 105–113.

Howell, P. & Rosen, S. (1983). Production and perception of rise time in the voiceless affricate/fricative distinction. *The Journal of the Acoustical Society of America*, 73(3), 976–984.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of english words. *The Journal of the Acoustical Society of America*, 29(2).

Hughes, G. W. & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, 28(2), 303–310.

Jakobson, R., Fant, C. G., & Halle, M. (1951). *Preliminaries to Speech Analysis: The distinctive features and their correlates*. Cambridge, MA: The MIT Press.

Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America*, 85(4), 1718–1725.

Jongman, A. & McMurray, B. (2017). On invariance: Acoustic input meets listener expectations. *The Speech Processing Lexicon: Neurocognitive and Behavioural Approaches*, 22, 21–51.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.*, 108(3), 1252–1263.

Kapelner, A. & Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4), 1–40.

Keller, E. F. (2005). Revisiting"scale-free" networks. *BioEssays*, 27(10), 1060–1068.

Kenyon, J. S. (1924). *American pronunciation*. University of Michigan Press.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4), 736.

King, R. D. (1967). Functional load and sound change. *Language*, (pp. 831–852).

Kirchner, R. M. (1998). *An effort-based approach to consonant lenition*. PhD thesis, UCLA.

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, 18(4), 686–706.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5), 1208–1221.

Klatt, D. H. & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2), 820–857.

Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics*.

Kluender, K. R. & Walsh, M. A. (1992). Amplitude rise time and the perception of the voiceless affricate/fricative distinction. *Perception & Psychophysics*, 51(4), 328–333.

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech, Language, and Hearing Research*.

Krámský, J. (1974). *The phoneme; introduction to the history and theories of a concept*. W. Fink.

Krull, D. (1989). Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *PERILUS*, 10, 87–108.

Kuehn, D. P. & Moll, K. L. (1976). A cineradiographic study of vc and cv articulatory velocities. *Journal of phonetics*, 4(4), 303–320.

Kuhl, P. K. (1979). The perception of speech in early infancy. *Speech and Language*, 1, 1–47.

Labov, W. (1981). Resolving the Neogrammarian controversy. *Language*, (pp. 267–308).

Labov, W. (1994). Principles of linguistic change: Internal factors.

Labov, W., Yaeger, M., & Steiner, R. (1972). *A quantitative study of sound change in progress*, volume 1. US Regional Survey.

Lahiri, A. (1999). Speech recognition with phonological features. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, volume 1 (pp. 715–718).

Lahiri, A., Gewirth, L., & Blumstein, S. E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America*, 76(2), 391–404.

Lahiri, A. & Reetz, H. (2002). Underspecified recognition. *Laboratory Phonology*, 7, 637–675.

Lehiste, I. & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *The Journal of the Acoustical Society of America*, 31(4), 428–435.

Lehmann, W. & Heffner, R. S. (1940). Notes on the length of vowels (III). *American Speech*, 15(4), 377–380.

Lehmann, W. & Heffner, R. S. (1943). Notes on the length of vowels (VI). *American Speech*, 18(3), 208–215.

Liberman, A. M. (1993). Some assumptions about speech and how they changed. *Haskins Laboratories Status Report on Speech Research*, 113, 1–32.

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, (pp. 497–516).

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8), 1–13.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358–368.

Lim, S.-J., Shinn-Cunningham, B. G., & Perrachione, T. K. (2019). Effects of talker continuity and speech rate on auditory working memory. *Attention, Perception, & Psychophysics*, 81(4), 1167–1177.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Springer.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 33(1), 42–49.

Lisker, L. (1970). Supraglottal air pressure in the production of English stops. *Language and Speech*, 13(4), 215–230.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, 29(1), 3–11.

Lisker, L. & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3), 384–422.

Lisker, L. & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, 10(1), 1–28.

Lloyd, H. (1936). Note on vowel length. *American Speech*, 11(2), 188–189.

Locke, W. & Heffner, R. S. (1940). Notes on the length of vowels (II). *American Speech*, 15(1), 74–79.

Löfqvist, A. & Gracco, V. L. (1994). Tongue body kinematics in velar stop production: Influences of consonant voicing and vowel context. *Phonetica*, 51(1-3), 52–67.

Lubker, J. F. & Parris, P. J. (1970). Simultaneous measurements of intraoral pressure, force of labial contact, and labial electromyographic activity during production of the stop consonant cognates /p/ and /b/. *The Journal of the Acoustical Society of America*, 47(2B), 625–633.

Luce, P. A. & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, 78(6), 1949–1957.

Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1), 1–36.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, 66(2), 81.

Mack, M. & Lieberman, P. (1985). Acoustic analysis of words produced by a child from 46 to 149 weeks. *Journal of Child Language*, 12(3), 527–550.

Maddieson, I. (1992). UCLA Phonological Segment Inventory Database.

Malécot, A. (1966). The effectiveness of intra-oral air-pressure-pulse parameters in distinguishing between stop cognates. *Phonetica*, 14(2), 65–81.

Marcus, S. M. (1978). Distinguishing “slit” and “split”–an invariant timing cue in speech perception. *Perception & Psychophysics*, 23(1), 58–60.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1), 71–102.

Martinet, A. (1952). Function, structure, and sound change. *Word*, 8(1), 1–32.

Massaro, D. W. & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. *Journal of the Acoustical Society of America*, 60(3), 704–717.

Massaro, D. W. & Oden, G. C. (1980). Evaluation and integration of acoustic features in speech perception. *The Journal of the Acoustical Society of America*, 67(3), 996–1013.

Matisoff, J. A. (1973). Tonogenesis in southeast asia. *Southern California Occasional Papers in Linguistics: Consonant types and tone*, 1.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746.

McMurray, B. & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2), 219–246.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.

Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.*, 27(2), 338–352.

Miller, J. L. & Grosjean, F. (1981). How the components of speaking rate influence perception of phonetic segments. *Journal of Experimental Psychology: Human Perception and Performance*, 7(1), 208.

Mugdan, J. (1985). The origins of the phoneme: Farewell to a myth. *Historiographia Linguistica*, 38(1), 85–110.

Munson, B. & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, 47(5), 1048–1058.

Myers, S. (2002). Gaps in factorial typology: The case of voicing in consonant clusters. Master's thesis, University of Texas at Austin.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, 101(6), 3241–3254.

Nearey, T. M. & Shammass, S. E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics*, 15(4), 17–24.

Newman, M. E. (2011). Complex systems: A survey. *arXiv preprint arXiv:1112.1440*.

Niyogi, P. & Sondhi, M. M. (2002). Detecting stop consonants in continuous speech. *the Journal of the Acoustical Society of America*, 111(2), 1063–1076.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234.

Norris, D. & McQueen, J. M. (2008). Shortlist b: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395.

Ohala, J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, 13(1-2).

Ohde, R. N. & Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America*, 74(3), 706–714.

Öhman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39(1), 151–168.

Parker, E. M., Diehl, R. L., & Kluender, K. R. (1986). Trading relations in speech and nonspeech. *Perception & Psychophysics*, 39(2), 129–142.

Peterson, G. E. & Lehiste, I. (1960). Duration of syllable nuclei in english. *The Journal of the Acoustical Society of America*, 32(6), 693–703.

Pike, K. L. (1972). *Phonetics: A Critical Analysis of Phonetic Theory and a Technic for the Practical Description of Sounds*. ERIC.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release)[www. buckeyecorpus. osu. edu] columbus, oh: Department of psychology. *Ohio State University (Distributor)*.

Port, R. F. (1976). *The influence of tempo on stop closure duration as a cue for voicing and place*. PhD thesis, University of Connecticut.

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, 69(1), 262–274.

Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, (pp. 510–515).

Rakerd, B., Dechovitz, D. R., & Verbrugge, R. R. (1982). An effect of sentence finality on the phonetic significance of silence. *Language and Speech*, 25(3), 267–282.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America*, 51(4B), 1296–1303.

Raphael, L. J. (1975). The physiological control of durational differences between vowels preceding voiced and voiceless consonants in English. *Journal of Phonetics*, 3(1), 25–33.

Raphael, L. J. & Dorman, M. F. (1980). Silence as a cue to the perception of syllable-initial and syllable-final stop consonants. *Journal of Phonetics*, 8(3), 269–275.

Reetz, H. (1999). Converting speech signals to phonological features. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, volume 3 (pp. 1733–1736).

Repp, B. H. (1978). Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. *Perception & Psychophysics*, 24(5), 471–485.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92(1), 81–110.

Repp, B. H. (1984a). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and Speech*, 27(3), 245–254.

Repp, B. H. (1984b). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and Speech*, 27(3), 245–254.

Roach, P. (1989). Report on the 1989 kiel convention: International phonetic association. *Journal of the International Phonetic Association*, 19(2), 67–80.

Scarborough, R. A. (2004). *Coarticulation and the structure of the lexicon*. PhD thesis, University of California, Los Angeles.

Shadle, C. H. (1985). *The acoustics of fricative consonants*. PhD thesis, MIT.

Shadle, C. H. & Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. In *ICSLP 96, Proceedings*, volume 3 (pp. 1521–1524).: IEEE.

Shadle, C. H. & Scully, C. (1995). An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. *J. Phon.*, 23(1), 53–66.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423.

Sharf, D. J. (1962). Duration of post-stress intervocalic stops and preceding vowels. *Language and Speech*, 5(1), 26–30.

Smith, B. L. (1978). Temporal aspects of English speech production: A developmental perspective. *Journal of Phonetics*, 6(1), 37–67.

Smits, R., ten Bosch, L., & Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluation. *The Journal of the Acoustical Society of America*, 100(6), 3865–3881.

Stathopoulos, E. T. & Weismer, G. (1983). Closure duration of stop consonants. *Journal of Phonetics*, 11(4), 395–400.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *J. Acoust. Soc. Am.*, 50(4B), 1180–1192.

Stevens, K. N. (1985). Evidence for the role of acoustic boundaries in the perception of speech sounds. *Phonetic linguistics: Essays in honor of Peter Ladefoged*, (pp. 243–255).

Stevens, K. N. & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64(5), 1358–1368.

Strang, A., Haynes, O., Cahill, N. D., & Narayan, D. A. (2018). Generalized relationships between characteristic path length, efficiency, clustering coefficients, and density. *Social Network Analysis and Mining*, 8(1), 14.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, 3(1), 32–49.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), 1074–1095.

Surendran, D. & Niyogi, P. (2003). Measuring the usefulness (functional load) of phonological contrasts. *Technical Report TR-2003-12*.

Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *Journal of the Acoustical Society of America*, 94(3), 1256–1268.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90(3), 1309–1325.

Sussman, H. M. & Shore, J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception & Psychophysics*, 58(6), 936–946.

Tabain, M. (2001). Variability in fricative production and spectra: Implications for the hyper-and hypo- and quantal theories of speech production. *Language and Speech*, 44(1), 57–93.

Tan, M., Zhou, W., Zheng, L., & Wang, S. (2012). A scalable distributed syntactic, semantic, and lexical language model. *Computational Linguistics*, 38(3), 631–671.

Tauberer, J. I. (2010). *Learning [voice]*. PhD thesis, University of Pennsylvania.

Tucker, B. V., Brenner, D., Danielson, K. D., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The Massive Auditory Lexical Decision database: Toward reliable, generalizable speech research. *Behavioral Research Methods*, (pp. 1–18).

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America*, 61(3), 846–858.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, 51(2), 408–422.

Vitevitch, M. S., Ercal, G., & Adagarla, B. (2011). Simulating retrieval from a highly clustered network: Implications for spoken word recognition. *Frontiers in Psychology*, 2, 369.

Wang, W. S.-Y. (1969). Competing changes as a cause of residue. *Language*, (pp. 9–25).

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442.

Wedel, A. (2004). *Self-organization and the development of higher-order phonological patterns*. PhD thesis, University of California, Santa Cruz.

Wedel, A. (2007). Feedback and regularity in the lexicon. *Phonology*, 24(1), 147–185.

Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition*, 4(4), 319–355.

Wedel, A. & Fatkullin, I. (2017). Category competition as a driver of category contrast. *Journal of Language Evolution*, 2(1), 77–93.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2), 179–186.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, 35(1), 49–64.

Woods, D. L., Yund, E. W., & Herron, T. (2010). Measuring consonant identification in nonsense syllables, words, and sentences. *Journal of Rehabilitation Research & Development*, 47(3), 243–60.

Wright, R. (2004). Factors of lexical competition in vowel articulation. *Papers in Laboratory Phonology VI*, (pp. 75–87).

You, H.-Y. (1979). *An acoustic and perceptual study of English fricatives*. PhD thesis, University of Alberta.

Yuan, J. & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5), 3878.

Appendix A

Additional tables and figures

Comparison of Canadian and American listener perception in Experiments 0a and 0b

		Canadian Listeners			American Listeners		
SNR	**Stimulus**	**Accuracy**	**Error 1**	**Error 2**	**Accuracy**	**Error 1**	**Error 2**
high	acute	75%	accuse (10%)	NA	80%	cute (10%)	NA
high	aging	55%	agent (35%)	NA	75%	agent (10%)	NA
high	agreed	25%	greed (10%)	NA	60%	agree (15%)	degree (15%)
high	assessed	20%	obsessed (65%)	NA	40%	assist (15%)	obsessed (15%)
high	ballad	25%	ballot (50%)	balance (15%)	20%	valid (40%)	ballot (30%)
high	beige	10%	bathe (50%)	NA	25%	page (15%)	bathe (10%)
high	buggy	50%	bug (15%)	bugging (15%)	55%	bug (15%)	muggy (10%)
high	bull	0%	bowl (15%)	bold (10%)	25%	bolt (30%)	bold (10%)
high	calorie	20%	salary (60%)	celery (10%)	95%	calories (10%)	NA
high	carriage	65%	courage (15%)	carrot (10%)	65%	courage (15%)	carrot (10%)
high	chunk	15%	trunk (40%)	tongue (15%)	60%	trunk (35%)	NA
high	class	55%	glass (10%)	NA	55%	glass (15%)	fast (10%)
high	cold	80%	core (10%)	NA	60%	cool (15%)	NA
high	conceive	50%	see (10%)	concede (10%)	70%	deceive (20%)	concede (10%)
high	continued	15%	team (20%)	tea (15%)	35%	continue (45%)	NA
high	cop	10%	crop (20%)	cross (15%)	40%	cup (20%)	clock (10%)
high	coping	50%	coconut (25%)	open (15%)	65%	coconut (10%)	NA
high	cub	40%	cover (10%)	cowboy (10%)	60%	cup (25%)	NA
high	dairy	85%	daring (10%)	NA	65%	daring (20%)	NA
high	delayed	80%	glade (10%)	NA	60%	glade (25%)	blade (15%)
high	delegate	45%	delicate (20%)	NA	85%	delicate (20%)	NA
high	depend	25%	pen (40%)	depends (10%)	65%	depends (10%)	pen (10%)
high	displays	55%	display (15%)	NA	85%	displace (10%)	NA
high	disrupted	55%	disruptive (40%)	NA	45%	disruptive (40%)	NA
high	eagle	70%	legal (20%)	able (10%)	85%	evil (10%)	NA
high	ethic	80%	epic (10%)	NA	80%	epic (20%)	NA
high	faint	70%	fate (25%)	NA	75%	think (10%)	NA
high	fang	75%	van (20%)	NA	65%	bang (20%)	NA

high	faster	55%	fester (45%)	NA	85%	fester (15%)	NA
high	feed	55%	feet (15%)	food (15%)	60%	feet (20%)	NA
high	fiery	55%	firing (25%)	fire (20%)	35%	firing (30%)	fire (20%)
high	fleck	75%	reflect (10%)	NA	70%	flex (10%)	NA
high	fraud	40%	frog (20%)	NA	70%	frog (20%)	NA
high	fully	25%	full (20%)	void (10%)	10%	full (10%)	pull (10%)
high	gem	70%	jam (20%)	gym (10%)	70%	gym (10%)	NA
high	grazing	70%	crazy (20%)	NA	70%	crazy (10%)	NA
high	guarantees	0%	guarantee (95%)	NA	65%	guarantee (25%)	NA
high	hack	15%	pack (75%)	NA	65%	pack (25%)	NA
high	having	0%	panic (30%)	hammock (15%)	30%	heaven (30%)	havoc (10%)
high	havoc	20%	habit (70%)	NA	55%	habit (35%)	NA
high	hiss	0%	disk (25%)	kiss (20%)	45%	kiss (25%)	this (15%)
high	hood	0%	first (25%)	put (15%)	40%	put (30%)	could (10%)
high	hooky	15%	cookie (65%)	NA	45%	cookie (15%)	bookie (10%)
high	hoop	0%	poop (15%)	coop (15%)	40%	poop (15%)	coop (10%)
high	hovering	75%	hover (10%)	NA	90%	covering (10%)	NA
high	jumble	75%	jumbled (10%)	NA	75%	jumbled (10%)	NA
high	keen	75%	king (15%)	NA	80%	king (15%)	NA
high	king	80%	cave (10%)	cane (10%)	85%	kin (10%)	NA
high	labor	60%	paper (15%)	neighbor (10%)	75%	liver (20%)	NA
high	lash	55%	flash (15%)	laugh (15%)	45%	flash (15%)	laugh (15%)
high	leaking	0%	leaving (20%)	beefing (15%)	45%	lethal (20%)	leaving (15%)
high	led	60%	bled (15%)	sled (10%)	80%	bled (10%)	NA
high	legion	45%	bluejay (10%)	NA	70%	region (10%)	NA
high	looting	15%	beating (15%)	brooding (10%)	10%	booting (20%)	booing (15%)
high	lope	0%	loaf (15%)	serious (15%)	25%	local (20%)	bloke (10%)
high	magnifies	0%	magnify (90%)	NA	20%	magnify (55%)	magnified (20%)
high	mashing	30%	matching (45%)	NA	40%	matching (35%)	NA
high	mood	0%	move (80%)	smooth (20%)	55%	move (20%)	NA
high	myth	15%	net (25%)	left (15%)	65%	nest (10%)	NA
high	nightlife	45%	meclife (15%)	life (10%)	70%	nightlight (20%)	NA
high	odor	25%	holding (15%)	shoulder (15%)	60%	holder (10%)	NA
high	odyssey	10%	honesty (15%)	policy (10%)	50%	policy (10%)	NA
high	overcook	85%	overcooked (10%)	NA	75%	overcooked (25%)	NA
high	palette	15%	palettes (15%)	balance (10%)	35%	palace (10%)	pellet (15%)
high	paw	20%	cough (40%)	NA	50%	awe (10%)	cough (10%)
high	peg	15%	bag (20%)	pegged (20%)	65%	beg (15%)	NA
high	pill	40%	kill (30%)	NA	70%	kill (15%)	NA
high	pocket	40%	bucket (25%)	buckle (15%)	30%	bucket (30%)	faucet (10%)
high	polity	0%	quality (60%)	apology (15%)	5%	quality (70%)	policy (10%)

high	pose	15%	poles (35%)	pole (15%)	20%	poles (50%)	NA
high	preface	50%	crevice (15%)	prefaced (10%)	70%	crevice (10%)	NA
high	puppy	65%	puffy (15%)	NA	80%	puffy (10%)	NA
high	qualifies	0%	qualify (80%)	qualified (15%)	30%	qualify (45%)	qualified (15%)
high	raffle	0%	baffle (30%)	bathroom (30%)	80%	baffled (10%)	NA
high	raid	0%	grade (55%)	afraid (10%)	20%	grade (30%)	brave (10%)
high	razor	55%	laser (25%)	razer (10%)	75%	risen (20%)	NA
high	required	0%	require (55%)	choir (15%)	30%	require (40%)	acquire (30%)
high	resent	30%	present (40%)	represent (15%)	10%	present (70%)	NA
high	resin	10%	risen (35%)	driven (15%)	30%	risen (50%)	prison (10%)
high	revive	0%	goodbye (20%)	robot (15%)	10%	provide (25%)	survive (15%)
high	ribbon	0%	complain (10%)	driven (10%)	60%	driven (20%)	NA
high	rigor	20%	trigger (45%)	NA	55%	trigger (20%)	NA
high	rocker	25%	rock (40%)	rocket (10%)	50%	rock (20%)	NA
high	shale	50%	jail (10%)	shell (10%)	35%	chill (25%)	shell (20%)
high	siege	15%	seed (15%)	seize (15%)	55%	seize (30%)	NA
high	soot	65%	sit (15%)	NA	80%	certain (10%)	NA
high	soothe	70%	sue (20%)	NA	60%	soothed (15%)	NA
high	supper	70%	suburb (15%)	stubborn (10%)	75%	suburb (15%)	NA
high	supplied	10%	climb (25%)	supply (25%)	60%	supply (35%)	NA
high	tea	55%	key (20%)	tweed (10%)	85%	key (15%)	NA
high	terrify	45%	clarify (15%)	purify (15%)	70%	terrified (20%)	NA
high	thatch	25%	batch (30%)	catch (15%)	25%	catch (10%)	fat (10%)
high	there	55%	bear (40%)	NA	75%	dare (10%)	NA
high	these	20%	bees (65%)	NA	30%	bees (55%)	NA
high	thin	50%	fin (35%)	NA	75%	fin (20%)	NA
high	thought	30%	fought (50%)	NA	45%	fuck (25%)	fought (15%)
high	troop	55%	truth (15%)	true (10%)	65%	truth (10%)	NA
high	trough	35%	cough (50%)	NA	15%	cough (35%)	NA
high	unified	30%	unify (50%)	beautify (15%)	50%	unify (30%)	NA
high	varied	20%	very (35%)	scary (15%)	45%	very (35%)	buried (10%)
high	vault	0%	fault (30%)	bald (15%)	50%	volt (20%)	NA
high	volley	40%	bully (15%)	bali (10%)	20%	bully (15%)	balling (15%)
high	waiver	70%	quiver (10%)	NA	75%	quiver (15%)	NA
high	watch	10%	wash (65%)	squash (10%)	80%	wash (10%)	NA
high	widowed	70%	widow (20%)	NA	50%	whittled (15%)	widow (15%)
low	acute	50%	accuse (20%)	cute (10%)	40%	cute (15%)	NA
low	aging	20%	agent (20%)	NA	50%	agent (15%)	NA
low	agreed	0%	day (15%)	do (20%)	25%	degree (30%)	grieve (10%)
low	ahead	35%	head (50%)	NA	75%	head (20%)	NA
low	assessed	0%	obsessed (45%)	disaster (10%)	40%	obsessed (25%)	incest (10%)

low	attaching	15%	attachment (50%)	attach (15%)	60%	attachment (15%)	NA
low	bad	50%	sad (10%)	NA	65%	sad (25%)	NA
low	ballad	0%	ballet (50%)	balance (10%)	15%	valid (40%)	ballot (20%)
low	barrier	0%	bread (45%)	rare (10%)	70%	vary (10%)	NA
low	bases	30%	basis (20%)	braces (20%)	85%	basis (5%)	NA
low	bathing	15%	favour (15%)	baby (10%)	65%	baby (10%)	NA
low	beacon	0%	taken (35%)	denim (15%)	55%	beaten (25%)	NA
low	beige	0%	bathe (20%)	babe (15%)	15%	age (15%)	NA
low	budding	30%	button (20%)	butting (15%)	70%	button (15%)	NA
low	buggy	0%	bug (20%)	budding (10%)	20%	muggy (10%)	NA
low	bull	0%	collar (25%)	NA	10%	bold (10%)	NA
low	calorie	0%	salary (50%)	celery (40%)	60%	calories (10%)	salary (10%)
low	camp	55%	cab (10%)	NA	55%	count (10%)	NA
low	carriage	40%	courage (20%)	carrot (15%)	15%	courage (25%)	NA
low	case	65%	face (20%)	NA	60%	taste (20%)	face (10%)
low	chunk	0%	trunk (20%)	tongue (15%)	30%	tongue (20%)	trunk (20%)
low	circus	50%	surface (25%)	circles (20%)	85%	surface (10%)	NA
low	class	55%	fast (10%)	NA	40%	glass (15%)	spots (10%)
low	clogging	0%	closet (10%)	quadrant (10%)	30%	clog (10%)	NA
low	coffee	40%	cough (15%)	cuff (10%)	45%	cough (10%)	NA
low	cold	40%	core (15%)	cord (10%)	45%	cool (20%)	NA
low	composure	0%	gorgeous (15%)	bridges (10%)	25%	closure (20%)	over (10%)
low	conceive	30%	see (35%)	concede (15%)	30%	concede (15%)	see (10%)
low	continued	0%	tea (65%)	NA	25%	continue (45%)	continuing (10%)
low	cooler	0%	cool (50%)	glue (10%)	60%	cool (10%)	NA
low	cop	0%	cross (30%)	crossing (10%)	15%	fuck (10%)	NA
low	coping	0%	coconut (15%)	focus (10%)	65%	culprit (10%)	NA
low	creek	30%	cooking (15%)	cricket (10%)	70%	creep (10%)	NA
low	cub	0%	cup (15%)	NA	25%	cup (45%)	come (10%)
low	cube	45%	cubed (20%)	IQ (10%)	65%	cute (20%)	queue (10%)
low	dairy	50%	dirty (10%)	scary (10%)	40%	daring (35%)	scary (20%)
low	delayed	15%	blade (30%)	blame (10%)	25%	blade (35%)	glade (20%)
low	delegate	10%	gully (10%)	NA	55%	delicate (10%)	NA
low	depend	0%	ten (25%)	pen (15%)	50%	pen (25%)	NA
low	dishwasher	60%	dishwashing (15%)	dishwashes (10%)	95%	dishwash (5%)	NA
low	displays	15%	display (15%)	serious (15%)	55%	plays (10%)	NA
low	disrupted	40%	disruptive (35%)	NA	35%	disruptive (50%)	constructive (10%)
low	dock	40%	dog (20%)	doc (10%)	45%	dog (10%)	gawk (10%)
low	dose	40%	ghost (30%)	gross (10%)	55%	ghost (15%)	NA
low	dual	0%	pool (45%)	cool (20%)	35%	cool (25%)	tool (25%)
low	eagle	20%	legal (20%)	able (10%)	50%	angle (10%)	NA

low	ethic	30%	effect (10%)	fake (10%)	80%	epic (10%)	NA
low	faint	25%	fate (35%)	fake (15%)	50%	fate (30%)	face (10%)
low	fang	45%	van (20%)	than (10%)	25%	bang (25%)	fan (10%)
low	fashion	75%	facism (10%)	NA	45%	fascist (20%)	NA
low	faster	35%	festive (20%)	fester (10%)	65%	fester (25%)	NA
low	feed	20%	fear (15%)	feel (10%)	25%	feet (30%)	fee (10%)
low	fiery	20%	fire (65%)	firing (10%)	35%	fire (25%)	firing (10%)
low	fleck	70%	reflect (10%)	NA	65%	flex (20%)	NA
low	fraud	10%	frog (15%)	garage (10%)	60%	frog (30%)	NA
low	fully	0%	full (45%)	fold (10%)	0%	pull (15%)	poor (10%)
low	fuse	30%	fume (15%)	sear (15%)	70%	feud (10%)	NA
low	gem	30%	jam (30%)	jab (10%)	70%	gym (10%)	NA
low	gender	10%	tender (30%)	kinder (20%)	75%	gentle (10%)	NA
low	genetic	10%	snow (20%)	index (10%)	65%	medic (10%)	phonetic (10%)
low	globe	0%	clove (10%)	grove (10%)	10%	gold (10%)	rolled (10%)
low	graph	45%	grab (10%)	NA	50%	grab (15%)	graft (10%)
low	grazing	30%	crazy (35%)	grazer (15%)	50%	raisin (20%)	crazy (10%)
low	guarantees	0%	guarantee (45%)	tea (15%)	55%	guarenteed (15%)	guarantee (15%)
low	hack	0%	pack (55%)	cat (10%)	55%	pack (25%)	NA
low	having	0%	hammock (20%)	seven (20%)	10%	heaven (30%)	NA
low	havoc	0%	adult (15%)	apple (10%)	55%	habit (30%)	NA
low	hiss	10%	gist (20%)	kiss (20%)	50%	this (15%)	kiss (10%)
low	hood	0%	purse (15%)	close (10%)	20%	foot (25%)	put (25%)
low	hooky	10%	cookie (35%)	NA	30%	cookie (25%)	bookie (10%)
low	hoop	0%	deep (10%)	teeth (10%)	10%	poop (15%)	NA
low	hovering	0%	hover (20%)	husband (10%)	40%	covering (15%)	NA
low	interface	0%	face (75%)	NA	80%	face (10%)	NA
low	issues	75%	tissues (10%)	NA	80%	tissues (10%)	NA
low	keen	40%	key (35%)	keep (10%)	60%	key (25%)	king (10%)
low	king	20%	cave (15%)	pain (10%)	45%	kin (15%)	NA
low	labor	0%	serious (20%)	maple (10%)	45%	river (10%)	NA
low	lash	30%	clash (15%)	flash (15%)	30%	laugh (20%)	clash (10%)
low	leaking	0%	briefing (10%)	lethal (10%)	35%	lethal (20%)	NA
low	led	15%	past (10%)	NA	75%	bled (20%)	NA
low	lethal	25%	refill (10%)	roof (10%)	40%	roof (10%)	NA
low	lope	0%	mission (10%)	serial (10%)	10%	elope (15%)	bolt (10%)
low	magnifies	0%	magnify (55%)	NA	15%	magnify (55%)	magnified (15%)
low	mashing	0%	map (10%)	math (10%)	40%	matching (25%)	nothing (10%)
low	monitors	0%	she (10%)	NA	10%	monitor (10%)	NA
low	mood	0%	move (50%)	smooth (15%)	15%	move (15%)	news (10%)
low	mug	0%	man (10%)	NA	20%	month (10%)	munch (10%)

low	muzzle	75%	muzzled (10%)	NA	70%	muzzled (15%)	mussel (10%)
low	myth	0%	enough (10%)	magnificent (10%)	45%	niff (10%)	sniff (10%)
low	nightlife	0%	nightlight (20%)	backlight (10%)	45%	nightlight (30%)	NA
low	odor	0%	show (20%)	showing (10%)	30%	holder (15%)	colder (10%)
low	odyssey	0%	buzzing (10%)	fussy (10%)	15%	artsy (15%)	NA
low	offensive	30%	expensive (25%)	senses (10%)	75%	defensive (10%)	NA
low	overcook	60%	overcooked (10%)	NA	60%	overcooked (30%)	NA
low	page	45%	paid (15%)	NA	55%	paid (15%)	cage (10%)
low	palette	0%	balance (10%)	ellis (10%)	30%	pellet (15%)	hell (10%)
low	paw	10%	cough (25%)	fog (10%)	30%	cough (15%)	aw (10%)
low	peg	10%	bag (25%)	stag (15%)	35%	beg (10%)	pegged (10%)
low	pill	30%	kill (15%)	till (10%)	80%	kill (15%)	NA
low	pity	20%	paid (25%)	pay (10%)	75%	date (10%)	NA
low	pocket	25%	buckle (10%)	talkative (10%)	30%	coffee (10%)	NA
low	polity	0%	quality (20%)	apologies (10%)	0%	quality (55%)	party (10%)
low	pony	60%	point (10%)	pulley (10%)	60%	point (10%)	NA
low	pose	0%	pole (15%)	poles (15%)	20%	poles (20%)	pole (10%)
low	preface	10%	crevice (30%)	NA	65%	surface (10%)	NA
low	puddle	50%	cuddle (20%)	NA	50%	cuddle (15%)	pedal (15%)
low	puppy	55%	bucket (15%)	cooking (10%)	75%	puffy (10%)	NA
low	qualifies	0%	qualify (95%)	NA	15%	qualify (55%)	qualified (25%)
low	raffle	0%	baffle (15%)	after (10%)	25%	baffle (15%)	NA
low	raid	0%	grade (70%)	NA	55%	rave (15%)	brave (10%)
low	razor	35%	lizard (20%)	laser (15%)	45%	lizard (10%)	risen (10%)
low	required	0%	quiet (20%)	choir (15%)	0%	choir (55%)	acquire (15%)
low	resent	15%	represent (15%)	NA	10%	present (55%)	NA
low	resin	10%	risen (20%)	driven (10%)	15%	reserve (15%)	risen (15%)
low	revel	0%	trouble (20%)	NA	0%	oval (10%)	NA
low	revive	0%	goodbye (40%)	buy (20%)	0%	survive (20%)	bye (10%)
low	ribbon	0%	complaint (15%)	lift (10%)	30%	driven (10%)	given (10%)
low	rigor	0%	trigger (15%)	scissor (10%)	35%	trigger (35%)	bigger (10%)
low	rocker	0%	rock (15%)	abrupt (10%)	15%	rock (35%)	NA
low	sauce	10%	socks (25%)	cross (10%)	70%	socks (20%)	NA
low	saving	20%	save (15%)	savior (15%)	70%	stable (10%)	NA
low	seizure	10%	susan (30%)	loser (10%)	75%	caesar (10%)	NA
low	selective	50%	selected (35%)	NA	75%	selected (10%)	NA
low	shadows	70%	shadow (25%)	NA	80%	shadow (10%)	NA
low	shady	10%	stadium (25%)	NA	40%	shaving (15%)	city (10%)
low	shale	0%	share (20%)	shave (15%)	20%	chill (25%)	fail (10%)
low	siege	0%	seige (20%)	seizure (15%)	25%	seige (35%)	sees (20%)
low	soot	30%	sit (15%)	foot (10%)	50%	foot (25%)	suit (10%)

low	soothe	75%	sued (10%)	NA	65%	sued (10%)	sues (10%)
low	suing	45%	sewage (10%)	sewing (10%)	45%	sewer (15%)	ceiling (10%)
low	suits	50%	seats (20%)	NA	75%	soups (10%)	NA
low	supper	35%	suburb (15%)	stubborn (10%)	65%	suffer (15%)	separate (10%)
low	supplied	0%	reply (15%)	fly (10%)	20%	supply (35%)	ply (10%)
low	sure	10%	pure (20%)	unsure (20%)	40%	tour (15%)	NA
low	tea	10%	key (25%)	keep (10%)	60%	key (15%)	NA
low	terrify	25%	clarify (15%)	horrify (10%)	50%	terrified (30%)	NA
low	thatch	15%	sash (10%)	NA	25%	fat (25%)	batch (10%)
low	there	30%	bear (45%)	NA	35%	bear (30%)	dare (10%)
low	these	10%	bees (45%)	bee (15%)	15%	bees (55%)	please (10%)
low	thin	30%	fin (30%)	dim (10%)	55%	fin (25%)	sin (10%)
low	thought	20%	fought (25%)	fault (15%)	40%	fought (25%)	NA
low	troop	45%	truth (15%)	NA	30%	truth (15%)	coop (10%)
low	trough	0%	cough (50%)	prof (10%)	0%	cough (60%)	cross (10%)
low	uneven	0%	illegal (10%)	leader (10%)	20%	anemic (10%)	NA
low	unified	0%	unify (35%)	five (10%)	45%	unify (35%)	beautify (15%)
low	varied	0%	carries (20%)	carry (20%)	15%	very (45%)	fairy (10%)
low	vault	15%	bald (30%)	NA	25%	fault (15%)	bolt (10%)
low	volley	10%	quality (15%)	balling (10%)	10%	falling (20%)	folly (10%)
low	watch	0%	walk (35%)	wash (30%)	40%	wash (50%)	NA
low	weaving	35%	leaving (10%)	NA	35%	leaving (25%)	grieving (10%)
low	widowed	15%	widow (20%)	will (20%)	40%	widow (35%)	NA
NA	NA	NA%	NA (NA%)	NA	NA%	NA (NA%)	NA

Table A.1: Stimulus accuracies and two most common errors from Canadian and American listeners in Experiments 0a and 0b. Only responses given by two or more participants are included in the above data, and in cases where only a single error meets this criterion (or where responses are divided completely between the correct response and a single incorrect response) the secondary error is listed as missing (NA). Items with no errors or no incorrect responses meeting the above criterion are excluded from this list.

Chapter 3: Lexical contrast perception
Supplementary tables and figures

Target phone accuracy in Exp. 1a/b (CV)

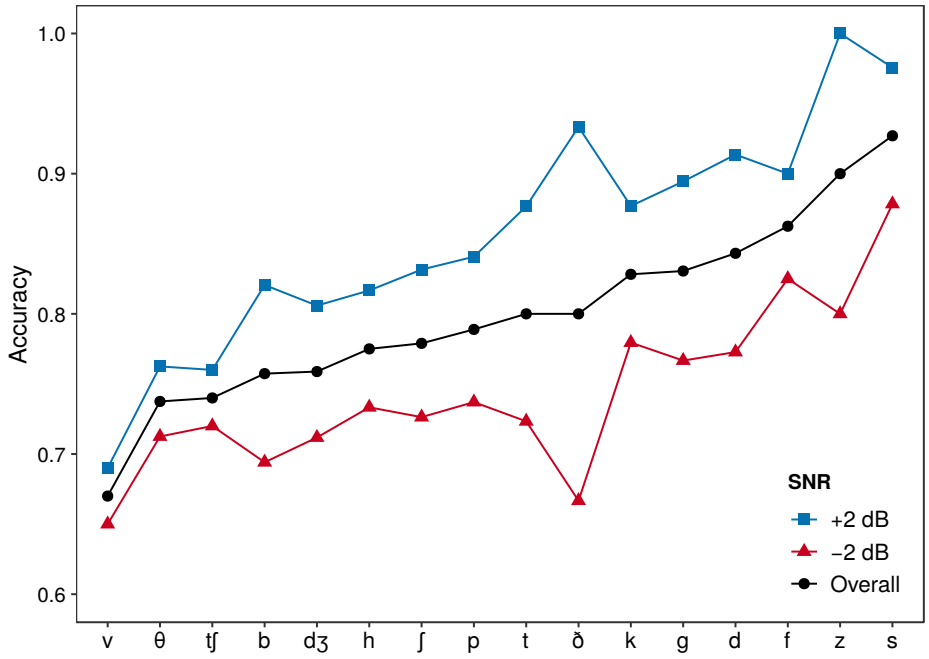

Figure A.1: Target phone accuracies in CV position in Exp. 1a, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

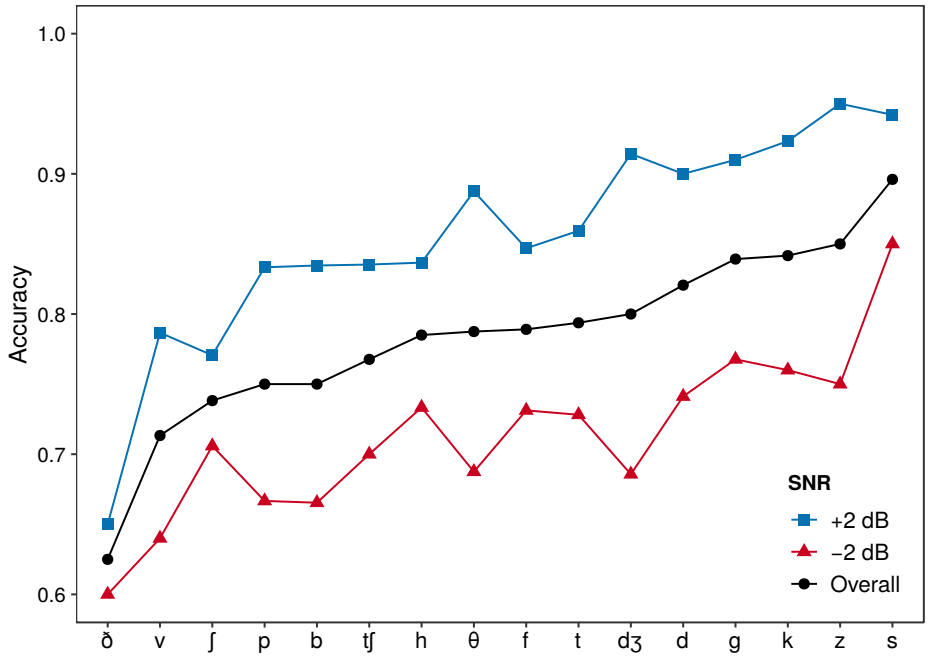

Figure A.2: Target phone accuracies in CV position in Exp. 1b, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

Target phone accuracy in Exp. 1a/b (VCV)

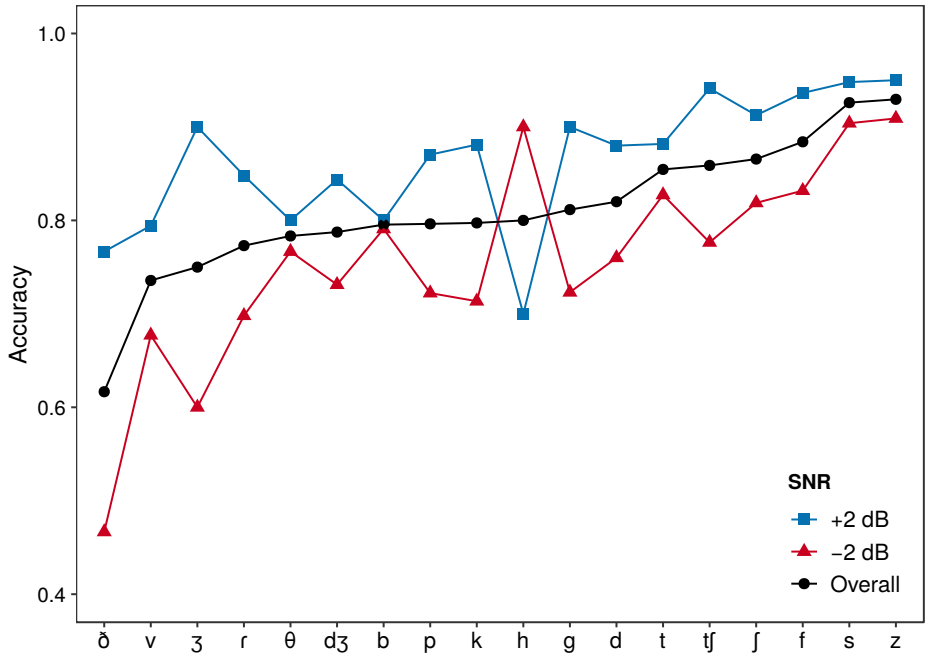

Figure A.3: Target phone accuracies in VCV position in Exp. 1a, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

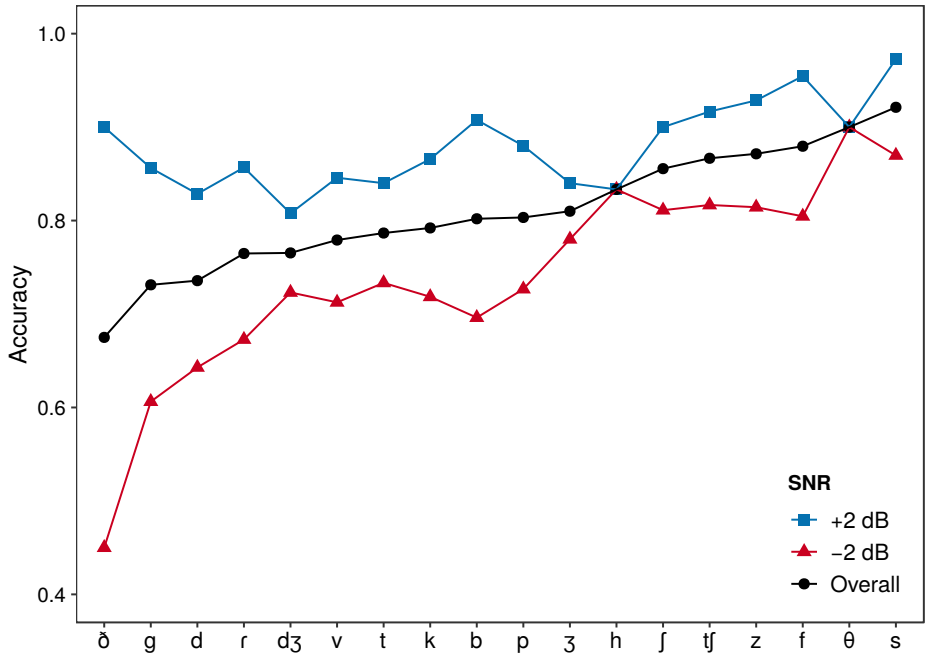

Figure A.4: Target phone accuracies in VCV position in Exp. 1b, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

Target phone accuracy in Exp. 1a/b (VC)

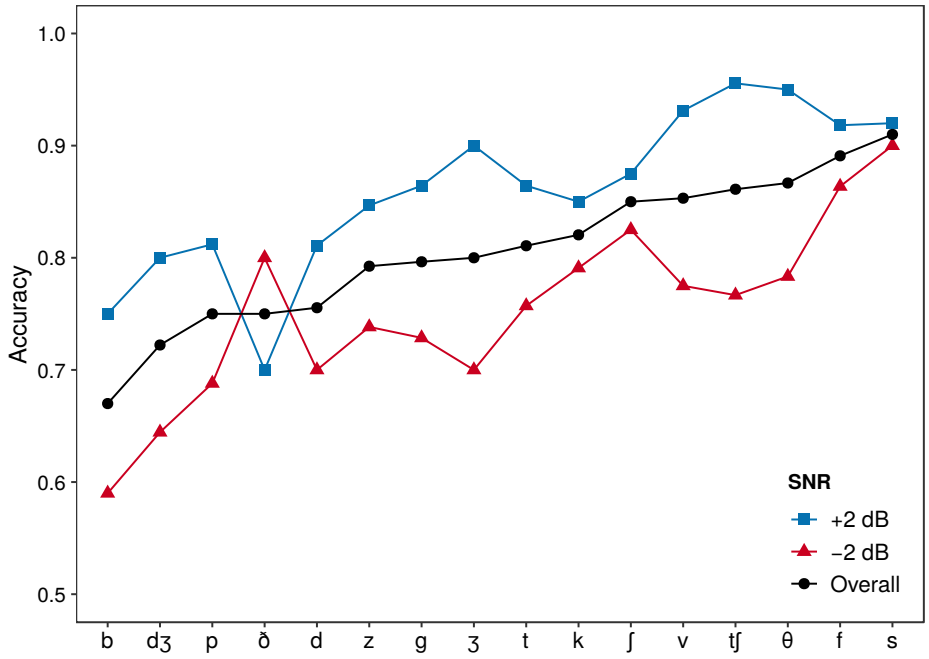

Figure A.5: Target phone accuracies in VC position in Exp. 1a, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

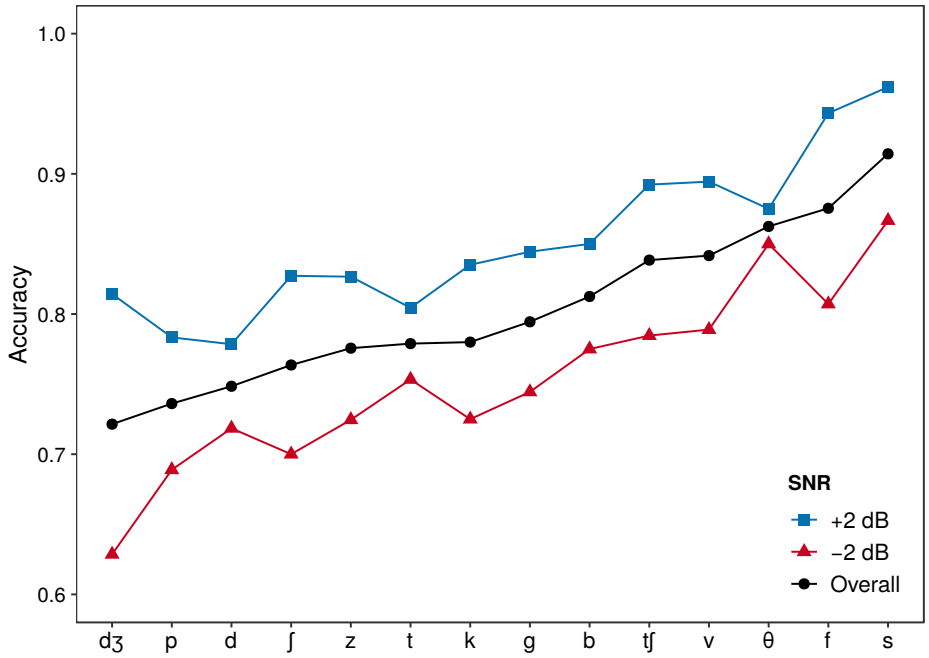

Figure A.6: Target phone accuracies in VC position in Exp. 1b, presented in ranked order overall (black circle) and in matched order for each SNR (blue square = +2 dB, red triangle = −2 dB).

Target feature accuracy in Exp. 1a (CV)

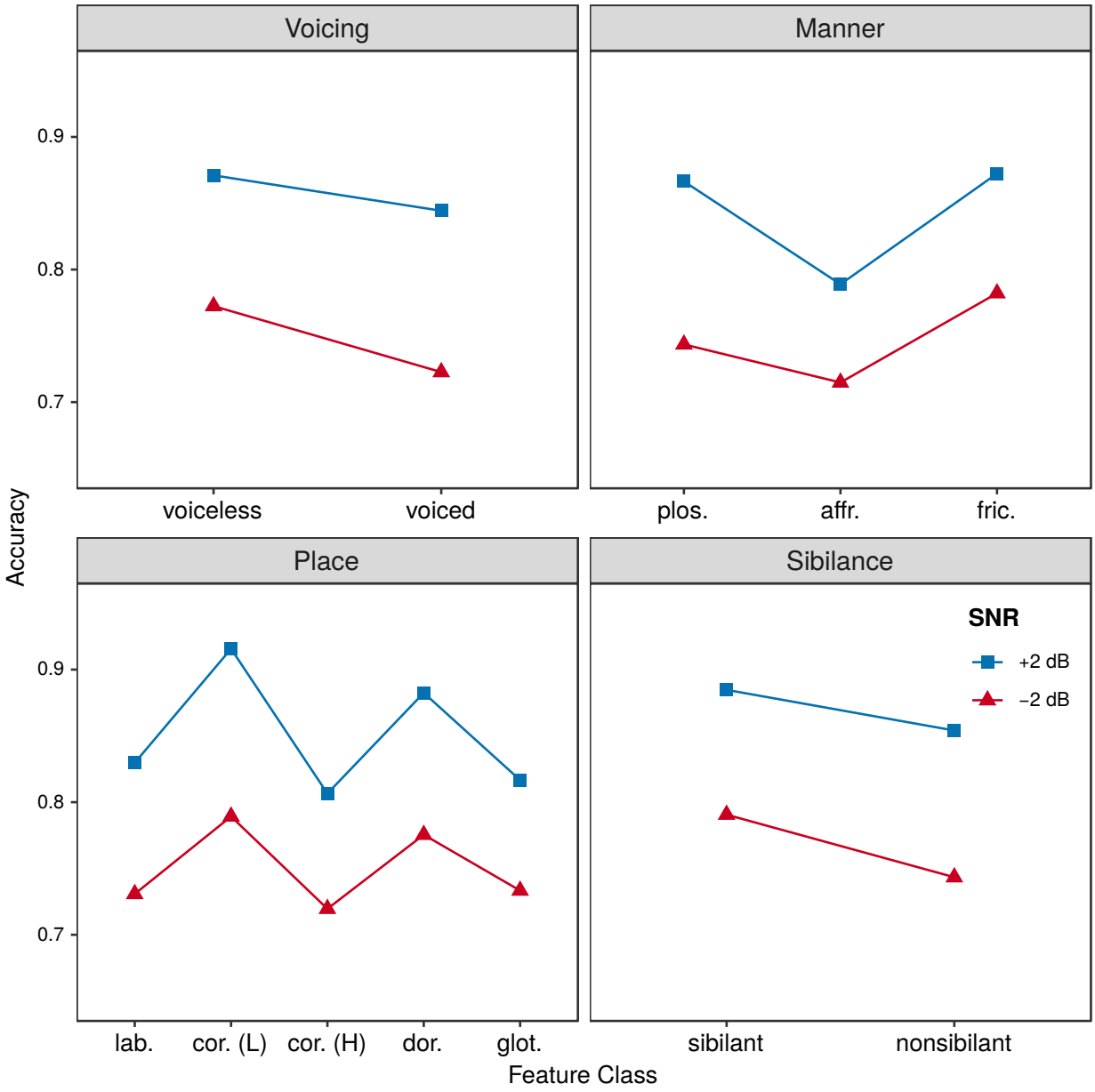

Figure A.7: Target feature accuracies by SNR in CV position in Experiment 1a.

Target feature accuracy in Exp. 1b (CV)

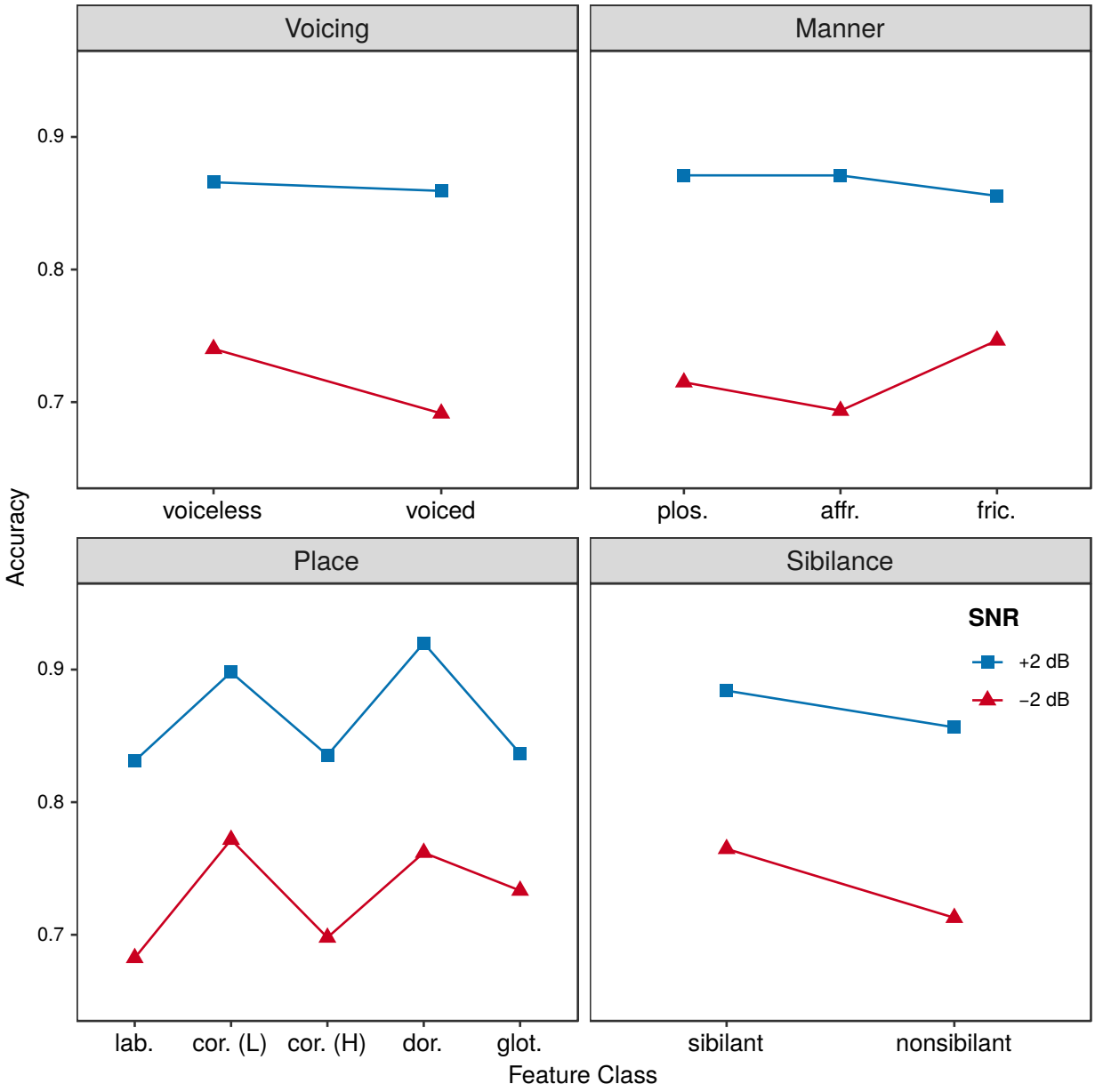

Figure A.8: Target feature accuracies by SNR in CV position in Experiment 1b.

Target feature accuracy in Exp. 1a (VCV)

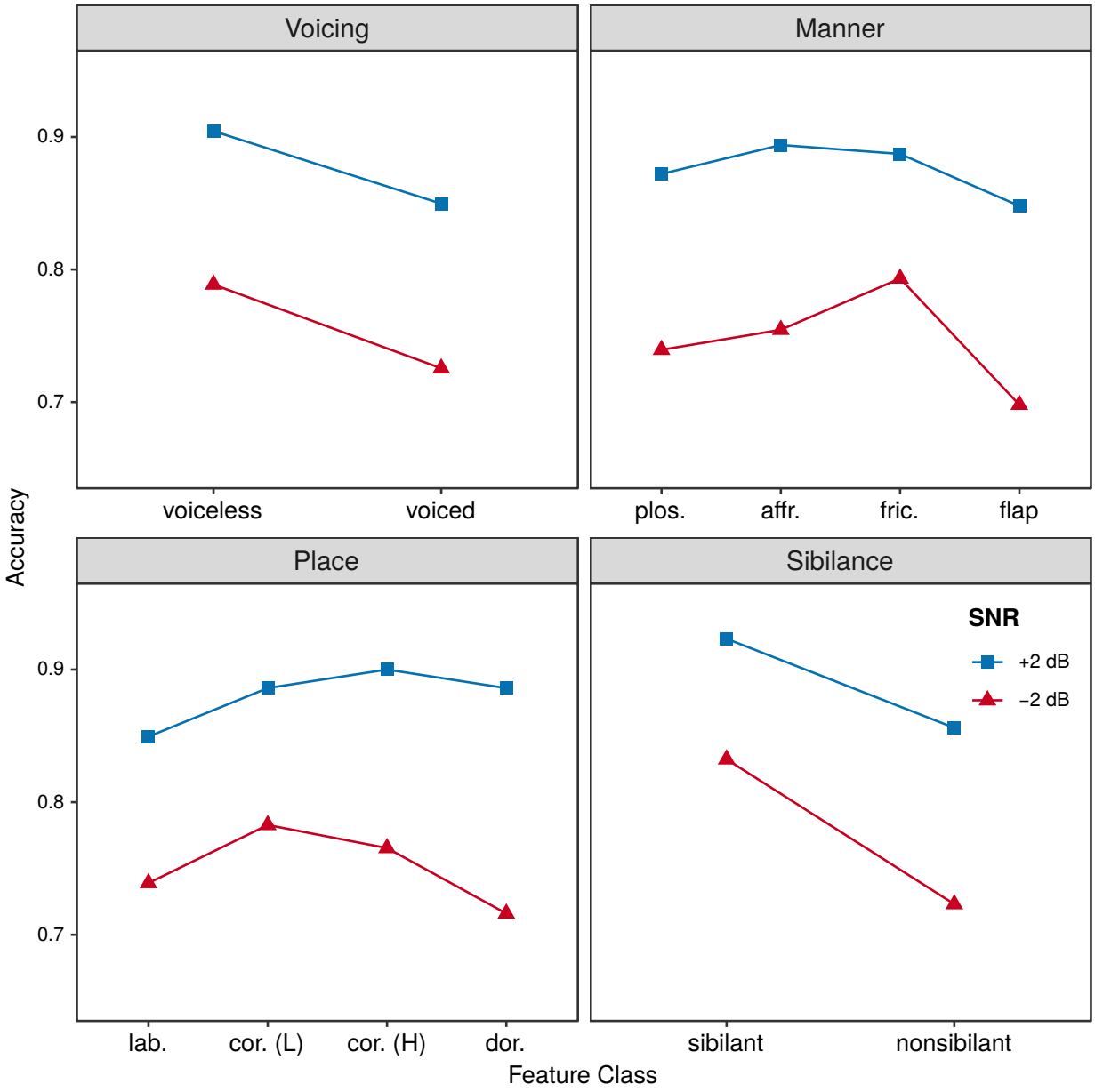

Figure A.9: Target feature accuracies by SNR in VCV position in Experiment 1a.

Target feature accuracy in Exp. 1b (VCV)

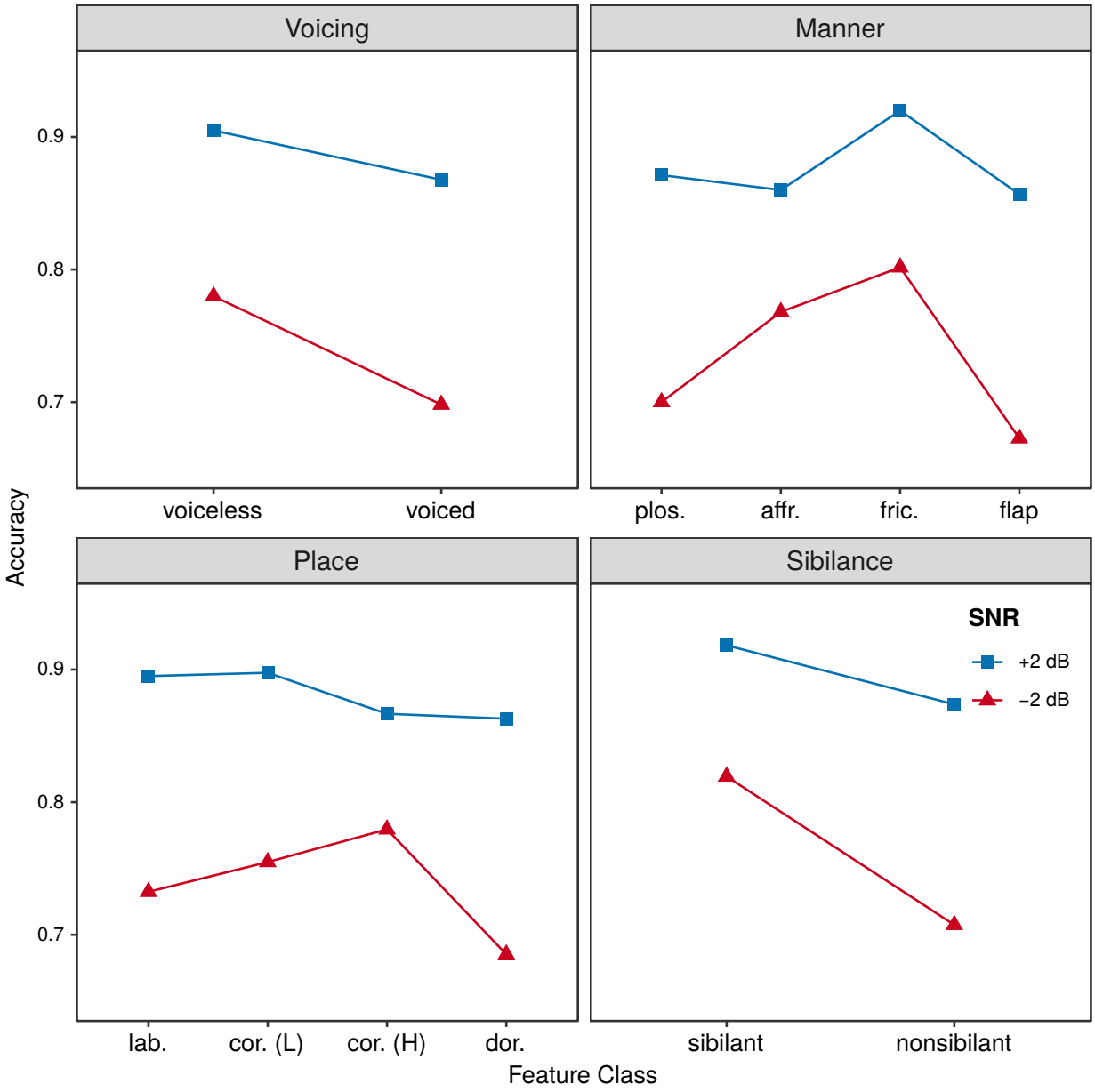

Figure A.10: Target feature accuracies by SNR in VCV position in Experiment 1b.

Target feature accuracy in Exp. 1a (VC)

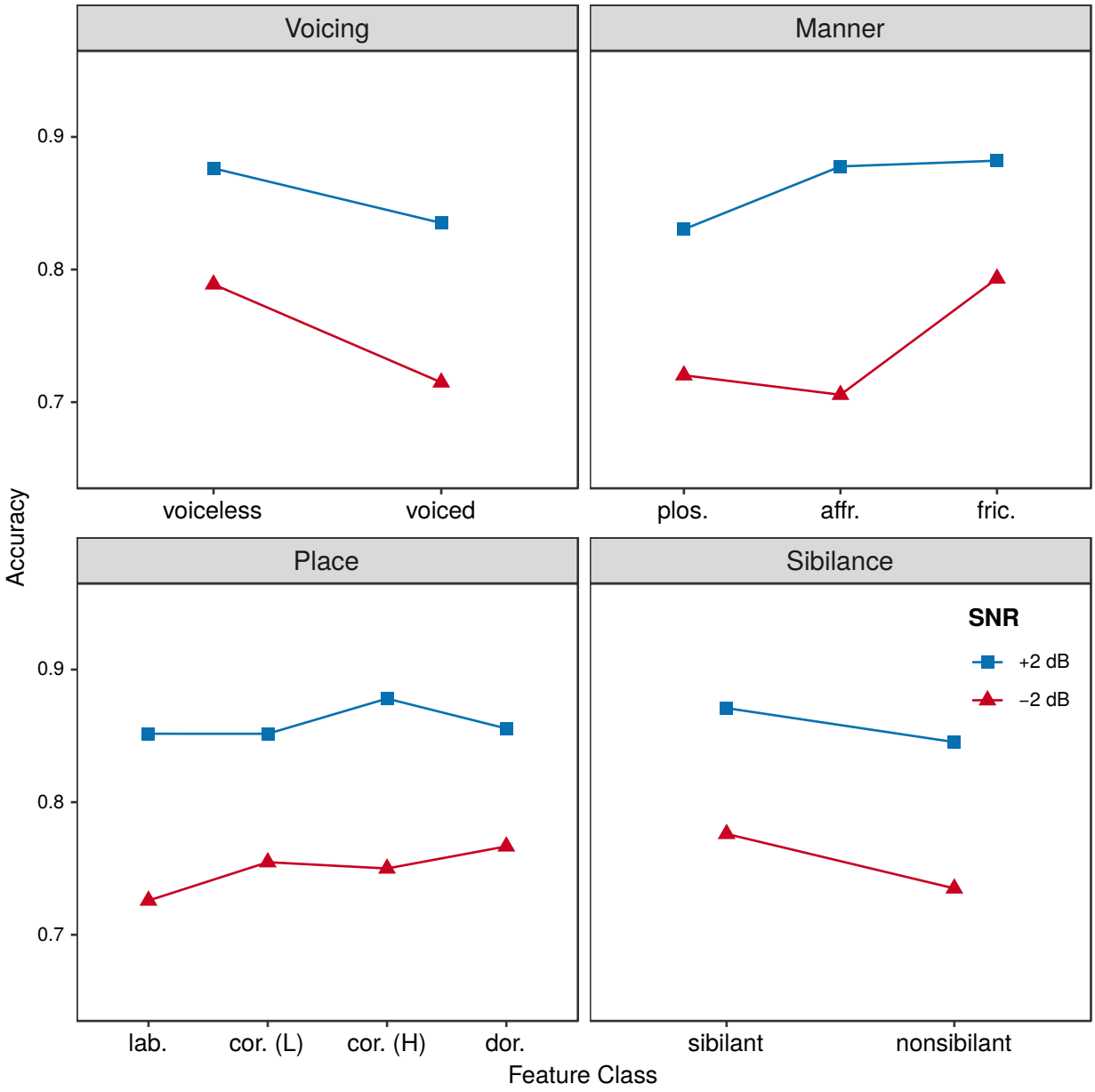

Figure A.11: Target feature accuracies by SNR in VC position in Experiment 1a.

Target feature accuracy in Exp. 1b (VC)

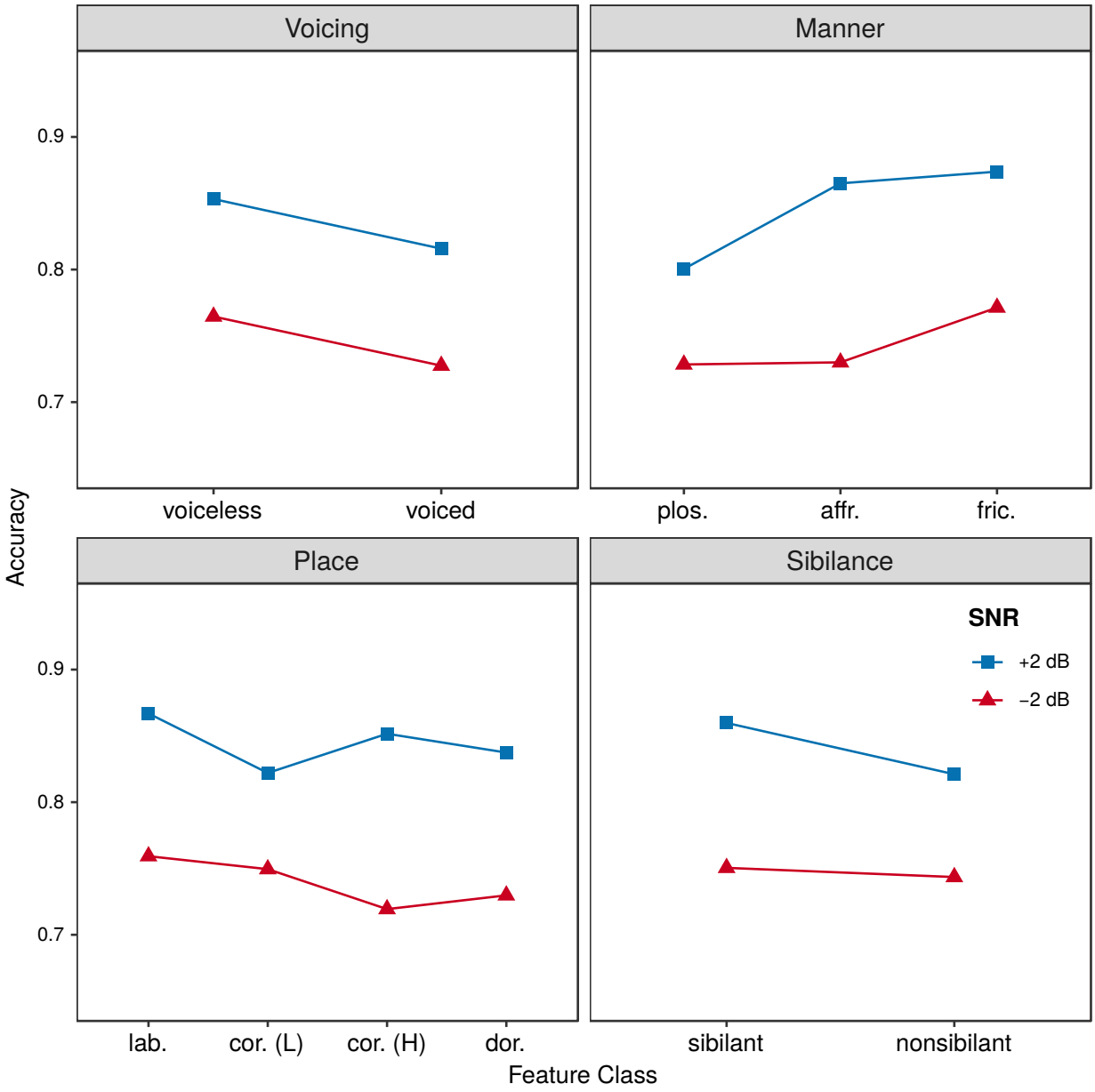

Figure A.12: Target feature accuracies by SNR in VC position in Experiment 1b.

Target feature accuracies by length and frequency in Exp. 1a (CV)

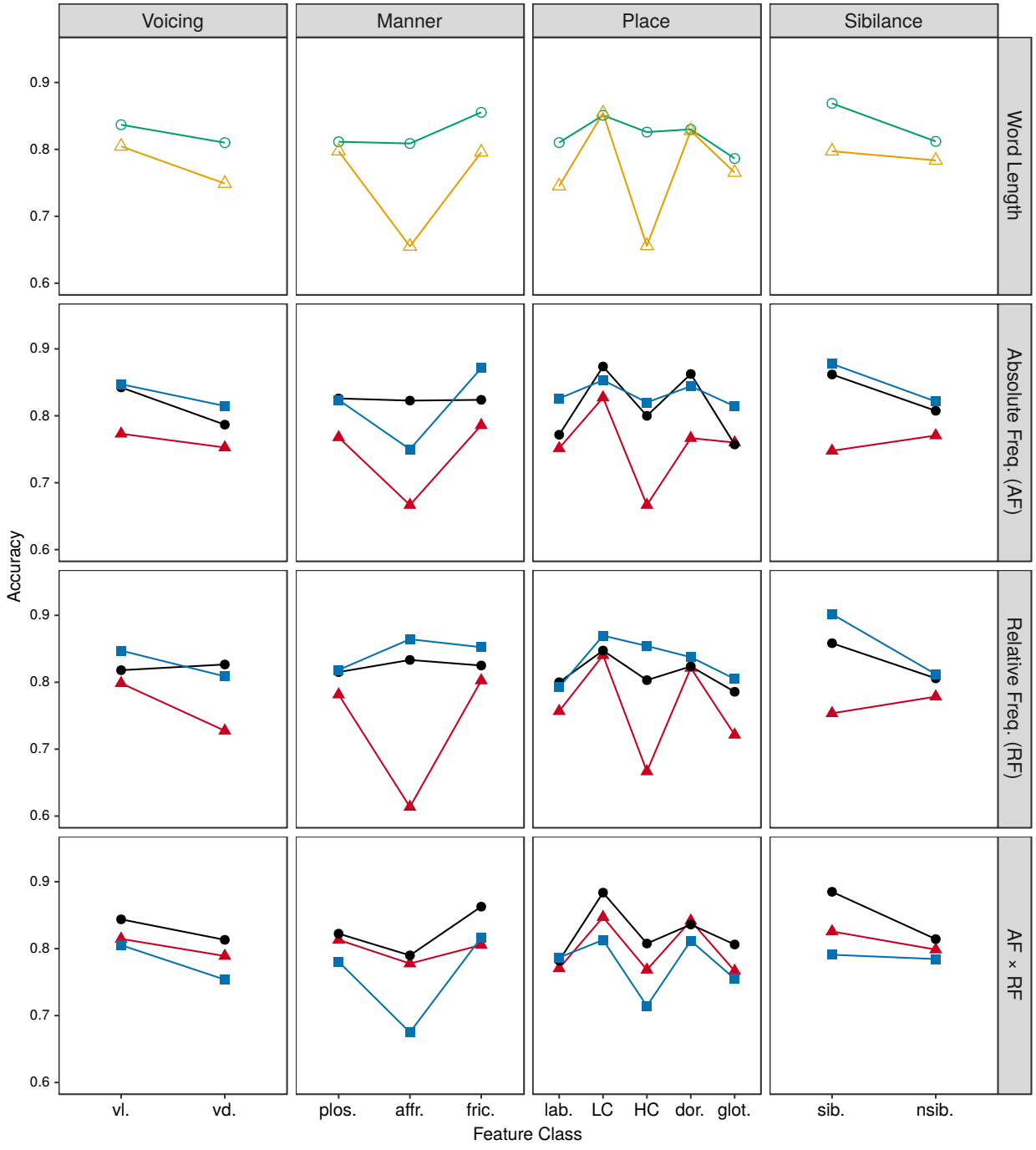

Figure A.13: Target feature accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in CV position in Experiment 1a. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Target feature accuracies by length and frequency in Exp. 1b (CV)

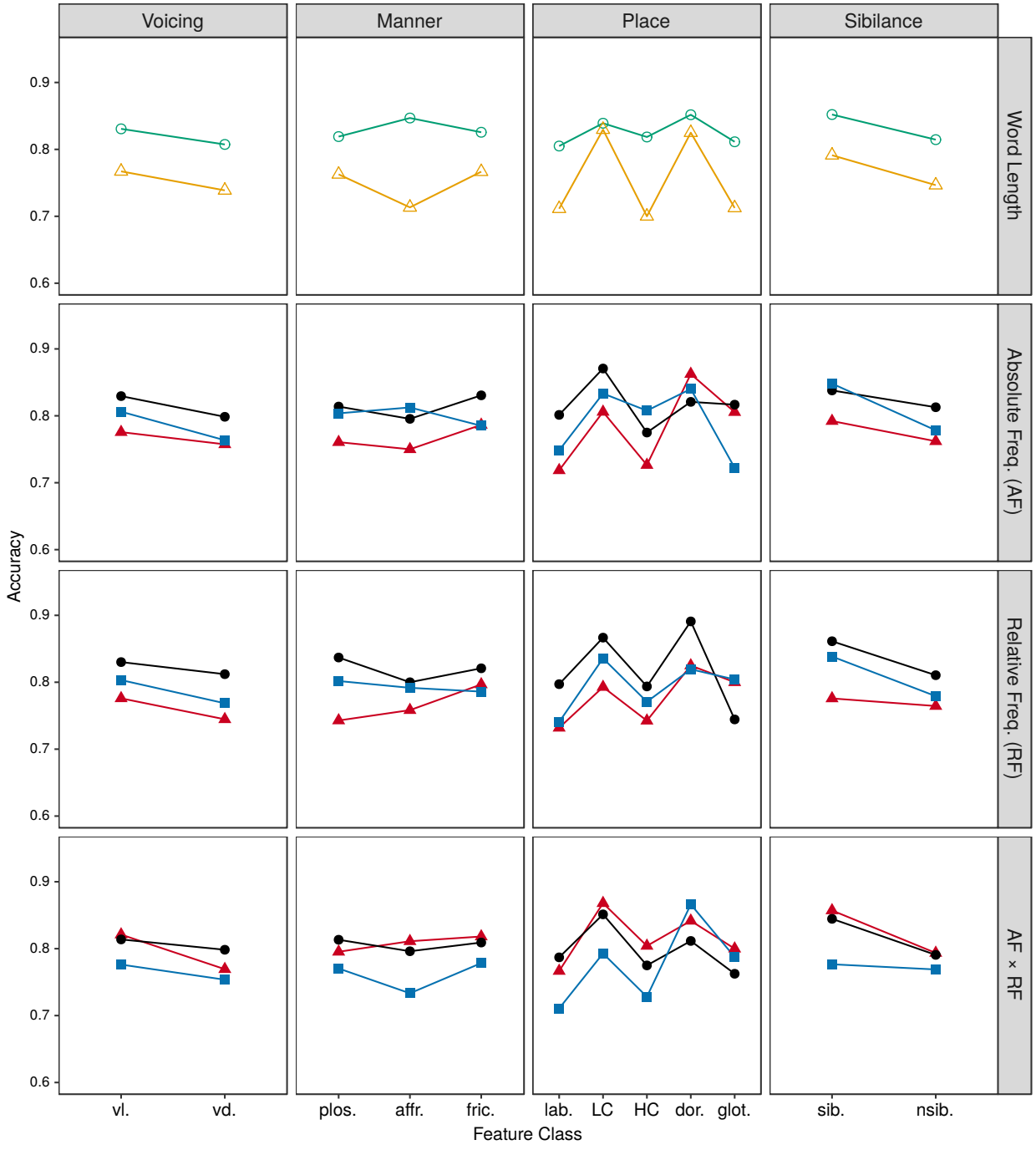

Figure A.14: Target feature accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in CV position in Experiment 1b. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Target feature accuracies by length and frequency in Exp. 1a (VCV)

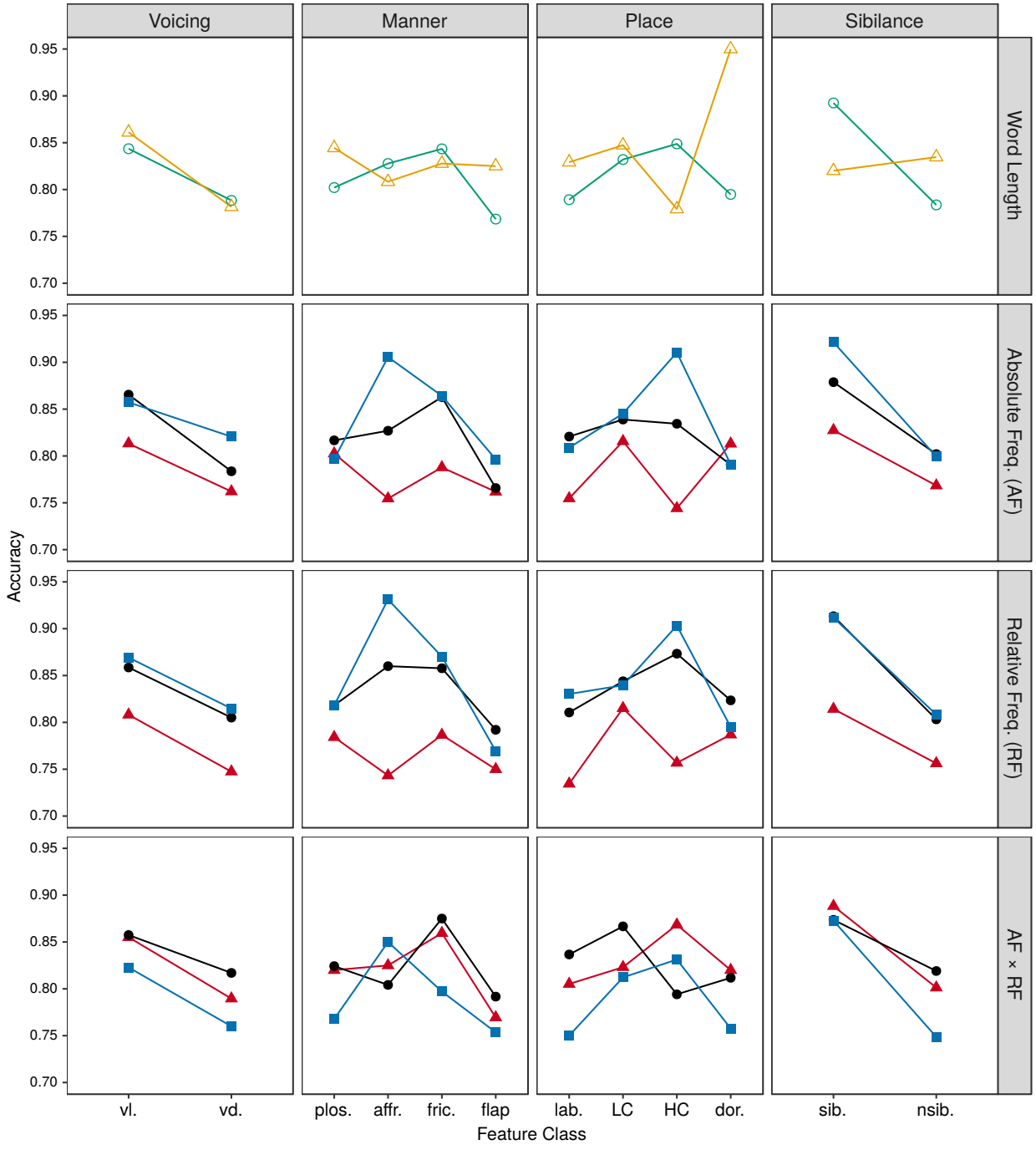

Figure A.15: Target feature accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VCV position in Experiment 1a. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Target feature accuracies by length and frequency in Exp. 1b (VCV)

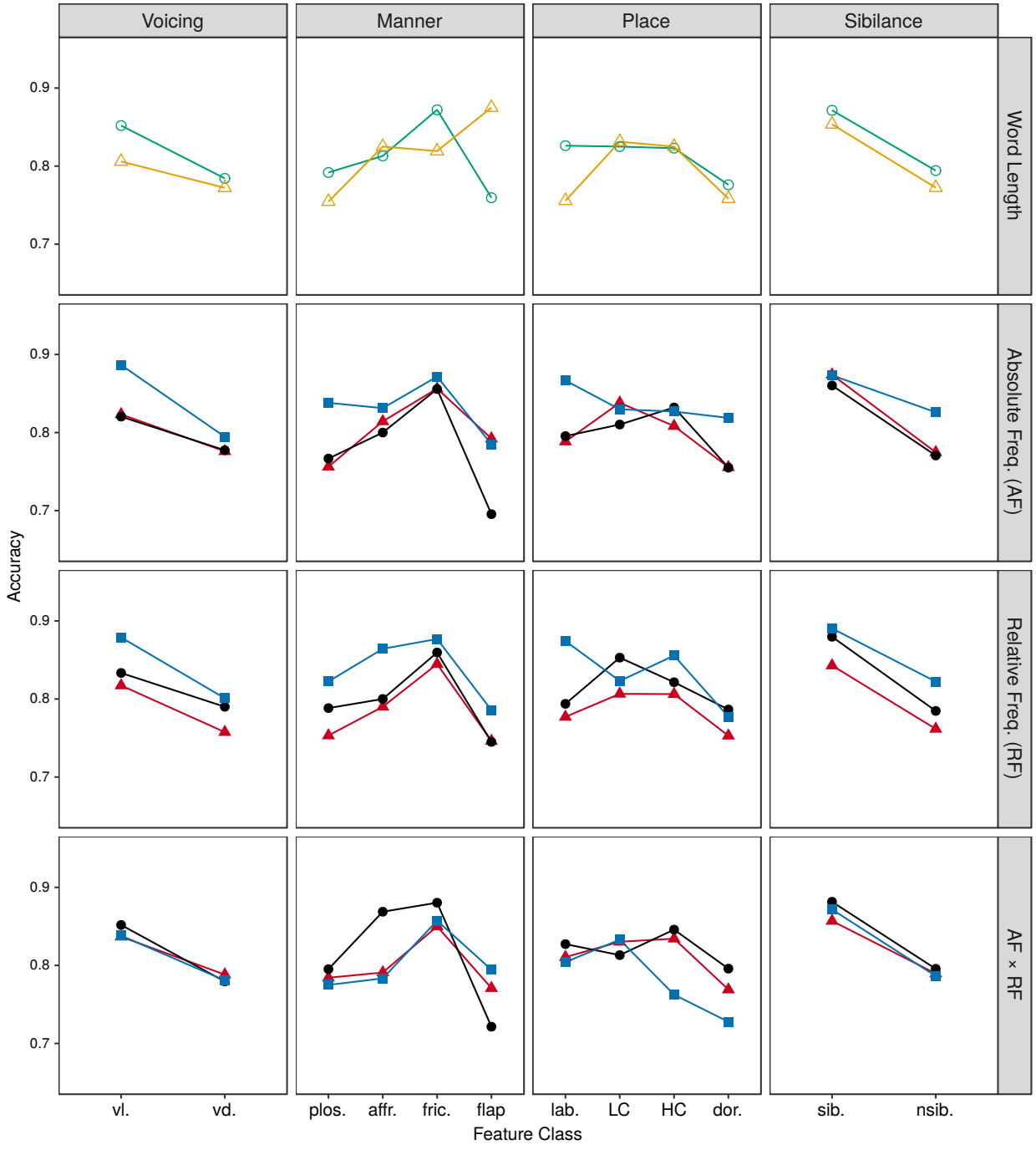

Figure A.16: Target feature accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VCV position in Experiment 1b. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Target feature accuracies by length and frequency in Exp. 1a (VC)

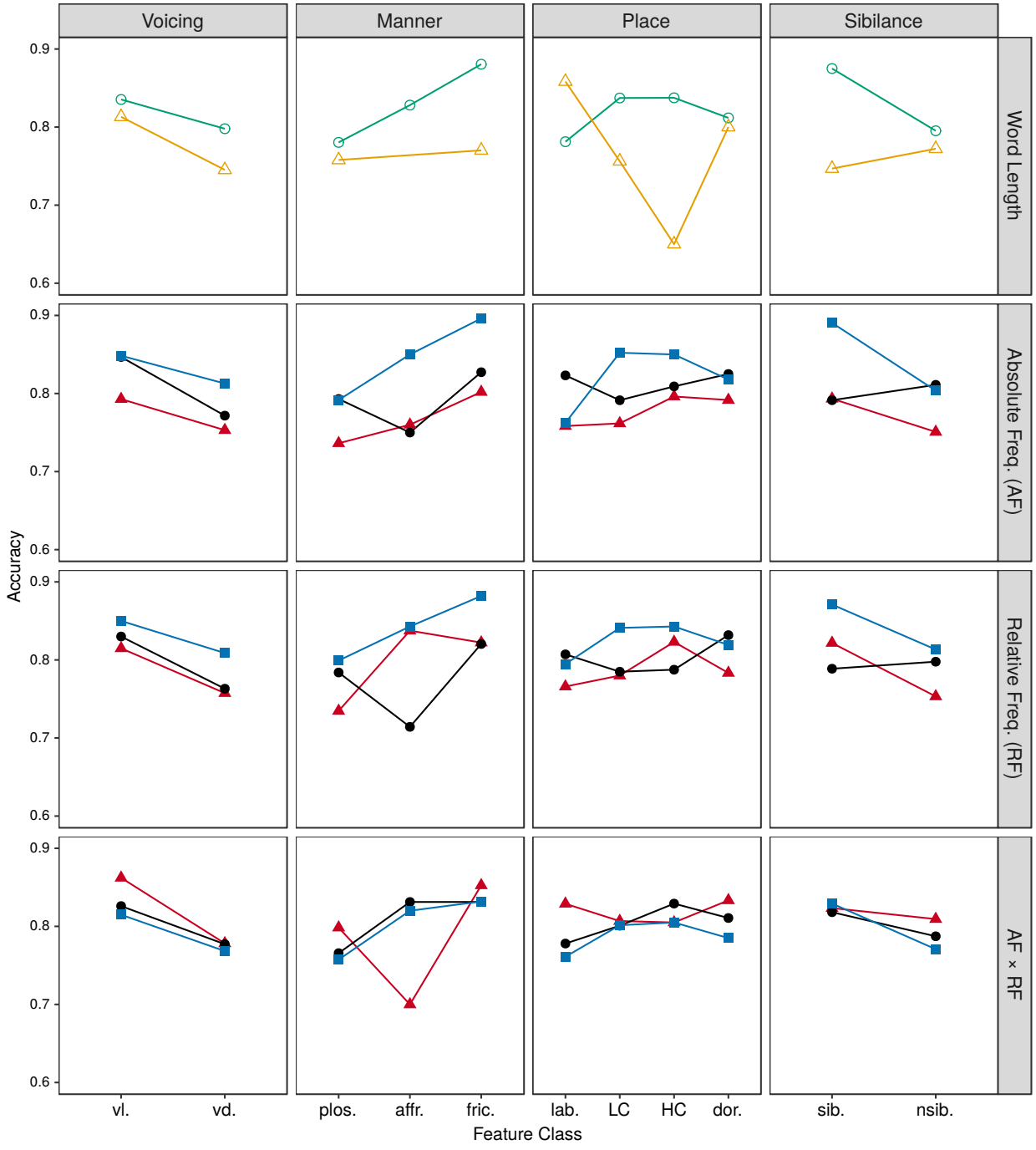

Figure A.17: Target feature accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VC position in Experiment 1a. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Target feature accuracies by length and frequency in Exp. 1b (VC)

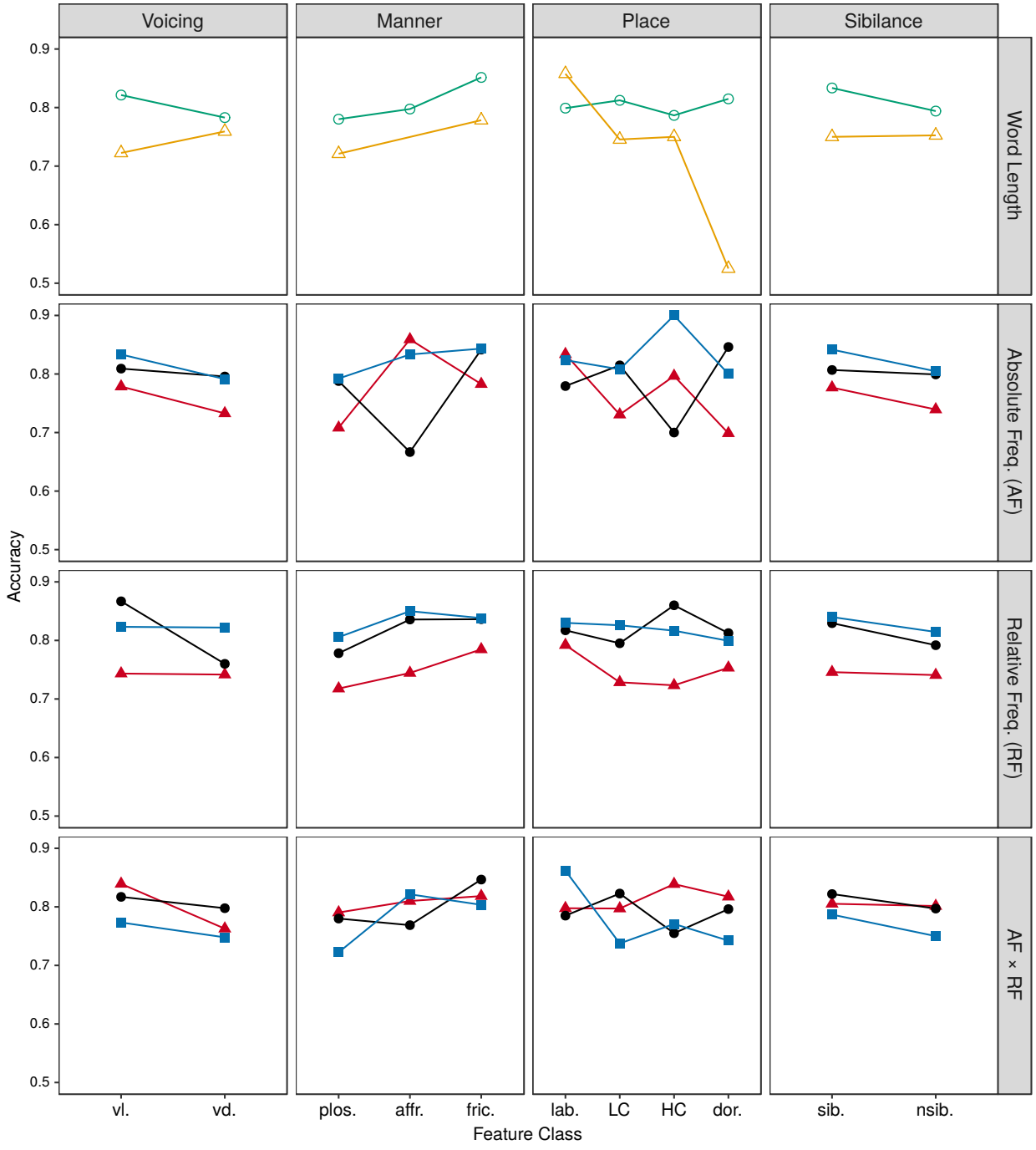

Figure A.18: Target feature accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VC position in Experiment 1b. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Word-initial contrast distributions in Experiments 1a and 1b

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**h**	**v**	**ð**	**z**
p	3	3	4	1	3	0	1	5	0	3	3	0	0	0	1
	t	4	4	2	2	0	3	1	1	3	3	3	1	0	0
		k	4	6	4	2	1	1	0	6	1	6	1	0	0
			b	3	2	0	1	3	5	4	1	2	1	0	0
				d	0	1	2	2	0	1	0	3	0	1	0
					g	1	1	1	0	1	2	1	0	0	0
						ʧ	0	1	0	3	0	2	0	0	0
							ʤ	2	0	4	0	0	1	1	0
								f	1	1	1	0	1	0	0
									θ	1	0	0	0	0	0
										s	6	2	2	0	0
											ʃ	1	0	1	0
												h	3	0	1
													v	0	0
														ð	0

Table A.2: Distribution of CV contrasts in minimal pair stimuli in Experiment 1a.

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**h**	**v**	**ð**	**z**
p	1	4	3	1	2	5	2	6	1	2	0	3	0	0	0
	t	4	5	3	0	0	3	3	3	4	1	2	2	0	1
		k	6	2	0	1	0	2	0	3	1	4	2	1	0
			b	1	0	1	1	3	0	2	0	3	1	0	0
				d	0	2	1	3	0	2	1	1	0	0	0
					g	1	3	0	1	1	1	0	1	0	0
						ʧ	0	1	0	0	2	2	2	0	0
							ʤ	1	0	2	0	1	0	0	0
								f	0	6	3	2	2	0	0
									θ	2	0	1	0	0	0
										s	4	7	3	0	0
											ʃ	3	1	0	0
												h	0	0	1
													v	1	0
														ð	0

Table A.3: Distribution of CV contrasts in minimal pair stimuli in Experiment 1b.

Word-medial contrast distributions in Experiments 1a and 1b

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**h**	**v**	**ð**	**z**	**ʒ**	**ɾ**
p	1	8	0	1	2	0	1	5	0	1	1	0	1	0	1	0	5
	t	0	2	0	0	0	2	1	0	3	1	0	0	0	1	0	0
		k	0	0	1	5	1	3	0	4	3	0	3	1	2	0	6
			b	0	3	1	0	0	0	0	1	0	2	0	0	0	2
				d	0	1	0	1	0	2	0	0	0	0	0	0	0
					g	0	0	0	1	1	1	0	2	0	0	0	2
						ʧ	1	0	0	1	1	0	1	0	0	1	5
							ʤ	1	1	2	0	0	1	0	2	0	4
								f	1	2	1	0	5	0	0	0	2
									θ	0	0	0	0	0	0	0	0
										s	1	0	4	0	2	0	2
											ʃ	0	1	1	0	1	3
												h	1	0	0	0	0
													v	1	6	0	7
														ð	0	0	3
															z	1	7
																ʒ	0

Table A.4: Distribution of VCV contrasts in minimal pair stimuli in Experiment 1a.

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**h**	**v**	**ð**	**z**	**ʒ**	**ɾ**
p	0	6	3	2	3	1	0	2	1	2	2	0	4	0	2	0	2
	t	2	2	0	1	0	2	4	0	1	1	0	1	0	1	0	0
		k	5	0	3	1	1	3	0	4	0	0	2	0	3	1	7
			b	1	0	2	0	2	0	0	1	1	0	0	1	0	8
				d	1	1	0	0	0	0	0	0	0	0	1	1	0
					g	0	0	1	0	1	1	0	0	0	1	0	4
						ʧ	0	2	0	4	0	0	0	0	0	0	1
							ʤ	0	0	2	0	0	1	0	1	1	5
								f	0	1	1	0	4	0	1	0	1
									θ	0	0	0	0	0	0	0	1
										s	1	0	4	0	4	2	7
											ʃ	0	2	0	0	0	0
												h	2	0	0	0	0
													v	1	1	0	2
														ð	0	0	1
															z	0	5
																ʒ	0

Table A.5: Distribution of VCV contrasts in minimal pair stimuli in Experiment 1b.

Word-final contrast distributions in Experiments 1a and 1b

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**v**	**ð**	**z**	**ʒ**
p	6	1	2	1	2	4	0	2	0	2	0	1	1	3	0
	t	4	1	5	4	0	2	3	1	4	1	5	0	6	0
		k	1	4	1	1	1	1	1	2	1	1	0	3	0
			b	2	0	0	1	1	0	1	1	0	0	0	0
				d	0	0	1	2	0	3	3	1	1	31	1
					g	2	0	0	0	1	2	1	0	1	0
						ʧ	1	0	0	0	1	0	0	0	0
							ʤ	1	1	0	0	0	0	1	0
								f	0	0	0	0	0	1	0
									θ	1	1	0	0	1	0
										s	2	2	0	7	0
											ʃ	0	0	0	0
												v	0	5	0
													ð	0	0
														z	1

Table A.6: Distribution of VC contrasts in minimal pair stimuli in Experiment 1a.

	t	**k**	**b**	**d**	**g**	**ʧ**	**ʤ**	**f**	**θ**	**s**	**ʃ**	**v**	**ð**	**z**	**ʒ**
p	4	7	0	0	0	0	1	0	1	1	0	2	0	2	0
	t	7	0	11	2	3	1	2	1	2	2	3	0	7	0
		k	1	4	0	3	0	0	1	3	1	0	0	1	0
			b	0	0	0	0	2	0	1	0	0	0	0	0
				d	1	1	2	1	2	1	2	4	0	36	0
					g	1	1	0	0	1	1	1	0	1	0
						ʧ	1	2	0	0	1	0	0	1	0
							ʤ	0	0	1	0	0	0	0	0
								f	0	2	1	3	0	1	0
									θ	2	1	0	0	0	0
										s	2	1	0	4	0
											ʃ	0	0	0	0
												v	0	4	0
													ð	0	0
														z	0

Table A.7: Distribution of VC contrasts in minimal pair stimuli in Experiment 1b.

Contrast accuracies in Experiment 1a (CV)

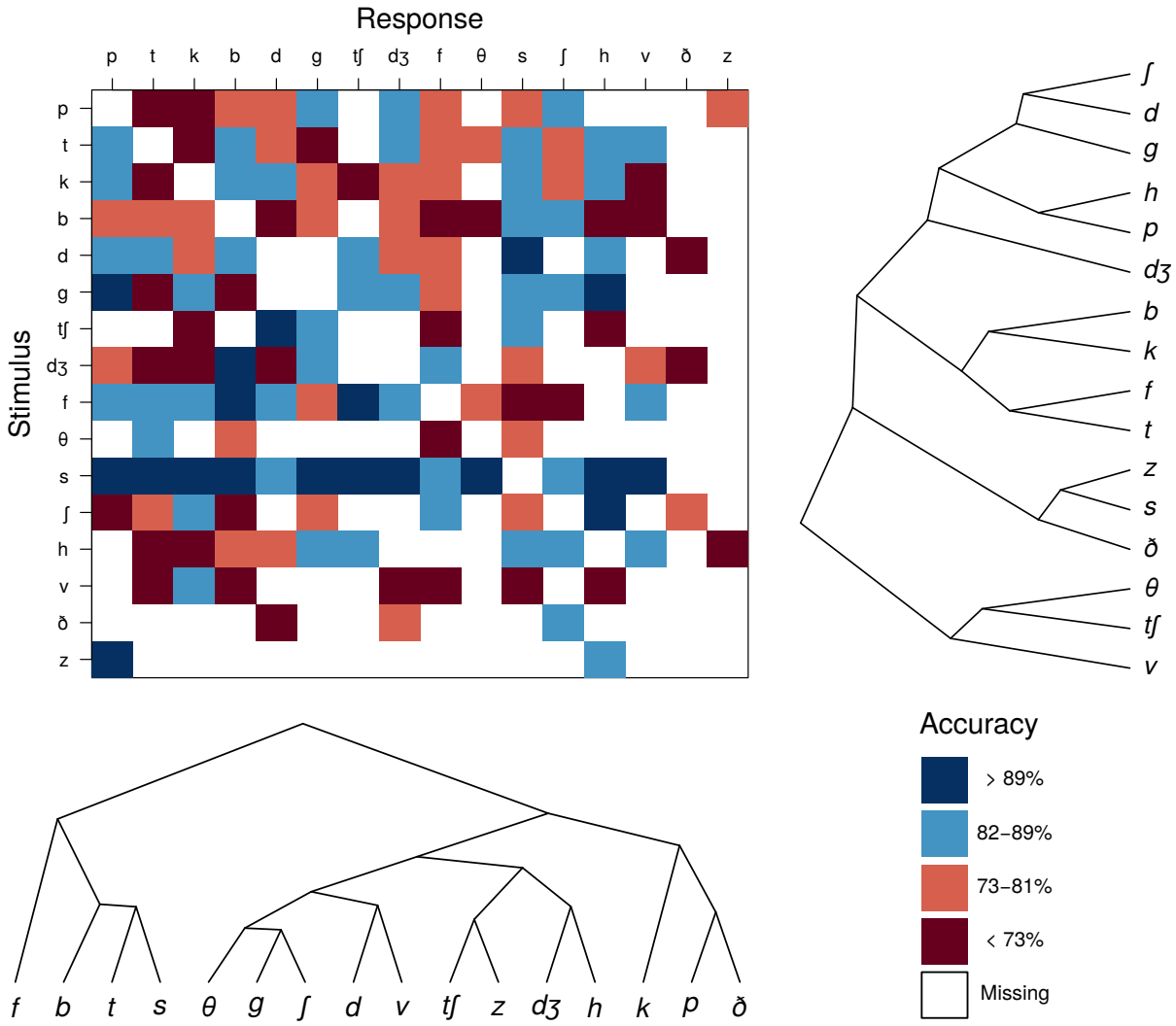

Figure A.19: Listener accuracies by contrast in Experiment 1a (CV). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

Contrast accuracies in Experiment 1b (CV)

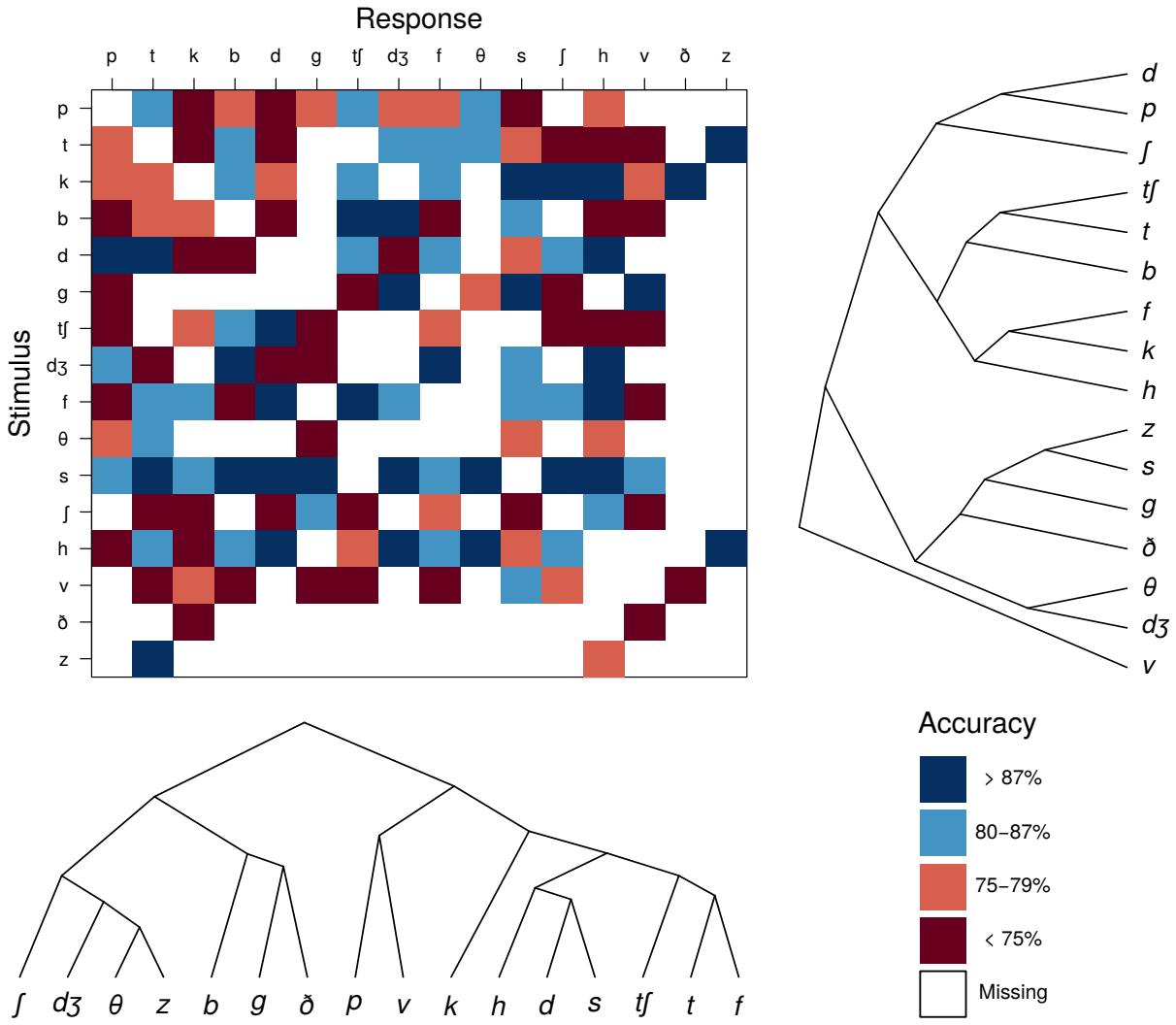

Figure A.20: Listener accuracies by contrast in Experiment 1b (CV). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

Contrast accuracies in Experiment 1a (VCV)

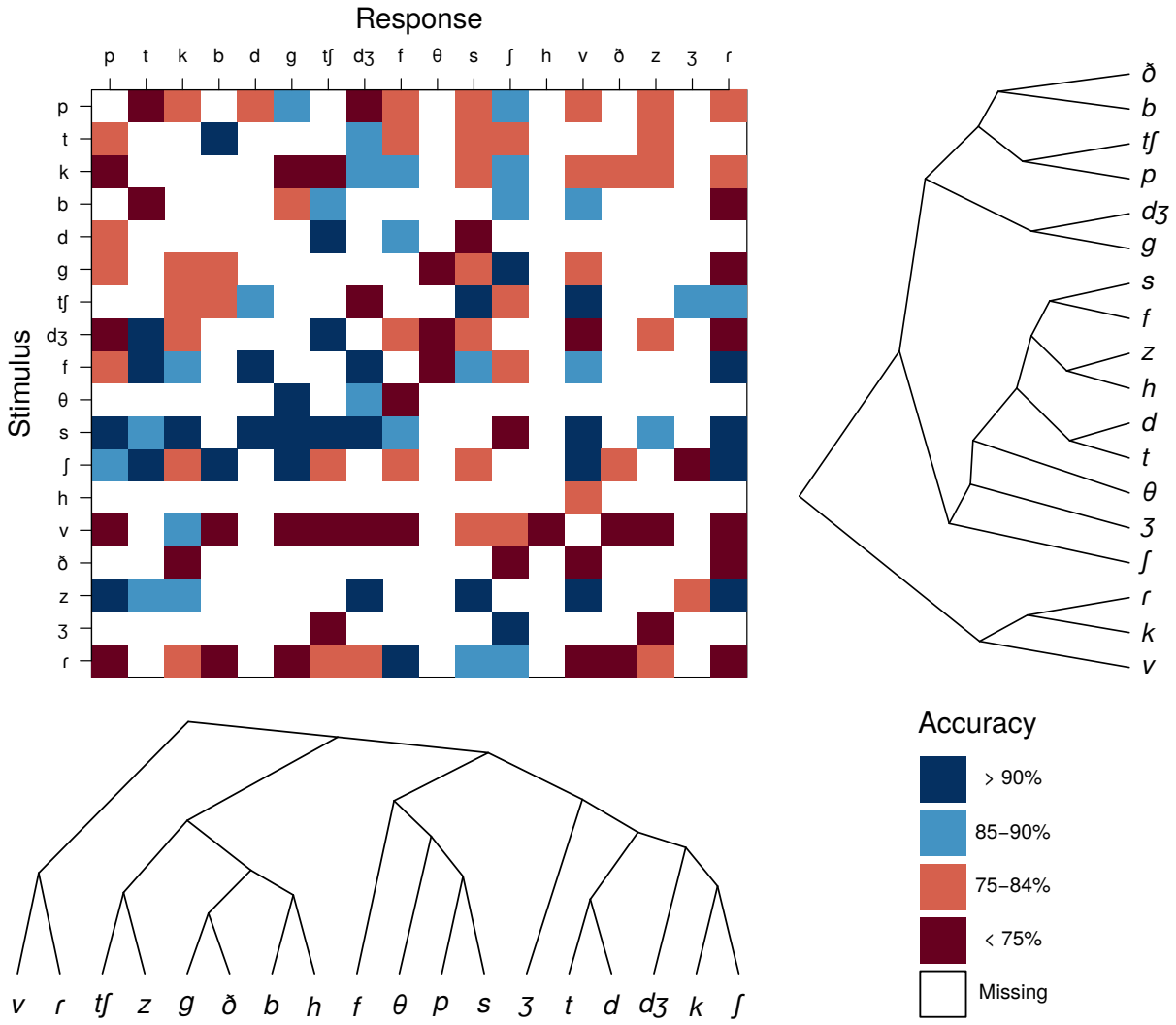

Figure A.21: Listener accuracies by contrast in Experiment 1a (VCV). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

Contrast accuracies in Experiment 1b (VCV)

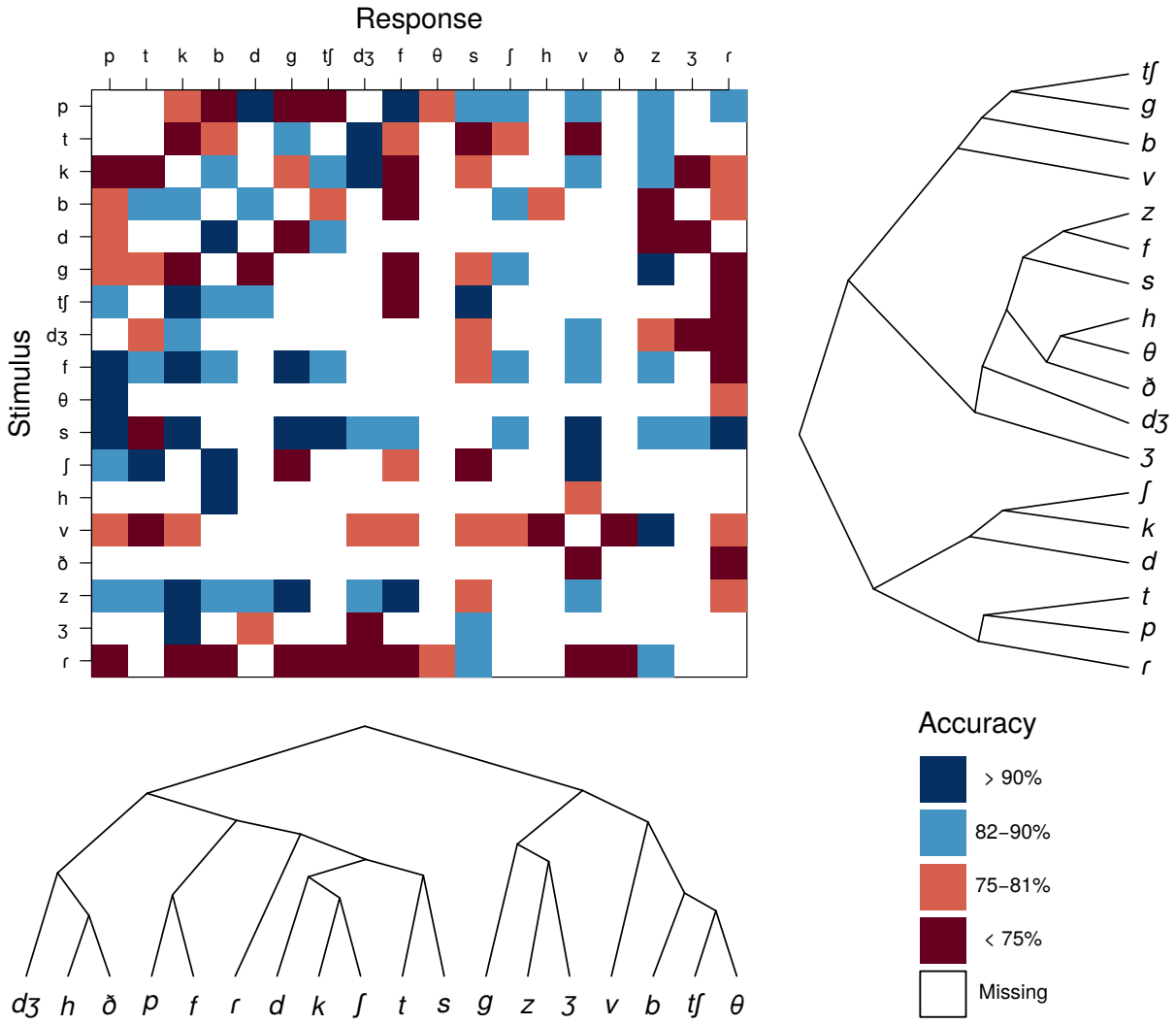

Figure A.22: Listener accuracies by contrast in Experiment 1b (VCV). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

Contrast accuracies in Experiment 1a (VC)

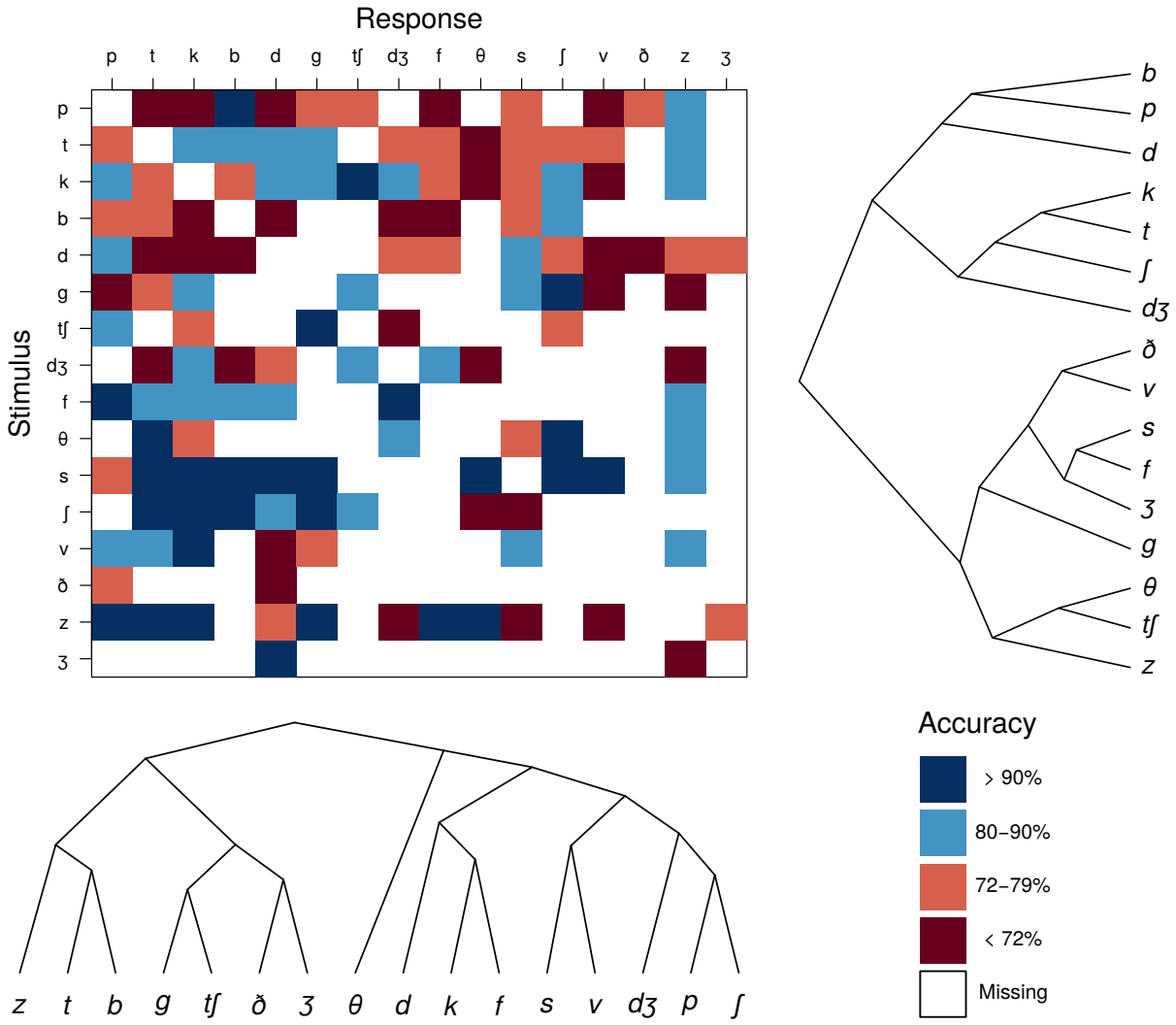

Figure A.23: Listener accuracies by contrast in Experiment 1a (VC). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

Contrast accuracies in Experiment 1b (VC)

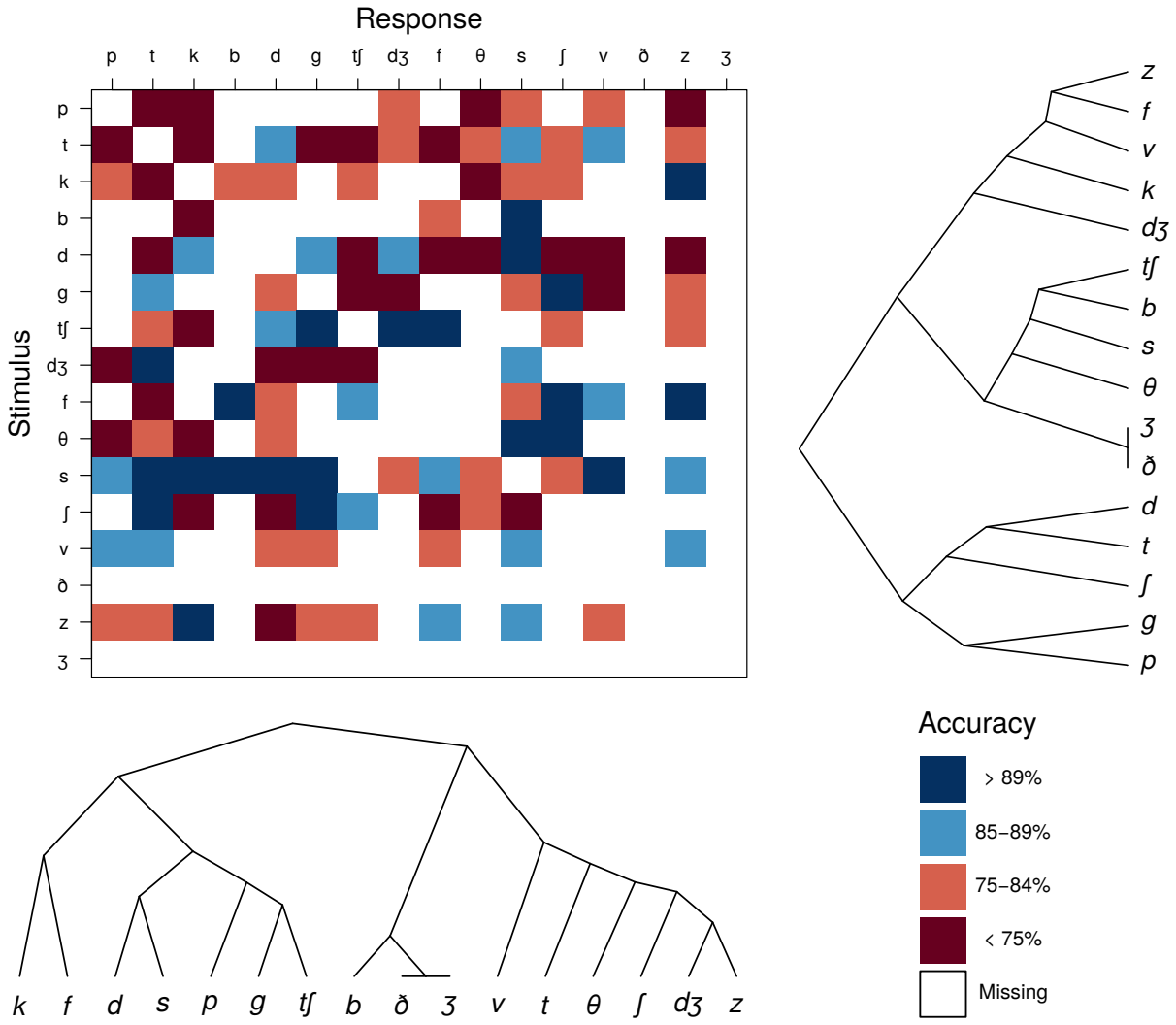

Figure A.24: Listener accuracies by contrast in Experiment 1b (VC). The upper left panel shows accuracies in a confusion matrix layout, with consonants ordered according to features, and each cell indicating listener accuracy on the contrast corresponding to that row-column combination (see the bottom-right panel for the legend). The upper right panel shows a dendrogram of a hierarchical clustering (using Ward's method) of the stimulus error patterns (i.e., clustering the row vectors in the *error matrix*, the complement of the accuracy matrix, where missing contrasts are coded as 0 errors). The bottom-left panel shows a mirror hierarchical clustering of the response patterns along the columns the error matrix. For both clustering solutions, phones with similar error distributions are grouped together, with increasing dissimilarity between phones *a* and *b* represented in a greater number of nodes between *a* and *b* in the dendrogram.

Featural contrast accuracy in Exp. 1a (CV)

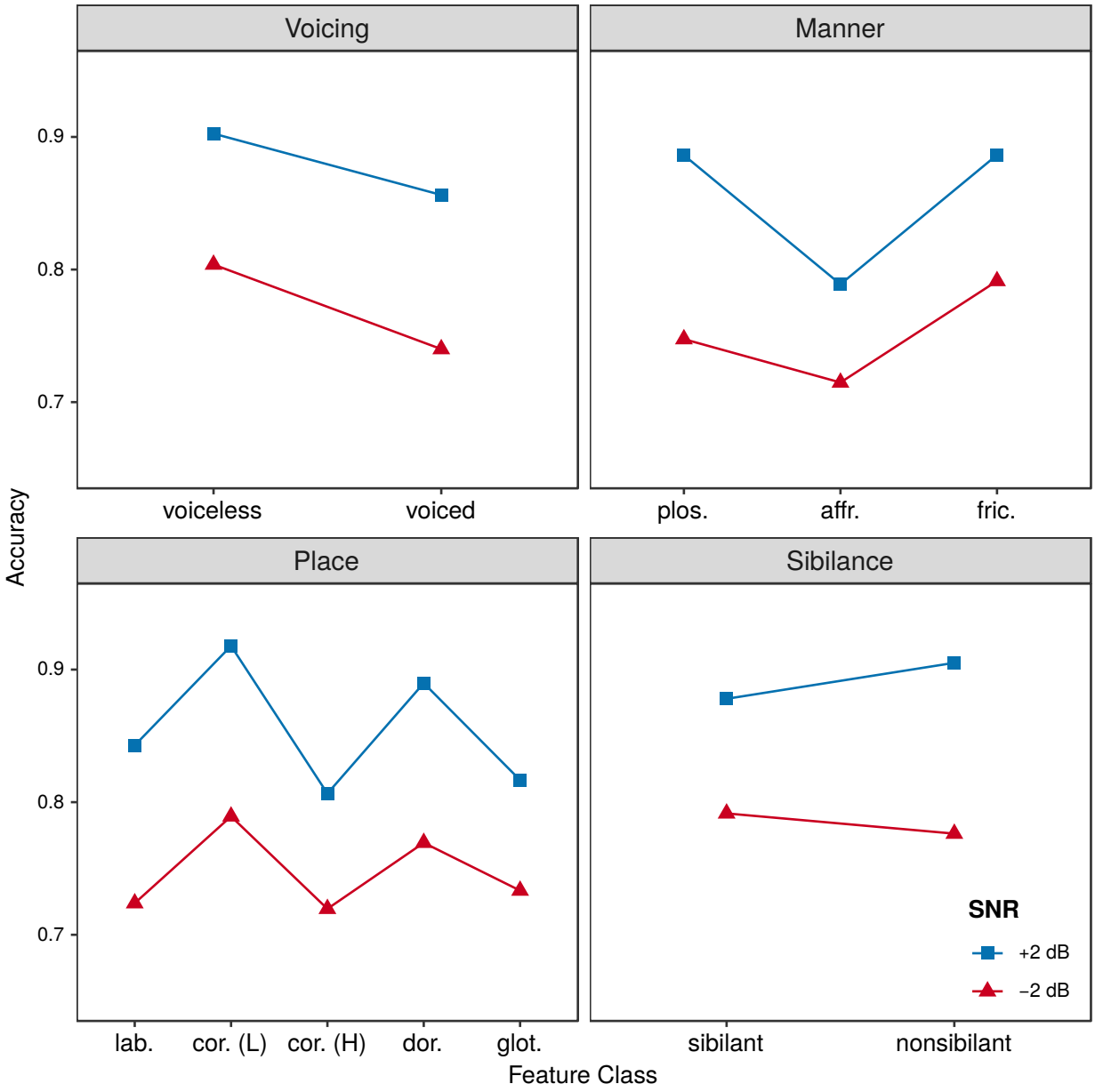

Figure A.25: Featural contrast accuracies by target feature and SNR in CV position in Experiment 1a.

Featural contrast accuracy in Exp. 1b (CV)

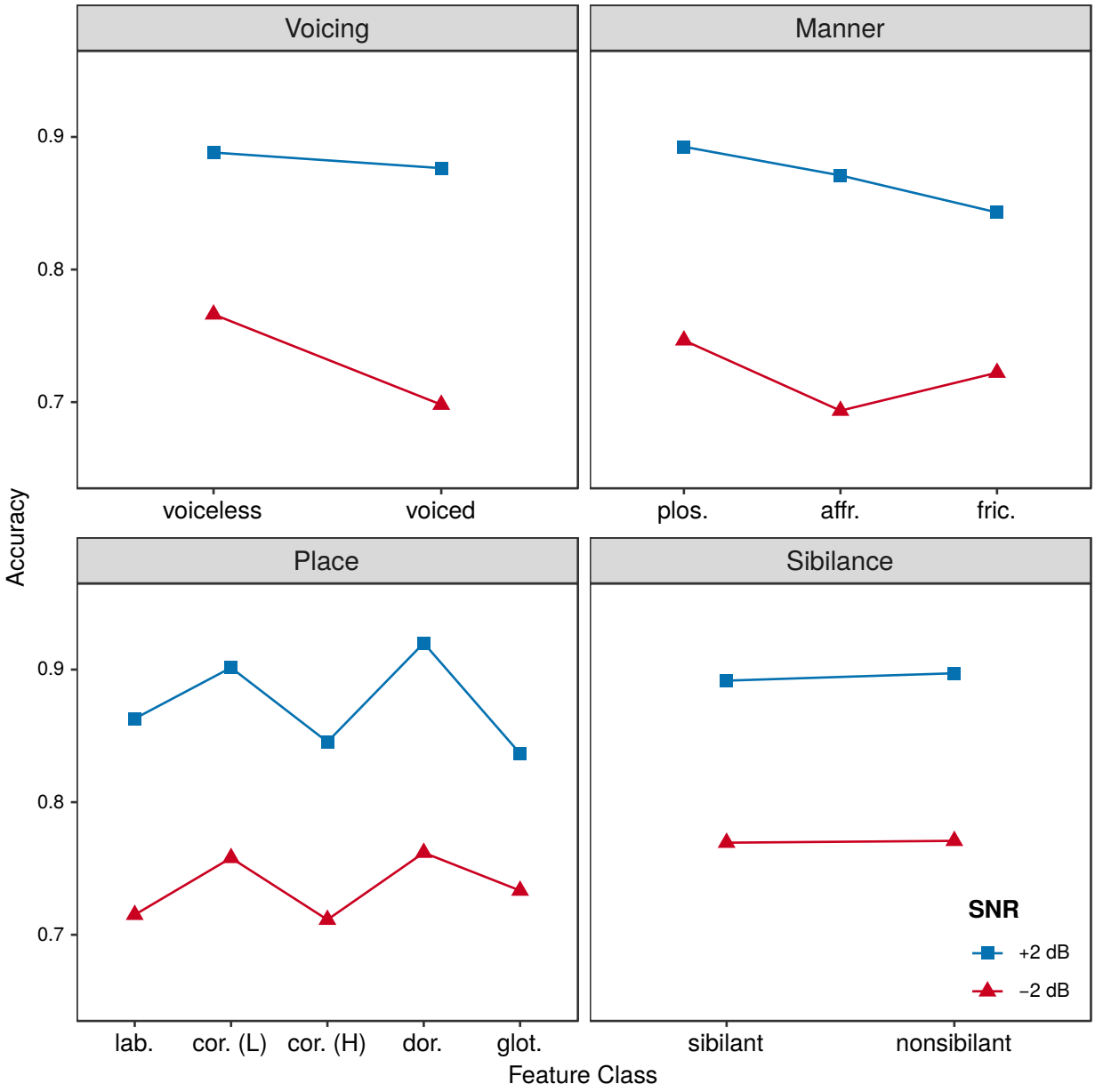

Figure A.26: Featural contrast accuracies by target feature and SNR in CV position in Experiment 1b.

Featural contrast accuracy in Exp. 1a (VCV)

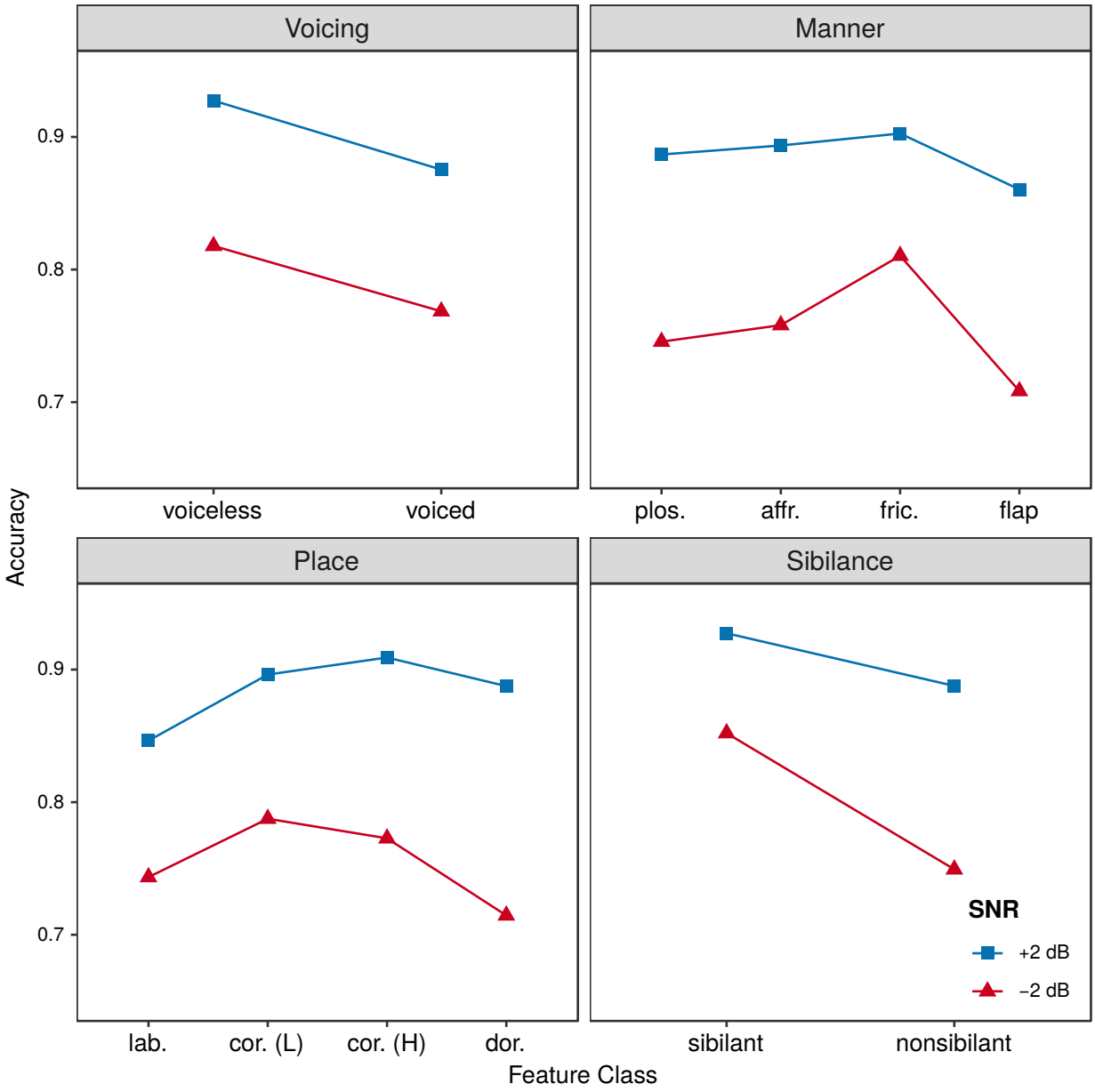

Figure A.27: Featural contrast accuracies by target feature and SNR in VCV position in Experiment 1a.

Featural contrast accuracy in Exp. 1b (VCV)

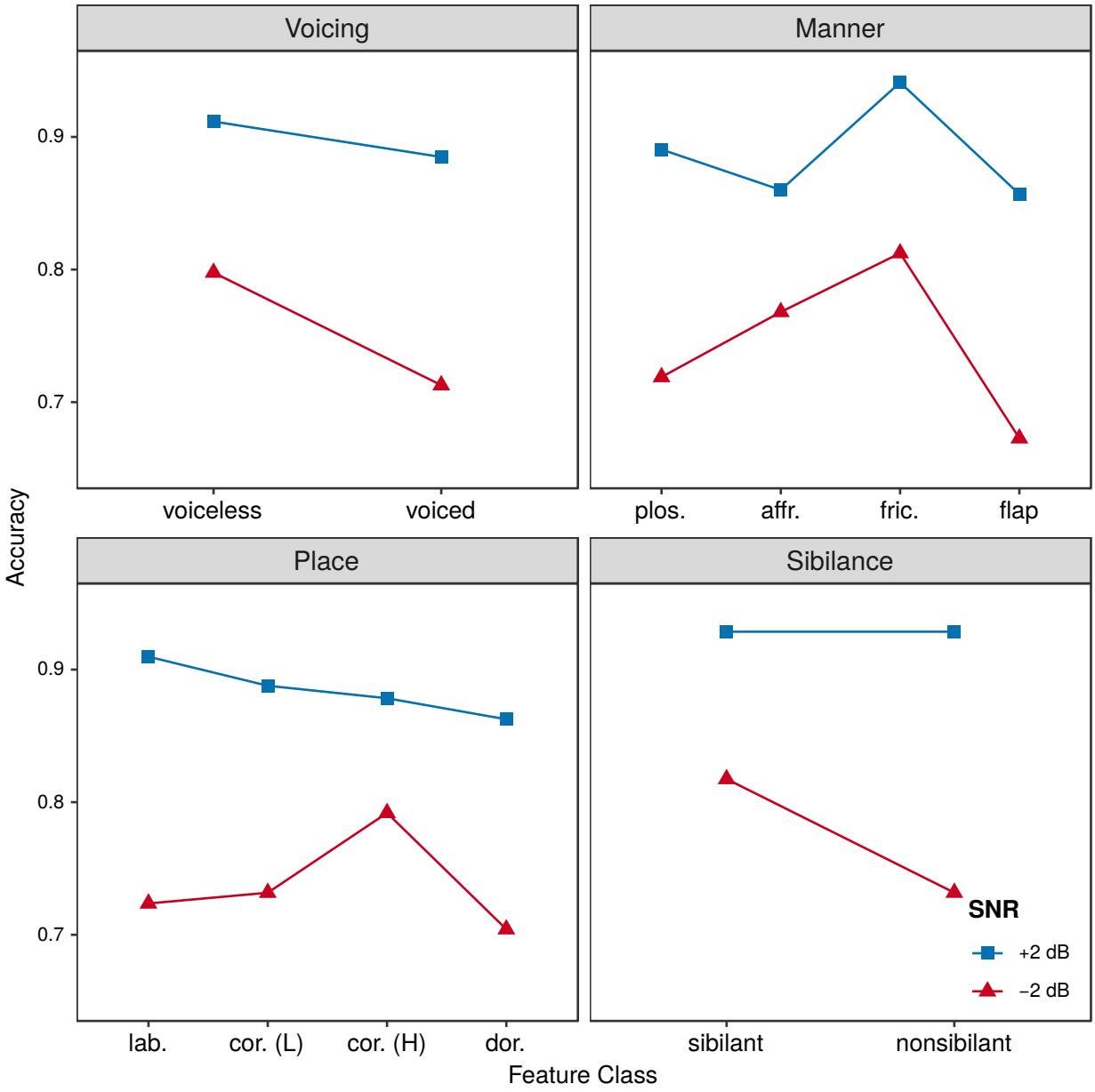

Figure A.28: Featural contrast accuracies by target feature and SNR in VCV position in Experiment 1b.

Featural contrast accuracy in Exp. 1a (VC)

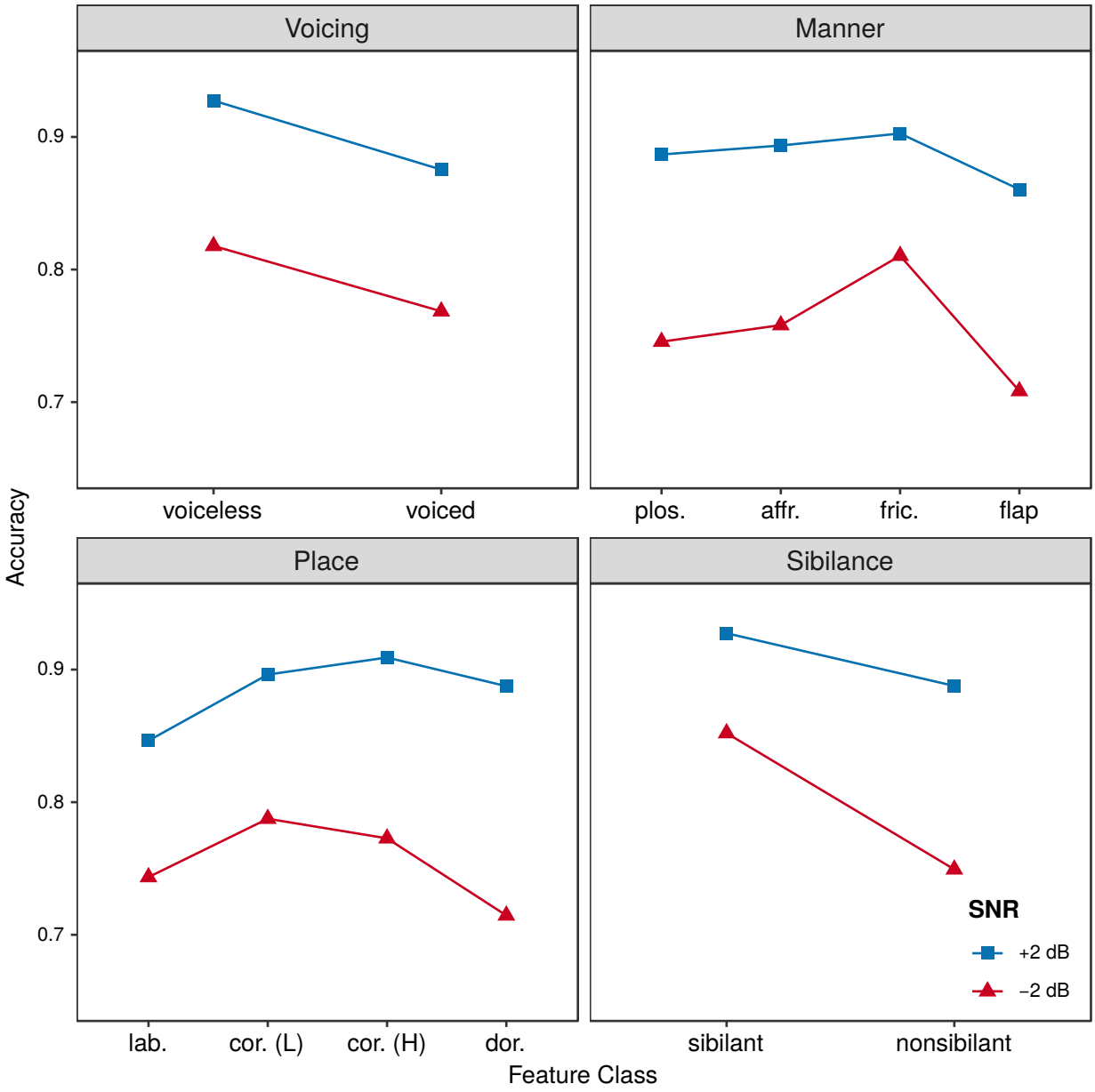

Figure A.29: Featural contrast accuracies by target feature and SNR in VC position in Experiment 1a.

Featural contrast accuracy in Exp. 1b (VC)

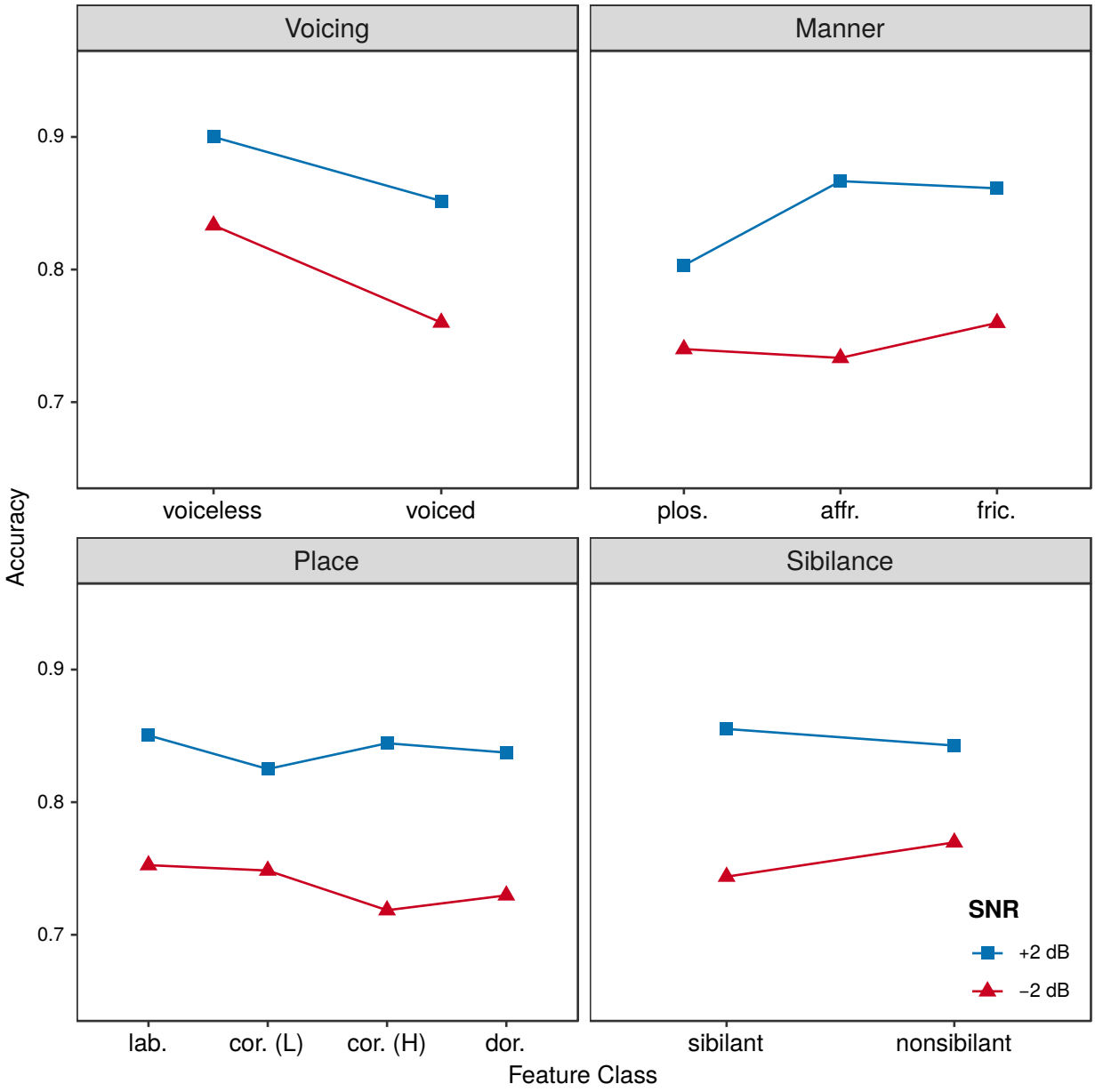

Figure A.30: Featural contrast accuracies by target feature and SNR in VC position in Experiment 1b.

Featural contrast accuracies by length and frequency in Exp. 1a (CV)

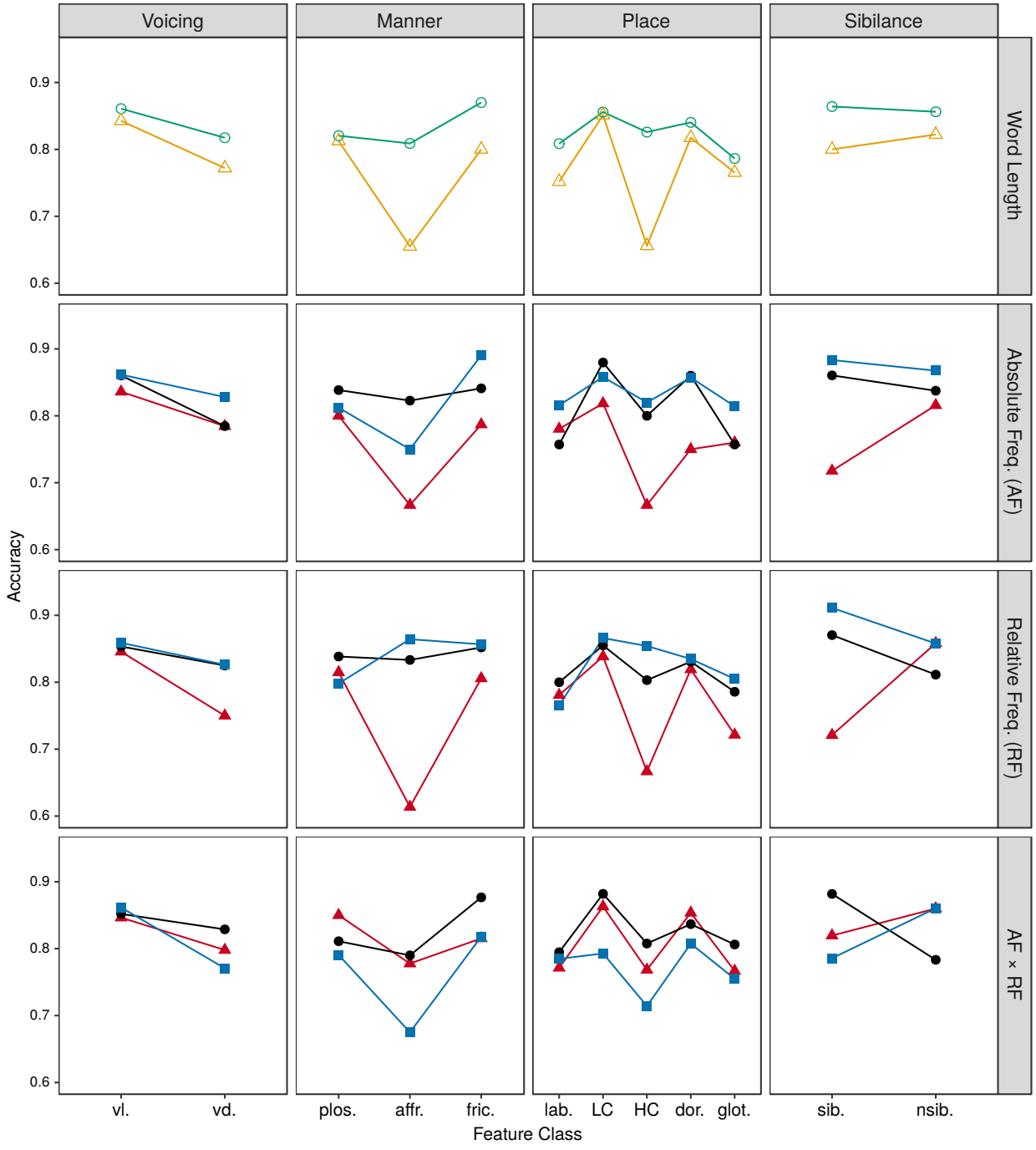

Figure A.31: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in CV position in Experiment 1a. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Featural contrast accuracies by length and frequency in Exp. 1b (CV)

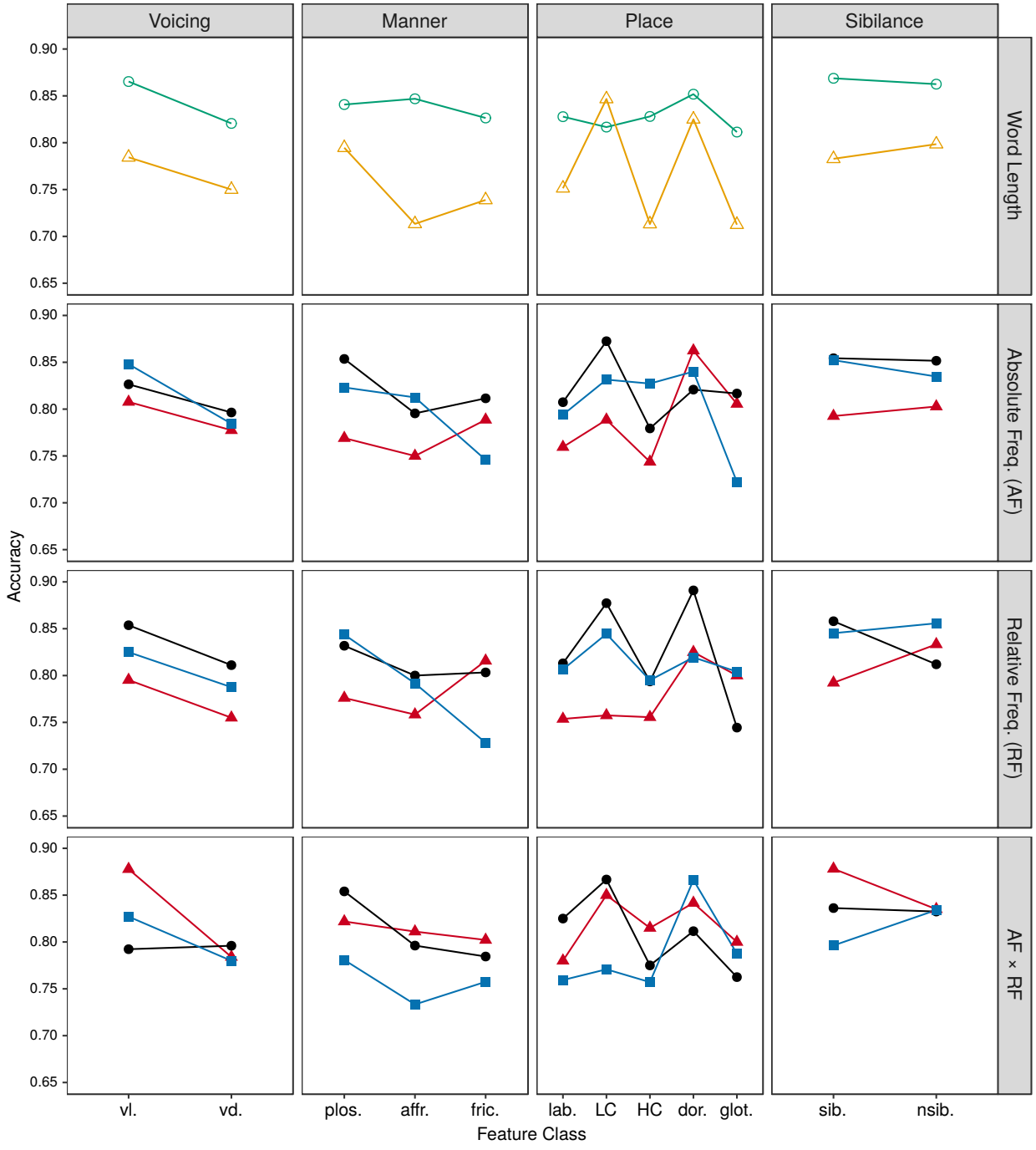

Figure A.32: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in CV position in Experiment 1b. For Length, monosyllables are shown in green open circles, and polysyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Featural contrast accuracies by length and frequency in Exp. 1a (VCV)

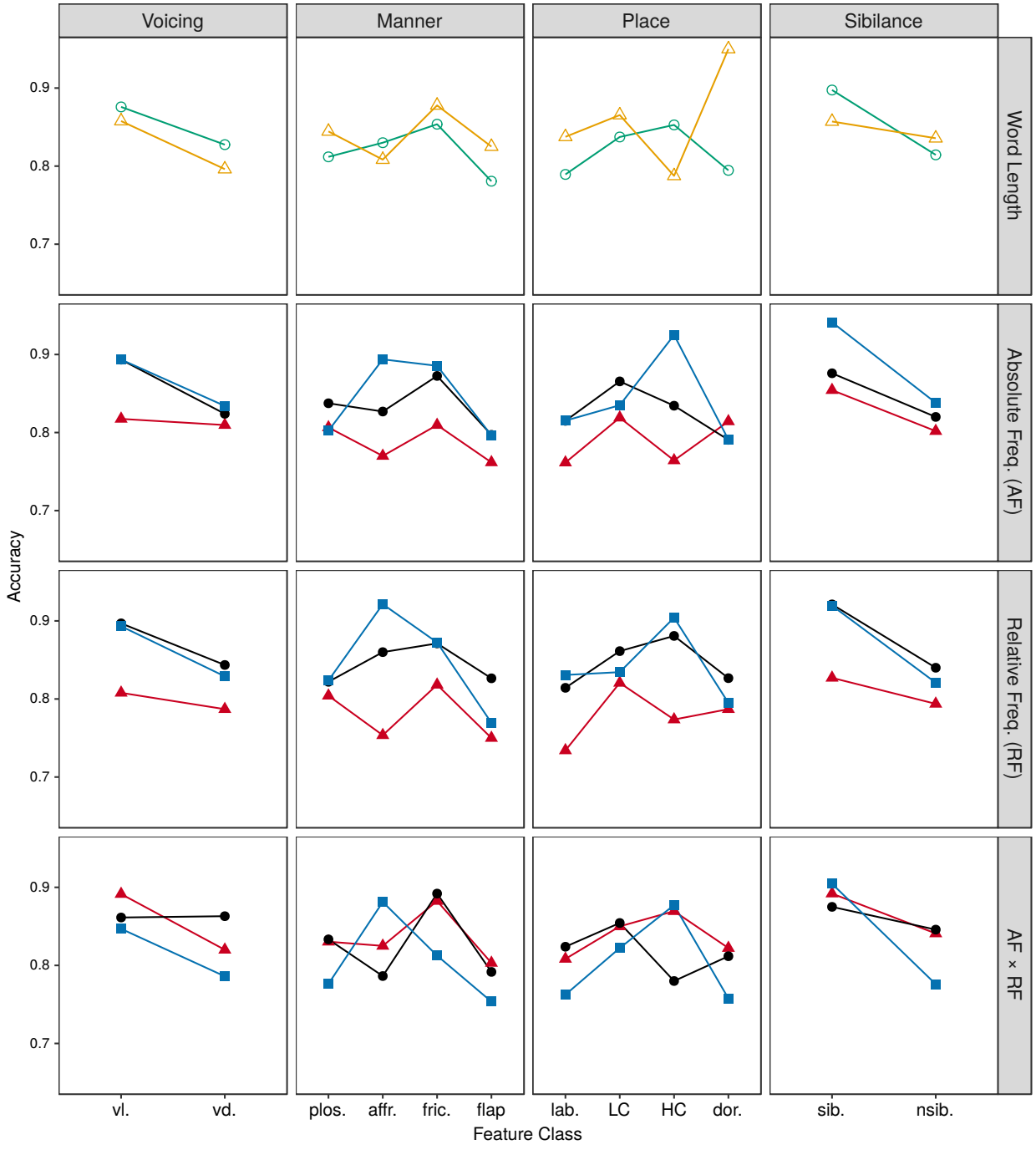

Figure A.33: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VCV position in Experiment 1a. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Featural contrast accuracies by length and frequency in Exp. 1b (VCV)

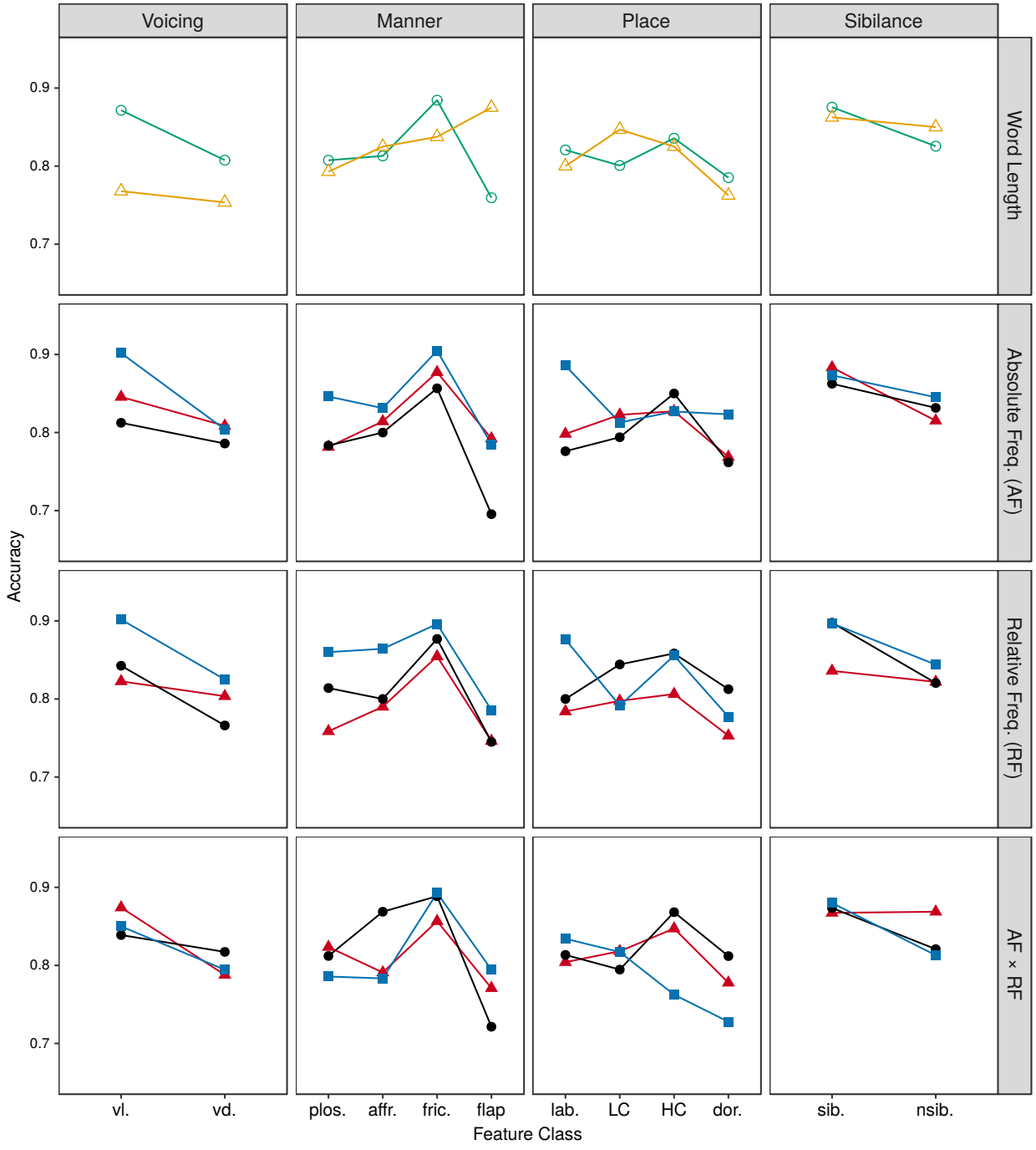

Figure A.34: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VCV position in Experiment 1b. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Featural contrast accuracies by length and frequency in Exp. 1a (VC)

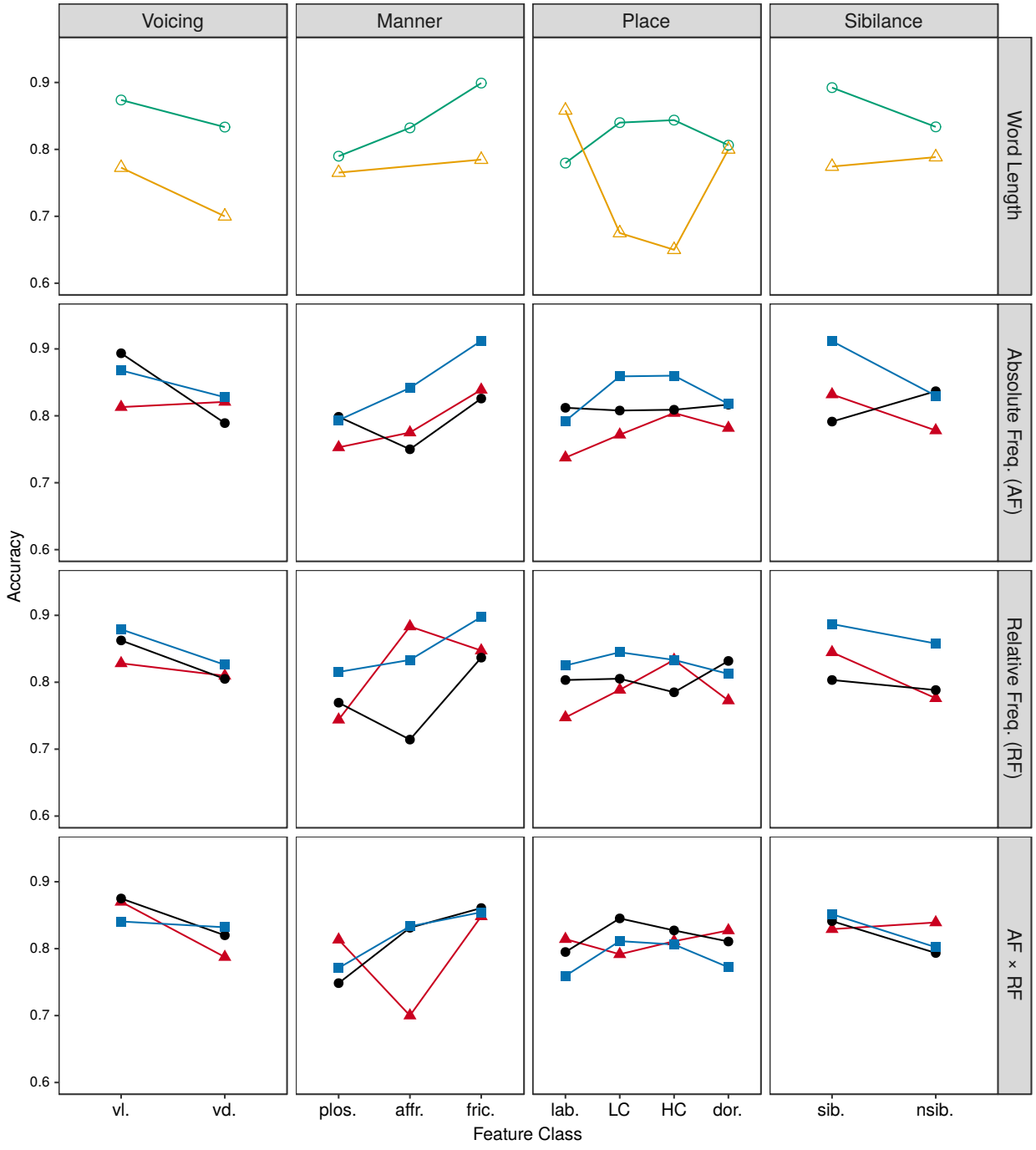

Figure A.35: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VC position in Experiment 1a. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Featural contrast accuracies by length and frequency in Exp. 1b (VC)

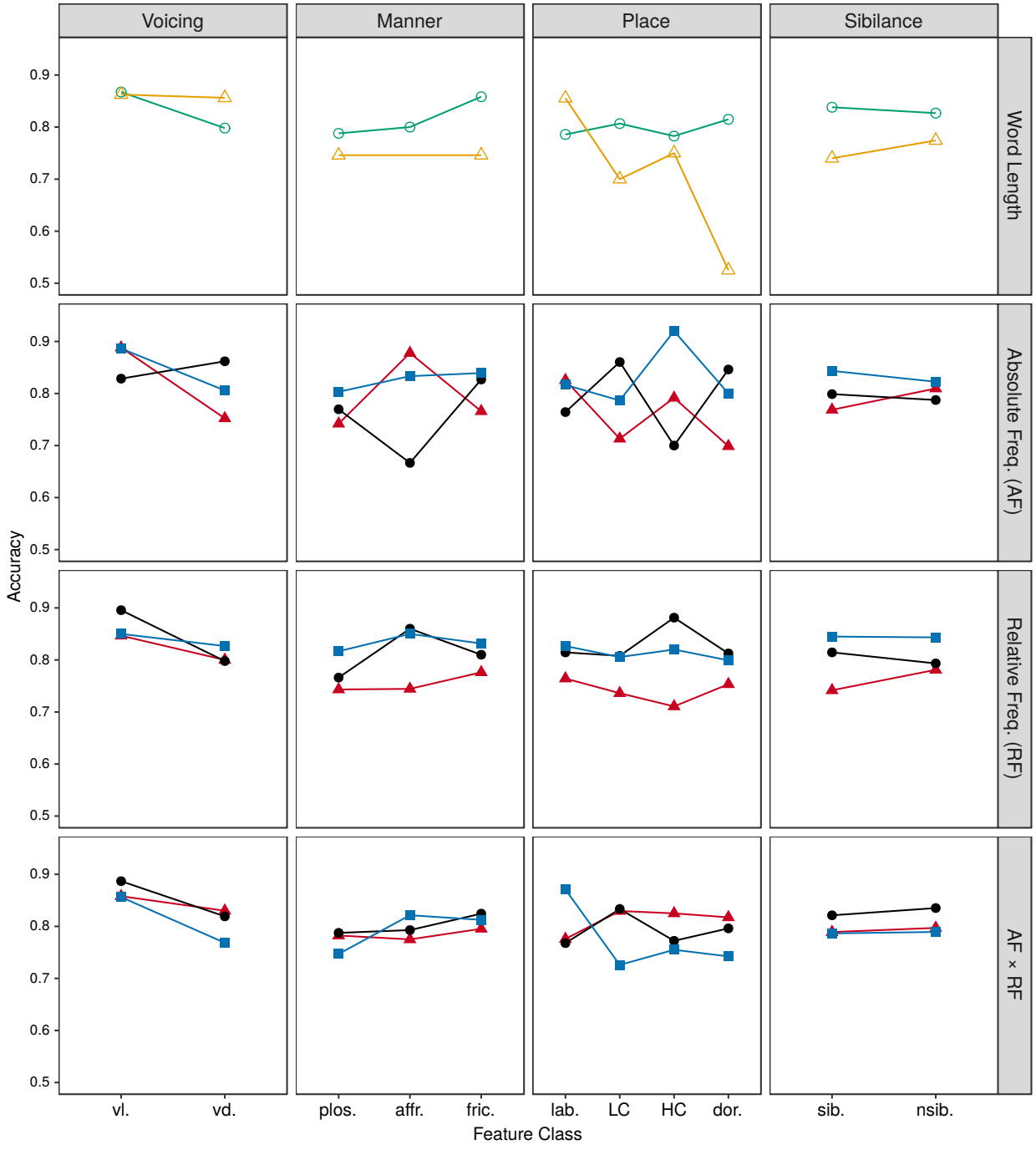

Figure A.36: Featural contrast accuracies by Word Length and Word Frequency (AF, RF, AF×RF) in VC position in Experiment 1b. For Length, disyllables are shown in green open circles, and trisyllables are shown in orange open triangles. For the Frequency variables, which are measured on a continuous scale, lower, middle, and upper terciles (< 0.33, $0.33 - 0.67$, > 0.67) of each variable are shown in red triangles, black circles, and blue squares, respectively.

Target phone error proportions by SNR in Exp. 1a/b (CV)

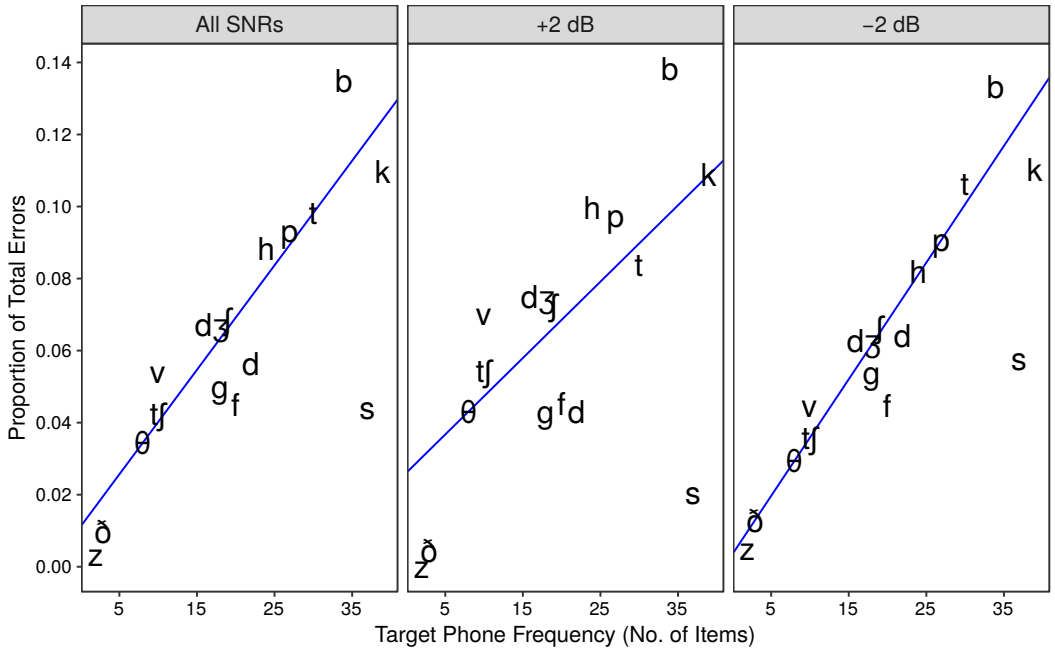

Figure A.37: Proportion of errors in CV position in Exp. 1a (overall and by SNR) attributable to each stimulus phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

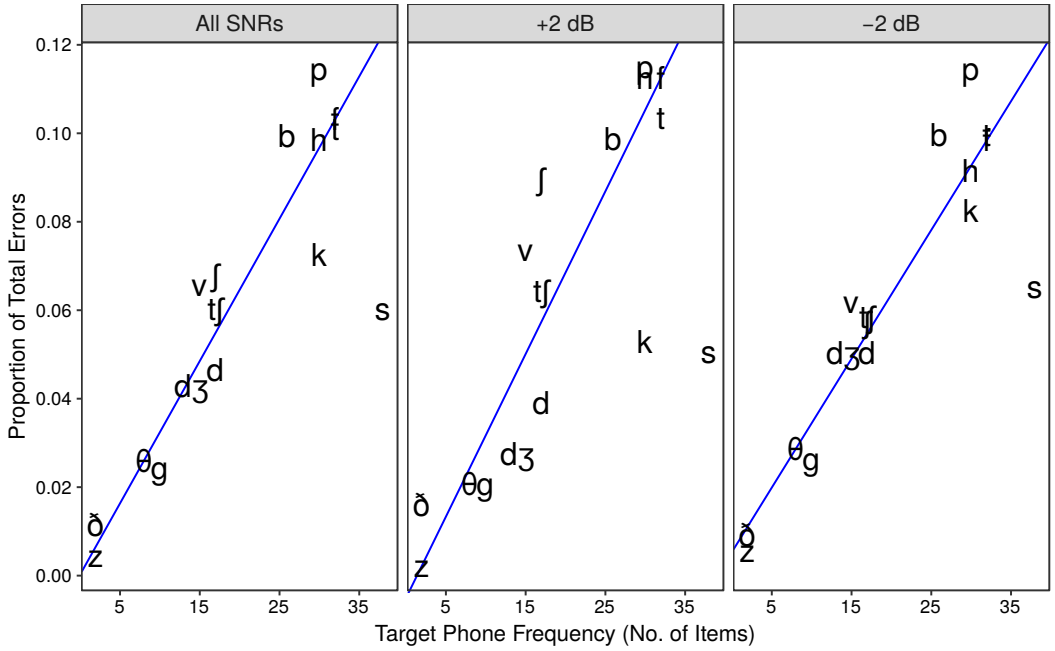

Figure A.38: Proportion of errors in CV position in Exp. 1b (overall and by SNR) attributable to each stimulus phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

Target phone error proportions by SNR in Exp. 1a/b (VCV)

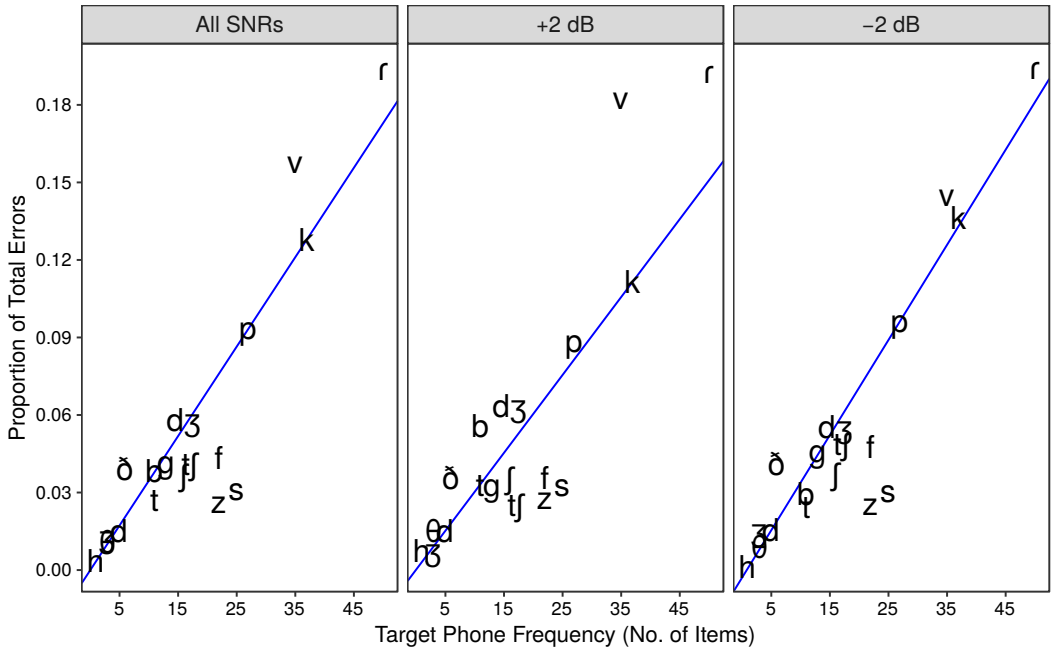

Figure A.39: Proportion of errors in VCV position in Exp. 1a (overall and by SNR) attributable to each stimulus phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

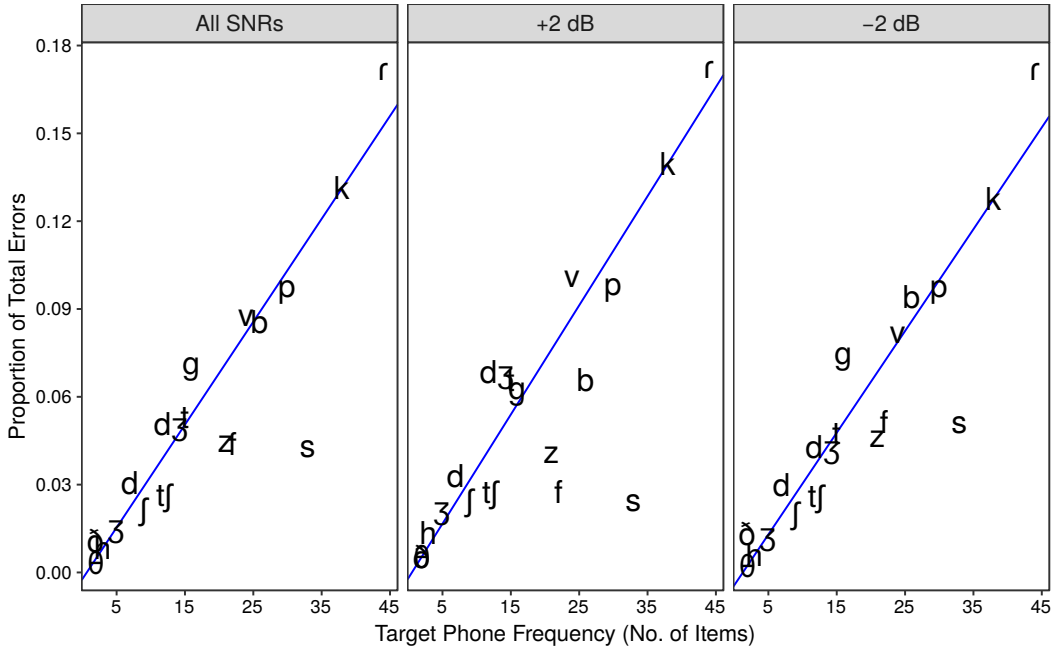

Figure A.40: Proportion of errors in VCV position in Exp. 1b (overall and by SNR) attributable to each stimulus phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

Target phone error proportions by SNR in Exp. 1a/b (VC)

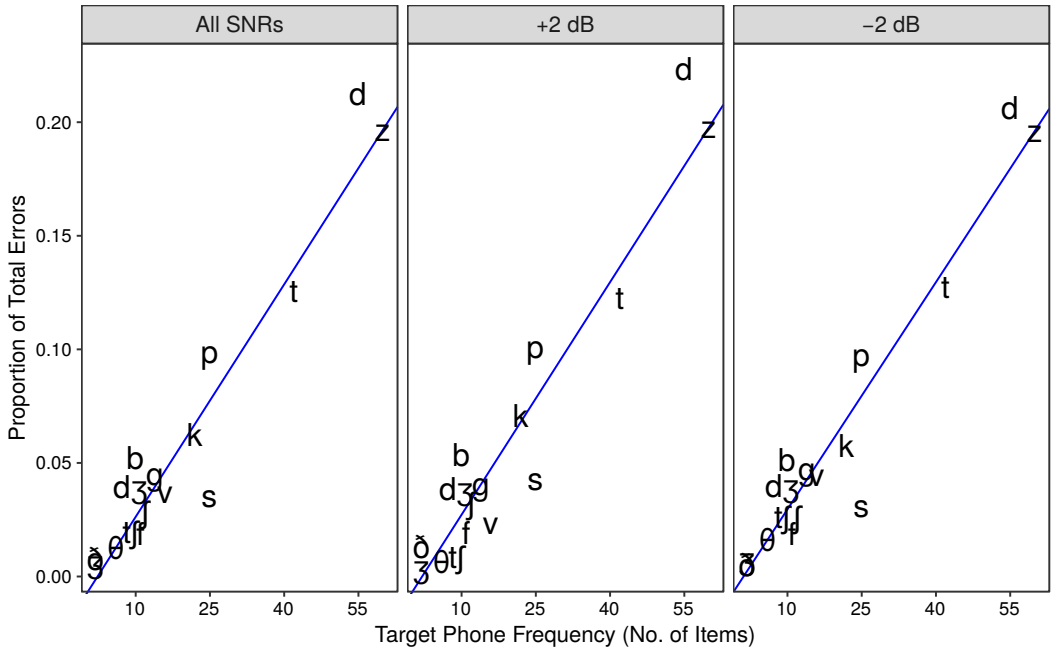

Figure A.41: Proportion of errors in VC position in Exp. 1a (overall and by SNR) attributable to each stimulus phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

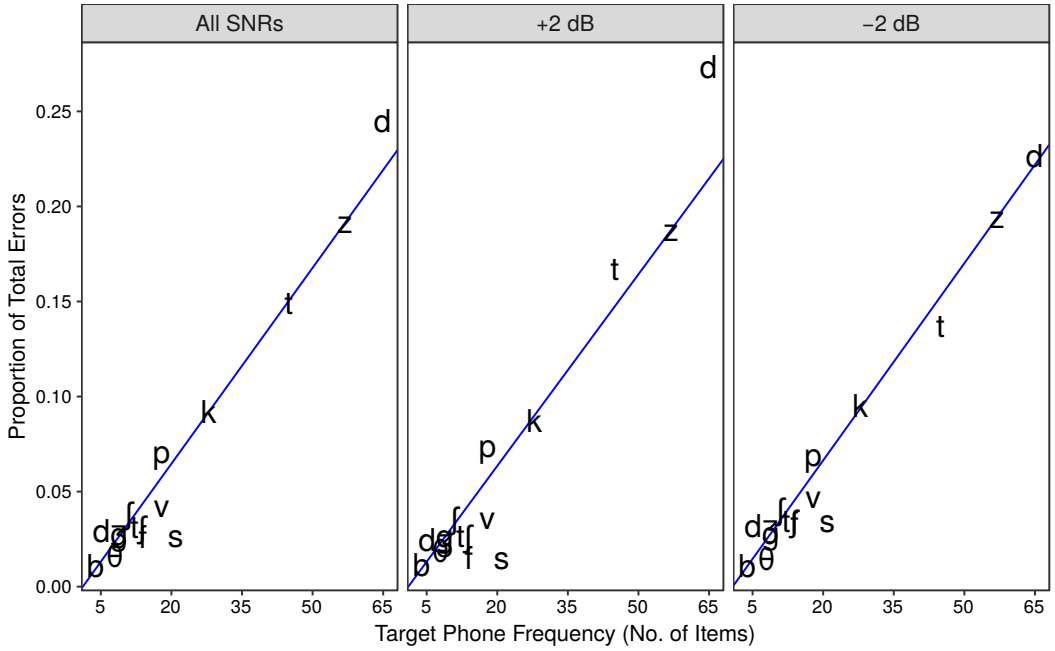

Figure A.42: Proportion of errors in VC position in Exp. 1b (overall and by SNR) attributable to each stimulus phone as a function of the number of items exhibiting that phone in the critical minimal contrast in the 2AFC task. Lines indicate median regression fits.

Target phone error proportions by Length in Exp. 1a/b (CV)

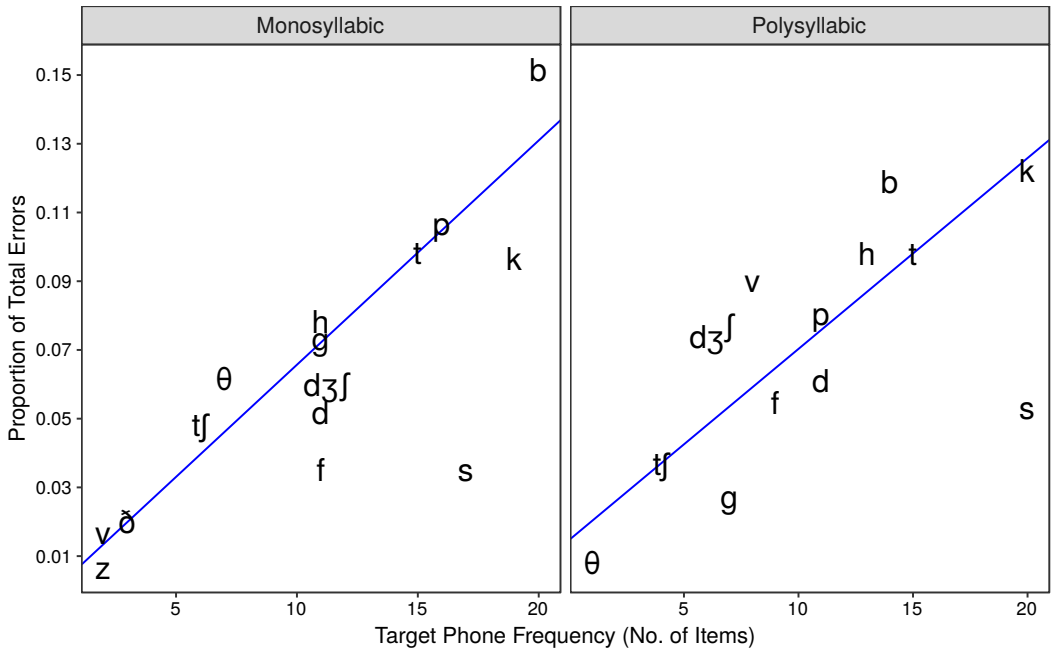

Figure A.43: Proportion of errors in CV position in Experiment 1a attributable to each target phone as a function of word length. Lines indicate median regression fits.

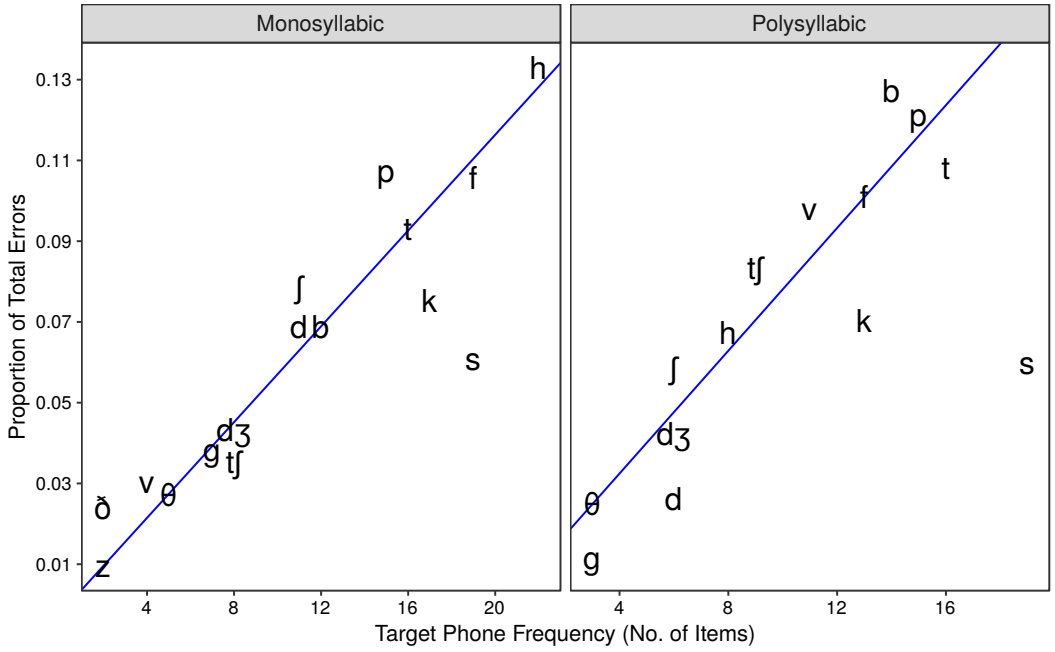

Figure A.44: Proportion of errors in CV position in Experiment 1b attributable to each target phone as a function of word length. Lines indicate median regression fits.

Target phone error proportions by Length in Exp. 1a/b (VCV)

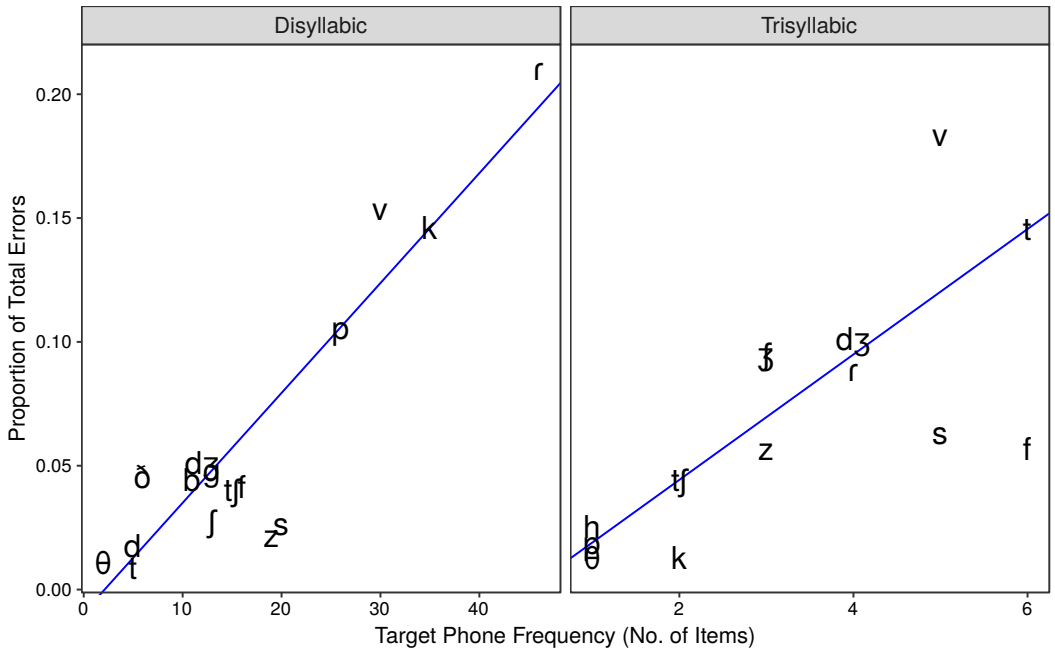

Figure A.45: Proportion of errors in VCV position in Experiment 1a attributable to each target phone as a function of word length. Lines indicate median regression fits.

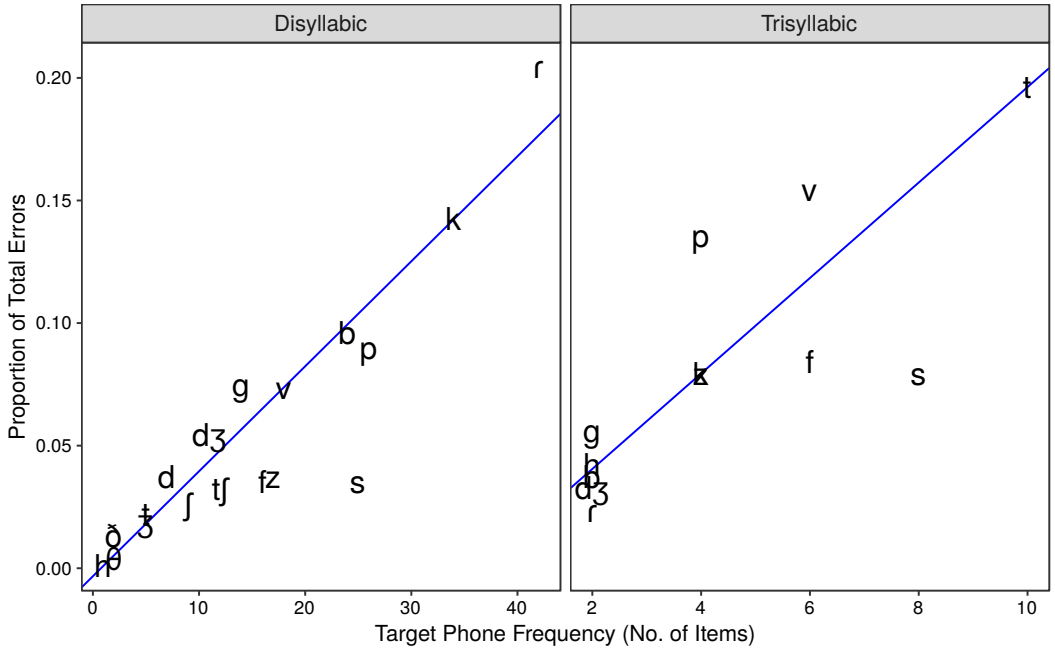

Figure A.46: Proportion of errors in VCV position in Experiment 1b attributable to each target phone as a function of word length. Lines indicate median regression fits.

Target phone error proportions by Length in Exp. 1a/b (VC)

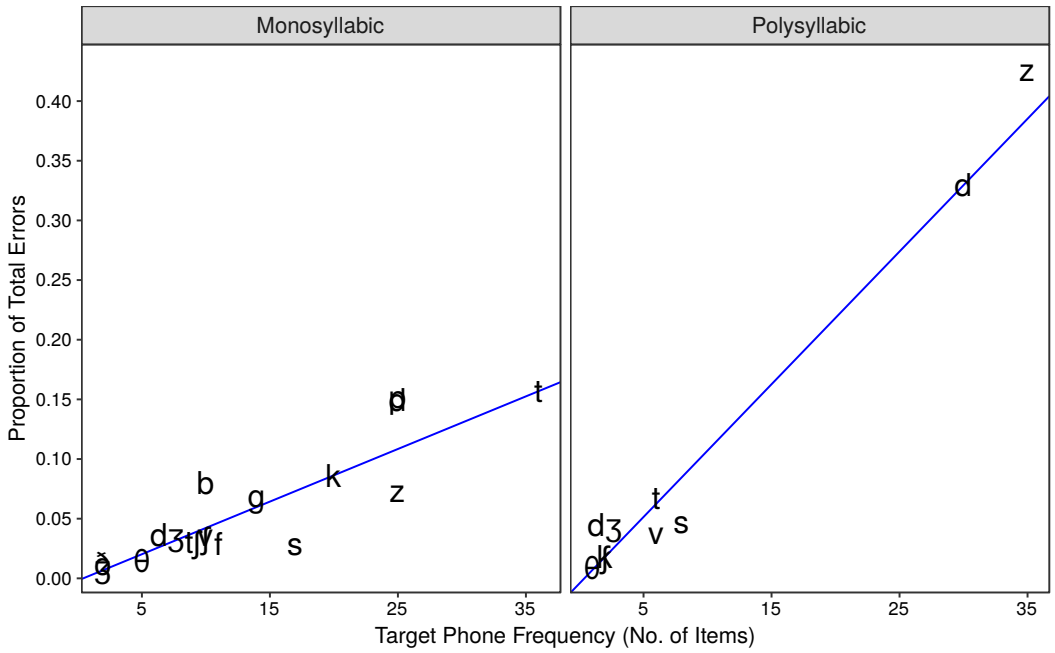

Figure A.47: Proportion of errors in VC position in Experiment 1a attributable to each target phone as a function of word length. Lines indicate median regression fits.

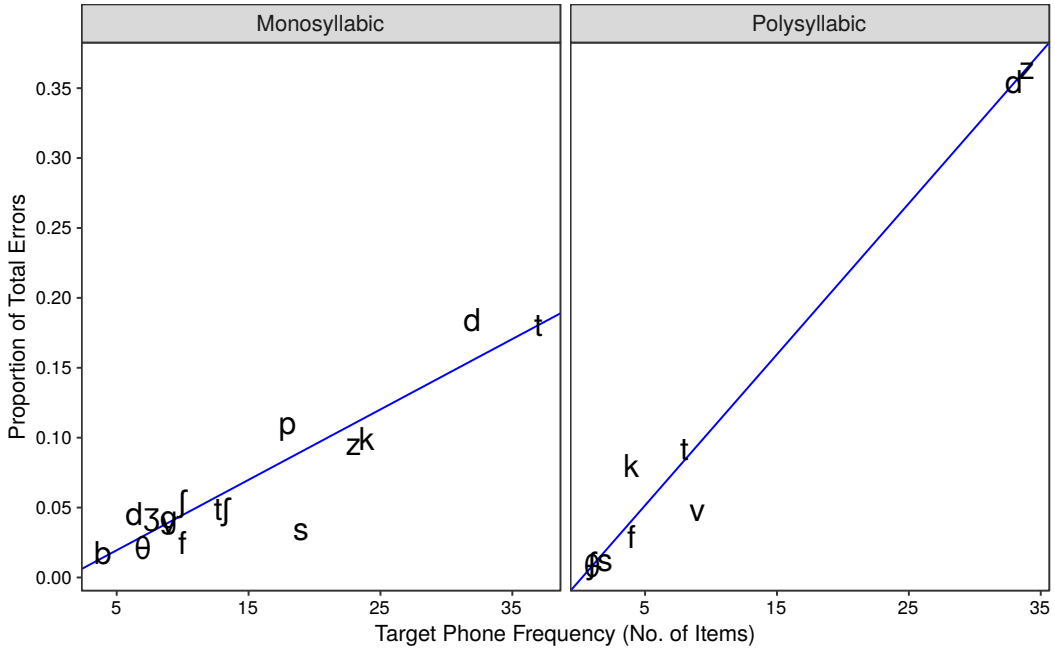

Figure A.48: Proportion of errors in VC position in Experiment 1b attributable to each target phone as a function of word length. Lines indicate median regression fits.

Target phone error proportions by Frequency in Exp. 1a/b (CV)

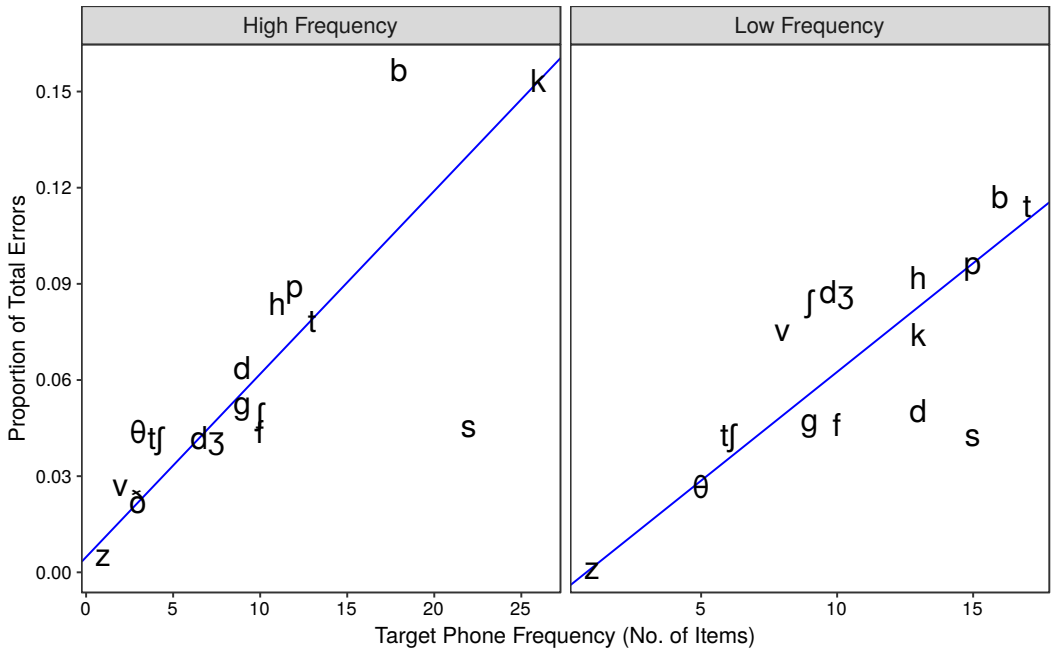

Figure A.49: Proportion of errors in CV position in Experiment 1a attributable to each target phone as a function of word frequency. Lines indicate median regression fits.

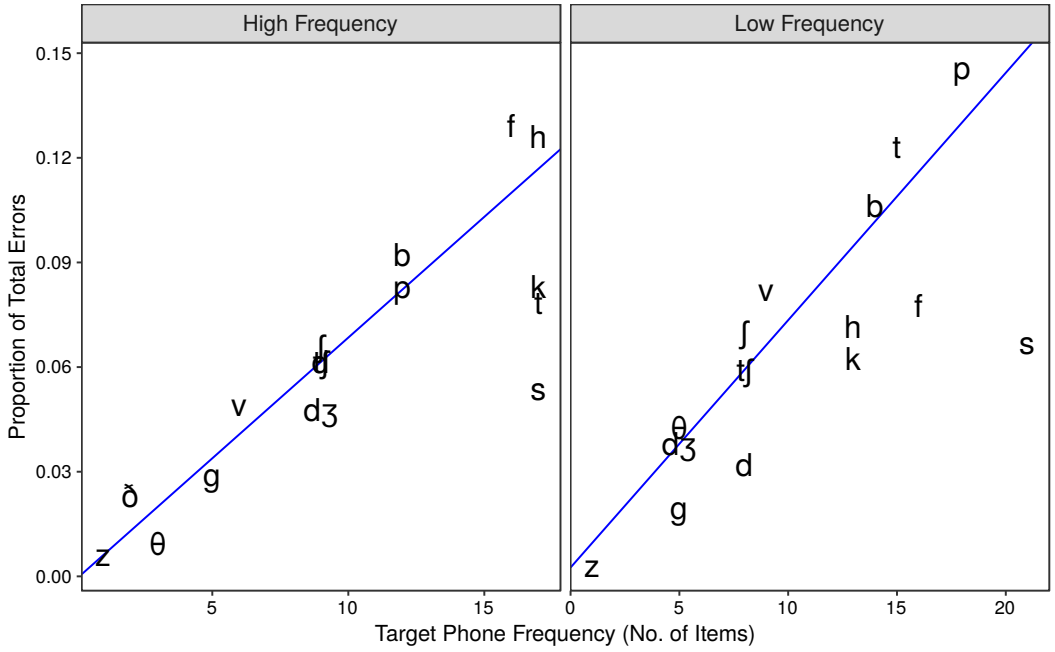

Figure A.50: Proportion of errors in CV position in Experiment 1b attributable to each target phone as a function of word frequency. Lines indicate median regression fits.

Target phone error proportions by Frequency in Exp. 1a/b (VCV)

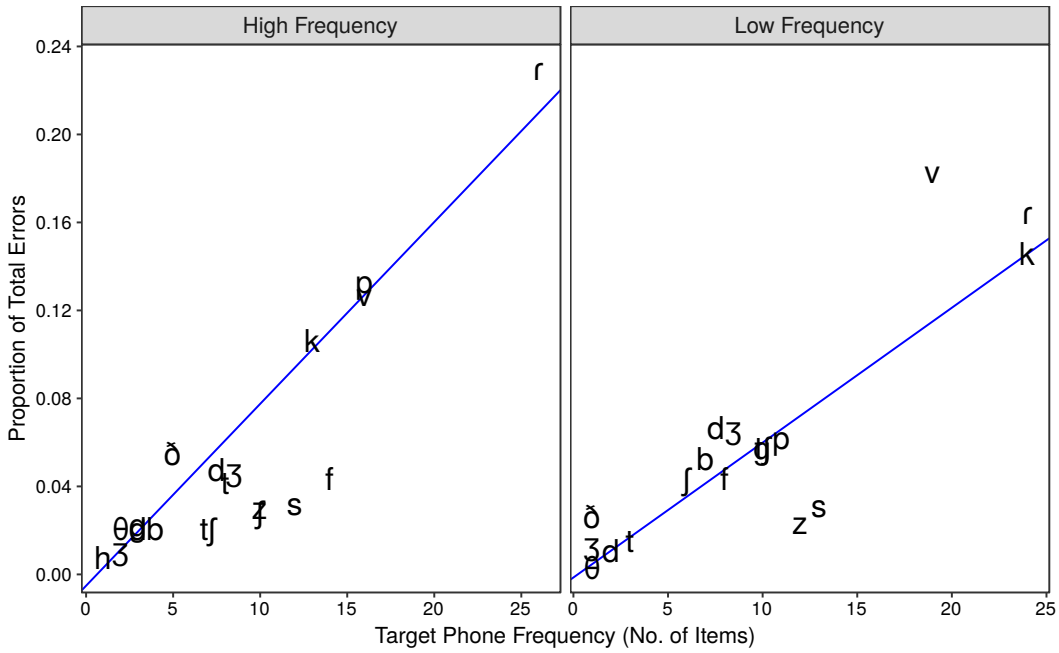

Figure A.51: Proportion of errors in VCV position in Experiment 1a attributable to each target phone as a function of word frequency. Lines indicate median regression fits.

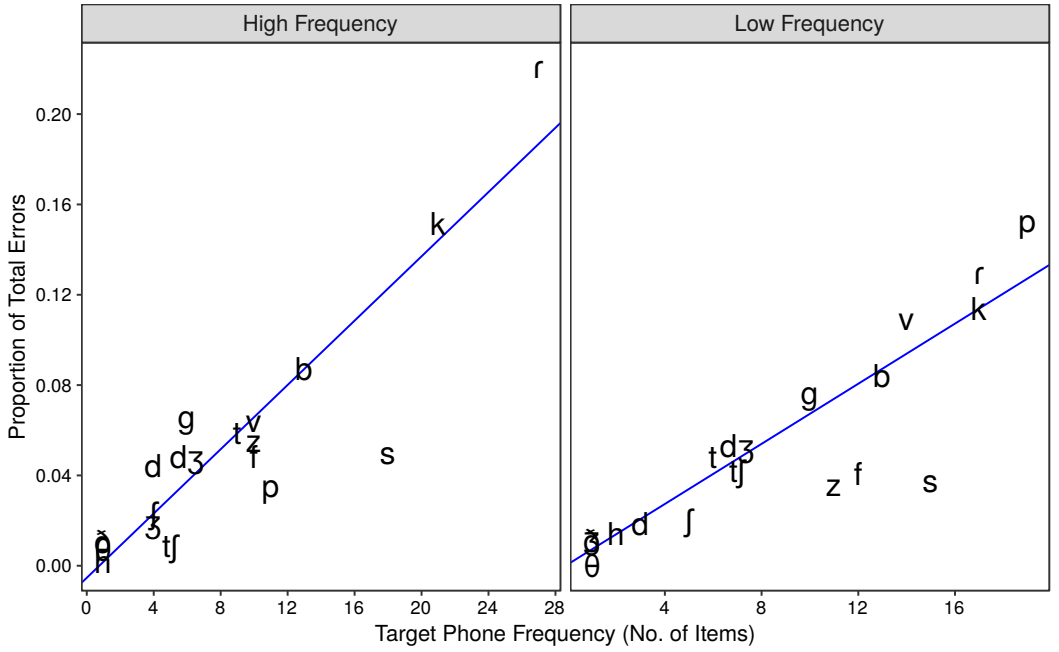

Figure A.52: Proportion of errors in VCV position in Experiment 1b attributable to each target phone as a function of word frequency. Lines indicate median regression fits.

Target phone error proportions by Frequency in Exp. 1a/b (VC)

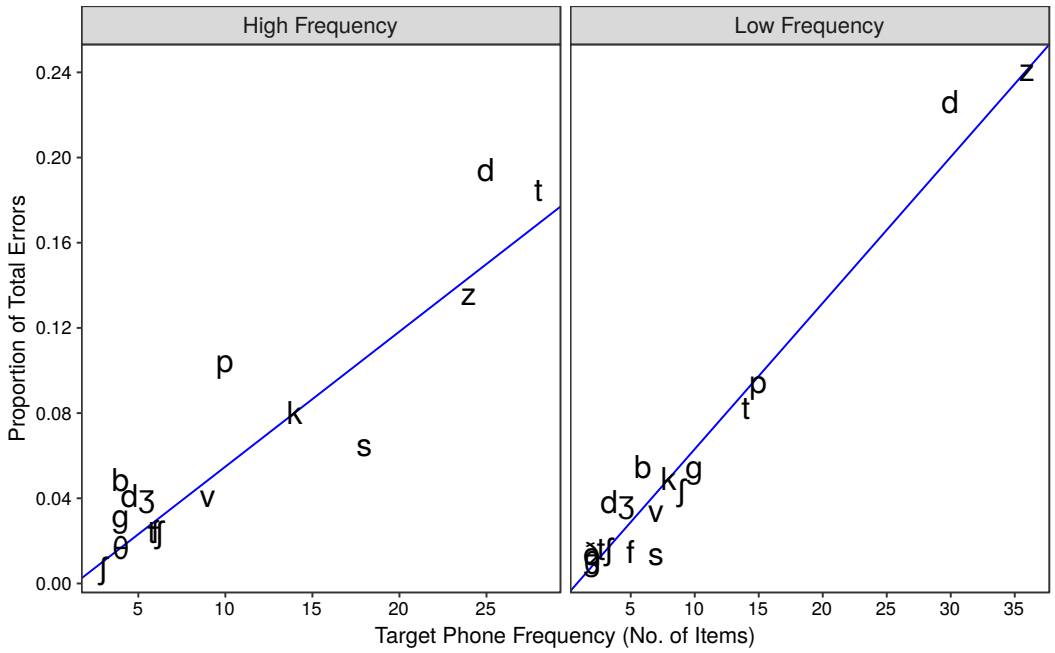

Figure A.53: Proportion of errors in VC position in Experiment 1a attributable to each target phone as a function of word frequency. Lines indicate median regression fits.

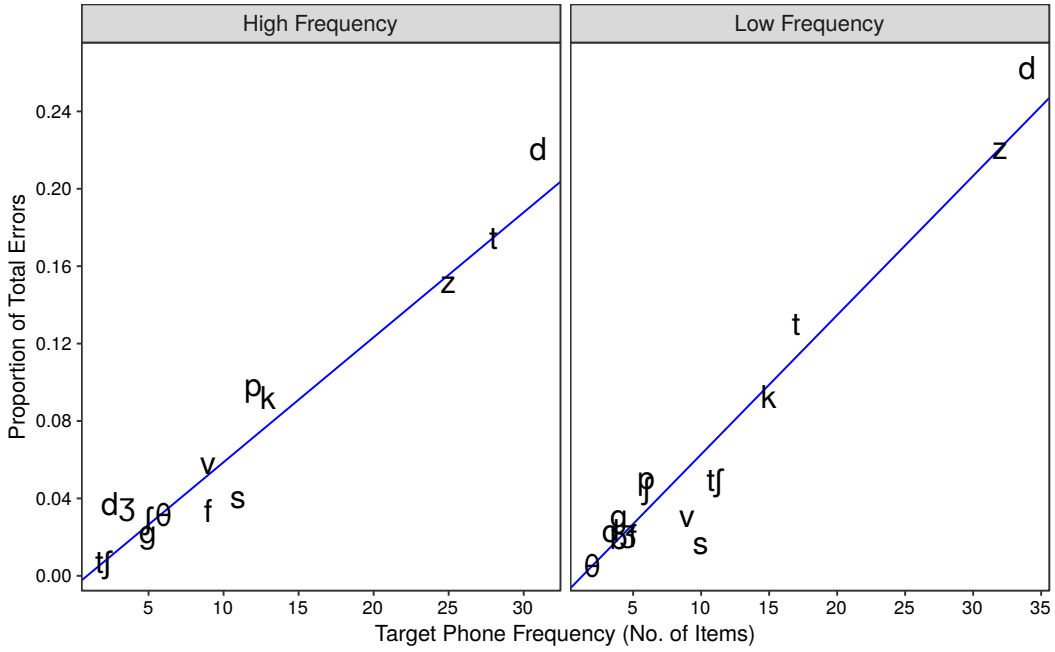

Figure A.54: Proportion of errors in VC position in Experiment 1b attributable to each target phone as a function of word frequency. Lines indicate median regression fits.

Contrast error proportions in Exp. 1a (CV)

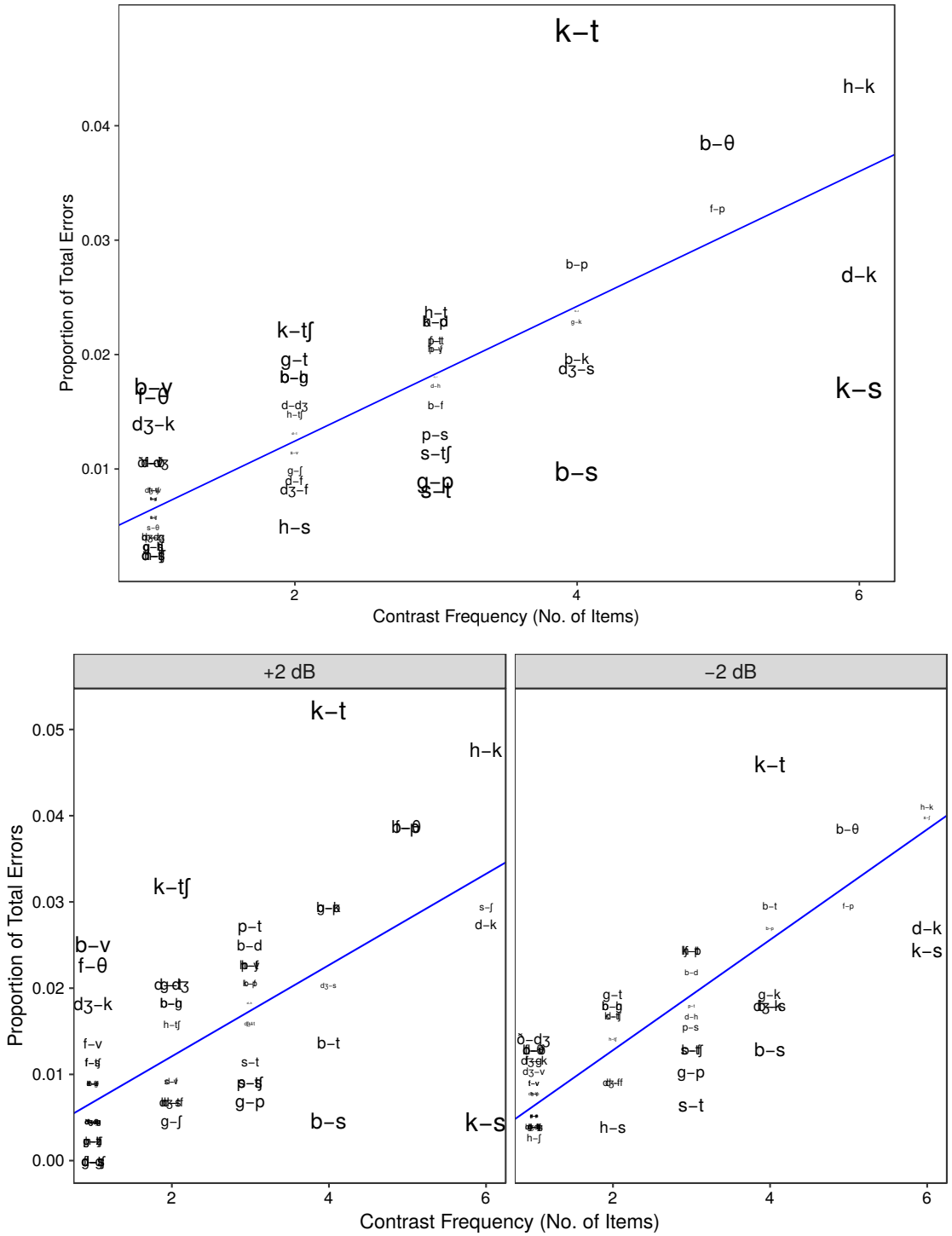

Figure A.55: Proportion of errors in CV position in Exp. 1a attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

Contrast error proportions in Exp. 1b (CV)

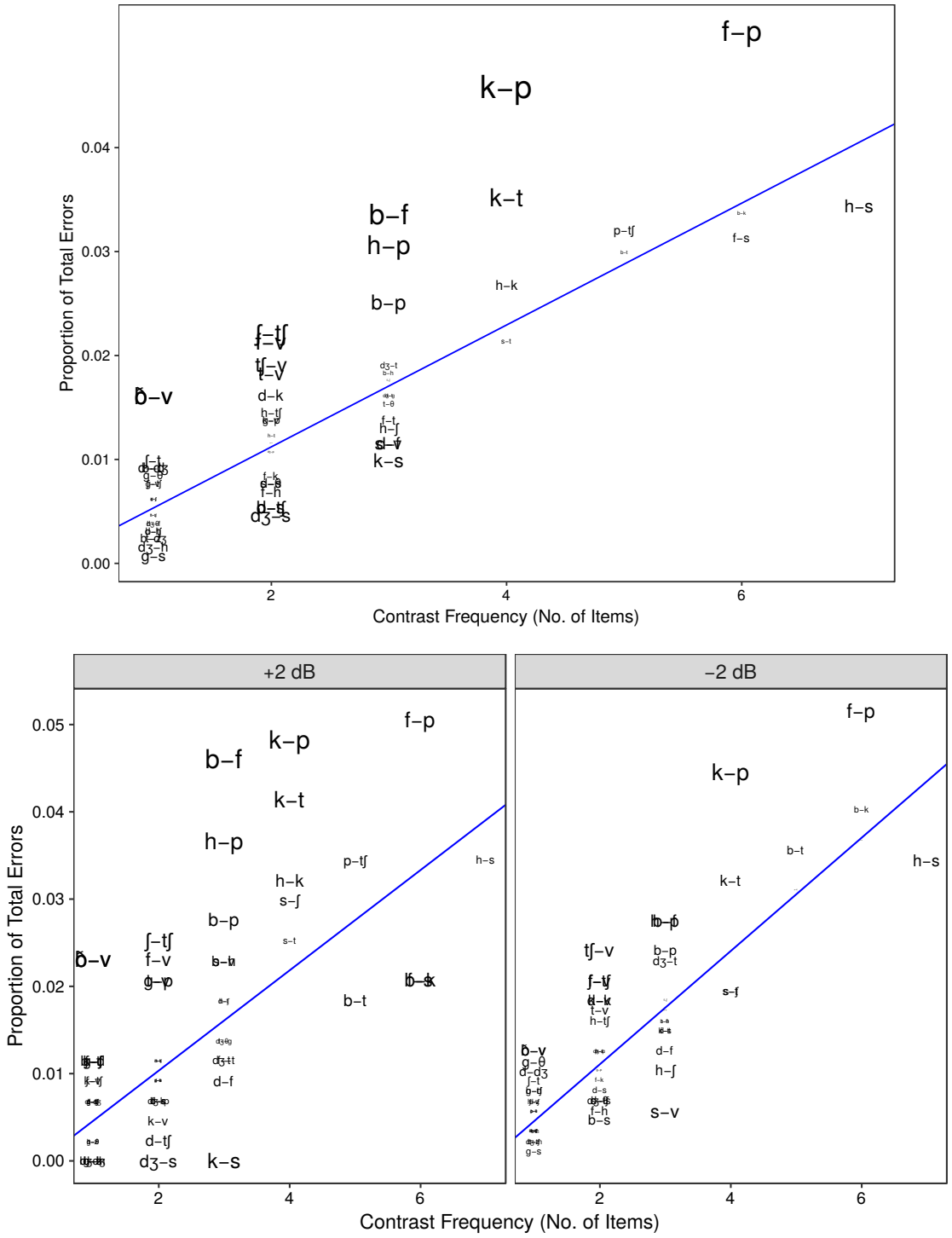

Figure A.56: Proportion of errors in CV position in Exp. 1b attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

Contrast error proportions by word length in Exp. 1a (CV)

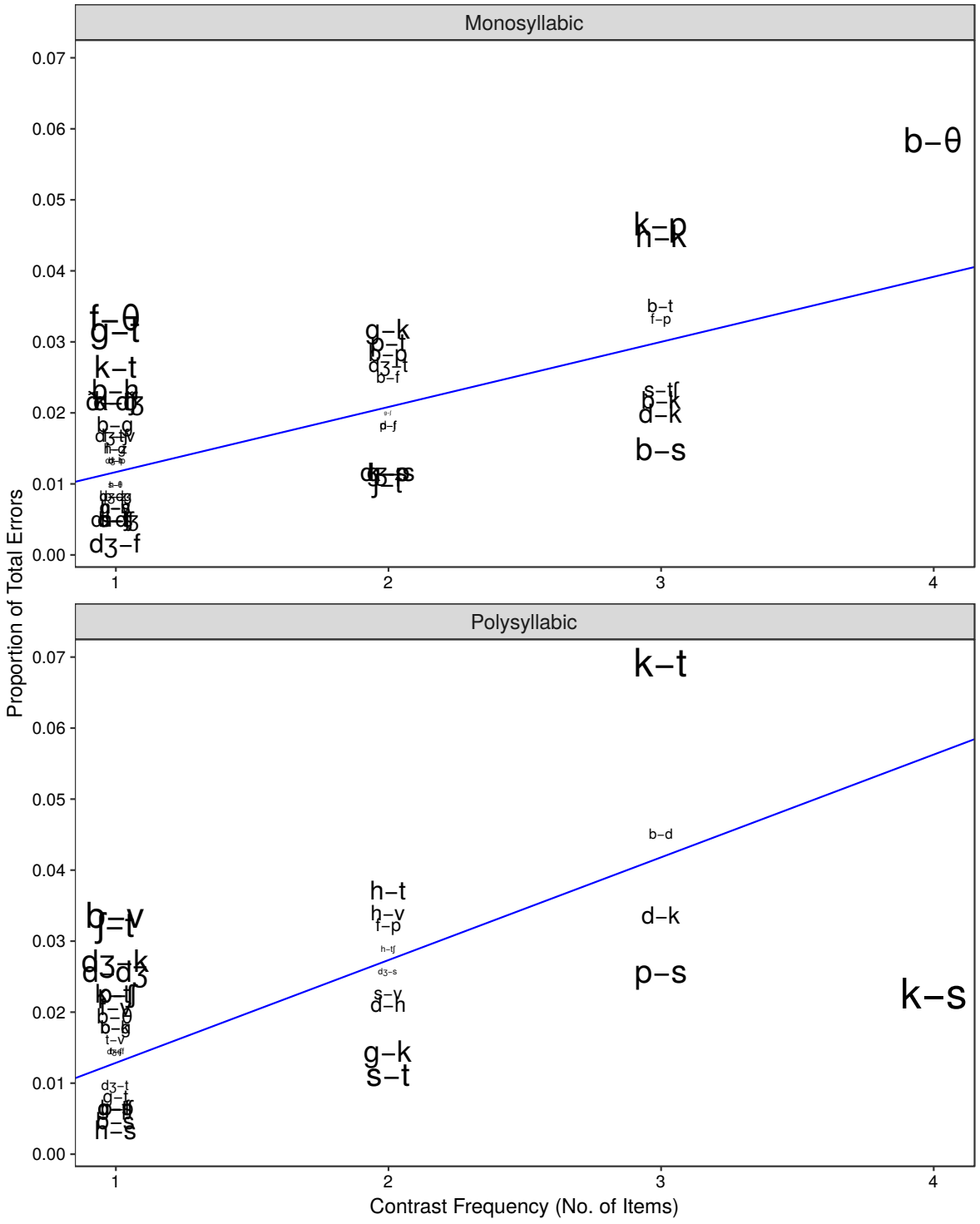

Figure A.57: Proportion of errors in CV position in Exp. 1a attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

Contrast error proportions by word length in Exp. 1b (CV)

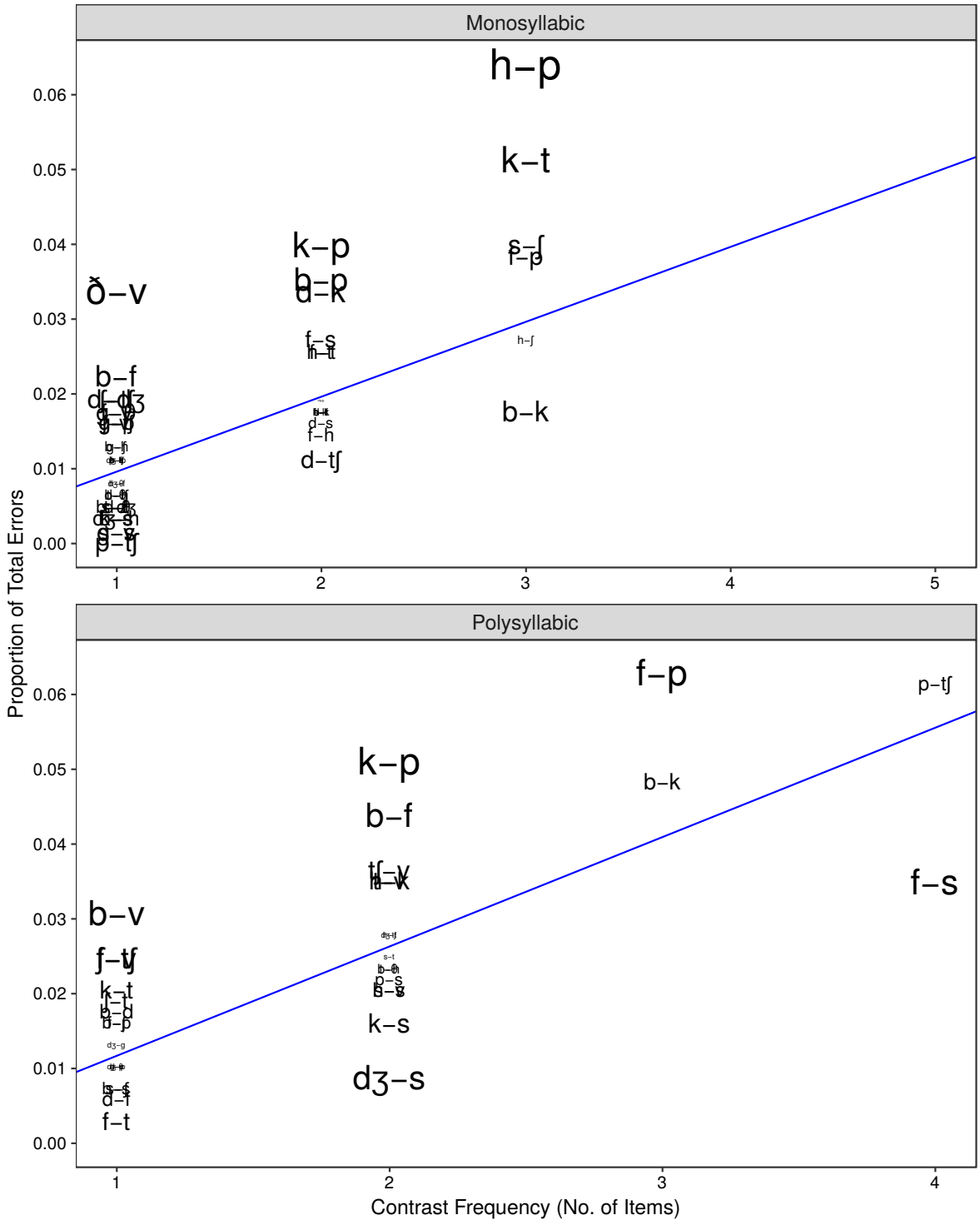

Figure A.58: Proportion of errors in CV position in Exp. 1b attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

Contrast error proportions by word frequency in Exp. 1a (CV)

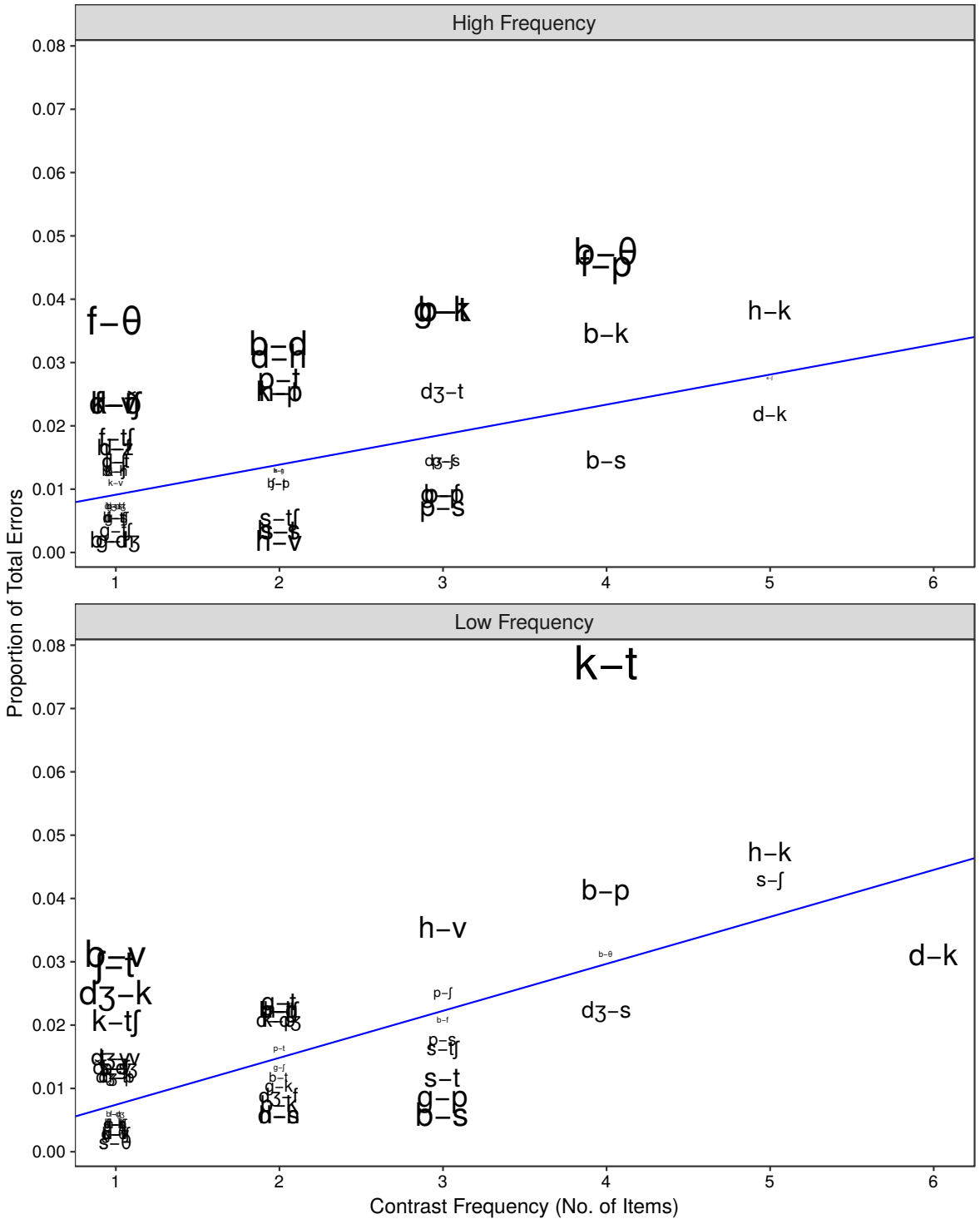

Figure A.59: Proportion of errors in CV position in Exp. 1a attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

Contrast error proportions by word frequency in Exp. 1b (CV)

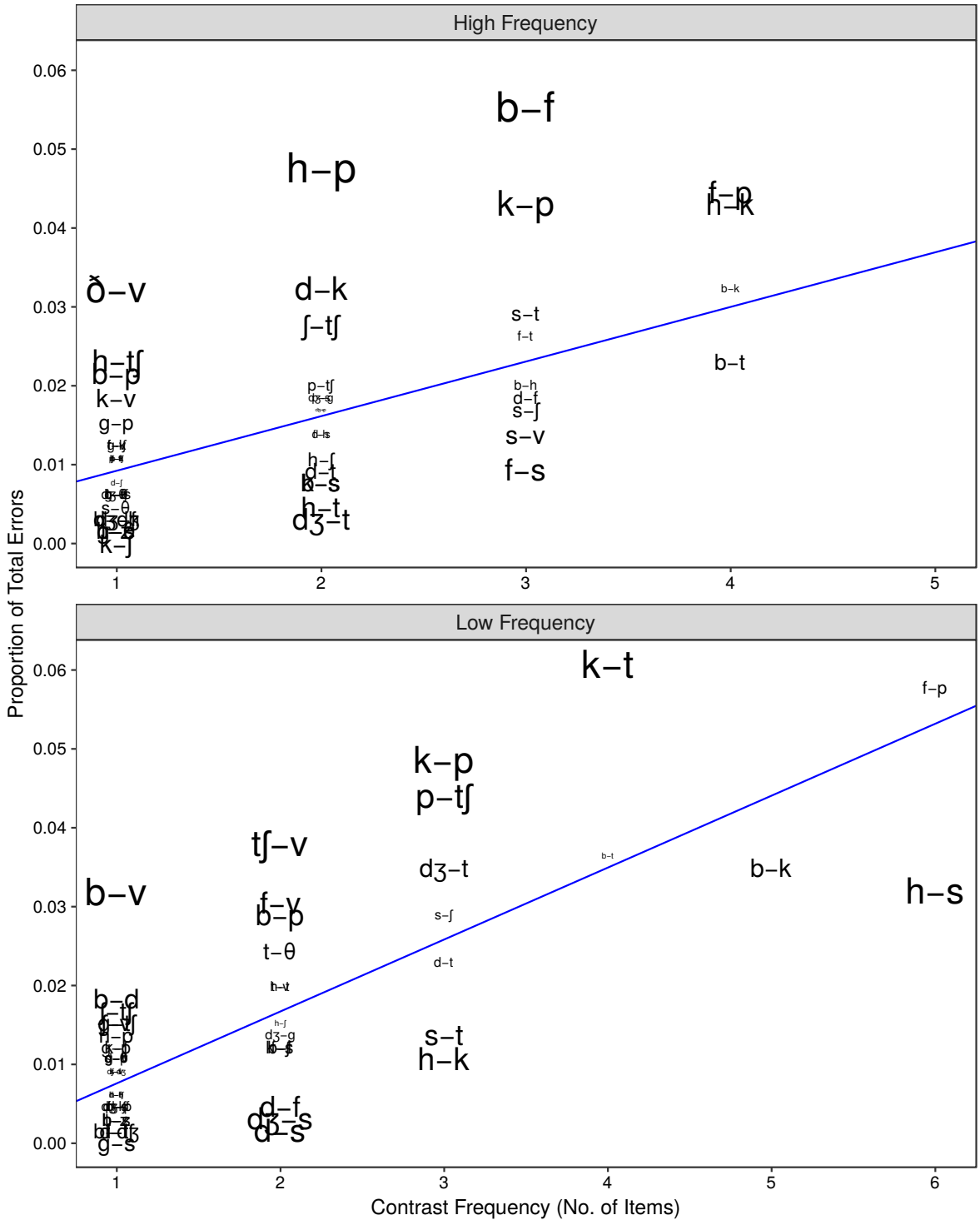

Figure A.60: Proportion of errors in CV position in Exp. 1b attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

Contrast error proportions in Exp. 1a (VCV)

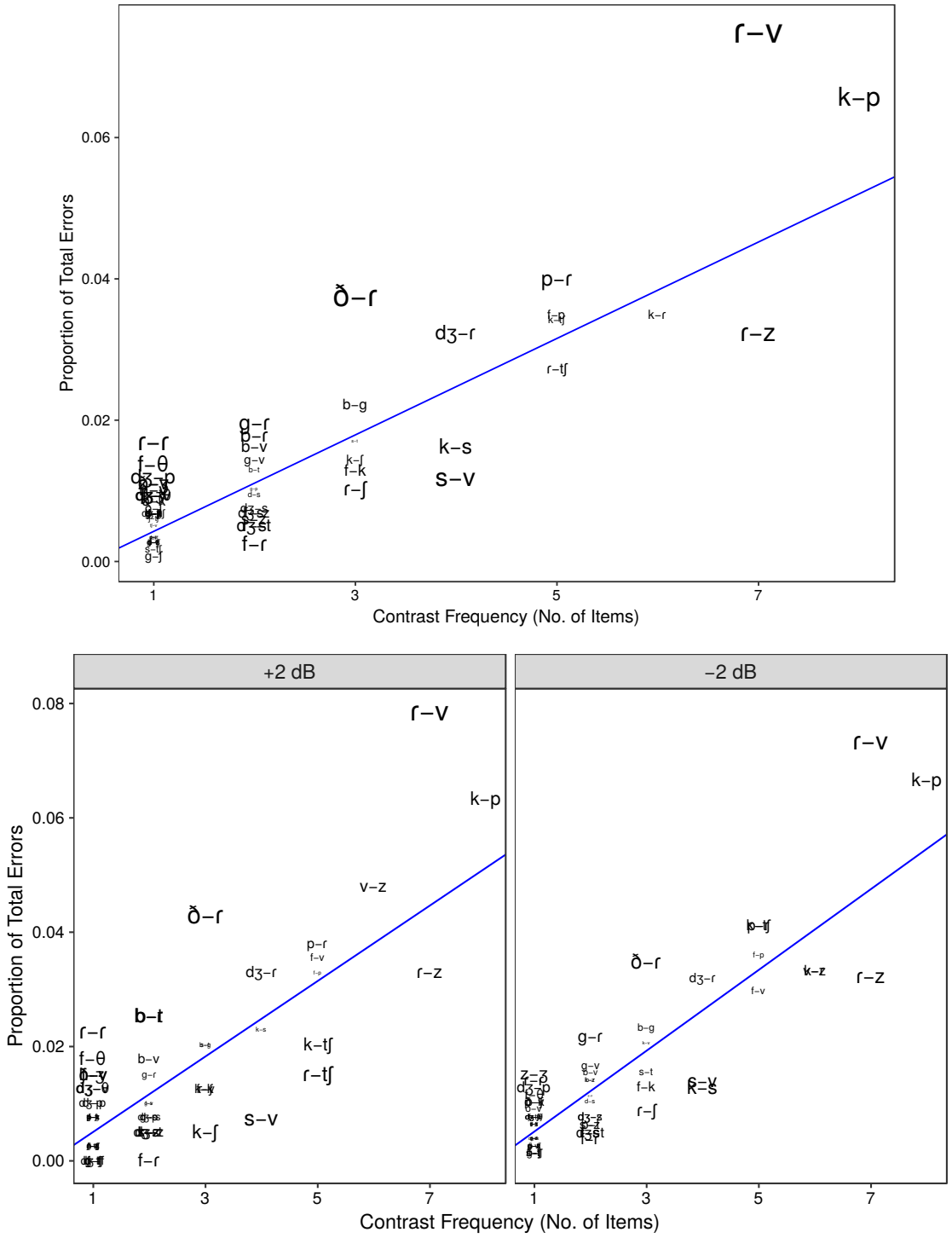

Figure A.61: Proportion of errors in VCV position in Exp. 1a attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

Contrast error proportions in Exp. 1b (VCV)

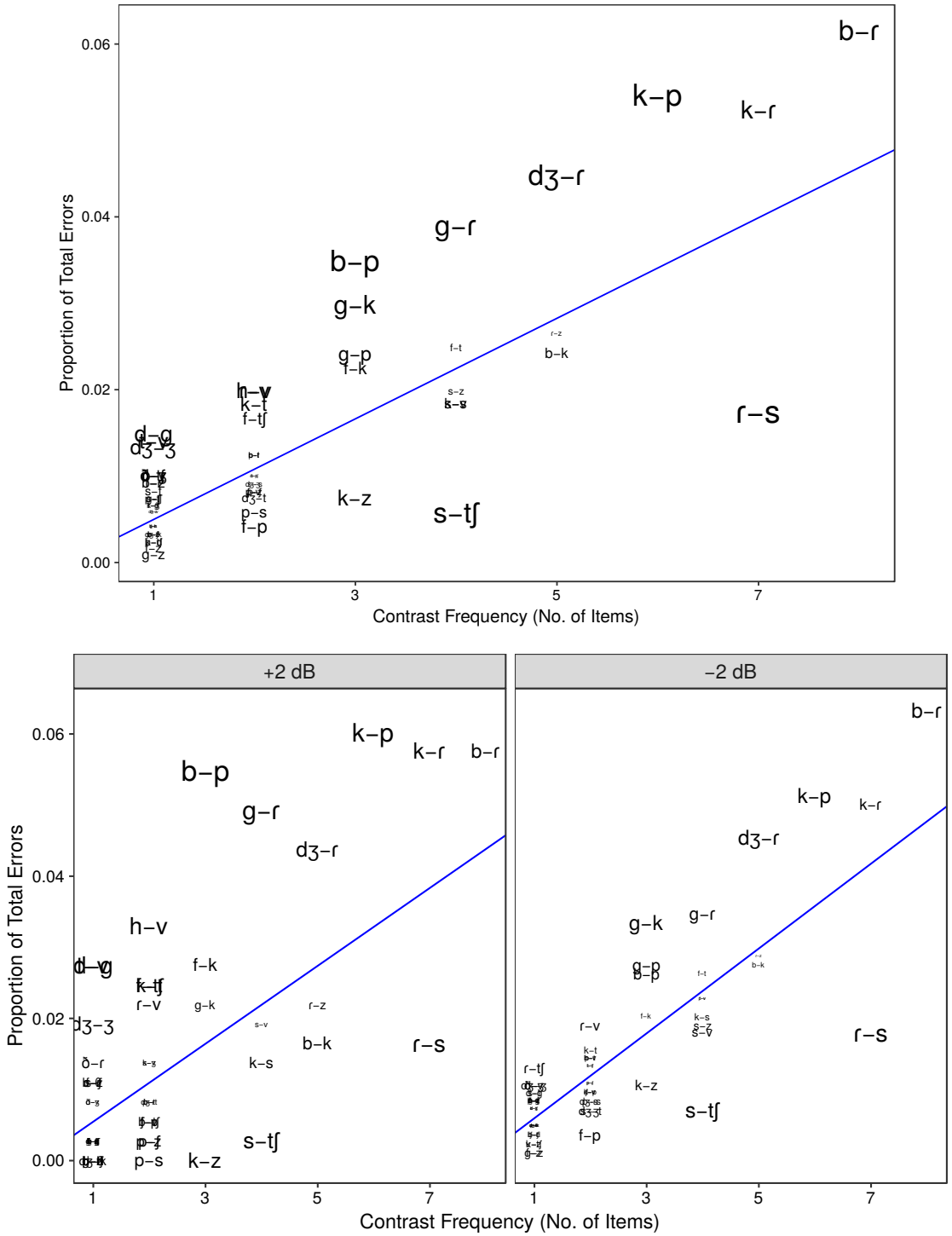

Figure A.62: Proportion of errors in VCV position in Exp. 1b attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

Contrast error proportions by word length in Exp. 1a (VCV)

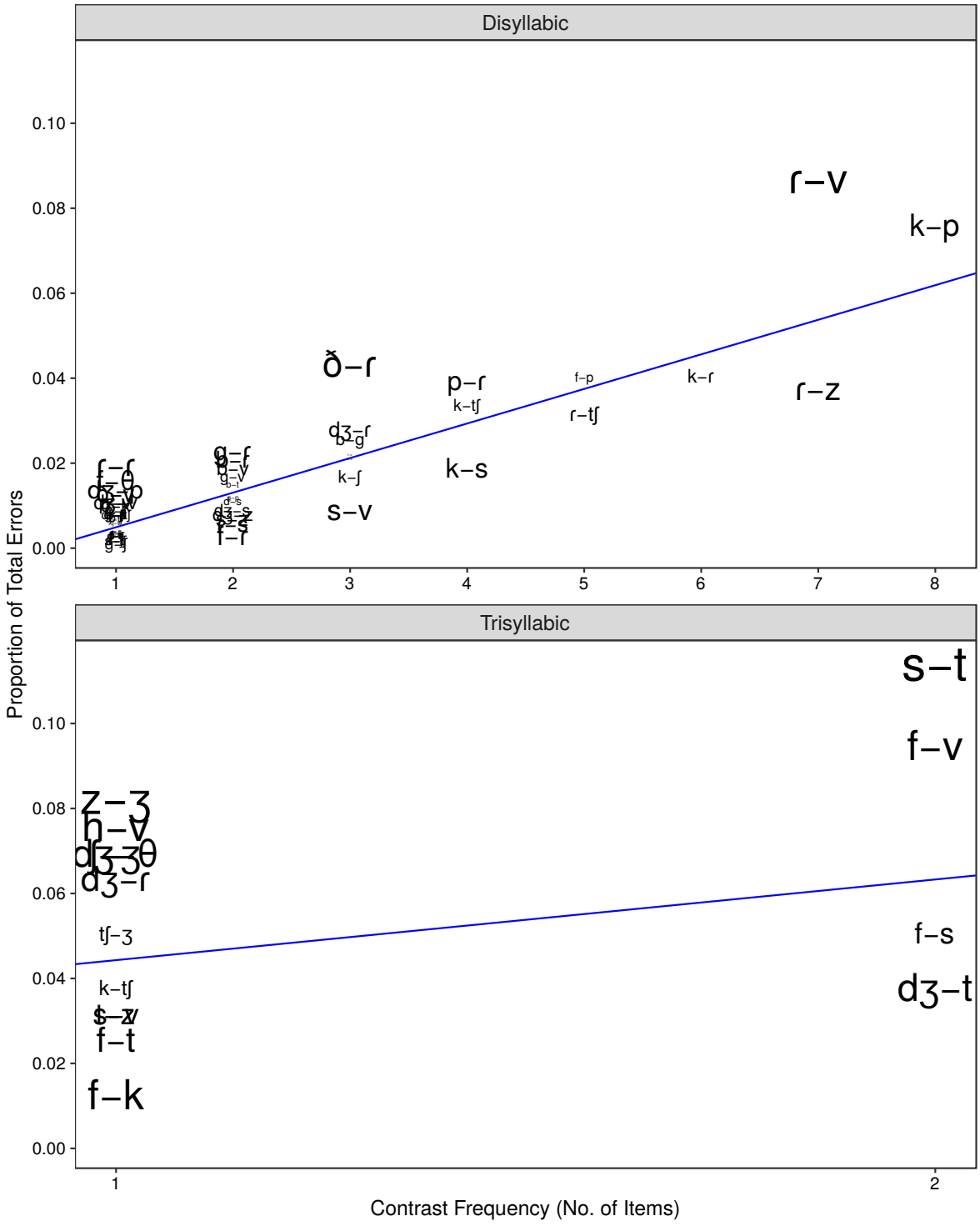

Figure A.63: Proportion of errors in VCV position in Exp. 1a attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

Contrast error proportions by word length in Exp. 1b (VCV)

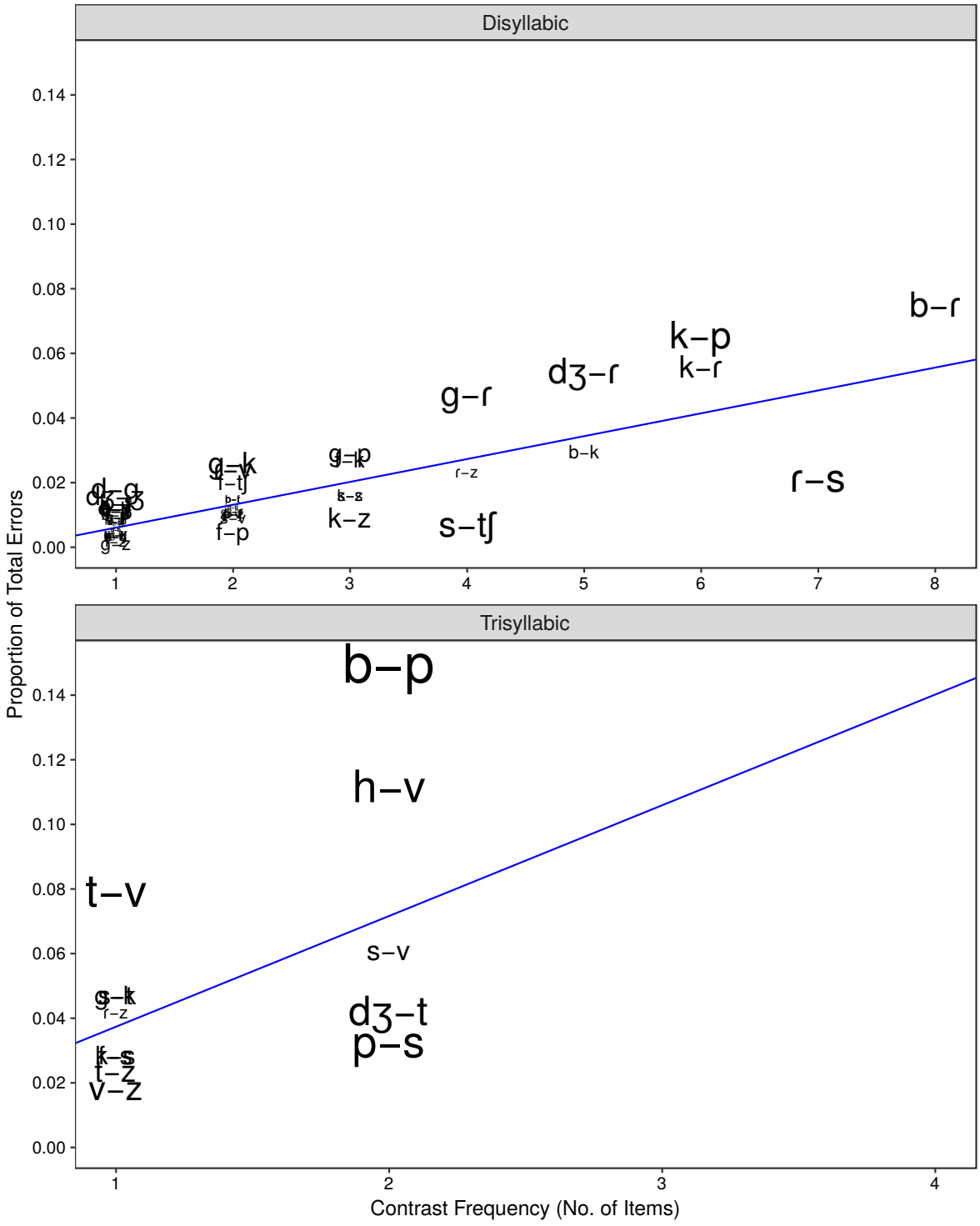

Figure A.64: Proportion of errors in VCV position in Exp. 1b attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

Contrast error proportions by word frequency in Exp. 1a (VCV)

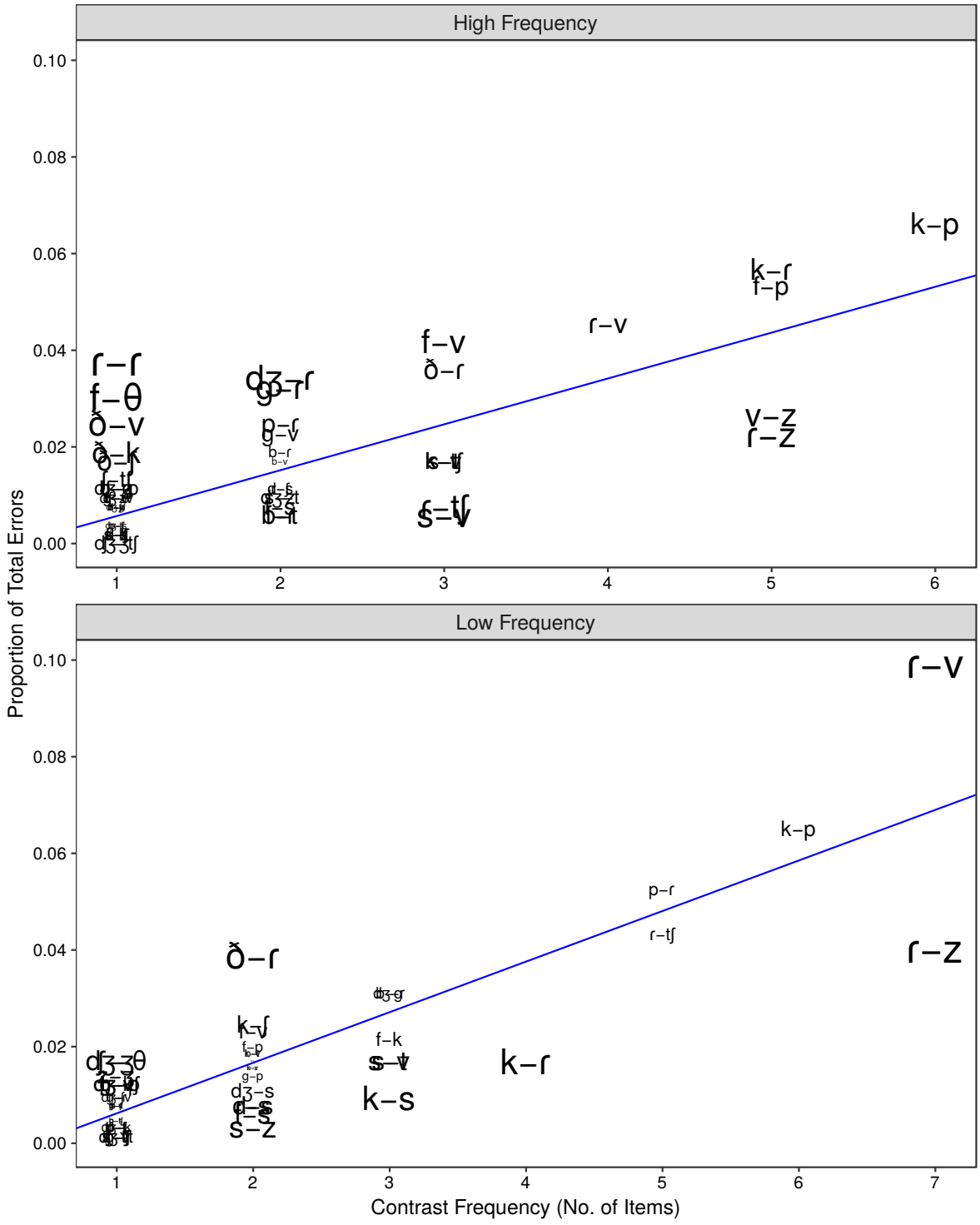

Figure A.65: Proportion of errors in VCV position in Exp. 1a attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

Contrast error proportions by word frequency in Exp. 1b (VCV)

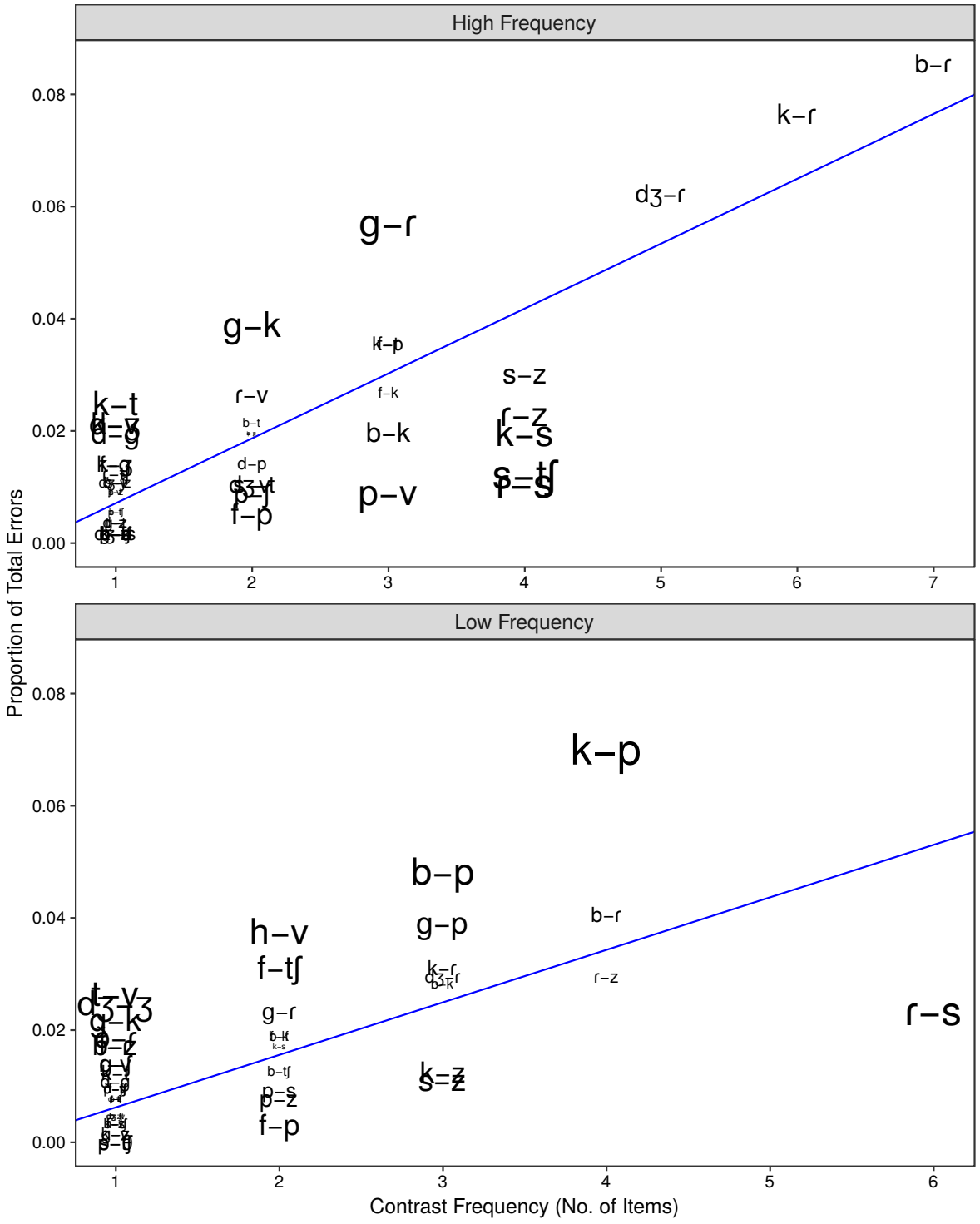

Figure A.66: Proportion of errors in VCV position in Exp. 1b attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

Contrast error proportions in Exp. 1a (VC)

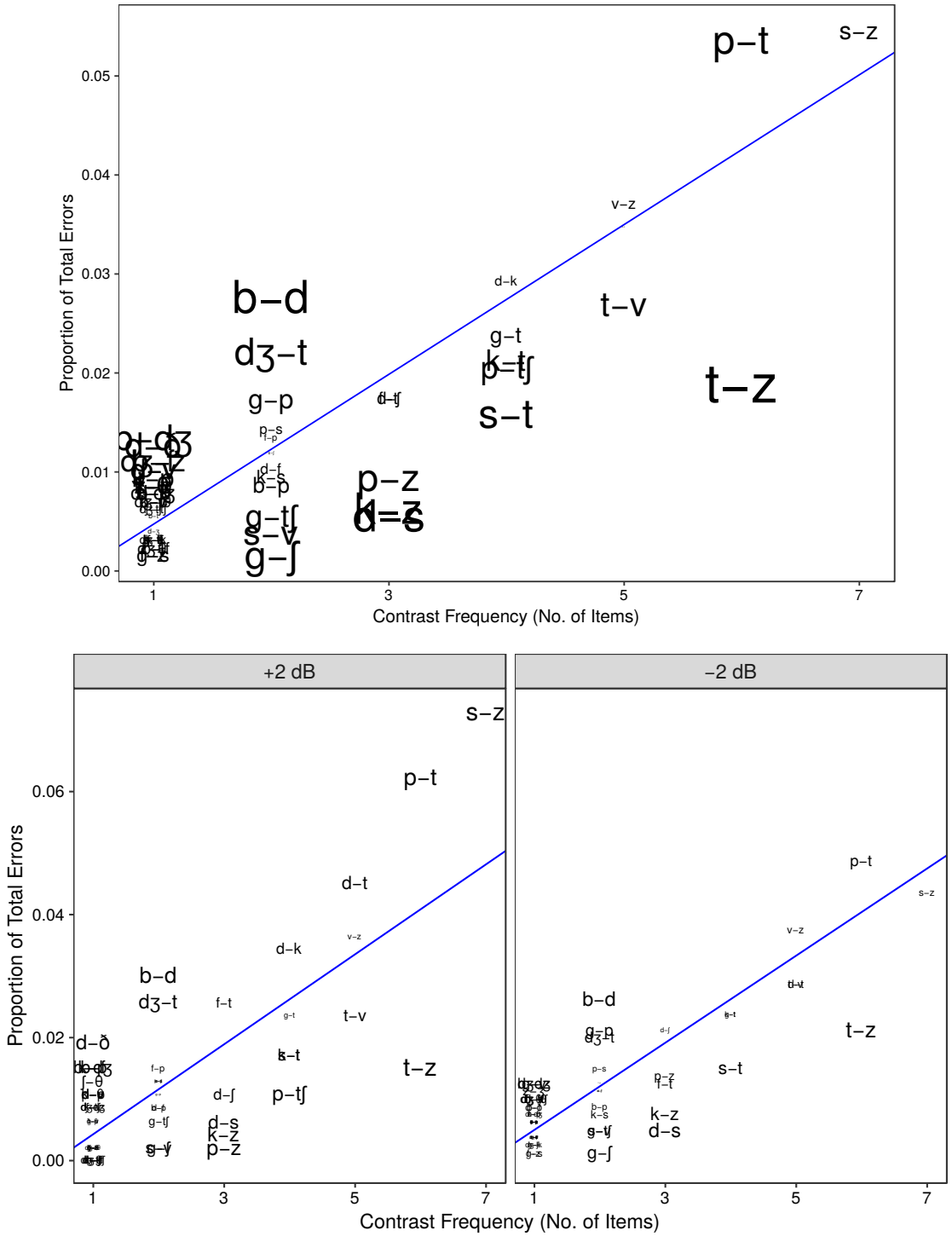

Figure A.67: Proportion of errors in VC position in Exp. 1a attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

Contrast error proportions in Exp. 1b (VC)

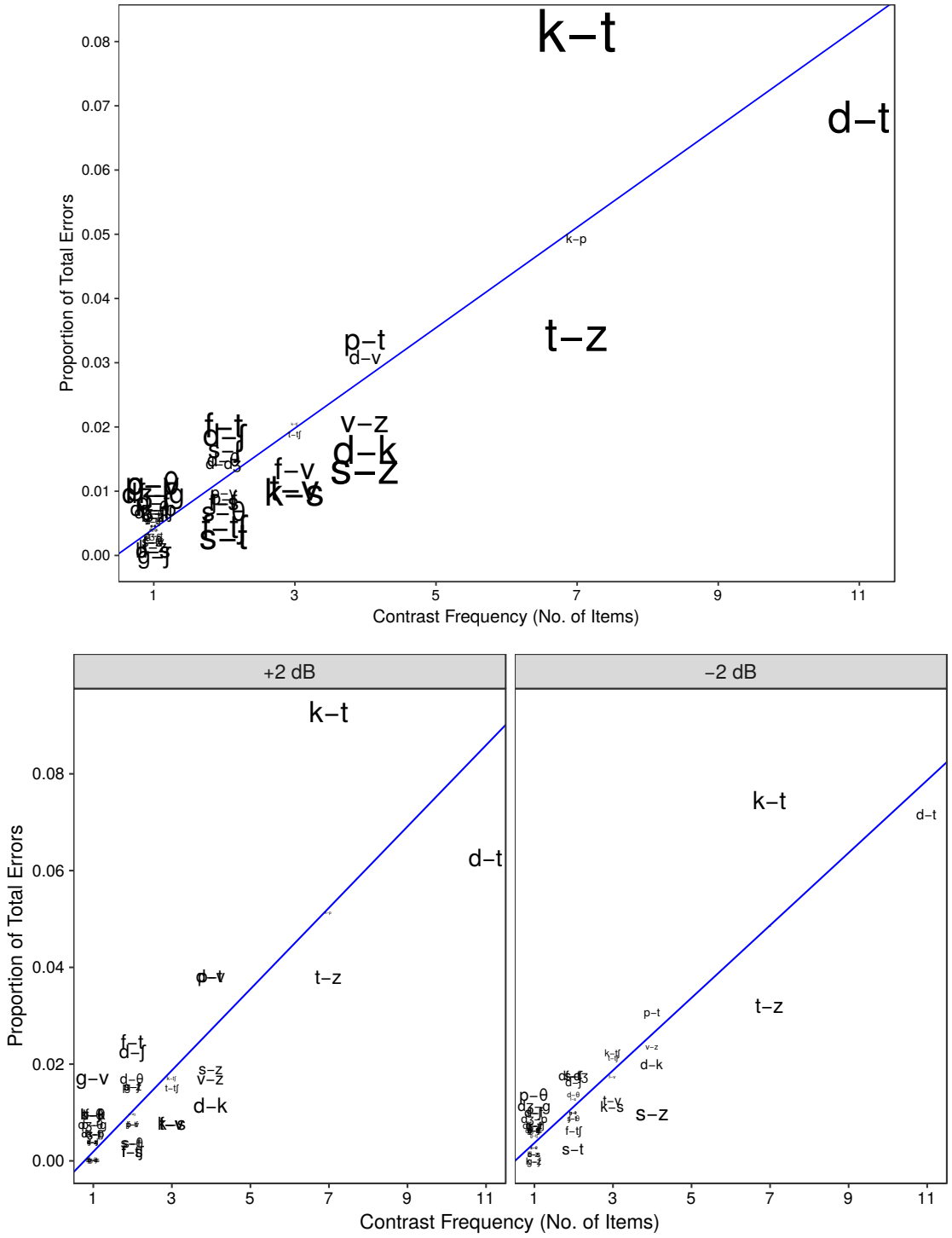

Figure A.68: Proportion of errors in VC position in Exp. 1b attributable to each contrast as a function of item count, both overall (upper panel) and by SNR (lower panels). Lines indicate median regression fits.

Contrast error proportions by word length in Exp. 1a (VC)

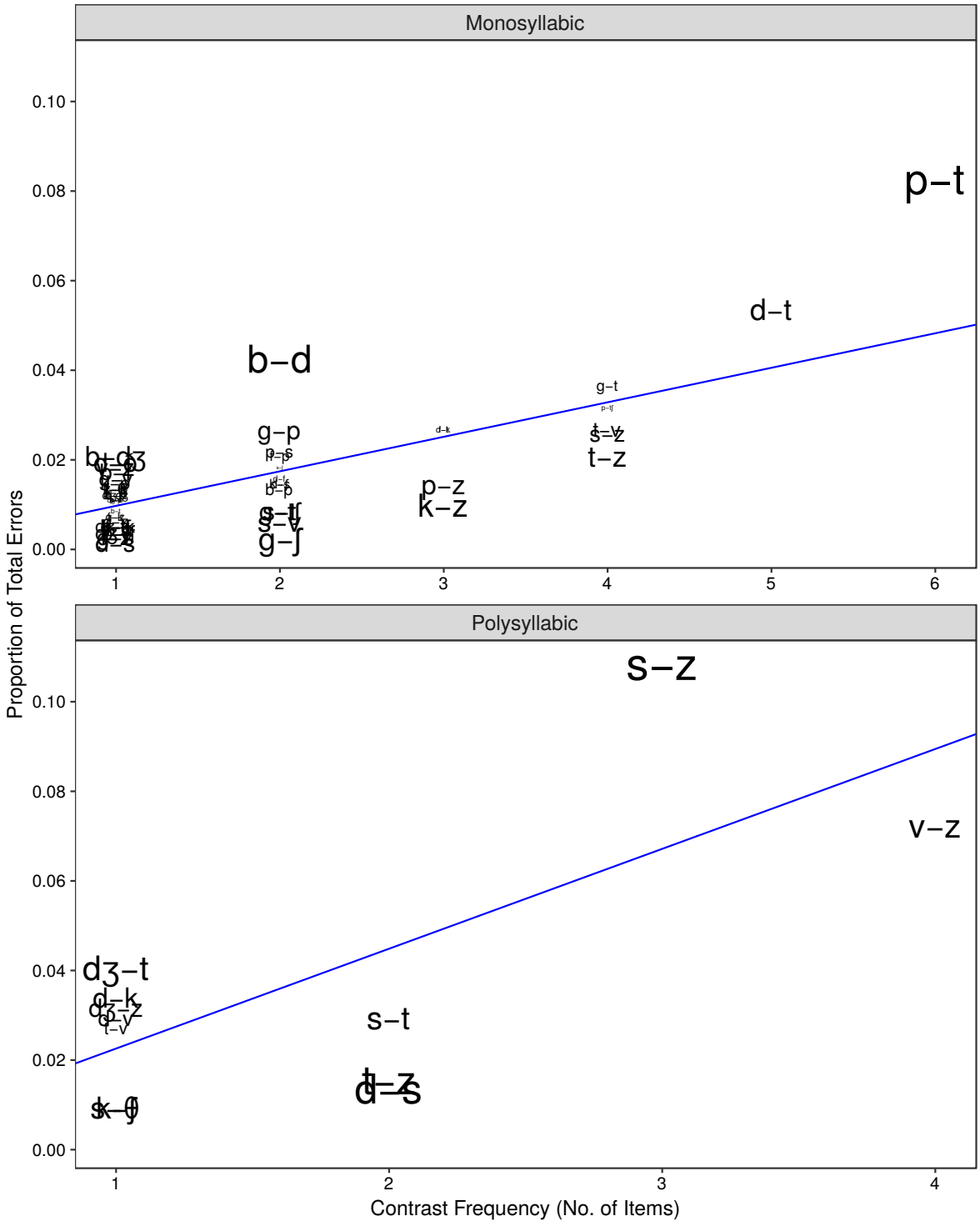

Figure A.69: Proportion of errors in VC position in Exp. 1a attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

Contrast error proportions by word length in Exp. 1b (VC)

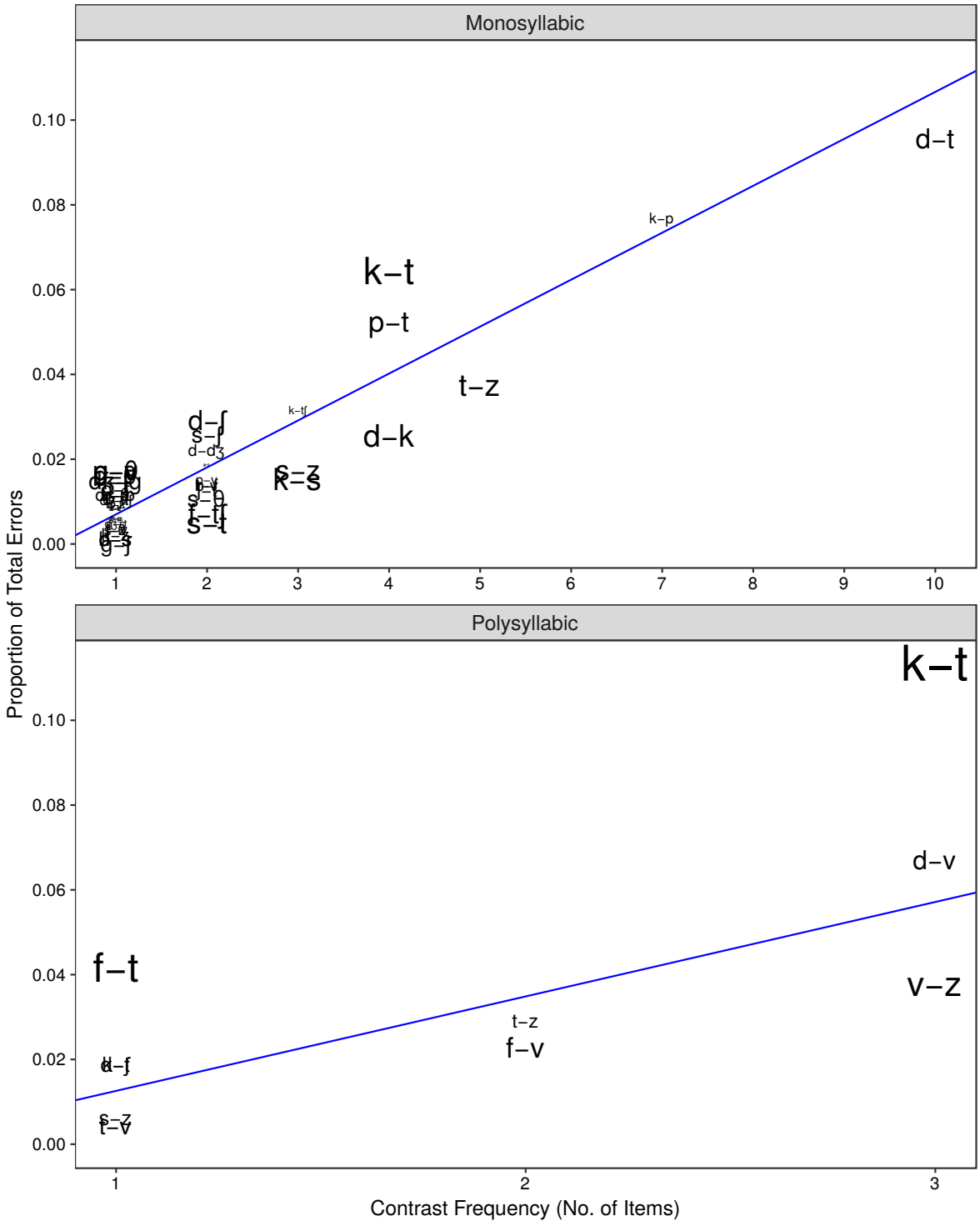

Figure A.70: Proportion of errors in VC position in Exp. 1b attributable to each contrast as a function of item count and word length. Lines indicate median regression fits.

Contrast error proportions by word frequency in Exp. 1a (VC)

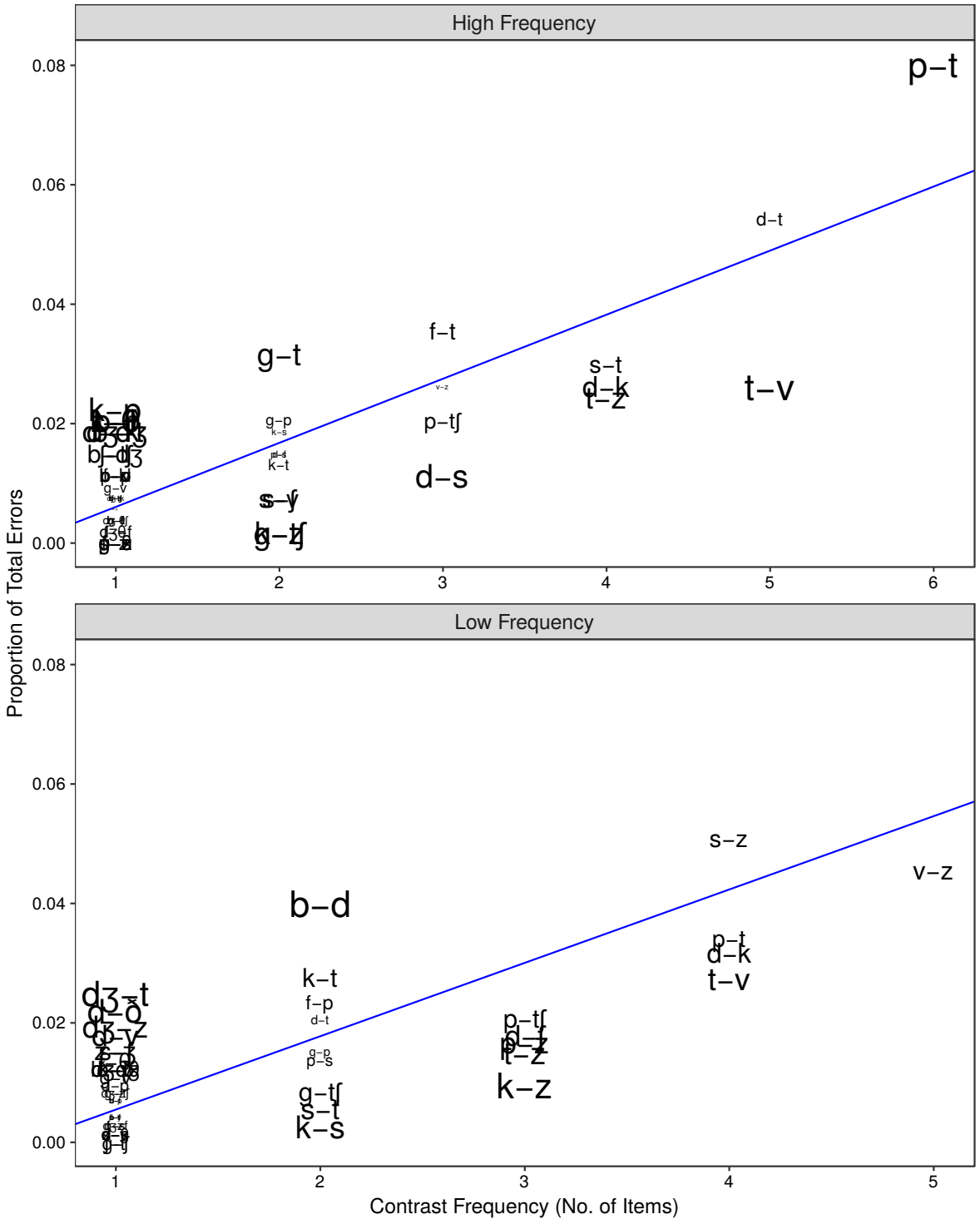

Figure A.71: Proportion of errors in VC position in Exp. 1a attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

Contrast error proportions by word frequency in Exp. 1b (VC)

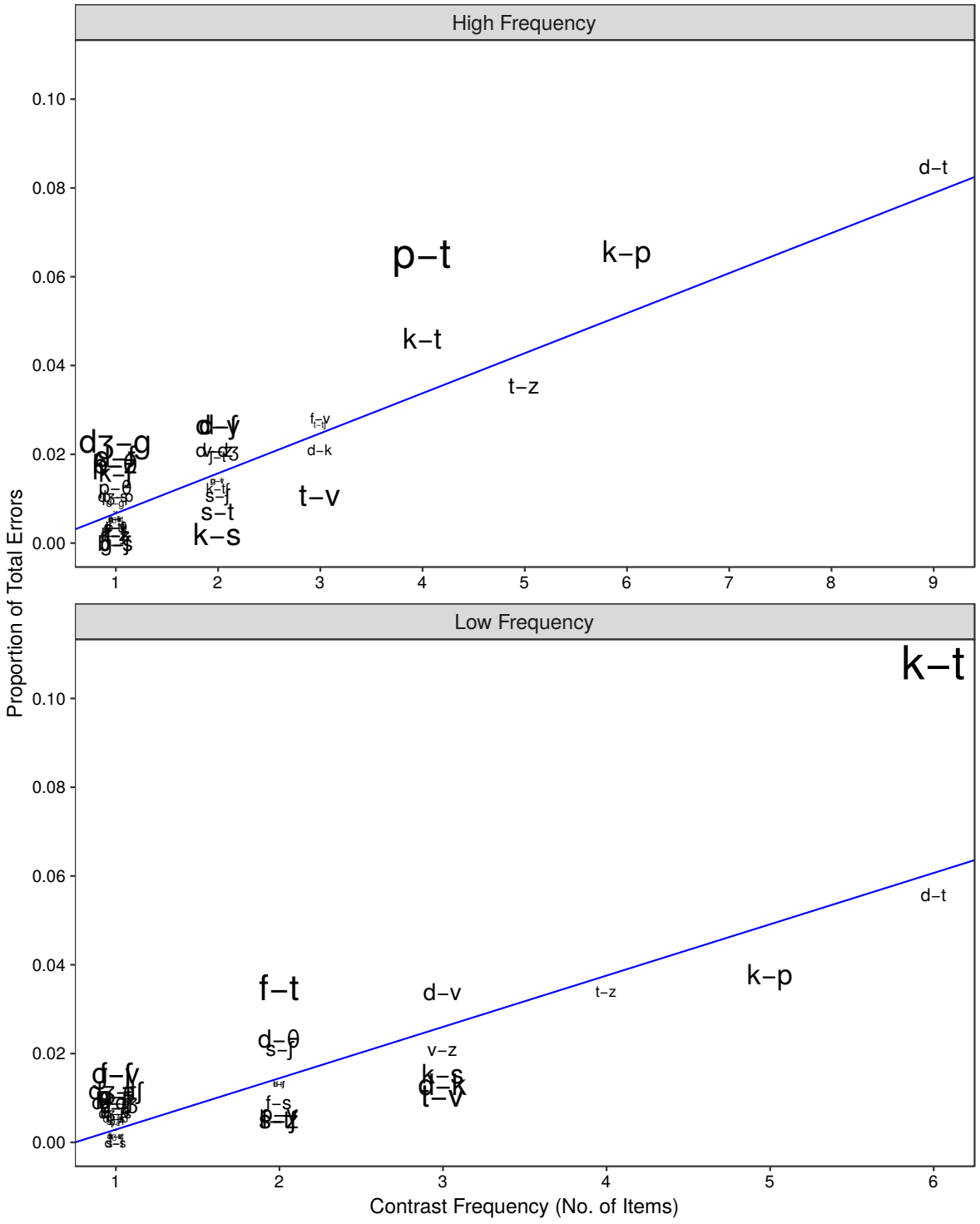

Figure A.72: Proportion of errors in VC position in Exp. 1b attributable to each contrast as a function of item count and word frequency. Lines indicate median regression fits.

Chapter 4: Cue integration
Supplementary tables and figures

Target parameter ranks in Exp. 1a (CV)

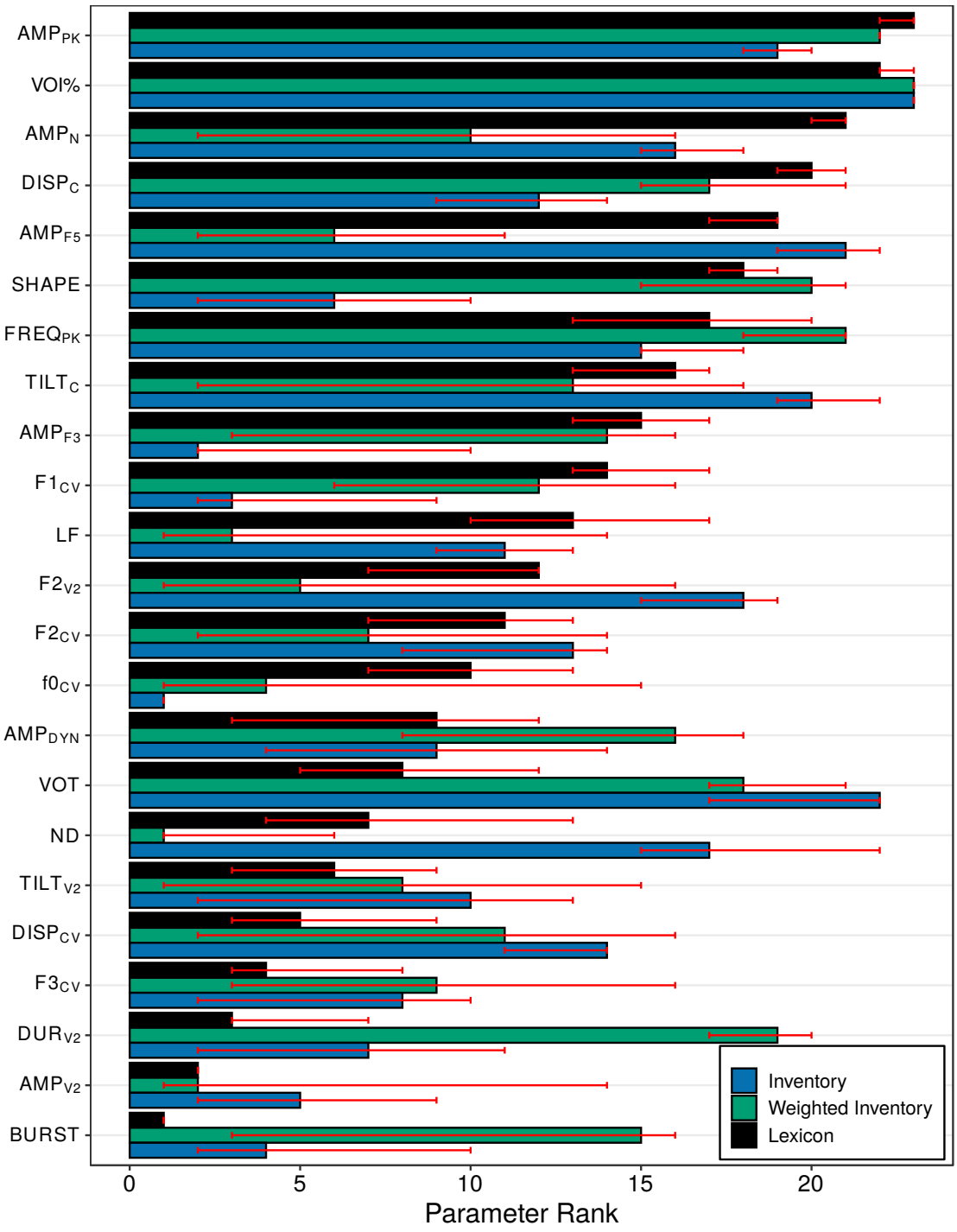

Figure A.73: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-initial contrasts fit to listener recognition data in Exp. 1a. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

Target parameter ranks in Exp. 1b (CV)

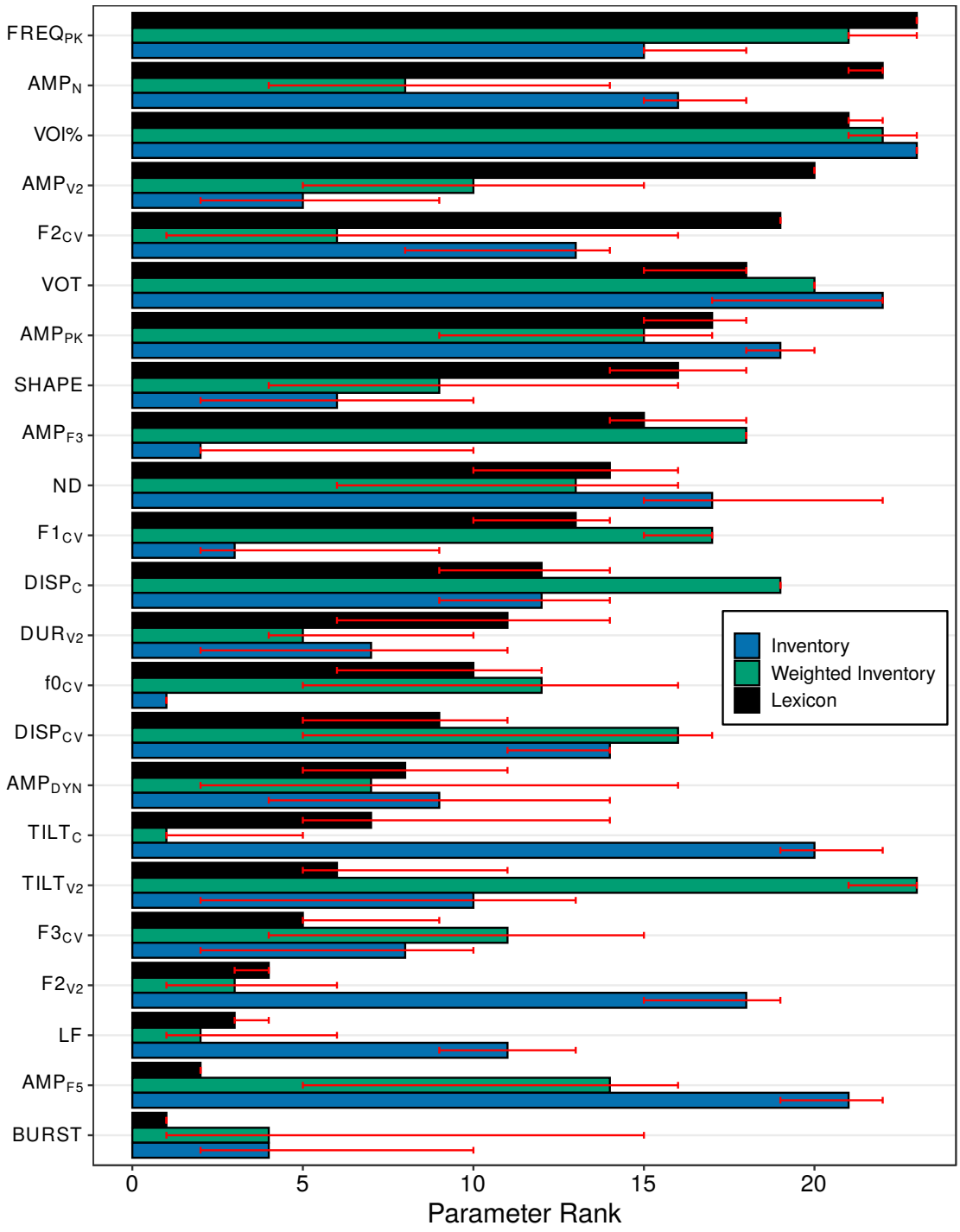

Figure A.74: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-initial contrasts fit to listener recognition data in Exp. 1b. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter ranks in Exp. 1a (CV)

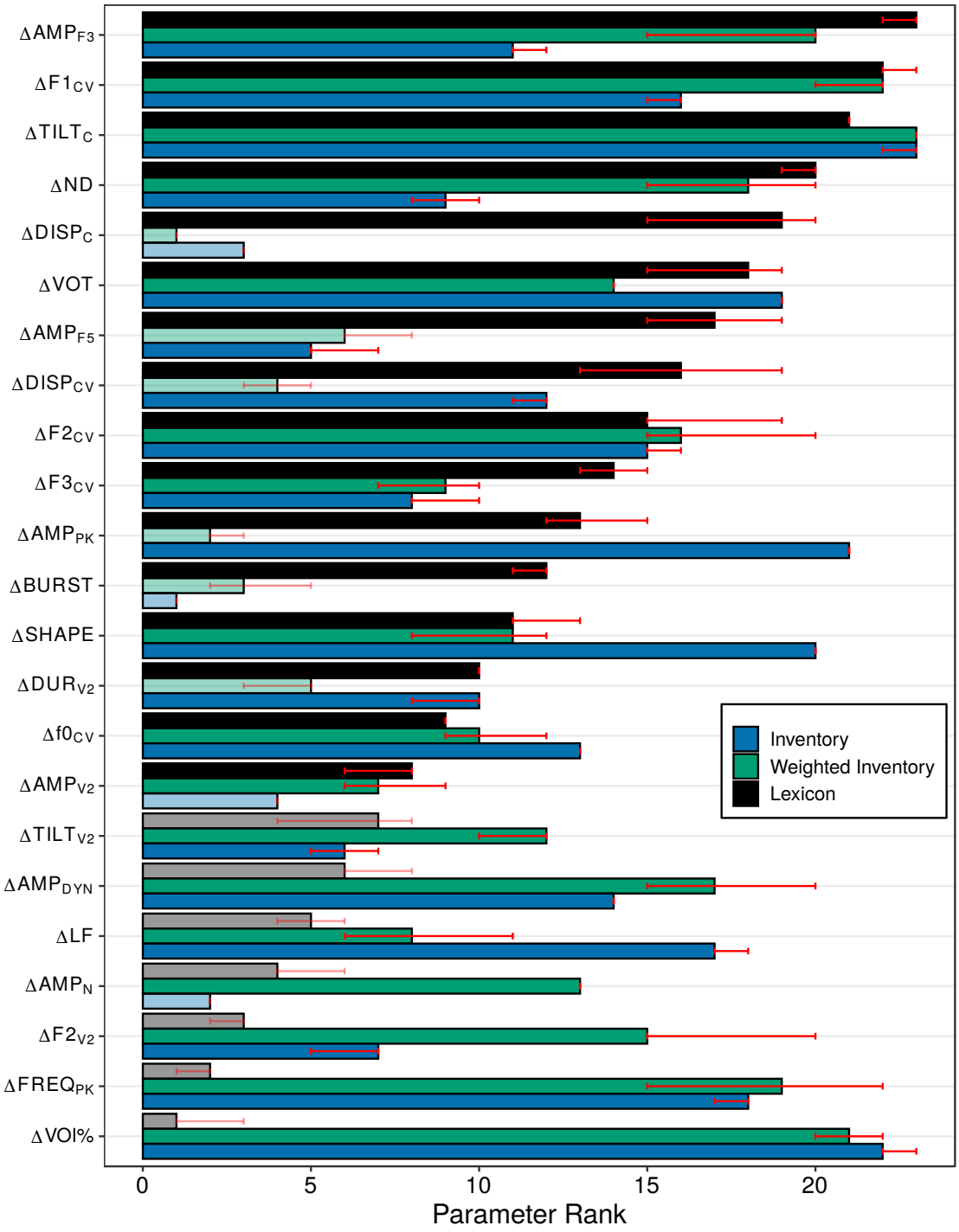

Figure A.75: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-initial contrasts fit to listener recognition data in Exp. 1a. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter ranks in Exp. 1b (CV)

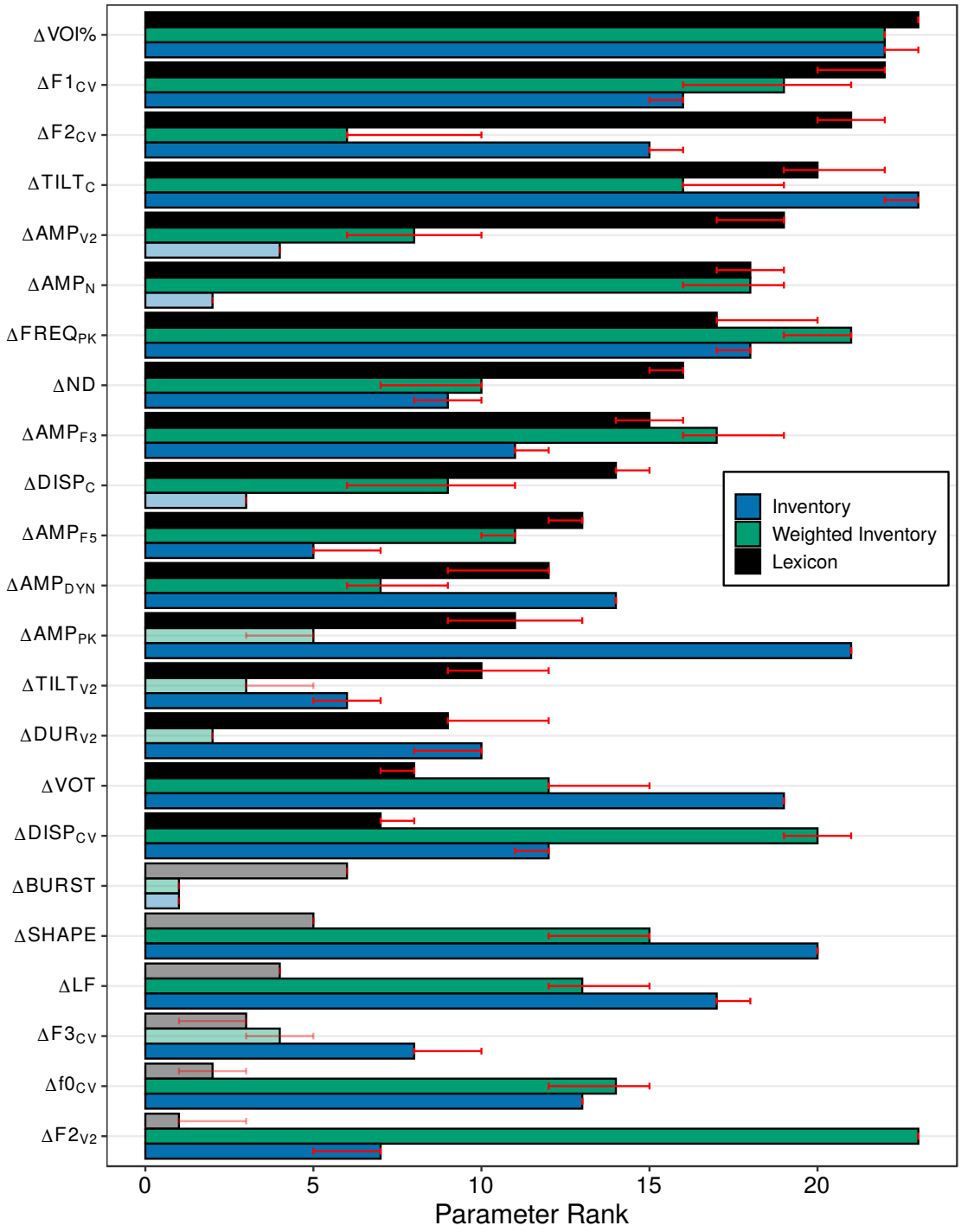

Figure A.76: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-initial contrasts fit to listener recognition data in Exp. 1b. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter correlations in Exp. 1a (CV)

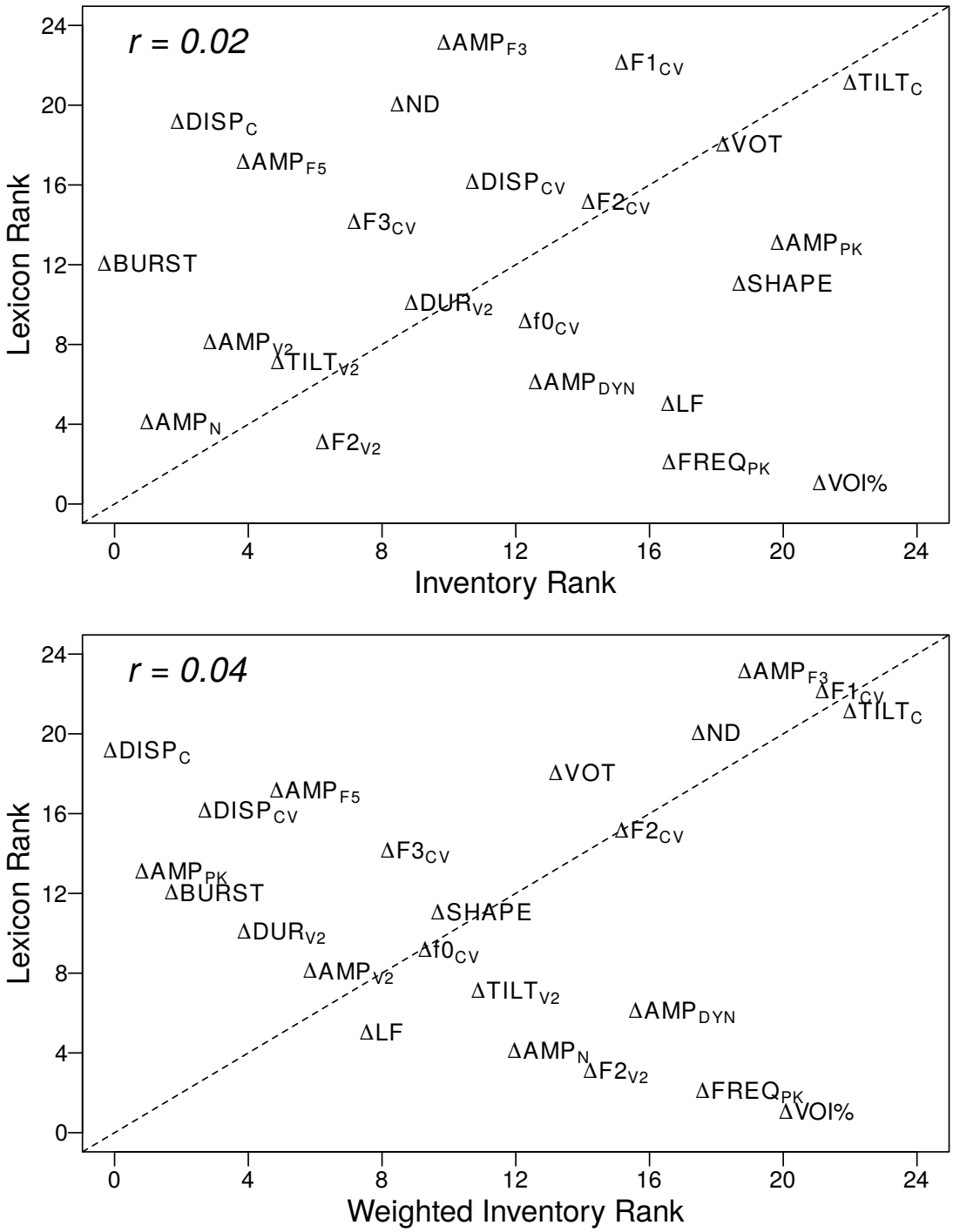

Figure A.77: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in CV position in Exp. 1a. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

Contrast parameter correlations in Exp. 1b (CV)

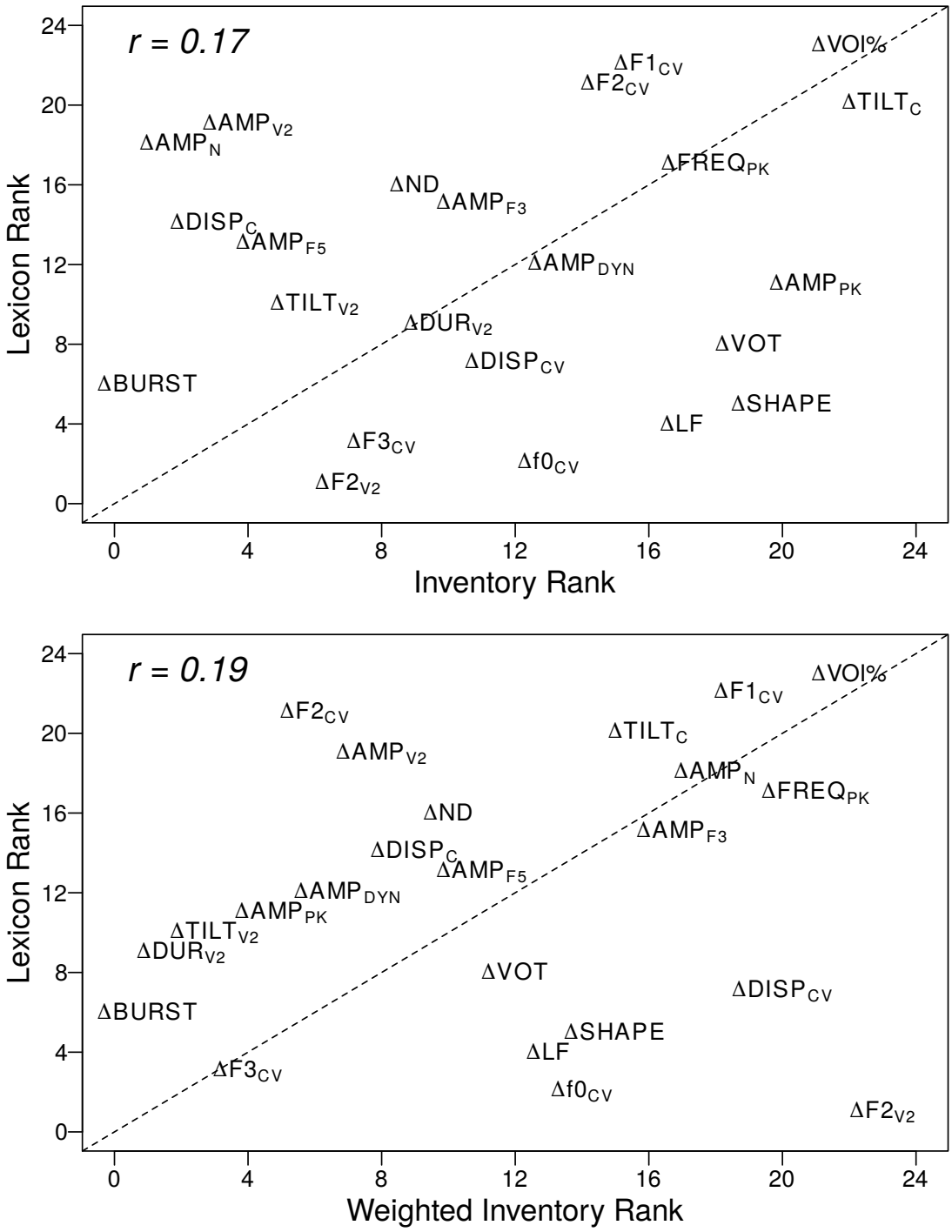

Figure A.78: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in CV position in Exp. 1b. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

Contrast parameter differences in Exp. 1a (CV)

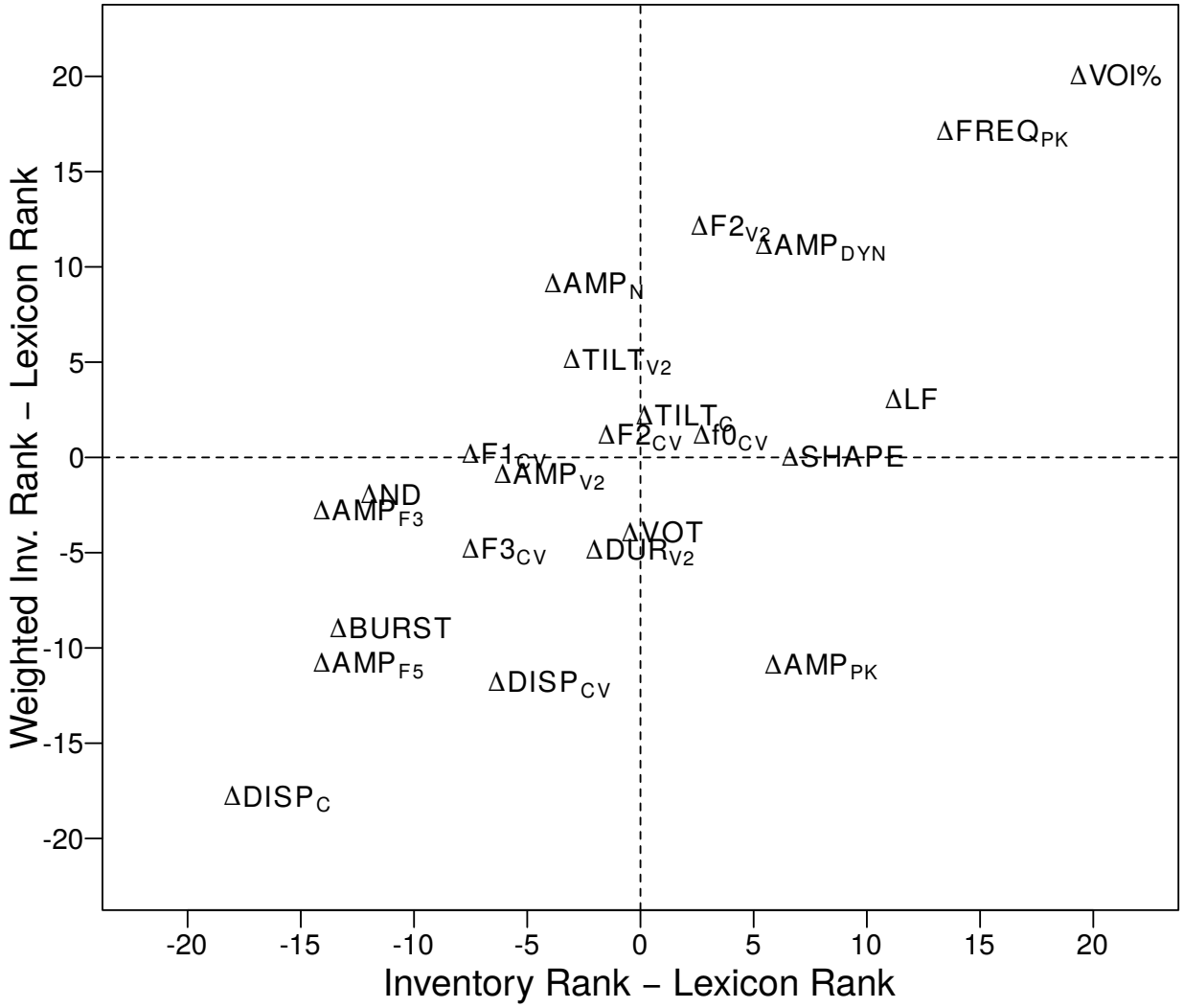

Figure A.79: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in CV position in Exp. 1a. Dashed lines indicate equivalence relations between each pair of models.

Contrast parameter differences in Exp. 1b (CV)

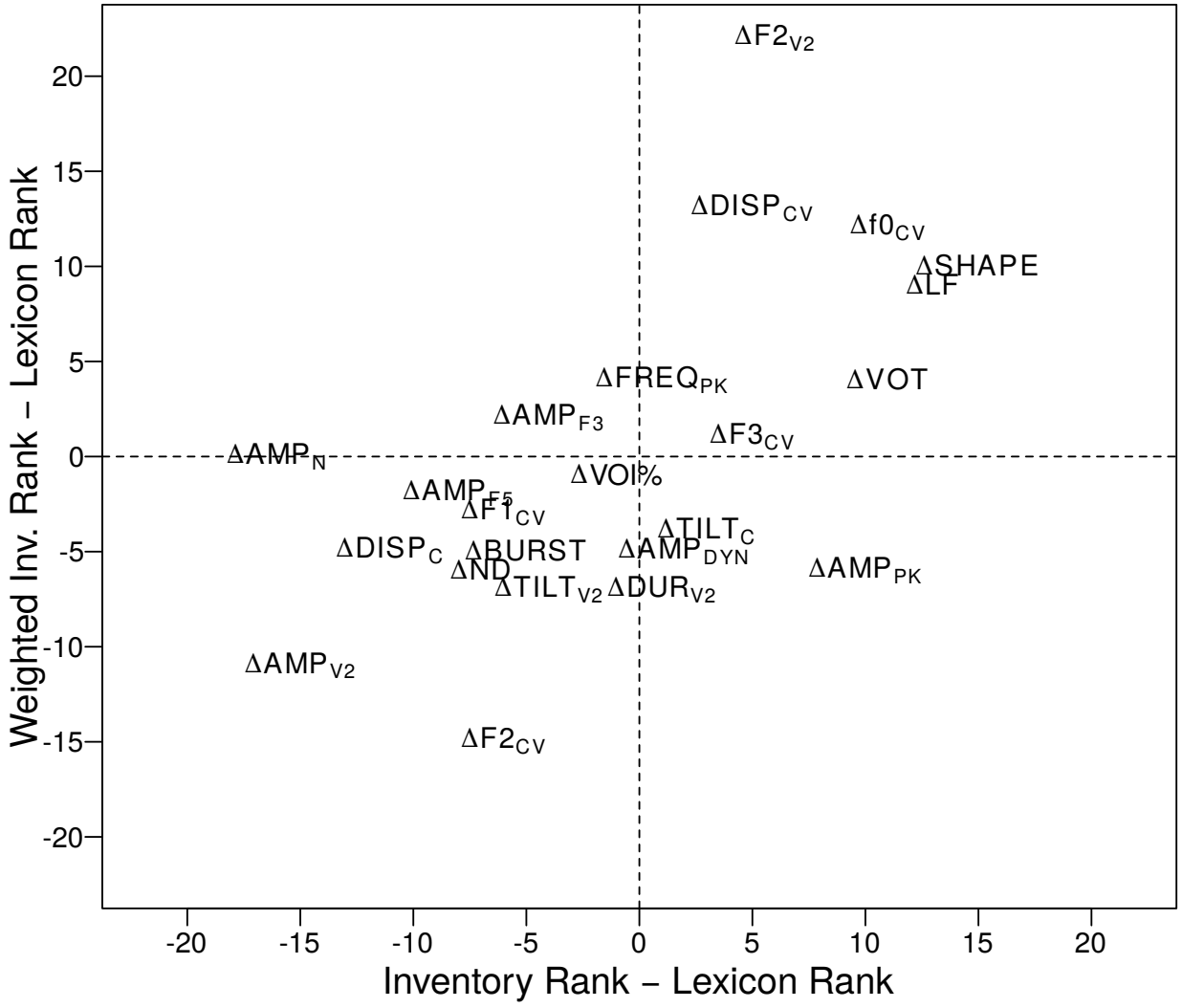

Figure A.80: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in CV position in Exp. 1b. Dashed lines indicate equivalence relations between each pair of models.

Target parameter ranks in Exp. 1a (VCV)

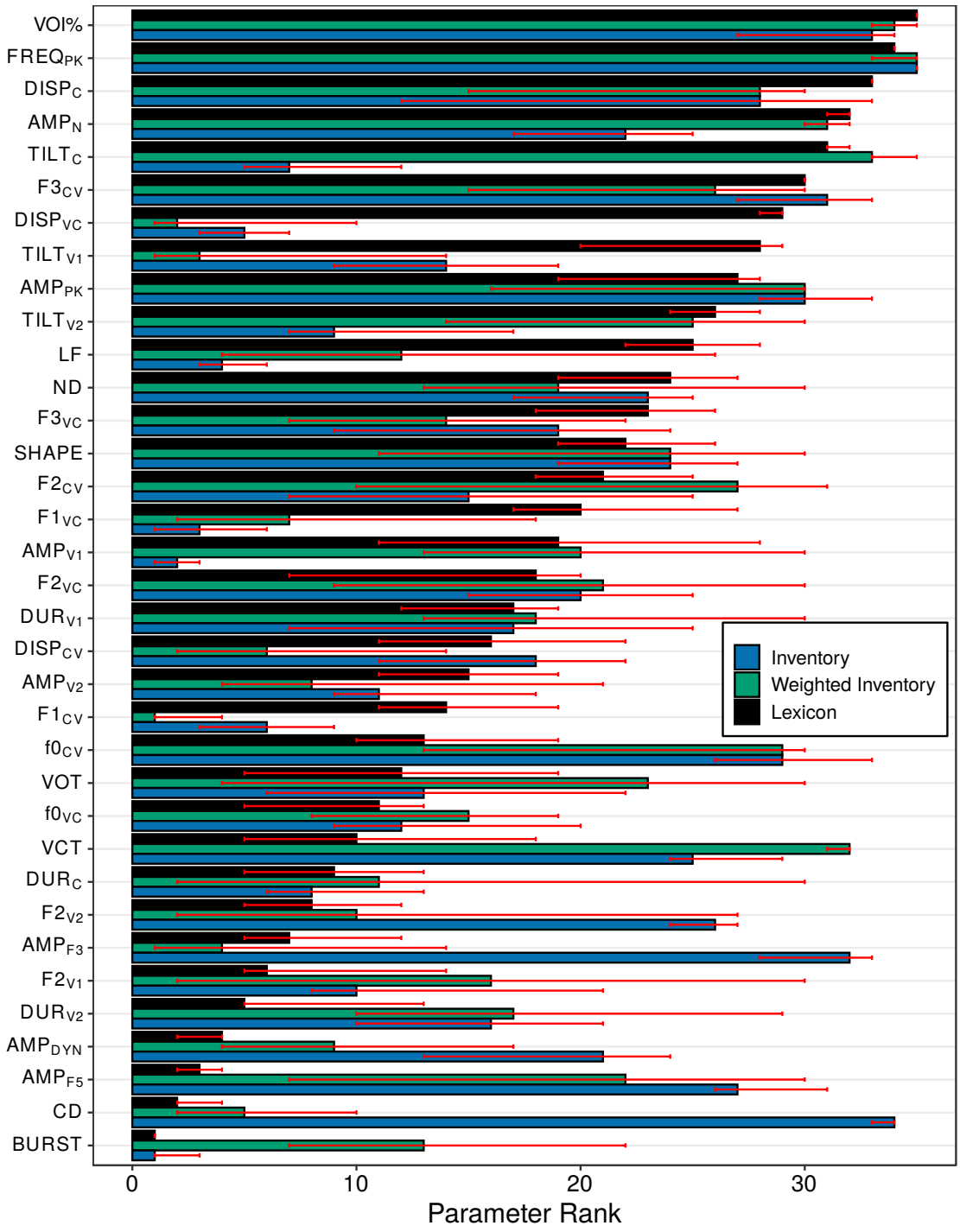

Figure A.81: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-medial contrasts fit to listener recognition data in Exp. 1a. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

Target parameter ranks in Exp. 1b (VCV)

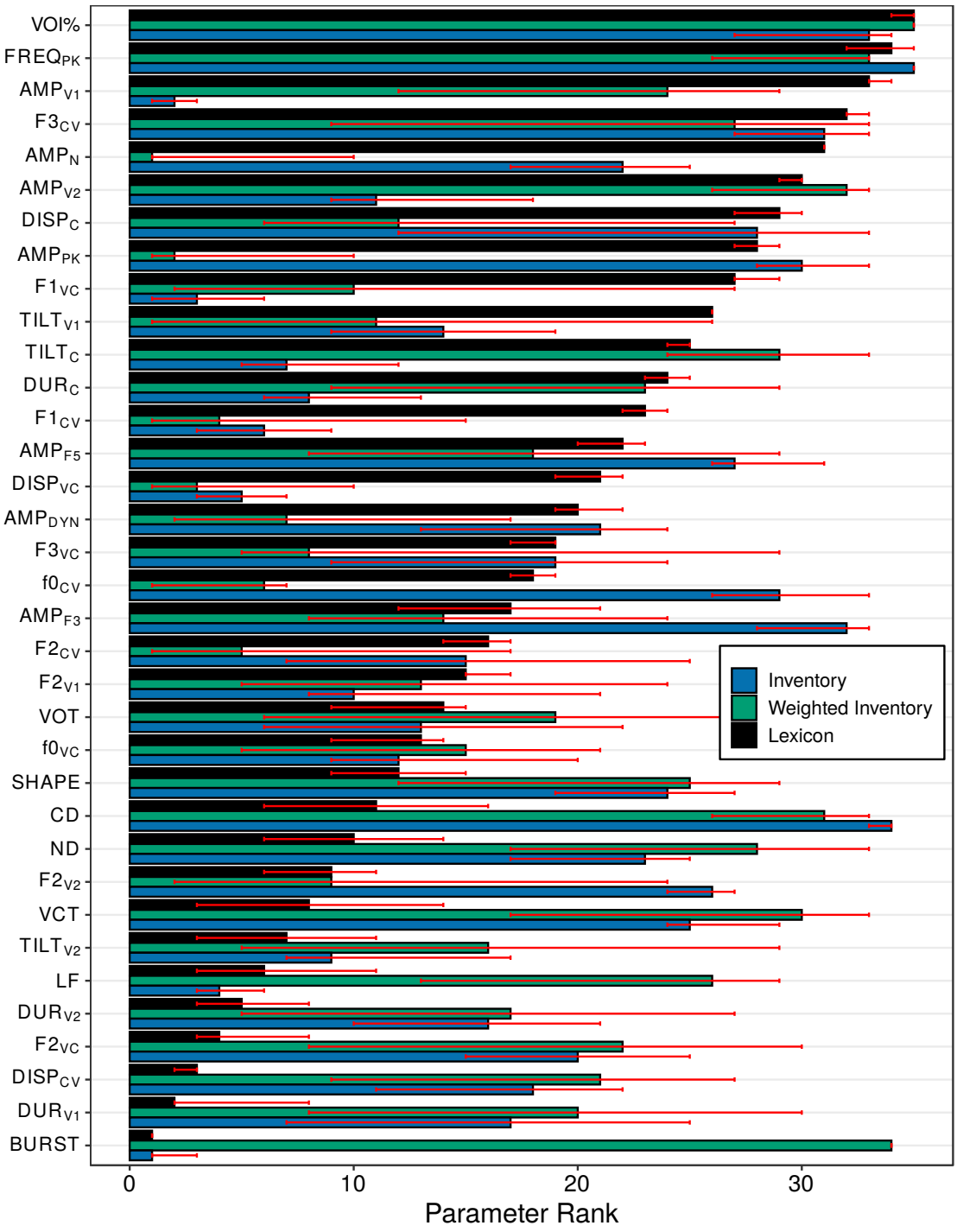

Figure A.82: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-medial contrasts fit to listener recognition data in Exp. 1b. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter ranks in Exp. 1a (VCV)

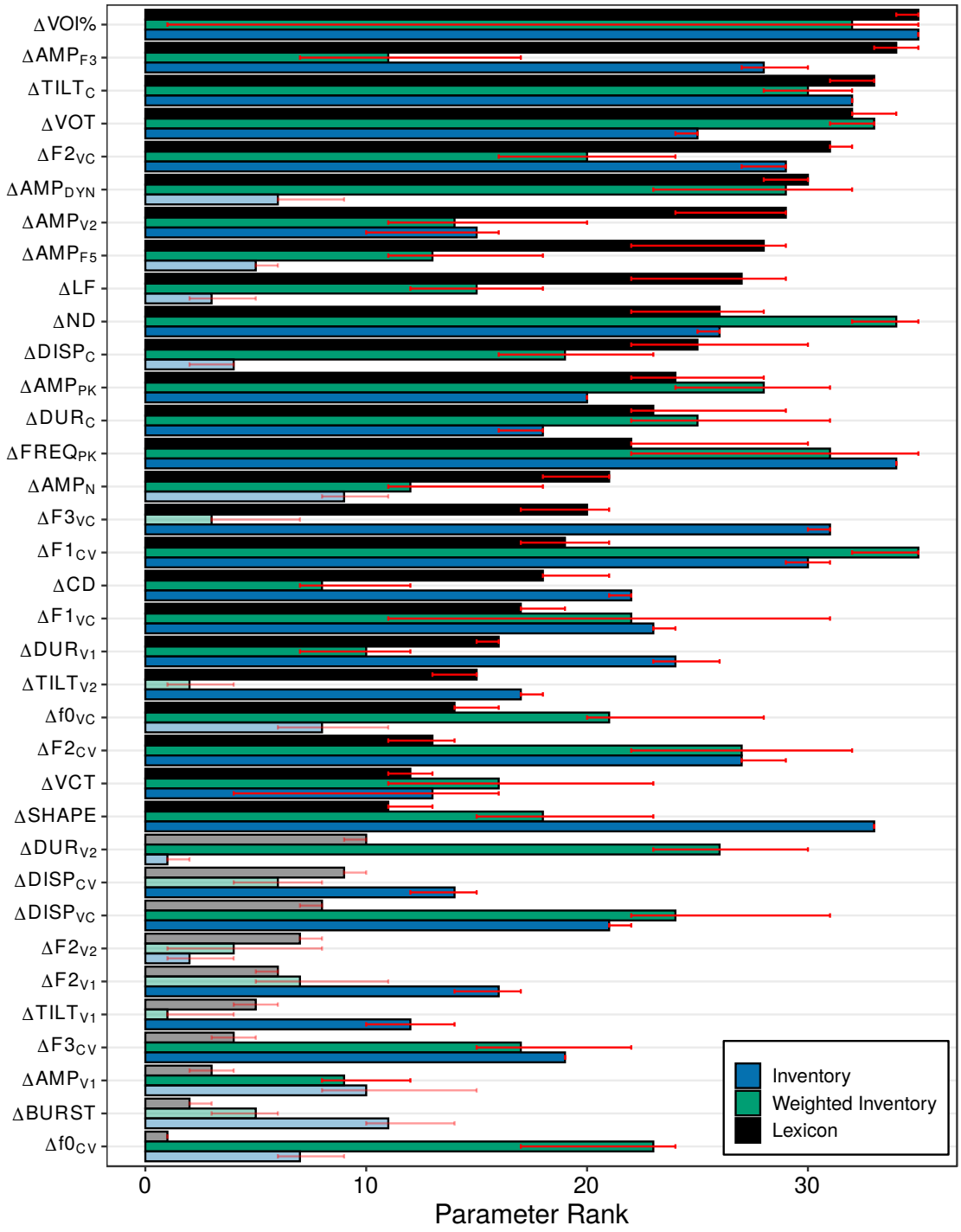

Figure A.83: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-medial contrasts fit to listener recognition data in Exp. 1a. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter ranks in Exp. 1b (VCV)

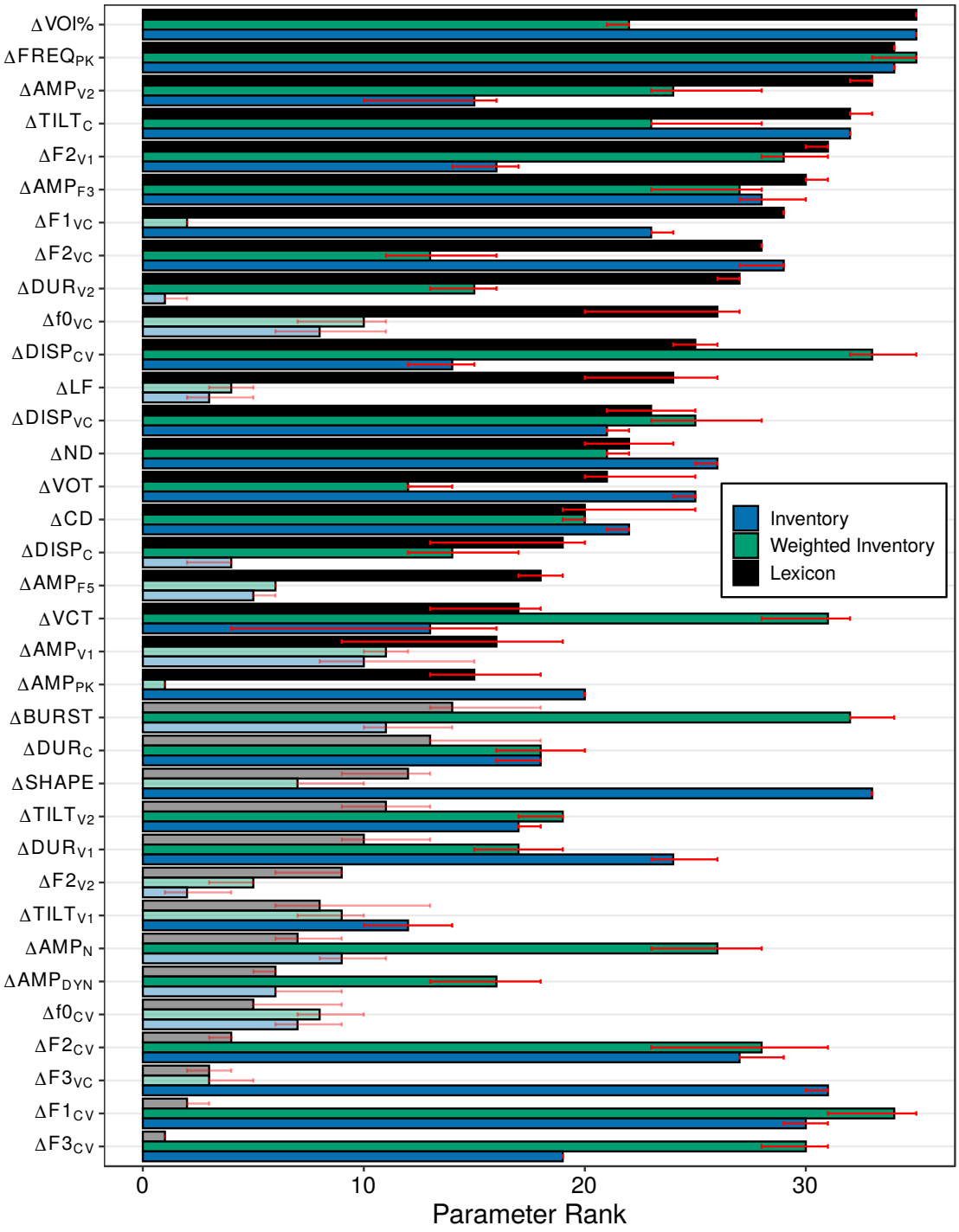

Figure A.84: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-medial contrasts fit to listener recognition data in Exp. 1b. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter correlations in Exp. 1a (VCV)

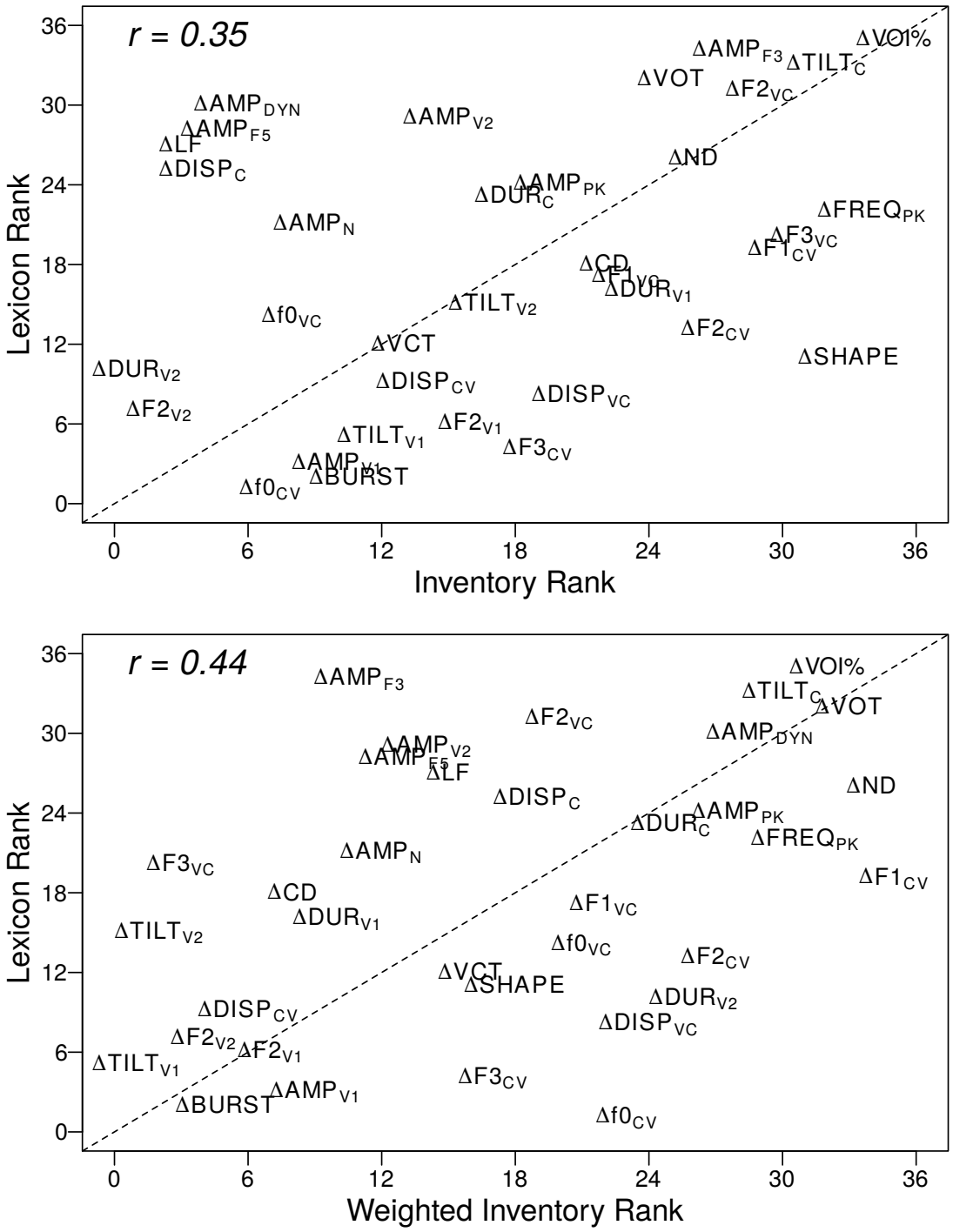

Figure A.85: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VCV position in Exp. 1a. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

Contrast parameter correlations in Exp. 1b (VCV)

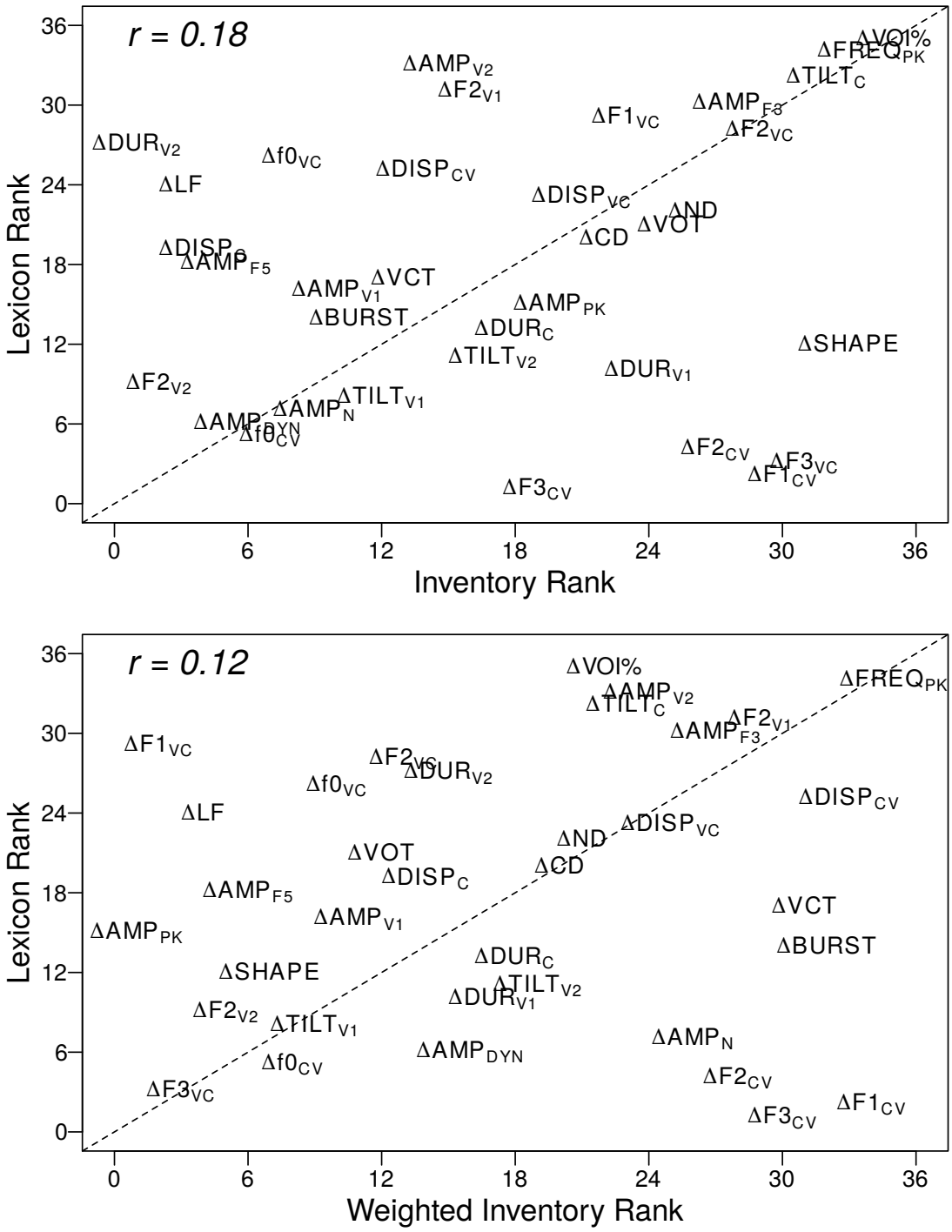

Figure A.86: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VCV position in Exp. 1b. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

Contrast parameter rank differences in Exp. 1a (VCV)

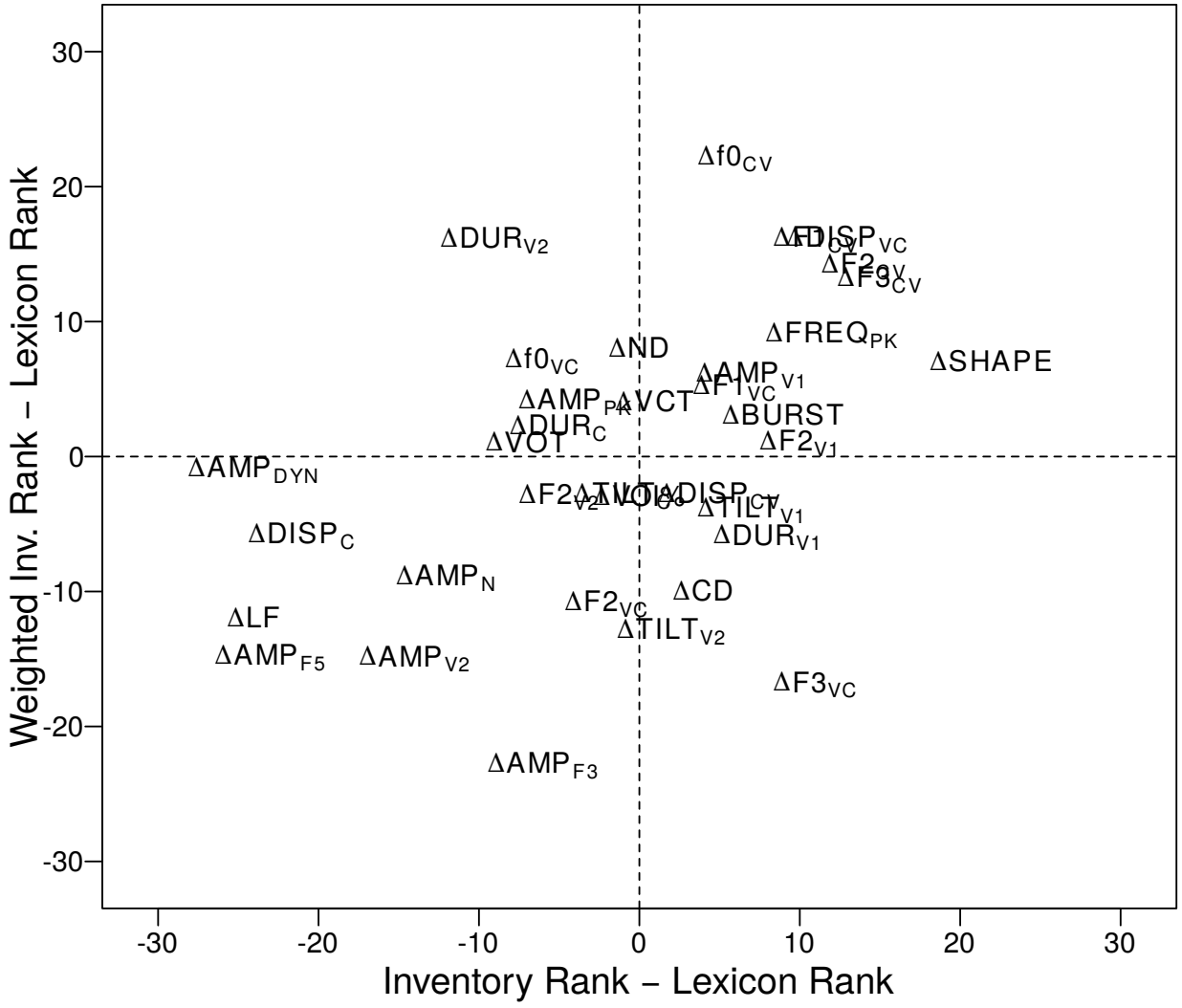

Figure A.87: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VCV position in Exp. 1a. Dashed lines indicate equivalence relations between each pair of models.

Contrast parameter rank differences in Exp. 1b (VCV)

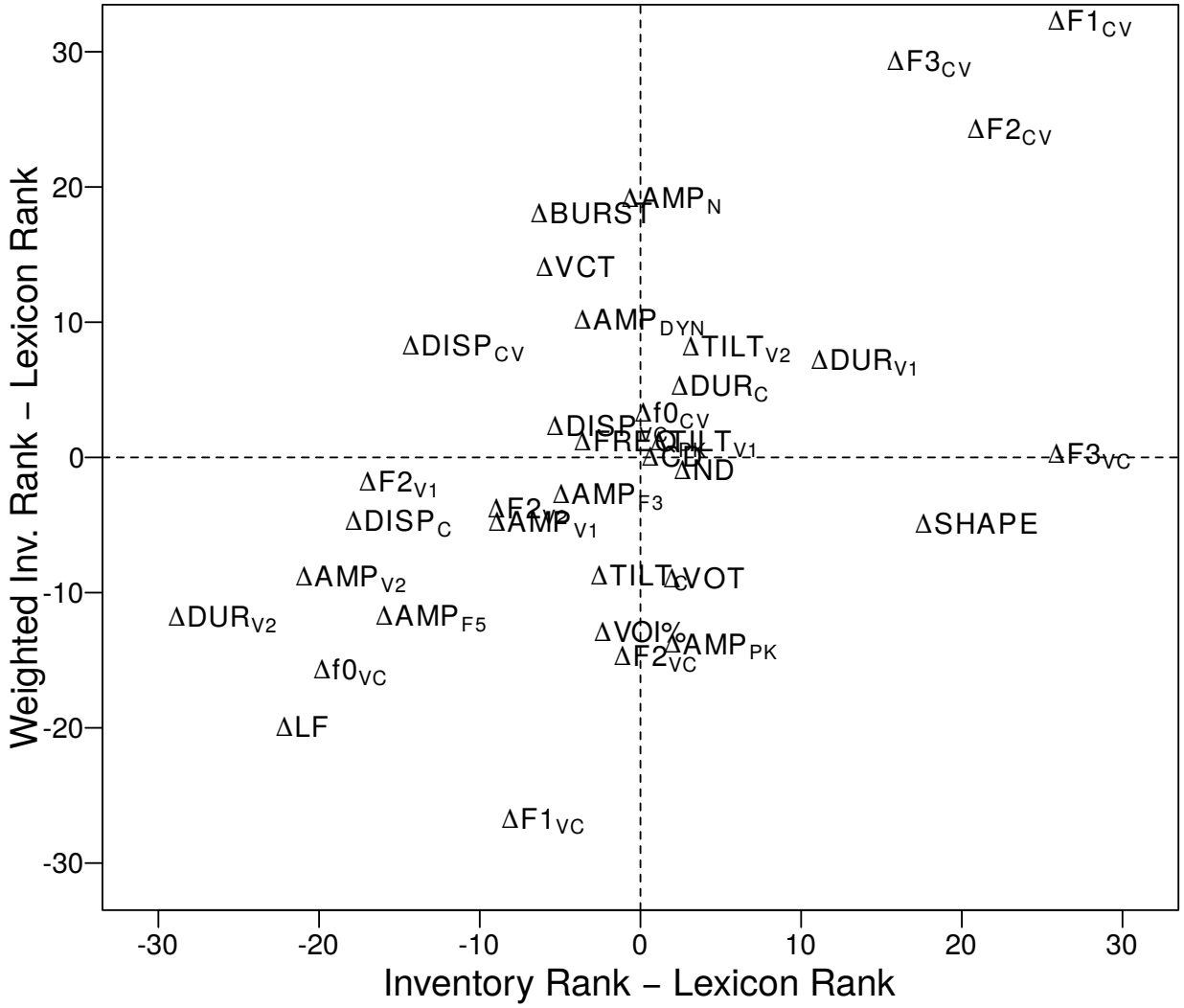

Figure A.88: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VCV position in Exp. 1b. Dashed lines indicate equivalence relations between each pair of models.

Target parameter ranks in Exp. 1a (VC)

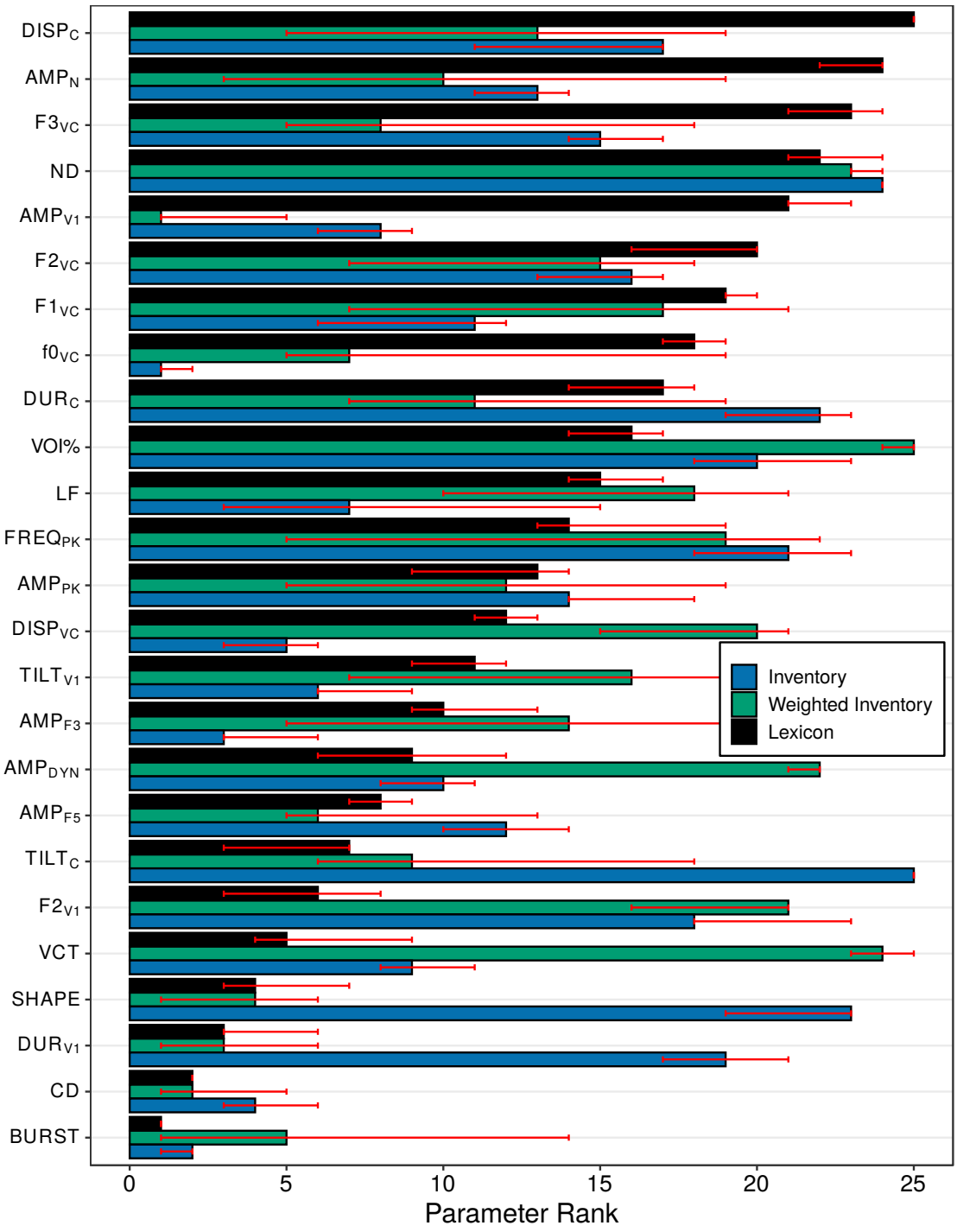

Figure A.89: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-final contrasts fit to listener recognition data in Exp. 1a. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

Target parameter ranks in Exp. 1b (VC)

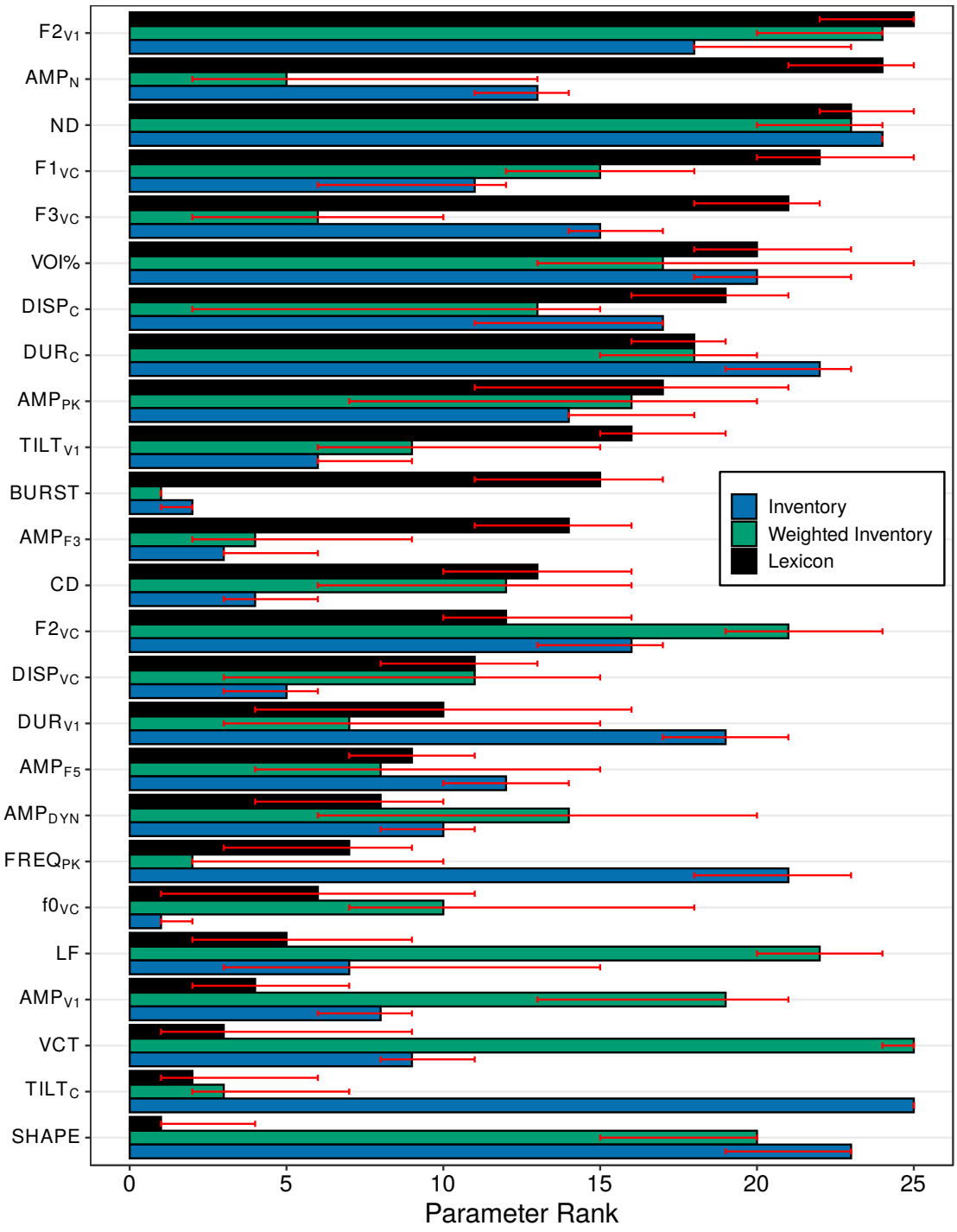

Figure A.90: Target parameter ranks in the lexicon, inventory, and weighted inventory models of word-final contrasts fit to listener recognition data in Exp. 1b. Ranks are based on the weights derived from the posterior median. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter ranks in Exp. 1a (VC)

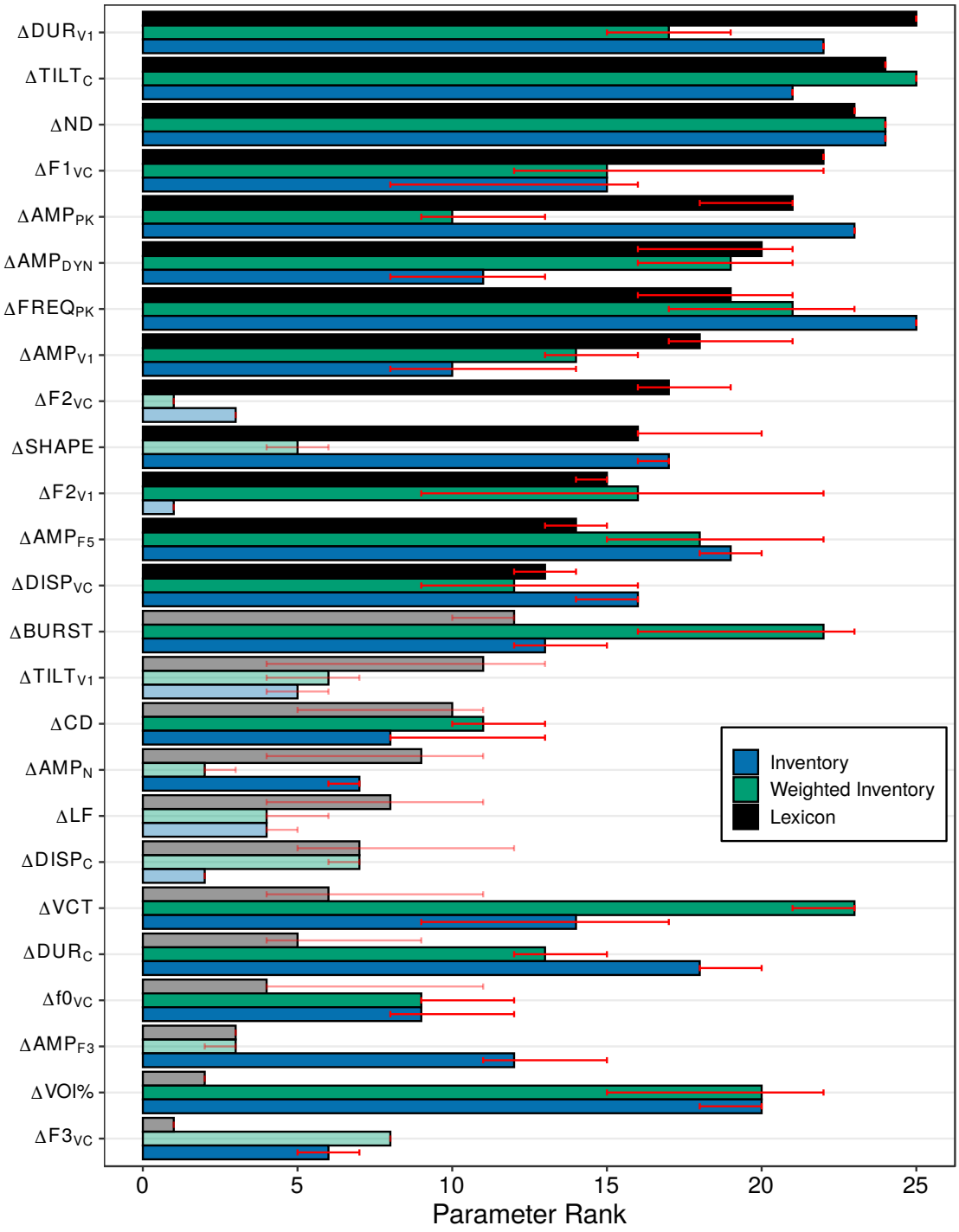

Figure A.91: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-final contrasts fit to listener recognition data in Exp. 1a. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter ranks in Exp. 1a (VC)

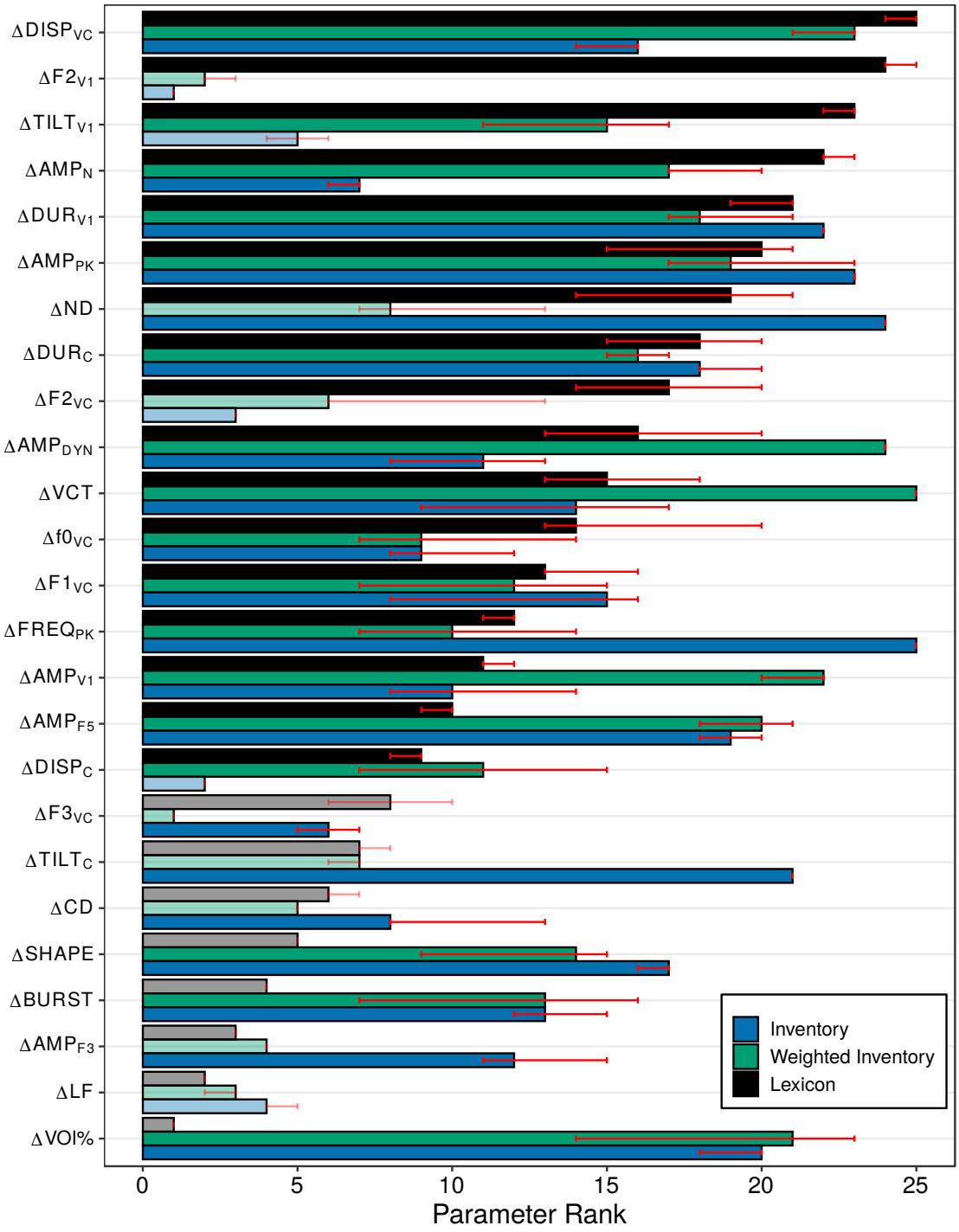

Figure A.92: Contrast parameter ranks in the lexicon, inventory, and weighted inventory models of word-final contrasts fit to listener recognition data in Exp. 1b. Ranks are based on the weights derived from the posterior median. Ranks derived from negative weights are displayed with translucent colors. Error bars indicate rank changes due to overlap in parameter weight distributions.

Contrast parameter correlations in Exp. 1a (VC)

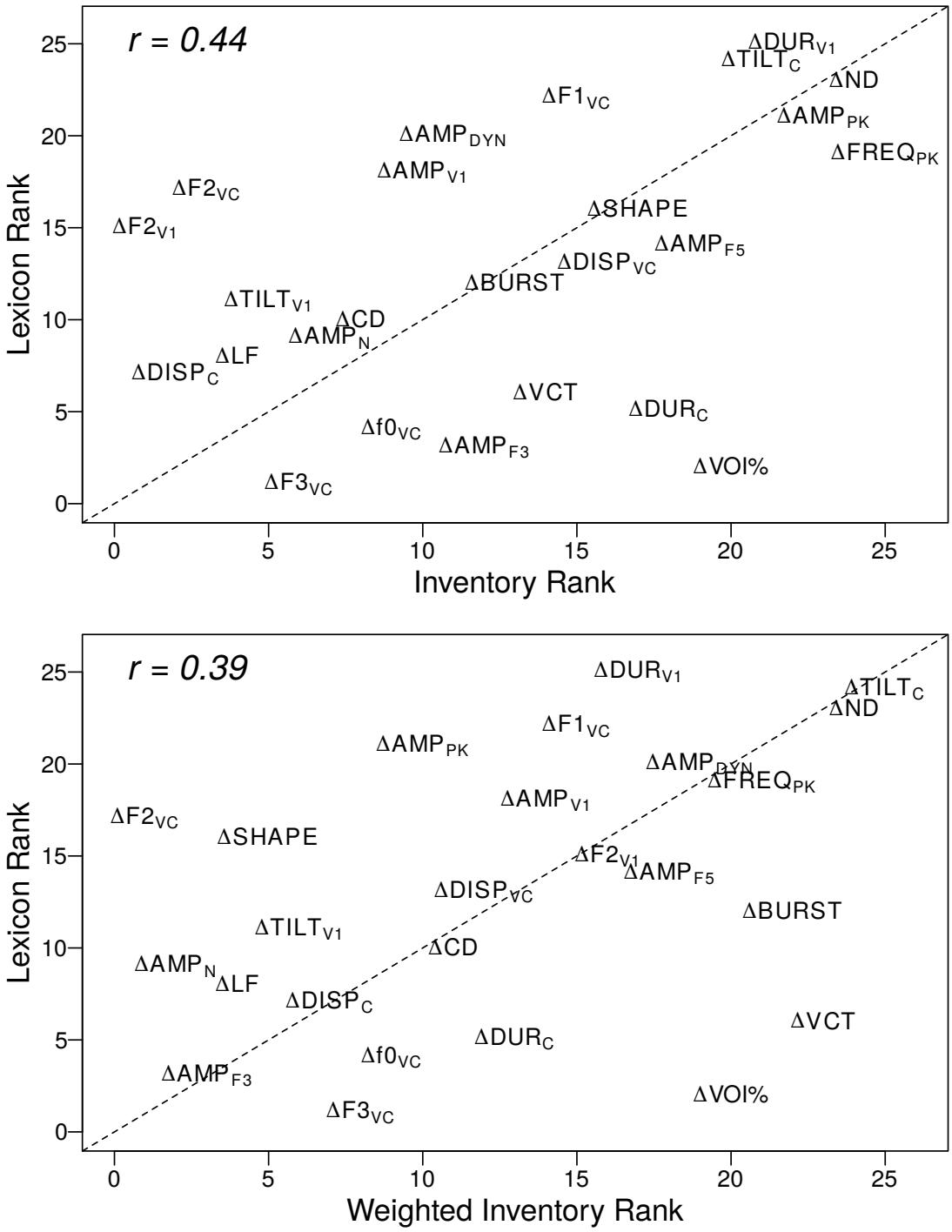

Figure A.93: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VC position in Exp. 1a. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

Contrast parameter correlations in Exp. 1b (VC)

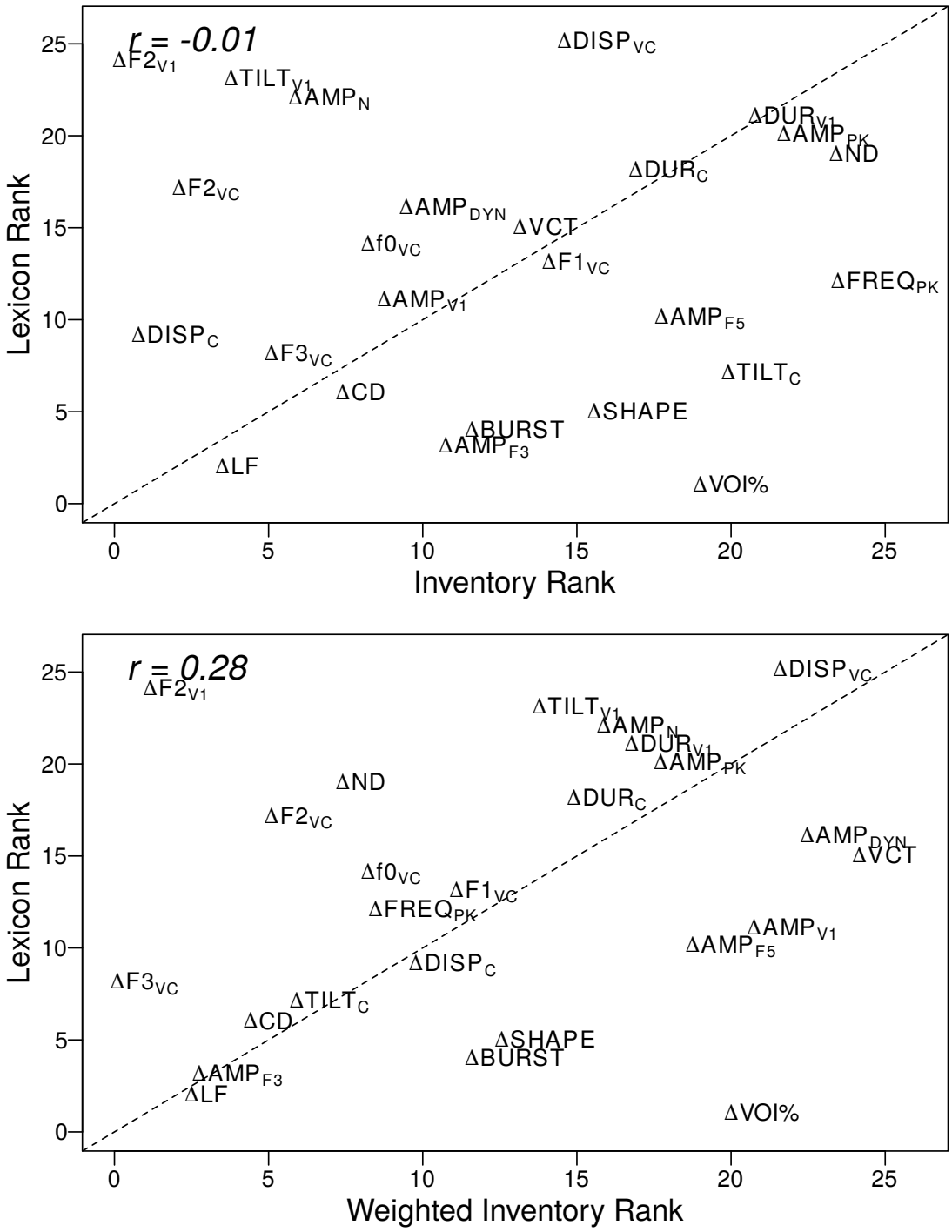

Figure A.94: Relations between parameter ranks in the lexicon model and ranks in the inventory (upper panel) and weighted inventory (lower panel) models in VC position in Exp. 1b. Correlations between parameter ranks in each pair of models are displayed in the upper left corner of each panel.

Contrast parameter rank differences in Exp. 1a (VC)

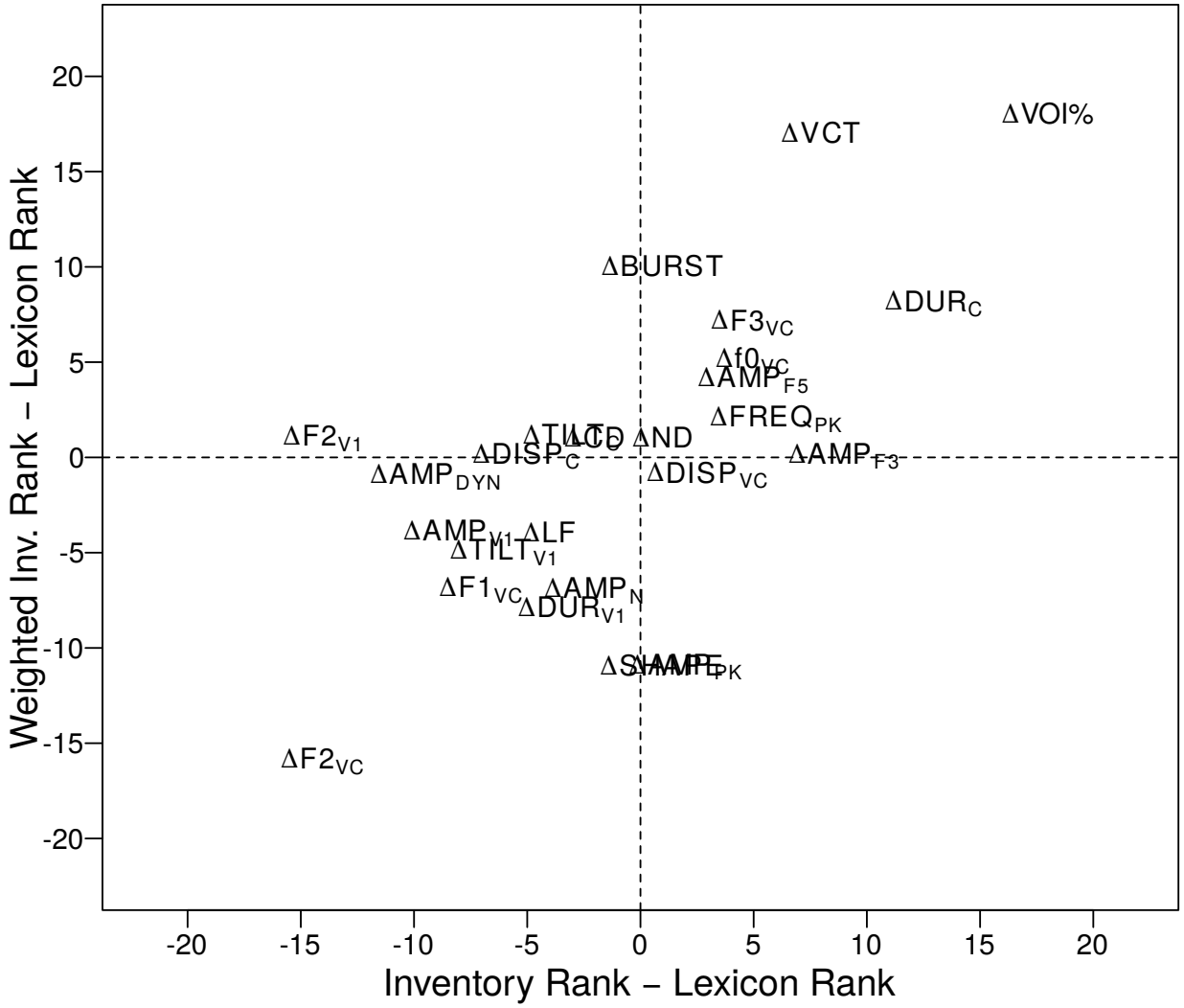

Figure A.95: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VC position in Exp. 1a. Dashed lines indicate equivalence relations between each pair of models.

Contrast parameter rank differences in Exp. 1b (VC)

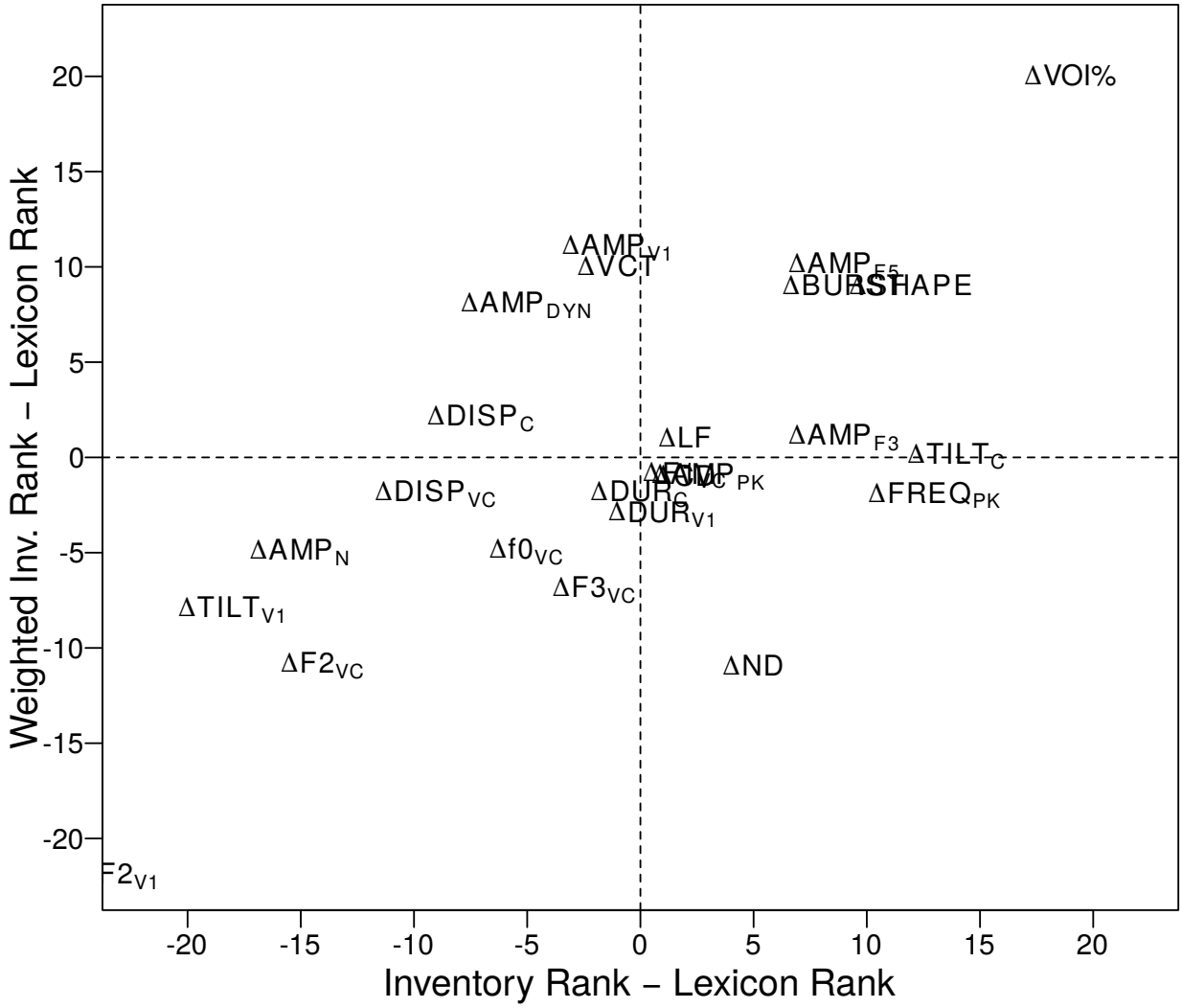

Figure A.96: Differences in parameter ranks between the lexicon model and the inventory/weighted-inventory models in VC position in Exp. 1b. Dashed lines indicate equivalence relations between each pair of models.