What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations

by Bob McMurray and Allard Jongman,

2011

McMurray, B., and 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, 219-246.

Published version: http://www.dx.doi.org/10.1037/a0022325

What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations.

Bob McMurray
Dept. of Psychology
University of Iowa

and

Allard Jongman
Dept. of Linguistics
University of Kansas

Running Head: Statistical models of fricative perception

In Press: *Psychological Review*

Corresponding Author
Bob McMurray
E11 SSH
Dept. of Psychology
University of Iowa
Iowa City, IA 52240
319-335-2408 (voice)
319-335-0191 (fax)

Abstract

Most theories of categorization emphasize how continuous perceptual information is mapped to categories. However, equally important is the informational assumptions of a model, the type of information subserving this mapping. This is crucial in speech perception where the signal is variable and context-dependent. This study assessed the informational assumptions of several models of speech categorization, in particular, the number of cues that are the basis of categorization and whether these cues represent the input veridically or have undergone compensation. We collected a corpus of 2880 fricative productions (Jongman, Wayland & Wong, 2000) spanning many talker- and vowel-contexts and measured 24 cues for each. A subset was also presented to listeners in an 8AFC phoneme categorization task. We then trained a common classification model based on logistic regression to categorize the fricative from the cue values, and manipulated the information in the training set to contrast 1) models based on a small number of invariant cues; 2) models using all cues without compensation, and 3) models in which cues underwent compensation for contextual factors. Compensation was modeled by Computing Cues Relative to Expectations (C-CuRE), a new approach to compensation that preserves fine-grained detail in the signal. Only the compensation model achieved a similar accuracy to listeners, and showed the same effects of context. Thus, even simple categorization metrics can overcome the variability in speech when sufficient information is available and compensation schemes like C-CuRE are employed.

What information is necessary for speech categorization? Assessing the informational assumptions of models of speech perception with a corpus of fricatives.

1. Categorization and Information

Work on perceptual categorization encompasses diverse domains like speech perception, object identification, music perception and face recognition. These are unified by the insight that categorization requires mapping from one or more continuous perceptual dimensions to a set of meaningful categories, and it is often assumed that the principles governing this may be common across domains (e.g., Goldstone & Kersten, 2003; though see Medin, Lynch & Solomon, 2000).

The most important debates concern the memory representations used to distinguish categories, contrasting accounts based on boundaries (e.g., Ashby & Perrin, 1988), prototypes (e.g., Homa, 1984; Posner & Keele, 1968; Reed, 1972), and sets of exemplars (Hintzman, 1986; Medin & Schafer, 1978; Nosofsky, 1986). Such representations are used to map individual exemplars, described by continuous perceptual cues, onto discrete categories. But these are only part of the story. Equally important for any specific type of categorization (e.g. speech categorization) is the nature of the perceptual cues.

There has been little work on this within research on categorization. What has been done emphasizes the effect of categories on perceptual encoding. We know that participants' categories can alter how individual cue-values along dimensions like hue are encoded (Hansen, Olkkonen, Walter & Gegenfurtner, 2006; Goldstone, 1995). For example, a color is perceived as more yellow in the context of a banana than a pear. Categories may also warp distance within a dimension as in categorical perception (e.g. Liberman, Harris, Hofffman & Griffith, 1957; Goldstone, Lippa & Shiffrin, 2001) though this has been controversial (Massaro & Cohen, 1983; Schouten, Gerrits & Van Hessen, 2003; Roberson, Hanley & Pak, 2009; Toscano, McMurray, Dennhardt & Luck, 2010). Finally, the acquisition of categories can influence the primitives or dimensions used for categorization (Schyns & Rodet, 1997; Oliva & Schyns, 1997).

While there has been some work examining how categories affect continuous perceptual processing, there has been little work examining the other direction, whether the type of information that serves as input to categorization matters. Crucially, does the nature of the perceptual dimensions constrain or distinguish theories of categorization? In fact, some approaches (e.g. Soto & Wasserman, 2010) argue that we can understand much about categorization by abstracting away from the specific perceptual dimensions.

Nonetheless, we cannot ignore this altogether. Smits, Jongman and Sereno (2006), for example, taught participants auditory categories along either resonance-frequency or duration dimensions. The distribution of the exemplars was manipulated to contrast boundary-based, prototype and statistical accounts. While boundaries fit well for frequency categories, duration categories required a hybrid of boundary and statistical accounts. Thus, the nature of the perceptual dimension may matter for distinguishing theoretical accounts of categorization.

Beyond the matter of the cue being encoded, a second issue, and the focus of this study, is whether and how perceptual cues are normalized during categorization. Perceptual cues are affected by multiple factors, and it is widely, though not universally, accepted that that perceptual systems compensate for these sources of variance. For example, in vision, to correctly perceive hue, observers compensate for light-source (McCann, McKee & Taylor, 1976); in music, pitch is computed relative to a tonic note (relative pitch); and in speech, temporal cues like duration may be calibrated to the speaking rate (Summerfield, 1981), while pitch is computed relative to the talker's pitch range (Honorof & Whalen, 2005).

Many theories of categorization do not address the relationship between compensation and categorization. Compensation is often assumed to be a low-level autonomous process occurring prior to and independently of categorization (though see Mitterer & de Ruiter, 2008). Moreover, in laboratory learning studies, it doesn't matter whether perceptual cues undergo compensation. Boundaries, prototypes and exemplars can be constructed with either type of information, and most experiments control for factors that demand compensation like lighting.

However, there are conditions under which such assumptions are unwarranted. First, if the input dimensions were context-dependent and categorization was difficult, compensation could make a difference in whether a particular model of categorization could classify the stimuli with the same accuracy as humans. This is unlikely to matter in laboratory-learning tasks where categorization is relatively unambiguous, but it may be crucial for real-life category systems like speech in which tokens can not always be unambiguously identified. Here, a model's accuracy may be as much a product of the information in the input as the nature of the mappings.

Second, if compensation is a function of categorization we cannot assume autonomy. Color constancy, for example, is stronger at hue values near category prototypes (Kulikowski & Vaitkevicius, 1997). In speech, phonemes surrounding a target phoneme affect the same cues, such that the interpretation of a cue for one phoneme may depend on the category assigned to others (Pisoni & Sawusch, 1974; Whalen, 1989; Smits, 2001a,b; Cole, Linebaugh, Munson & McMurray, 2010; though see Nearey, 1990). Such bidirectional effects imply that categorization and compensation are not independent and models of categorization must account for both, something that few models in any domain have considered (though see Smits, 2001a,b).

Finally, and perhaps most importantly, some theories of categorization make explicit claims about the nature of the information leading to categorization. In vision, Gibsonian approaches (Gibson, 1966) and Geon theory (e.g. Biederman, 1995) posit invariant cues for object recognition; while in speech perception, the theory of acoustic invariance (Stevens & Blumstein, 1978; Blumstein & Stevens, 1980; Lahiri, Gewirth & Blumstein, 1984) and quantal theory (Stevens, 2002; Stevens & Keyser, 2010) posit invariant cues for some phonetic distinctions (cf., Sussman, Fruchter, Hilbert& Sirosh, 1998). Other approaches such as the version of exemplar theory posited in speech explicitly claim that while there may be no invariant perceptual cues, category representations can cope with this without normalization (e.g. Pisoni, 1997). In such theories, normalization prior to categorization is not necessary, and this raises the possibility that normalization does not occur at all as part of the categorization process.

Any theory of categorization can be evaluated on two levels: the mechanisms which partition the perceptual space, and the nature of the perceptual space. This latter construct refers to the *informational assumptions* of a theory and is logically independent from the categorization architecture. For example, one could build a prototype theory on either raw or normalized inputs, and exemplars could be represented in either format. In Marr's (1982) levels of analysis, the informational assumptions of a theory can be seen as part of the first, computational level of analysis, where the problem is defined in terms of input/output relationships, while the mechanism of categorization may be best described at the algorithmic level. However, when the aforementioned conditions are met, understanding the informational assumptions of a theory may be crucial for evaluating it. Contrasting with Marr, we argue that levels of analysis may constrain each other: one must properly characterize a problem to distinguish solutions.

Speech perception presents a compelling domain in which to examine these issues, as all of the above conditions are met. Thus, the purpose of this study is to evaluate the informational assumptions of several approaches to speech categorization, and to ask what kind of information is necessary to support listener-like categorization. This was done by collecting a large dataset of measurements on a corpus of speech tokens, and manipulating the input to a series of categorization models to determine what informational structure is necessary to obtain listener-like performance. While our emphasis is on theories of speech perception, consistent with a history of work relating speech categories to general principles of categorization[1], this may also uncover principles that are relevant to categorization more broadly.

The remainder this introduction will discuss the classic problems in speech perception, and the debate over normalization. We then present a new approach to compensation which addresses concerns about whether fine-grained acoustic detail is preserved. Finally, we describe the speech categories we investigated, the eight fricatives of English. Section 2 presents the empirical work that is the basis of our modeling: a corpus of 2880 fricatives, with measurements of 24 cues for each, and listeners' categorization for a subset of them. Sections 3 and 4 then present a series of investigations in which the input to a standard categorization model is manipulated to determine which informational account best yields listener-like performance.

1.1 The Information Necessary for Speech Perception

Speech perception is increasingly being described as a problem of mapping from continuous acoustic cues to categories (e.g. Oden & Massaro, 1978; Nearey, 1997; Holt & Lotto, 2010). We take a relatively theory-neutral approach to what a cue is, defining a cue as a specific measurable property of the speech signal that can potentially be used to identify a useful characteristic like the phoneme category or the talker. Our use of this term is not meant to imply that a specific cue is actually used in speech perception, or that a given cue is the fundamental property that is encoded during perception. Cues are merely a convenient way to measure and describe the input.

A classic framing in speech perception is the problem of lack of invariant cues in the signal for categorical distinctions like phonemes. Most speech cues are context-dependent and there are few, if any, that invariantly signal a given phonetic category. There is debate on how to solve this problem (e.g. Fowler, 1996; Ohala, 1996; Lindblom, 1996) and about the availability of invariant cues (e.g. Blumstein & Stevens, 1980; Lahiri et al., 1984; Sussman et al, 1998). But there is little question that this is a fundamental issue that theories of speech perception must address. Thus, the information in the signal to support categorization is of fundamental importance to theories of speech perception.

As a result of this, a common benchmark for theories of speech perception accuracy, the ability to separate categories. This benchmark is applied to complete theories (e.g. Johnson, 1997; Nearey, 1990; Maddox, Molis & Diehl, 2002; Smits, 2001a; Hillenbrand & Houde, 2003), and even to phonetic analyses of particular phonemic distinctions (e.g., Stevens & Blumstein, 1978; Blumstein & Stevens, 1980, 1981; Forrest, Weismer, Milenkovic & Dougall, 1988; Jongman, Wayland & Wong, 2000; Werker et al, 2007). The difficulty attaining this benchmark means that sometimes accuracy is all that is needed to validate a theory. Thus, speech meets our first condition: categorization is difficult and the information available to it matters.

Classic approaches to the lack of invariance problem posited normalization or compensation processes for coping with specific sources of variability. In speech, normalization is typically defined as a process that factors out systematic but phonologically non-distinctive acoustic variability (e.g. systematic variability that does not distinguish phonemes) for the purposes of identifying phonemes or words. Normalization is presumed to operate on the perceptual encoding prior to categorization and classic normalization processes include rate (e.g., Summerfield, 1981) and talker (Nearey, 1978; see chapters in Johnson & Mullenix, 1997) normalization. Not all systematic variability is non-phonological, however: the acoustic signal at any point in time is always affected by the preceding and subsequent phonemes, due to a phenomenon known as coarticulation (e.g., Delattre, Liberman & Cooper, 1955; Öhman, 1966;

Fowler & Smith, 1986; Cole et al., 2010). As a result, the term compensation has often been invoked as a more general term to describe both normalization (e.g. compensating for speaking rate) and processes that cope with coarticulation (e.g. Mann & Repp, 1981). While normalization generally describes processes at the level of perceptual encoding, compensation can be accomplished either pre-categorically or as part of the categorization process (e.g. Smits, 2001b). Due to this greater generality, we use the term compensation throughout this paper.

A number of studies show that listeners compensate for coarticulation in various domains (e.g. Mann & Repp, 1981; Pardo & Fowler, 1997; Fowler & Brown, 2000). Crucially, the fact that portions of the signal are affected by multiple phonemes raises the possibility that how listeners categorize one phoneme may affect how subsequent or preceding cues are interpreted. For example, consonants can alter the formant frequencies listeners use to categorize vowels (Öhman, 1966). Do listeners compensate for this variability, by categorizing the consonant and then interpret the formant frequencies differently on the basis of the consonant? Or, do they compensate for coarticulation by tracking low-level contingencies between the cues for consonants and vowels or higher level contingencies between phonemes? Studies on this offer conflicting results (Mermelstein, 1978; Whalen, 1989; Nearey, 1990; Smits, 2001a).

Clearer evidence for such bidirectional processes comes from work on talker identification. Nygaard, Sommers and Pisoni (1994), for example, showed that learning to classify talkers improves speech perception, and a number of studies suggest that visual cues about a talker's gender affect how auditory cues are interpreted (Strand, 1999; Johnson, Strand & D'Imperio, 1999). Thus, interpretation of phonetic cues may be conditioned on judgments of talker identity. As a whole, then, there is ample interest, and some evidence that compensation and categorization are interactive, the second condition under which informational factors are important for categorization.

Compensation is not a given however. Some forms of compensation may not fully occur prior to lexical access. Talker voice, or indexical, properties of the signal (which do not contrast individual phonemes and words) affects a word's recognition (Creel, Aslin & Tanenhaus, 2008; McLennan & Luce, 2005) and memory (Palmeri, Goldinger & Pisoni, 1993; Bradlow, Nygaard & Pisoni, 1999). Perhaps most tellingly, speakers' productions gradually reflect indexical detail in auditory stimuli they are shadowing (Goldinger, 1998), suggesting that such detail is part of the representations that mediate perception and production. Thus, compensation for talker voice is not complete—indexical factors are not (completely) removed from the perceptual representations used for lexical access and may even be stored with lexical representations (Pisoni, 1997). This challenges the necessity of compensation as a precursor to categorization.

Thus, informational factors are essential to understanding speech categorization: the signal is variable and context-dependent; compensation may be dependent on categorization, but may also be incomplete. As a result, it is not surprising that some theories of speech perception make claims about the information necessary to support categorization.

On one extreme, although many researchers have abandoned hope of finding invariant cues (e.g. Ohala, 1996; Lindblom, 1996), for others, the search for invariance is ongoing. A variety of cues have been examined, such as burst onset spectra (Blumstein & Stevens, 1981; Kewley-Port & Luce, 1984), or locus equations for place of articulation in stop consonants, (Sussman et al, 1998; Sussman & Shore, 1996), and duration ratios for voicing (e.g. Port & Dalby, 1982; Pind, 1995). Most importantly, Quantal Theory (Stevens, 2002; Stevens & Keyser, 2010) posits that speech perception harnesses specific invariant cues for some contrasts (particularly manner of articulation, e.g., the *b/w* distinction). Invariance views, broadly construed, then, make the informational assumptions that 1) a small number of cues should suffice for many types of categorization; and 2) compensation is not required to harness them.

On the other extreme, exemplar approaches (e.g., Johnson, 1997; Goldinger, 1998; Pierrehumbert, 2001, 2003; Hawkins, 2003) argue that invariant cues are neither necessary nor available. If the signal is represented faithfully and listeners store many exemplars of each word, context-dependencies can be overcome without compensation. Each exemplar in memory is a holistic chunk containing both the contextually conditioned variance and the context, and is matched in its entirety to incoming speech. Because of this, compensation is not needed and may impede listeners by eliminating fine-grained detail that helps sort things out (Pisoni, 1997). Broadly construed, then, exemplar approaches make the informational assumptions that 1) input must be encoded in fine-grained detail with all available cues; and 2) compensation or normalization does not occur.

Finally, in the middle, lie a range of theoretical approaches that do not make strong informational claims. For lack of a better term, we call these cue-integration approaches, and they include the Fuzzy Logical Model, FLMP (Oden, 1978; Oden & Massaro, 1978), the Normalized a Posteriori Probability model, NAPP (Nearey, 1990), the Hierarchical Categorization of coarticulated phonemes, HICAT (Smits, 2001a,b), statistical learning models (McMurray, Aslin & Toscano, 2009a; Toscano & McMurray, 2010), and connectionist models like TRACE (Elman & McClelland, 1986). Most of these can be characterized as prototype models, though they are also sensitive to the range of variation. All assume that multiple (perhaps many) cues participate in categorization, and that these cues must be represented more or less veridically. However, few make strong claims about whether explicit compensation of some form occurs (although many implementations use raw cue-values for convenience). In fact, given the high-dimensional input, normalization may not be needed—categories may be separable with a high-dimensional boundary in raw cue-space (e.g., Nearey, 1997) and these models have been in the forefront of debates as to whether compensation for coarticulation is dependent on categorization (e.g., Nearey, 1990, 1992, 1997; Smits, 2001a,b). Thus it is an open question whether compensation is needed in such models.

Across theories, two factors describe the range of informational assumptions. Invariance accounts can be distinguished from exemplar and cue-integration accounts on the basis of number of cues (and their invariance). The other factor is whether cues undergo compensation or not. On this, exemplar and invariance accounts argue that cues do not undergo explicit compensation, while cue-integration models appear more agnostic. Our goal is to contrast these informational assumptions using a common categorization model. However, this requires a formal approach to compensation, which is not currently available. Thus, the next section describes several approaches to compensation and elaborates a new, generalized approach which builds on their strengths to offer a more general and formally well-specified approach based on Computing Cues Relative to Expectations (C-CuRE). .

1.2 Normalization, Compensation and C-CuRE

Classic normalization schemes posit interactions between cues that allow the perceptual system to remove the effects of confounding factors like speaker and rate. These are bottom-up processes motivated by articulatory relationships and signal processing. Such accounts are most associated with work on vowel categorization (e.g., Rosner & Pickering, 1994; Hillenbrand & Houde, 2003), though to some extent complex cue-combinations like locus equations (Sussman et al., 1998) or CV ratios (Port & Dalby, 1982), also fall under this framework. Such approaches offer concrete algorithms for processing the acoustic signal, but they have not led to broader psychological principles for compensation.

Other approaches emphasize principles at the expense of computational specificity. Fowler's (1984; Fowler & Smith, 1986; Pardo & Fowler, 1997) gestural parsing posits that speech is coded in terms of articulatory gestures, and that overlapping gestures are parsed into underlying causes. So for example, when a partially nasalized vowel precedes a nasal stop, the nasality gesture is assigned to the stop (a result of anticipatory coarticulation), since English does not use nasalized vowels contrastively (as does French), and the vowel is perceived as more oral (Fowler & Brown, 2000). As part of direct realist accounts, gestural parsing only compensates for coarticulation—the initial gestural encoding overcomes variation due to talker and rate.

Gow (2003) argues that parsing need not be gestural. His feature-cue parsing suggests that similar results can be achieved by grouping principles operating over acoustic features. This too has been primarily associated with coarticulation—variation in talker and/or rate is not discussed. However, the principle captured by both accounts is that by grouping overlapping acoustic cues or gestures, the underlying properties of the signal can be revealed (Ohala, 1981).

In contrast, Kluender and colleagues argue that low-level auditory mechanisms may do some of the work of compensation. Acoustic properties (like frequency) may be interpreted relative to other portions of the signal: a 1000 Hz tone played after a 500 Hz tone will sound higher than after an 800 Hz tone. This is supported by findings that non-speech events (e.g. pure tones) can create seemingly compensatory effects on speech (e.g. Lotto & Kluender, 1998, Holt, 2006; Kluender, Coady & Kiefte, 2003; though see Viswanathan, Fowler, & Magnuson, 2009). Thus, auditory contrast, either from other events in the signal or from long-term expectations about cues (Kluender et al, 2003) may alter the information available in the signal.

Parsing and contrast accounts offer principles that apply across many acoustic cues, categories and sources of variation. However, they have not been formalized in a way that permits a test of the sufficiency of such mechanisms to support categorization of speech input. All three accounts also make strong representational claims (articulatory vs. auditory), and a more general approach to compensation may be more useful (Ohala, 1981).

In developing a principled, yet computationally specific approach to compensation, one final concern is the role of fine phonetic detail. Traditional approaches to normalization assumed bottom-up processes that operate autonomously to clean up the signal before categorization, stripping away factors like talker or speaking rate[2]. However, research shows that such seemingly irrelevant detail is useful to phonetic categorization and word recognition. Word recognition is sensitive to within-category variation in voice onset time (Andruski, Blumstein & Burton, 1994; McMurray et al, 2002, 2008a), indexical detail (Creel et al, 2005, Goldinger, 1998), word-level prosody (Salverda, Dahan & McQueen, 2003), coarticulation (Marslen-Wilson & Warren, 1994; Dahan, Magnuson, Tanenhaus & Hogan, 2001), and alternations like reduction (Connine, 2004; Connine, Ronbom & Patterson, 2008) and assimilation (Gow, 2003). In many of these cases, such detail facilitates processing by allowing listeners to anticipate upcoming material (Martin & Bunnel, 1981, 1982; Gow, 2001, 2003), resolve prior ambiguity (Gow, 2003; McMurray, Tanenhaus & Aslin, 2009b) and disambiguate words faster (Salverda et al, 2003). Such evidence has led some to reject normalization altogether in favor of exemplar approaches (e.g. Pisoni, 1997; Port, 2007) which preserve continuous detail.

What is needed is a compensation scheme which is applicable across different cues and sources of variance, is computationally well-specified, and can retain and harness fine-grained acoustic detail. Cole et al. (2010; McMurray, Cole & Munson, in press) introduced such a scheme in an analysis of vowel coarticulation; we develop it further as a more complete account of compensation. This account, Computing Cues Relative to Expectations (C-CuRE), combines grouping principles from parsing accounts with the relativity of contrast accounts.

Under C-CuRE, incoming acoustic cues are initially encoded veridically, but as different sources of variance are categorized, cues are recoded in terms of their difference from expected values. Consider a stop-vowel syllable. The fundamental frequency (F0) at the onset of the vowel is a secondary cue for voicing. In the dataset we describe here, F0 at vowel onset had a mean of 149 Hz for voiced sounds and 163 Hz for voiceless ones, though it was variable (SD_{voiced}=43.2; $SD_{voiceless}$=49.8). Thus F0 is informative for voicing, but any given F0 is difficult to interpret. An F0 of 154 Hz, for example, could be high for a voiced sound or low for a voiceless one. However once the talker is identified (on the basis of other cues or portions of the signal) this cue may become more useful. If the talker's average F0 was 128 Hz, then the current 154 Hz is 26 Hz higher than expected and likely the result of a voiceless segment. Such an operation removes the effects of talker on F0 by recoding F0 in terms of its difference from the expected F0 for that talker, making it more useful for voicing judgments.

C-CuRE is similar to both versions of parsing in that it partials out influences on the signal at any time point. It also builds on auditory-contrast approaches by positing that acoustic cues are coded as the difference from expectations. However, it is also more general than these theories. Unlike gestural and feature-cue parsing, talker is parsed from acoustic cues the same way coarticulation is; unlike contrast accounts, expectations can be based on abstractions and there is an explicit role for categorization (phonetic categories, talkers, etc.) in compensation.

C-CuRE is straightforward to implement using linear regression. To do this, first a regression equation is estimated predicting the cue-value from the factor(s) being parsed out, for example, F0 as a function of talker gender. Next, for any incoming speech token, this formula is used to generate the expected cue-value given what is known (e.g., if the speaker is known), and the actual cue value is subtracted from it. The residual becomes the estimate of contrast or deviation from expectations. In terms of linear regression, then, parsing in the C-CuRE framework separates the variance in any cue into components and uses the regression formula to generate expectations on which the remaining variance can be used to do perceptual work.

When the independent factors are dichotomous (e.g. male/female), the regression predictions will be based on the cell means of each factor. This could lead to computational intractability if the regression had to capture all combinations of factors. For example, a vowel's first formant frequency (F1) is influenced by the talker's gender, the voicing of neighboring consonants, and the height of the subsequent vowel. If listeners required cell-means of the four-way <talker> × <initial voicing> × <final voicing> × <vowel height> contrast to generate expectations it is unlikely that they could track all of the possible combinations of influences on a cue. However, Cole et al (2010; McMurray et al, in press) demonstrated that by performing the regression hierarchically (e.g., first partialing out the simple effect of talker, then the simple effect of voicing, then vowel height, and so on), substantial improvements can be made in the utility of the signal using only simple effects, without needing higher-order interactions.

In sum, C-CuRE offers a somewhat new approach to compensation that is computationally well specified, yet principled. It maintains a continuous representation of cue values and does not discard variation due to talker, coarticulation and the like. Rather C-CuRE capitalizes on this variation to build representations for other categories. It is neutral with respect to whether speech is auditory or gestural, but consistent with principles from both parsing approaches, and with the notion of contrast in auditory contrast accounts. Finally, C-CuRE explicitly demands a categorization framework: compensation occurs as the informational content of the signal is interpreted relative to expectations driven by categories.

The goal of this project is to evaluate the informational assumptions of theories of speech categorization in terms of compensated vs. uncompensated inputs, a question at the computational level (Marr, 1982). Testing this quantitatively requires that we assume a particular form of compensation, a solution properly described at the algorithmic level, and existing forms of compensation do not have the generality or computational specificity to be applied. C-CuRE offers such a general, yet implementable compensation scheme and as a result, our test of compensation at the information level also tests this specific processing approach.

Such a test has only been examined in limited form by Cole et al (2010). This study used parsing in the C-CuRE framework to examine information used to anticipate upcoming vowels, and did not examine compensation for variance in a target segment. It showed that the recoding of formant values as the difference from expectations on the basis of talker and intervening consonant was necessary to leverage coarticulatory information for anticipating upcoming vowels. This validated C-CuRE's ability to harness fine-grained detail, but did not address its generality as only two cues were examined (F1 and F2); these cues were similar in form (both frequencies); only a small number of vowels were used; and results were not compared to listener data. Thus, it is an open question as to how well C-CuRE scales to dozens of cues representing different signal components (e.g. amplitudes, durations and frequencies) in the context of a larger set of categories, and it is unclear whether its predictions match listeners.

1.3 Logic and Overview

Our goal is to contrast three informational accounts: 1) that a small number of invariant cues distinguishes speech categories; 2) that a large number of cues is sufficient without compensation; and 3) that compensation must be applied. To accomplish this, we measured a set of cues from a corpus of speech sounds, and used them to train a generic categorization model. This was compared to listener performance on a subset of that corpus. By manipulating the cues available to the model and whether or not compensation was applied, we assessed the information required to yield listener performance.

One question that arises is which phonemes to use. Ideally, they should be difficult to classify, as accuracy is likely to be a distinguishing factor. There should also be a large number of categories for a more realistic test. For a fair test, the categories should have a mix of cues in which some have been posited to be invariant and others more contextually determined. Finally, C-CuRE suggests that the ability to identify context (e.g. the neighboring phoneme) underlies compensation. Thus, during perceptual testing, it would be useful to be able to separate the portion of the stimulus that primarily cues the phoneme categories of interest from portions that primarily cue contextual factors. The fricatives of English meet these criteria.

1.4 Phonetics of Fricatives.

English has eight fricatives created by partially obstructing the airflow through the mouth (see Table 1). They are commonly defined by three phonological features: sibilance, place of articulation and voicing. There are four places of articulation, each of which can be either voiced or voiceless. Fricatives produced at the alveolar ridge (/s/ or /z/ as in *sip* and *zip*) or at the post-alveolar position (/ʃ/ as in *ship* or /ʒ/ as in *genre*) are known as sibilants due to their high-frequency spectra; labiodental (/f/ or /v/ as in *face* and *vase*) and interdental fricatives (/ɵ/ or /ð/ as in *think* and *this*) are non-sibilants. The fact that there are eight categories makes categorization a challenging but realistic problem for listeners and models. As a result, listeners are not at ceiling even for naturally produced unambiguous tokens (LaRiviere, Winitz & Herriman, 1975; You, 1979; Jongman, 1989; Tomiak, 1990; Balise & Diehl, 1994), particularly for the non-sibilants (/f, v, ɵ, ð/) where accuracy estimates range from 43% to 99%.

Fricatives are signaled by a large number of cues. Place of articulation can be

Table 1: The eight fricatives of English can be classified along two dimensions: voicing (whether the vocal folds are vibrating or not) and place of articulation. Fricatives produced with alveolar and post-alveolar places of articulation are known as sibilants, others are non-sibilants.

	Place of Articulation	**Voiceless**		**Voiced**	
		IPA	**Examples**	**IPA**	**Examples**
Non-sibilants	Labiodental	f	*fat, fork*	v	*van, vase*
	Interdental	ɵ	*think, thick*	ð	*this, those*
Sibilants	Alveolar	s	*sick, sun*	z	*zip, zoom*
	Post Alveolar	ʃ	*ship, shut*	ʒ	*Jacques, genre*

distinguished by the four spectral moments (mean, variance, skew and kurtosis of the frequency spectrum of the frication) (Forrest, et al., 1988; Jongman, et al., 2000), and by spectral changes in the onset of the subsequent vowel, particularly the second formant (Jongman et al, 2000; Fowler, 1994). Duration and amplitude of the frication are related to place of articulation, primarily distinguishing sibilants from non-sibilants (Behrens & Blumstein, 1988; Baum & Blumstein, 1987; Crystal & House, 1988; Jongman et al., 2000; Strevens, 1960). Voicing is marked by changes in duration (Behrens & Blumstein, 1988; Baum & Blumstein, 1987; Jongman et al., 2000) and spectral properties (Stevens, Blumstein, Glicksman, Burton, & Kurowski, 1992; Jongman et al, 2000). Thus, there are large numbers of potentially useful cues.

This offers fodder for distinguishing invariance and cue-integration approaches on the basis of the number of cues, and across fricatives we see differences in the utility of a small number of cues. The four sibilants (/s, z, ʃ, ʒ/ can be distinguished at 97% by just the four spectral moments (Forrest et al, 1988). Thus, an invariance approach may be sufficient for sibilants. On the other hand, most studies have failed to find any single cue that distinguishes non-sibilants (/f, v, ɵ, ð/) (Maniwa, Jongman & Wade, 2008, 2009; though see Nissen & Fox, 2005), and Jongman et al.'s (2000) discriminant analysis using 21 cues only achieved 66% correct. Thus, non-sibilants may require many information sources, and possibly compensation.

In this vein, virtually all fricative cues are dependent on context including talker identity (Hughes & Halle, 1956; Jongman et al, 2000), the adjacent vowel (Soli, 1981; Jongman et al, 2000; LaRiviere et al., 1975; Whalen, 1981) and socio-phonetic factors (Munson, 2007; Munson et al, 2006; Jongman, Wang & Sereno, 2000). This is true even for cues like spectral moments that have been posited to be relatively invariant.

Thus, fricatives represent an ideal platform for examining the informational assumptions of models of speech categorization. They are difficult to categorize, and listeners can potentially utilize a large number of cues to do so. Both invariant and cue-combination approaches may be appropriate for some fricatives, but the context dependence of many, if not all cues, raises the possibility that compensation is necessary. Given the large number of cues, it is currently uncertain what information will be required for successful categorization.

1.5 Research Design

We first collected a corpus of fricative productions and measured a large set of cues in both the frication and vocalic portion of each. Next, we presented a subset of this corpus to listeners in an identification experiment with and without the vocalic portion. While this portion contains a number of secondary cues to fricative identity, it is also necessary for accurate identification of the talker (Lee, Dutton & Ram, 2010) and vowel, which is necessary for compensating in the C-CuRE framework. The complete corpus of measurements, including the perceptual results, is available in the online supplement. Finally, we implemented a generic categorization model using logistic regression (see also Cole et al, 2010; McMurray et al, in press), which is inspired by Nearey's (1990, 1997) NAPP model and Smits' (2001a,b) HICAT model. This model was trained to predict the intended production (not listeners' categorizations, as in NAPP and some versions of HICAT) from particular sets of cues in either raw form or after parsing. The model's performance was then compared to listeners' to determine what informational structure is needed to create their pattern of responding. This was used to contrast three informational accounts distinguished by the number of cues and the presence or absence of compensation.

2.0 Empirical Work

2.1 The Corpus

The corpus of fricatives for this study was based on the recordings and measurements of Jongman et al (2000) with additional measurements of ten new cues on these tokens.

2.1.1 Methods and Measurements

Jongman et al (2000) analyzed 2873 recordings of the 8 English fricatives /f, v, ɵ, ð, s, z, ʃ, ʒ/. Fricatives were produced in the initial position of a CVC syllable in which the vowel was /i, e, æ, ɑ, o, u/, and the final consonant was /p/. Twenty speakers (10 female) produced each CVC three times in the carrier phrase "Say ____ again". This led to 8 (fricatives) × 6 (vowels) × 3 (repetitions) × 20 (speakers), or 2880 tokens, of which 2873 were analyzed. All recordings were sampled at 22 kHz (16 bit quantization, 11 kHz low-pass filter). The measurements reported here are all of the original measurements of Jongman et al. (2000, the JWW database)[3], although some cues were collapsed (e.g. spectral moments at two locations). We also measured 10 new cues from these tokens to yield a set of 24 cues for each fricative. A complete list is shown in Table 2, and Figure 1 shows a labeled waveform and spectrogram of a typical fricative recording. Details on the measurements of individual cues and the transformations applied to them can be found in Appendix A.

We deliberately left out compound or relative cues (based on two measured values) like locus equations or duration ratios to avoid introducing additional forms of compensation into our dataset. We did include the independent measurements that contribute to such cues (e.g. duration of the vowel and consonant separately). Compound cues are discussed (and modeled in a similar framework) in the Online Supplement, Note #6.

The final set of 24 cues represents to the best of our knowledge all simple cues that have been proposed for

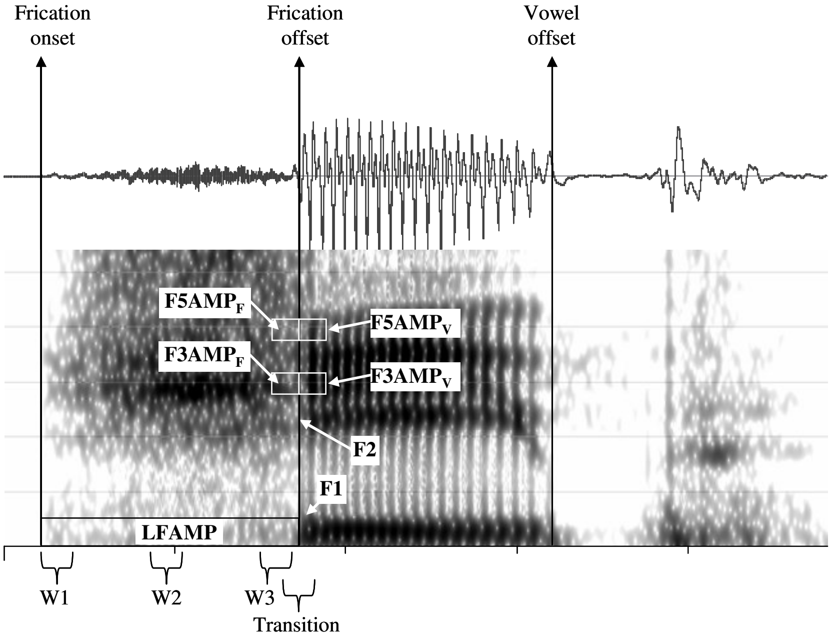

Figure 1: Annotated waveform and spectrogram of a typical sample in the corpus, /ʃip/ 'sheep'. Annotations indicate a subset of the cues that are described in Table 2.

Table 2: Summary of the cues included in the present study. JWW indicates cues that were previously reported by Jongman et al (2000). Also shown are several derived cues included in a subset of the analysis. The "cue for" column indicates the phonological feature typically associated with each cue.

Cue	**Variable**	**Noise Cue**	**Description**	**Cue for**	**Source**
Peak Frequency	MaxPF	*	Frequency with highest amplitude.	Place	JWW
Frication Duration	DUR_F	*	Duration of frication	Voicing	JWW
Vowel Duration	DUR_V		Duration of vocalic portion	Voicing	JWW
Frication RMS	RMS_F	*	Amplitude of frication	Sibilance	JWW
Vowel RMS	RMS_V		Amplitude of vocalic portion	Normalization	JWW
F3 narrow-band amplitude (frication).	$F3AMP_F$	*	Amplitude of frication at F3.	Place	New
F3 barrow band Amplitude (vowel)	$F3AMP_V$		Amplitude of vowel at F3.	Place	New
F5 narrow-band amplitude (frication).	$F5AMP_F$	*	Amplitude of frication at F5.	Place	New
F5 barrow band Amplitude (vowel)	$F5AMP_V$		Amplitude of vowel at F5.	Place	New
Low Frequency Energy	LF	*	Mean RMS below 500 Hz in frication	Voicing	New
Pitch	F_0		Fundamental frequency at vowel onset	Voicing	New
First Formant	F1		First formant frequency of vowel	Voicing	New
Second Formant	F2		Second formant frequency of vowel	Place	JWW
Third Formant	F3		Third formant frequency of vowel	Place	New
Fourth Formant	F4		Fourth formant frequency of vowel	Unknown	New
Fifth Formant	F5		Fifth formant frequency of vowel	Unknown	New
Spectral Mean	M1	*	Spectral mean at three windows in frication noise (onset, middle, offset)	Place / Voicing	JWW
Spectral Variance	M2	*	Spectral variance at three windows in frication noise	Place	JWW
Spectral Skewness	M3	*	Spectral skewness at three windows in frication noise	Place / Voicing	JWW
Spectral Kurtosis	M4	*	Spectral kurtosis at three windows in frication noise	Place	JWW
Transition Mean	$M1_{trans}$		Spectral mean in window including end of frication and vowel onset	Place	JWW
Transition Variance	$M2_{trans}$		Spectral variance in window including end of frication and vowel onset	Place	JWW
Transition Skewness	$M3_{trans}$		Spectral skewness in window including end of frication and vowel onset	Place	JWW
Transition Kurtosis	$M4_{trans}$		Spectral kurtosis in window including end of frication and vowel onset	Place	JWW

distinguishing place, voicing or sibilance in fricatives, and also includes a number of cues not previously considered (such as F3, F4 and F5, and low frequency energy).

2.1.2 Results

Since the purpose of this corpus is to examine the information available to categorization, we do not report a complete phonetic analysis. Instead, we offer a brief analysis that characterizes the information in this dataset, asking which cues could be useful for fricative categorization and the effect of context (talker and vowel) on them. A complete analysis is found in the Online Supplement, Note #1, and see Jongman et al. (2000) for extensive analyses of those measures.

Our analyses consisted of a series of hierarchical linear regressions. In each one, a single cue was the dependent variable, and the independent variables were a set of dummy codes[4] for fricative identity (7 variables). Each regression first partialed out the effect of talker (19 dummy codes) and vowel (5 dummy codes), before entering the fricative terms into the model. We also ran individual regressions breaking fricative identity down into phonetic features (sibilance, place of articulation and voicing). Table 3 displays a summary.

Table 3: Summary of regression analyses examining effects of speaker (20), vowel (6) and fricative (8) for each cue. Shown are R^2_{change} values. Missing values were not significant (p>.05). The final column shows secondary analyses examining individual contrasts. Each cue is given the appropriate letter code if the effect size was Medium or Large (R^2_{change}>.05). A few exceptions with smaller effect sizes are marked because there were few robust cues to non-sibilants. Sibilant vs. non-sibilant (/s, z, ʃ, ʒ/ vs. /f, v, ɵ, ð/) is coded as S; voicing is coded as V; place of articulation in non-sibilants (/f, v/ vs. /ɵ, ð/) is coded as P_n; and place of articulation in sibilants (/s, z/ vs. /ʃ, ʒ/) is coded as P_s.

Cue	***Contextual Factors***		***Fricative Identity df=7,2848***	***Cue for***
	Speaker df=19,2860	***Vowel df=5,2855***		
MaxPF	0.084*		0.493*	S, P_s
DUR_F	0.158*	0.021*	0.469*	S, V
DUR_V	0.475*	0.316*	0.060*	V
RMS_F	0.081*		0.657*	S, V
RMS_V	0.570*	0.043*	0.011*	
$F3AMP_F$	0.070*	0.028*	0.483*	S, P_s
$F3AMP_V$	0.140*	0.156*	0.076*	P_n^1, P_s
$F5AMP_F$	0.077*	0.012*	0.460*	S
$F5AMP_V$	0.203*	0.040*	0.046*	
LF	0.117*	0.004$^+$	0.607*	S, V
F_0	0.838*	0.007*	0.023*	
F1	0.064*	0.603*	0.082*	V^2
F2	0.109*	0.514*	0.119*	S, P_n, P_s
F3	0.341*	0.128*	0.054*	P_n
F4	0.428*	0.050*	0.121*	P_n, P_s
F5	0.294*	0.045*	0.117*	P_n, P_s
M1	0.122*		0.425*	V, P_s
M2	0.036*		0.678*	S, V, P_s
M3	0.064*		0.387*	S, P_s
M4	0.031*		0.262*	P_s
$M1_{trans}$	0.066*	0.043*	0.430*	S, V, P_s
$M2_{trans}$	0.084*	0.061*	0.164*	P_n^3, P_s
$M3_{trans}$	0.029*	0.079*	0.403*	S, V, P_n, P_s
$M4_{trans}$	0.031*	0.069*	0.192*	S, P_n, P_s

$^+$p<.05 [1] R^2_{change}=.043 [3] R^2_{change}=.038
*p<.0001 [2] R^2_{change}=.045

Every cue was affected by fricative identity. While effect sizes ranged from large (10 cues had R^2_{change}> .40) to very small (RMS_{vowel}, the smallest: R^2_{change} =.011), all were highly significant. Even cues that were originally measured to compensate for variance in other cues (e.g. vowel duration to normalize fricative duration) had significant effects. Interestingly, two of the new measures (F4 and F5) had surprisingly large effects.

A few cues could clearly be attributed to one feature over others, although none were associated with a single feature. The two duration measures and low frequency energy were largely associated with voicing; RMS_F and $F5AMP_F$ were largely affected by sibilance; and the formant frequencies, F2, F4 and F5 had moderate effects of place of articulation for sibilants and non-sibilants.

However, the bulk of the cues were correlated with multiple features.

While few cues were uniquely associated with one feature, most features had strong correlates. Many cues were sensitive to place of articulation in sibilants, suggesting an invariance approach may be successful for distinguishing sibilants. However, there were few cues for place in non-sibilants (F4, F5, and the 3rd and 4th moments in the transition). These showed only moderate to low effect sizes (none greater than .1), and were context-dependent. Thus, categorizing non-sibilants may require at least cue-integration, and potentially, compensation.

We next asked if any cues appeared more invariant than others. That is, are there cues that are correlated with a single feature (place of articulation, sibilance or voicing), but not with context? There is no standard for what statistically constitutes an invariant cue, so we adopted a simple criterion based on Cohen and Cohen's (1983) definition of effect sizes as small (R^2<.05), medium (.05<R^2<.15) and large (R^2>.15): a cue is invariant if it had a large effect of a single feature (sibilance, place of articulation, voicing) and at most, small effects of context.

No cue met this definition. Contextual factors (talker and vowel) accounted for a significant portion of the variance in every cue, particularly cues in the vocalic portion. However, relaxing this criterion to allow moderate context effects yielded several.

Peak frequency (MaxPF) was highly correlated with place of articulation (R^2_{change}=.483), less so with sibilance (R^2=.260) (it distinguishes /s/ from /ʃ/), and virtually uncorrelated with voicing (R^2=.004). While it was moderately related to talker (R^2=.084), it was not related to vowel. The narrow-band amplitudes in the frication ($F3AMP_F$ and $F5AMP_F$) showed a similar pattern. Amplitude at F3 had a strong relationship to place (R^2=.450; /s/ vs. /ʃ/), and a smaller relationship to sibilance (R^2=.239); while amplitude at F5 was related to sibilance (R^2_{change}=.394) but not place within either class (non-sibilants: R^2_{change}<.001; sibilants: R^2_{change}=.02). Neither was strongly related to voicing ($F3AMP_F$: R^2_{change}=.002; $F5AMP_F$=.024) and they were only moderately affected by context ($F3AMP_F$: R^2_{change}=.098; $F5AMP_F$=.089).

Finally, the upper spectral moments in the frication were strongly associated with fricative identity (M2: R^2_{change}=.68; M3: R^2_{change}=.39; M4: R^2_{change}=.26) (primarily place of articulation), and only moderately with context (M2: R^2_{change}=.04; M3: R^2_{change}=.07; M4: R^2_{change}=.03). This was true to a lesser extent for M1 (Fric.: R^2_{change}=.42; Context: R^2_{change}=.12).

In sum, every cue was useful for distinguishing fricatives, although most were related to multiple phonetic features, and every cue was affected by context. There were several highly predictive cues that met a liberal criterion for invariance. Together, they may be sufficient for categorization, particularly given the large number of potentially supporting cues.

2.2 Perceptual Experiment

The perceptual experiment probed listeners' categorization of a subset of the corpus. We assessed overall accuracy and variation in accuracy across talkers and vowels on the complete syllable and the frication alone. Excising the vocalic portion eliminates some secondary cues to fricatives, but also reduces the ability to categorize the vowel and talker, which is required for compensation in C-CuRE. Thus, the difference between the *frication-only* and *complete-syllable* conditions may offer a crucial platform for model comparison.

2.2.1 Methods

The 2880 fricatives in the corpus were too many for listeners to classify in a reasonable amount of time, so this was trimmed to include 10 talkers (5 female), 3 vowels (/i, ɑ, u/), and the second repetition. This left 240 stimuli which were identified twice by each listener. The presence or absence of the vocalic portion was manipulated between-subjects.

Procedure. Listeners were tested in groups of two to four. Stimuli were played from disk over Sony (MDR-7506) headphones, using BLISS (Mertus, 1989). Stimuli were presented in random order at 3-s intervals. Listeners responded by circling one of 9 alternatives f, v, th, dh, s, z, sh, zh, or 'other' on answer sheets. Participants were asked to repeat a few words with /ɵ, ð, ʃ, ʒ/ in initial position to ensure they were aware of the difference between these sounds.

Participants. Forty Cornell University students (20 females) participated. Twenty served in each condition (complete-syllable vs. frication-only). All were native speakers of English with no known speech or hearing impairments. Participants were paid for their participation.

2.2.2 Results

Figure 2 shows a summary of listeners' accuracy. In the ***complete-syllable*** condition, listeners were highly accurate overall (M=91.2%), particularly on the sibilants (M=97.4%), while in the ***frication-only*** condition, performance dropped substantially (M=76.3%). There were also

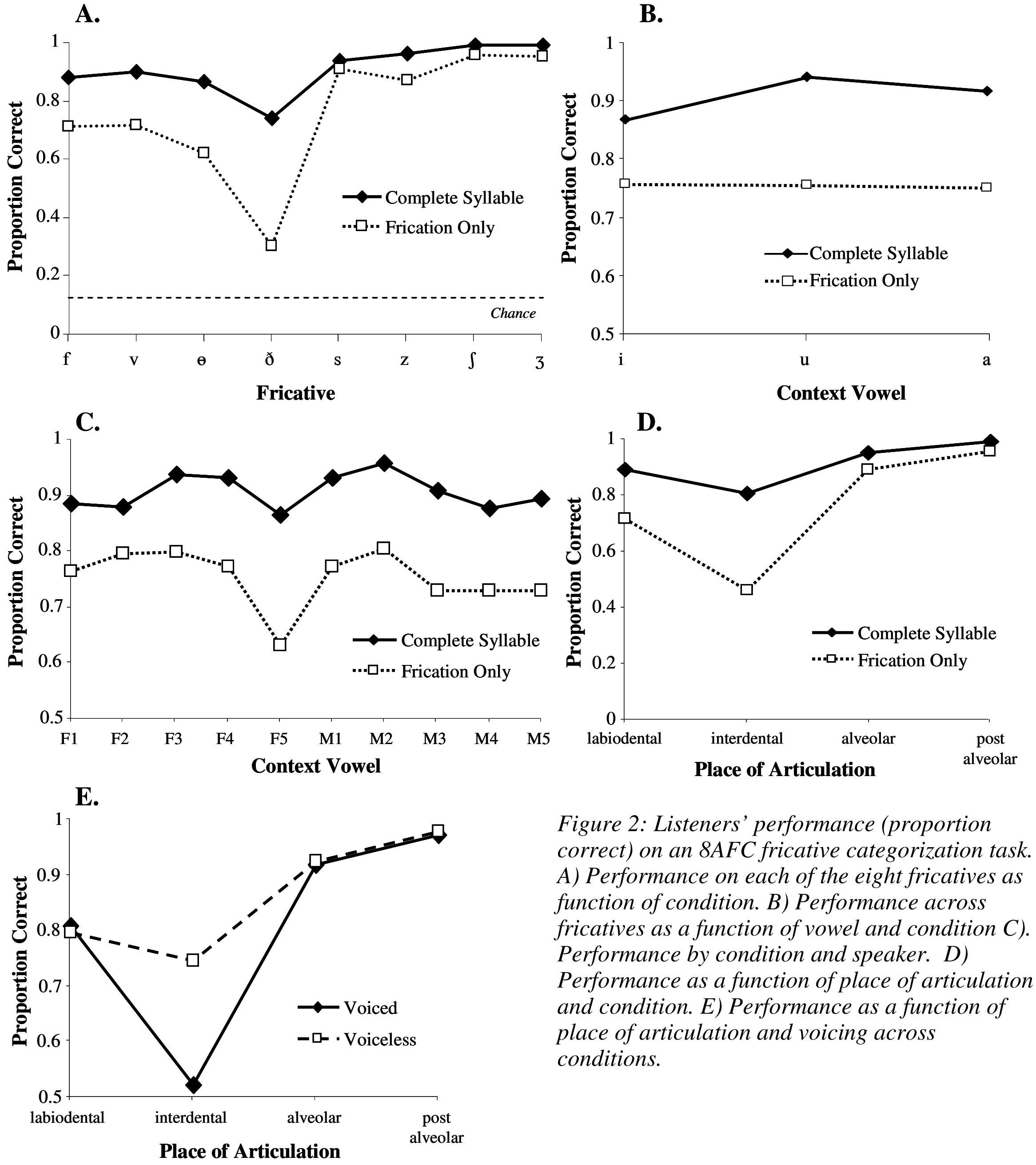

Figure 2: Listeners' performance (proportion correct) on an 8AFC fricative categorization task. A) Performance on each of the eight fricatives as function of condition. B) Performance across fricatives as a function of vowel and condition C). Performance by condition and speaker. D) Performance as a function of place of articulation and condition. E) Performance as a function of place of articulation and voicing across conditions.

systematic effects of vowel (Figure 2B) and talker (Figure 2C) on accuracy. It was necessary to characterize which of these effects were reliable to identify criteria for model evaluation. However, this proved challenging given that our dependent measure has eight possibilities (eight response categories), and the independent factors included condition, talker, vowel, place and voicing. Since we only needed to identify diagnostic patterns, we simplified this by focusing on accuracy and collapsing the dependent measure into a single binary variable: correct or incorrect (though see Supplementary Note 3 for a more descriptive analysis of the confusion matrices).

We used generalized estimating equations with a logistic linking function to conduct the equivalent of a repeated measures ANOVA (Lipsitz, Kim & Zhao, 2006). Talker, vowel, place and voicing were within-subjects factors; syllable-type (complete-syllable vs. frication-only) was between-subjects. Since we only had two repetitions of each stimulus per subject, the complete model (subject, five factors and interactions) was almost fully saturated. Thus, the context factors (talker and vowel) were included as main effects, but did not participate in interactions.

There was a significant main effect of syllable-type (Wald $\chi^2(1)=69.9$, $p<.0001$) with better performance in the complete-syllable condition (90.8% vs. 75.4%) for every fricative (Figure 2A). Vowel (Figure 2B) had a significant main effect (Wald $\chi^2(2)=12.1$, $p=.02$): fricatives preceding /i/ had the lowest performance followed by those preceding /ɑ/, and then /u/; and both /i/ and /ɑ/ were significantly different from /u/ (/i/: Wald $\chi^2(1)=12.1$, $p=.001$; /ɑ/: Wald $\chi^2(1)=5.5$, $p=.019$). Talker was also a significant source of variance (Talker: Wald $\chi^2(9)=278.1$, $p<.0001$), with performance by talker ranging from 75.0% to 88.2% (Figure 2C).

Place of articulation was highly significant (Wald $\chi^2(3)=312.5$, $p<.0001$). Individual comparisons against the postalveolars (which showed the best performance) showed that all three places of articulation were significantly worse (Labiodental: Wald $\chi^2(1)=72.2$, $p<.0001$; Interdental: Wald $\chi^2(1)=75.0$, $p<.0001$; Alveolar: Wald $\chi^2(1)=51.6$, $p<.0001$), though the large difference between sibilants and non-sibilants was the biggest component of this effect. A similar place effect was seen in both syllable-types (Figure 2D), though attenuated in complete-syllables, leading to a place × condition interaction (Wald $\chi^2(3)=12.0$, $p=.008$).

The main effect of voicing was significant (Wald $\chi^2(1)=6.2$, $p=.013$): voiceless fricatives were identified better than voiced fricatives. This was driven by the interdentals (Figure 2E), leading to a significant voicing × place interaction (Wald $\chi^2(3)=47.0$, $p<.0001$). The voicing effect was also enhanced in the noise-only condition, where voiceless sounds were 8.9% better, relative to the complete-syllable condition where the difference was 2.1%, a significant voicing × syllable-type interaction (Wald $\chi^2(1)=4.1$, $p=.042$). The three-way interaction (voicing × place × syllable type) was not significant (Wald $\chi^2(3)=5.9$, $p=.12$).

Follow-up analyses separated the data by syllable-type. Complete details are presented in the Online Note #2, but several key effects should be mentioned. First, talker was significant for both conditions (complete-syllable: Wald $\chi^2(9)=135.5$, $p<.0001$; frication-only: Wald $\chi^2(9)=196.9$, $p<.0001$), but vowel was only significant in complete-syllables (complete-syllable: Wald $\chi^2(1)=71.0$, $p<.0001$; frication-only: Wald $\chi^2(2)=.9$, $p=.6$). Place of articulation was significant in both conditions (complete-syllable: Wald $\chi^2(3)=180.5$, $p<.0001$; frication-only: Wald $\chi^2(3)=189.8$, $p<.0001$), although voicing was only significant in the frication-only condition (complete-syllable: Wald $\chi^2(1)=.08$, $p=.7$; frication-only: Wald $\chi^2(1)=15.0$, $p<.0001$).

To summarize, we found that 1) performance without the vocalic portion was substantially worse than with it, though performance in both cases was fairly good; 2) accuracy varied across talkers; 3) sibilants were easier to identify than non-sibilants but there were place differences even within sibilants; and 4) the vowel identity affected performance, but only in the complete-syllable condition. This may be due to two factors. First, particular vowels may alter secondary cues in the vocalic portion in a way that misleads listeners (for /ɑ/ and /i/) or helps them (for /u/). Alternatively, the identity of the vowel may cause subjects to treat the cues in the frication noise differently. This may be particularly important for /u/—its lip rounding has a strong effect on the frication. As a result, listeners' ability to identify the vowel (and thus account for these effects) may offer a benefit for /u/ that is not seen for the unrounded vowels.

2.3 Discussion

The acoustic analysis revealed that every cue was useful for categorizing fricatives, but all were affected by context. However, the handful of nearly invariant cues raises the possibility that uncompensated cues, particularly in combination, may be sufficient for separating categories. Our perceptual study also revealed consistent differences across talkers, vowels and fricatives in accuracy. The presence or absence of the vocalic portion had the largest effect. This hints at compensation using C-CuRE mechanisms since this difference may be both due to secondary cues to the fricative, and also to listeners' ability to identify the talker and the vowel as a basis of compensation. This may also account for the effect of context vowels on accuracy in the complete-syllable condition but not in the fricative-only condition.

3. Computational Approach

Our primary goal was to determine what information is needed to separate fricative categories at listener-like levels. We thus employed multinomial logistic regression as a simple, common model of phoneme categorization that is theoretically similar to several existing approaches (Oden & Massaro, 1978; Nearey, 1990; Smits, 2001; Cole et a., 2010). We varied its training set to examine three sets of informational assumptions:

1) ***Naïve Invariance***: This model used the small number of cues that were robustly correlated with fricative identify and less with context. Cues did not undergo compensation—if cues are invariant with respect to context, this should not be required.
2) ***Cue-Integration:*** This model used every cue available, without compensation. This is consistent with the informational assumptions of exemplar approaches, and is an unexamined assumption of cue-integration models like NAPP (Nearey, 1997).
3) ***Compensation:*** This model used every cue, but after the effects of talker and vowel on these cues had been accounted for using C-CuRE.

It may seem a forgone conclusion that compensation will yield the best performance–it has the most information and involves the most processing. However, our acoustic analysis suggests there is substantial information in the raw cues to support fricative categorization, and no one has tested the power of integrating 24 cues for supporting categorization. Thus, uncompensated cues may be sufficient. Moreover, compensation in C-CuRE is not optimized to fricative categorization – it could transform the input in ways that hurt categorization. Finally, the goal is not necessarily the best performance, but listener-like performance—none of the models are optimized to the listeners' responses and they may or may not show such effects

The next section describes the categorization model and its assumptions. Next, we describes how we instantiated each of our three hypotheses in terms of specific sets of cues.

3.1 Logistic Regression as a model of Phoneme Categorization

Our model is based on work by Nearey (1990, 1997; see also Smits, 2001a,b; Cole et al., 2010; McMurray, et al., in press) which uses logistic regression as a model of listeners' mappings between acoustic cues and categories. Logistic regression first weights and combines multiple cues linearly. This is transformed into a probability (e.g. the probability of an /s/ given the cues).

Weights are determined during training to optimally separate categories (Hosmer & Lemshow, 2000, for a tutorial). These parameters allow the model to alter both the location of the boundary in a multi-dimensional space and the amount each cue participates in categorization.

Logistic regression typically uses a binary dependent variable (e.g. /s/ vs. /ʃ/) as in (1).

$$P(s \mid x_1, x_2 ...) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ...}} \tag{1}$$

Here, the exponential term is a linear function of the independent factors (cues: $x_1 ... x_n$) weighted by their regression coefficients (*β's*). Multinomial logistic regression generalizes this to map cues to any number of categories. Consider, for example, a model built to distinguish /s/, /z/, /f/ and /v/. First, there are separate regression parameters for each of the four categories, except one, the reference category[5] (in this case, /v/). The exponential of each of these linear terms is in (2)

$$L(s) = e^{\beta_{s0} + \beta_{s1} x_1 + \beta_{s2} x_2 ...} \qquad L(z) = e^{\beta_{z0} + \beta_{z1} x_1 + \beta_{z2} x_2 ...} \qquad L(f) = e^{\beta_{f0} + \beta_{f1} x_1 + \beta_{f2} x_2 ...} \tag{2}$$

These are then combined to yield a probability for any given category.

$$P(s) = \frac{L(s)}{1 + L(s) + L(z) + L(f)} \tag{3}$$

The probability of the reference category is

$$P(v) = \frac{1}{1 + L(s) + L(z) + L(f)} \tag{4}$$

Thus, if there are 10 cues, the logistic regression requires 10 parameters plus an intercept for each category (minus one). Thus, a multinomial logistic regression mapping 10 cues to four categories requires 33 parameters. Typically these are estimated using gradient descent methods that maximize the likelihood of the data given the parameters.

The models used here were first trained to map the dataset of acoustic measurements to the intended production. This is overgenerous. The model knows both the cues in the acoustic signal, and the category the speaker intended to produce—something which learners may not always have access to. However, by training it on intended productions, not listener data, its match to listeners' performance must come from the information in the input, as the model is not trained to match listeners, only to achieve the best categorization it can with the input.

We held out from training the 240 tokens used in the perception experiment. Thus, training consisted of 2880 - 240 = 2640 tokens. After estimating the parameters, we used them to determine the likelihood of each category for each token in the perceptual experiment.

3.1.1 Evaluating the models

Evaluating logistic models is tricky. There is no agreed upon measure for model comparison (like R^2 in linear regression). Moreover, this study compares models that use only a handful of cues to those that use many, and therefore should compensate for the increased power of the more complex models. Finally, model fit (e.g. how well the model predicts the training data) is less important than its ability to yield listener-like performance (on which it was not trained). Thus, we evaluated our models in three ways: Bayesian Information Criterion (BIC); estimated performance by experimental condition; and the likelihood of the human data given the model. We discuss the first two here, and discuss the final measure in Section 4.4 where it is used.

Bayesian Information Criterion (BIC; Schwarz, 1978) is used for selecting among competing models. BIC is sensitive to the number of free parameters and the sample-size. It is usually computed using (5), which provides an asymptotic approximation for large sample sizes.

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n) \tag{5}$$

Here, L is the likelihood of the model, k is the number of free parameters, and n is the number of samples. Given two models, the one with the lower BIC is preferred.

BIC can be used in two ways. It is primarily used to compare two models' fit to the training data. Secondarily, it offers an omnibus test of model fit. To do this, the model is first estimated with no independent variables. This "intercept-only" model should have little predictive value, but if one response was *a priori* more likely, it could perform above chance. Next, the independent factors are added and the two models are compared using BIC to determine if the addition of the variables offers any real advantage.

Categorization performance can be computed from logistic regression models and is analogous to the listener data. This estimated listener performance can be compared as a function of experimental condition (e.g. as a function of fricative, talker or vowel), for a qualitative match to listeners. If one of the models shows similar effects of talker, vowel, or fricative this may offer a compelling case for this set of informational assumptions.

Crucially, this relies on the ability to generate data analogous to listener categorization from the logistic model. While, the logistic formula yields a probability of each of the categories for any given set of cues, there is debate about how best to map this to listener performance.

For any token, the optimal decision rule is to choose the most likely category as the response (Nearey & Hogan, 1986)[6]. This implies that listeners always choose the same category for repetitions of the same token (even if it is only marginally better). This seems unrealistic: in our experiment listeners responded identically to each repetition only 76.4% of the time in the frication-only condition and 90.5% in complete-syllables (close to the average accuracies). Thus, a more realistic approach is to use the probabilities generated by the model as the probability the listener chose each category (as in Nearey, 1990; Oden & Massaro, 1978).

The discrete-choice rule generally yields better performance than the probabilistic rule (typically about 10% in these models) and listeners likely lie between these extremes. This could be modeled with something like the Luce-choice rule (Luce, 1959), which includes a temperature parameter controlling how "winner-take-all" the decision is. However, we had no independent data on the listeners' decision criteria, and since models were fit to the intended production, not to the perceptual response, we could not estimate this during training. We thus report both the discrete-choice and probabilistic decision rules for each model as a range, with the discrete-choice as the upper limit and the probabilistic rule as the lower limit.

Finally, neither method offers a direct fit to the perceptual data. BIC is based on the training data, and performance-based measures are analogous to perceptual data but offer no way to quantitatively relate them. Thus, in Section 4.4, we describe a method of comparing models based on the likelihood that the perceptual data was generated by each model.

3.1.2 Theoretical Assumptions of logistic regression as a categorization model

As a model of the interface between continuous cues and phoneme categories, logistic regression makes a number of simplifications. First, it assumes linear boundaries in cue-space (unless interaction terms are included). However, Nearey (1990) has shown that this can be sufficient for some speech categories. Similarly, cue combination is treated as a linear process. However, weighting-by-reliability in vision (e.g. Jacobs, 2002; Ernst & Banks, 2002) also assumes linear combinations and this has been tested in speech as well (Toscano & McMurray, 2010). Given the wide-spread use of this assumption in similar models in speech (e.g. Oden & Massaro, 1978; Nearey, 1997; Smits, 2001b), this seems uncontroversial. Moreover, lacking hypotheses about particular nonlinearities or interaction terms, the use of a full complement of interactions and nonlinear transformations may add too many parameters to fit effectively.

Second, while there are more complicated ways to model categorization, many of these approaches are related to logistic regression. For example, a network that uses no hidden units and the softmax activation function is identical to logistic regression, and an exemplar model in which speech is compared to all available exemplars will be highly similar to our approach

Third, logistic regression can be seen as instantiating the outcome of statistical learning (e.g. Werker et al, 2007) as its categories are derived from the statistics of the cues in the input. However, many statistical learning approaches in speech perception (e.g. McMurray et al, 2009a; Maye, Werker & Gerken, 2002) assume unsupervised learning, while logistic regression is supervised–the learner has access to both cues and categories. We are not taking a strong stance on learning—likely both are at work in development. Logistic regression is just a useful tool for getting the maximum categorization value out of the input to compare informational hypotheses.

Finally, our use logistic regression as a common categorization platform intentionally simplifies the perceptual processes proposed in models of speech perception. However this allows us to test assumptions about the information that contribute to categorization. By modeling phoneme identification using the same framework, we can understand the unique contributions of these informational assumptions made by each class of models.

3.2 Hypotheses and datasets

3.2.1 Naïve Invariance Model

The ***naïve invariance*** model asked whether a small number of uncompensated cues are sufficient for classification. Prior studies have asked similar questions for fricatives (Forrest et al, 1988; Jongman et al, 2000) using discriminant analysis and results have been good, though imperfect. This has not yet been attempted with more powerful logistic regression; and we have a lot more cues (particularly for non-sibilants). Thus, it would be premature to rule out such hypotheses.

Section 2.1.2 suggested a handful of cues that are somewhat invariant with respect to context (Table 4). These nine cues distinguish voicing and sibilance in all fricatives, and place of articulation in sibilants. We did not find any cues that were even modestly invariant for place of articulation in non-sibilants. Thus, we added four additional cues: two with relatively high R^2's for place of articulation, but also context (F4 and F5), and two that were less associated with place but also with context ($M3_{trans}$ and $M4_{trans}$). These were located in the vocalic portion, offering a way for the naïve invariance model to account for the differences between the frication-only and complete-syllable conditions in the perceptual experiment—the loss of these cues should lead to bigger decrements for non-sibilants, and smaller decrements for sibilants. As our selection of this cue-set was made solely by statistical reliability (rather than a theory of production), and we did not use any compound cues, we term this a *naïve invariance* approach.

3.2.2 Cue-integration Model

The cue-integration hypothesis suggests that if sufficient cues are encoded in detail, their combination is sufficient to overcome variability in any one cue. This is reflected in the informational assumptions of exemplar approaches (e.g. Goldinger, 1998; Pierrehumbert, 2001, 2003) and it is an unexamined assumption in many cue-integration models. It is possible that in a high-dimensional input-space (24 cues), there are boundaries that distinguish the eight fricatives.

Our use of logistic regression as a categorizer could is a potentially problematic assessment of exemplar accounts, as it is clearly more akin to a prototype model than true exemplar matching. However, if the speech signal is compared with the entire "cloud" of exemplars, then the decision of an exemplar model for any input will reflect an aggregate of all the exemplars, a "Generic Echo" (Goldinger, 1998, p. 254), allowing it to show prototype-like effects (Pierrehumbert, 2003). In contrast, if the signal is compared to only a smaller number of tokens this breaks down. Ultimately, formal models will be required to determine the optimal decision rule for exemplar models in speech. However, given our emphasis on information, logistic regression is a reasonable test—it maps closely to both cue-integration models and some versions of exemplar theory and is sufficiently powerful to permit good categorization.

Thus, we instantiated the cue-integration assumptions by using all 24 cues with no compensation for talker or vowel. To account for the difference between the frication-only and complete-syllable condition, we eliminated the 14 cues found in the vocalic portion (Table 2), asking whether the difference in performance was due to the loss of additional cues.

Table 4: Summary of the 9 cues that were relatively invariant with respect to speaker and vowel as well as four non-invariant cues that were included because they provided the best information about place of articulation in non-sibilants. ***R^2*** *are change statistics taken from analyses presented in Section 2.*

Cue	**Cue for**	**Context Effects**
MaxPF	Place in sibilants (R^2=.504)	Speaker: Moderate (R^2=.084) Vowel: n.s.
DUR_F	Voicing (R^2=.40)	Speaker: Large (R^2=.16) Vowel: Small (R^2=.021)
RMS_F	Sibilance (R^2=.419)	Speaker: Moderate (R^2=.081) Vowel: n.s.
F3AMP_F	Place in sibilants (R^2=.44) Sibilance (R^2=.24)	Speaker: Moderate (R^2=.07) Vowel: Small (R^2=.028)
F5AMP_F	Sibilance (R^2=.39)	Speaker: Moderate (R^2=.07) Vowel: Small (R^2=.012)
LF	Voicing (R^2=.48)	Speaker: Moderate (R^2=.11) Vowel: Small (R^2=.004)
M1	Place in sibilants (R^2=.55)	Speaker: Moderate (R^2=.122) Vowel: n.s.
M2	Sibilance (R^2=.44) Place in sibilants (R^2=.34)	Speaker: Small (R^2=.036) Vowel: n.s.
M3	Place in sibilants (R^2=.37)	Speaker: Moderate (R^2=.064) Vowel: n.s.
Non-invariant cues to place in non-sibilants		
F4	Place in non-sibilants (R^2=.083)	Speaker: Large (R^2=.43) Vowel: Small (R^2=.05)
F5	Place in non-sibilants (R^2=.082)	Speaker: Large (R^2=.29) Vowel: Small (R^2=.045)
M3_trans	Place in non-sibilants (R^2=.061)	Speaker: Small (R^2=.029) Vowel: Moderate (R^2=.079)
M4_trans	Place in non-sibilants (R^2=.062)	Speaker: Small(R^2=.031) Vowel: Small (R^2=.069)

3.2.3 Compensation / C-CuRE Model

The final dataset tests the hypothesis that compensation is required to achieve listener-like performance. This is supported by our phonetic analysis suggesting that all of the cues were somewhat context-sensitive. To construct this dataset, all 24 cues were processed to compensate for the effects of the talker and vowel on each one. While the goal of this dataset was to test compensation in general, this was instantiated using C-CuRE because it is, to our knowledge, the only general purpose compensation scheme that is computationally specific and can be applied to any cue; and can accomplish compensation without discarding fine-grained detail in the signal (and may enable greater use of it: Cole et al, 2010).

To construct the C-CuRE model, all 24 cues were first subjected to individual regressions in which each cue was the dependent variable and talker and vowel were independent factors. These two factors were each represented by 19 and 5 dummy variables (respectively), one for each talker/vowel (minus one)[7]. After the regression equation was estimated, individual cue-values were then recoded as standardized residuals. This recodes each cue as the difference between the actual cue value and what would be predicted for that talker and vowel.

The use of context in C-CuRE offers an additional route to account for differences between the complete-syllable and frication-only condition. While excising the vocalic portion eliminates a lot of useful first order cues (as in the cue-integration model), it would also impair parsing. This is because it would be difficult to identify the talker or vowel from the frication alone. Indeed, Lee et al. (2010) showed that the vocalic portion supports much more robust identification of gender than the frication, even when only six pitch pulses were present. Thus, if fricative categorization is contingent on identifying the talker or vowel, we should see an additional decrement in the frication-only condition due to the absence of this information.

Our use of regression for compensation complicates model comparison using BIC. How do we count the additional parameters? In the cue-integration model, each cue corresponds to seven degrees of freedom (one parameter for each category minus one). In contrast, the compensation / C-CuRE model uses additional degrees of freedom in the regressions for each cue: 19 df for talkers, 5 for vowels, and an intercept. Thus, instead of 7x24 parameters, the complete C-CuRE model now has 32x24 parameters, suggesting a substantial penalty.

However, there are three reasons why such a penalty would be ill advised. First, the free parameters added by C-CuRE do not directly contribute to categorization, nor are they optimized when the categorization model is trained. These parameters are fit to a different problem (the relationship between contextual factors and cues), and are not manipulated to estimate the logistic regression model. So while they are parameters in the system, they are not optimized to improve categorization. In fact, it is possible that the transformations imposed by C-CuRE impede categorization or make categorization look less like listeners, as C-CuRE is removing the effect of factors like vowel and talker that we found to affect listener performance.

Second, the parameters for parsing with C-CuRE can be computed directly from the data. When the independent variables are discrete, the regression parameters are related to the combination of cell means. No complex optimization is needed to estimate these values.

Finally, any scaled up system, even without compensation, would need to identify vowels and talkers. Since the parameters used in C-CuRE are simply the mean values of each cue with respect to context, a cue-integration model that was trained to identify the context along with the fricative would be estimating similar parameters anyways—they just would not be used in fricative identification. The C-CuRE model simply reuses those parameters for compensation.

4. Results

Our analysis starts with a description of the results of each model. Next we describe a scheme for making quantitative model fits to the perceptual data and compare the three approaches. Finally, we address several assumptions of the Compensation / C-CuRE model.

4.1 Naïve Invariance

4.2.1 Complete Syllables

Overall, the naïve invariance model offered a good fit to the production data. BIC decreased substantially from the intercept-only model (Intercept-only: 11034; Invariance Model: 3399), and the χ^2 test of model fit was significant ($\chi^2(91)=8353$, p<.0001). Likelihood ratio tests showed that the model used all 13 cues (all $\chi^2(7)>28$, p<.0001). The model averaged 83.3%

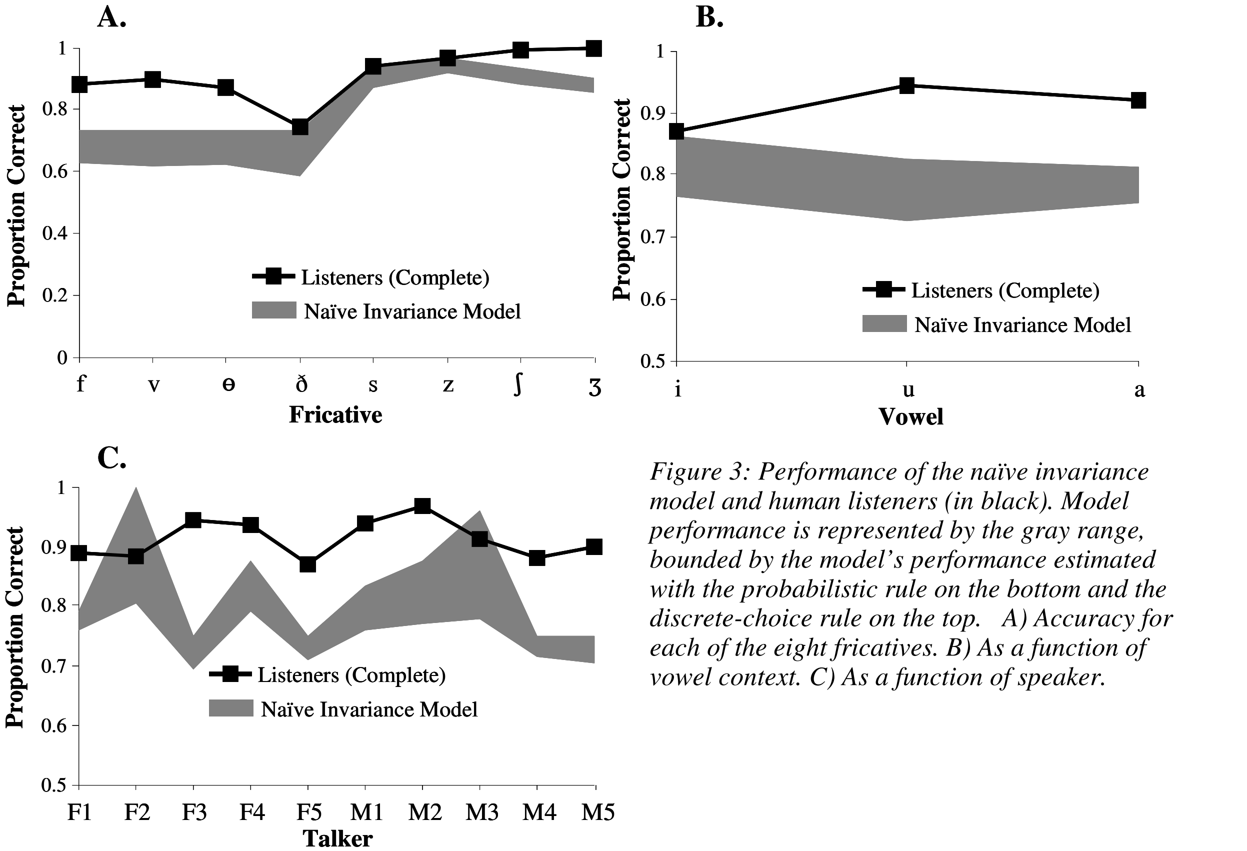

Figure 3: Performance of the naïve invariance model and human listeners (in black). Model performance is represented by the gray range, bounded by the model's performance estimated with the probabilistic rule on the bottom and the discrete-choice rule on the top. A) Accuracy for each of the eight fricatives. B) As a function of vowel context. C) As a function of speaker.

correct on the perception tokens using the discrete-choice rule, and 74.8% correct using the probabilistic rule. Thus, this handful of invariant cues supports fairly accurate identification, though less so than listeners.

However, it was not a good fit to listener performance. It is much poorer on non-sibilants than listeners, and there are no differences within them (Figure 3A). Similarly, within sibilants, it fails to capture listeners' slightly better performance on the post-alveolars (/ʃ, ʒ) than alveolars (/s, z/). Moreover, the breakdown of performance by both vowel (Figure 3B) and talker (Figure 3C) does not show the expected patterns, with the model showing the inverse effect of vowel, and little correlation with listeners' performance across talkers (R=.18).

Thus, this model undershoots listener performance by about 15% when measured with the more realistic probabilistic decision rule and by about 7.5% with the discrete choice rule. More importantly, it does not describe listeners' errors. The model showed the inverse effect of context vowel, and a different effect of talker. Together, this suggests that the information in this small set of cues does not fully capture the similarity relations that underlie listeners' categorization, nor is it sufficient to support listeners' levels of accuracy.

4.2.2 Frication Only

The invariant cues were largely in the frication portion of the stimulus – only four of the 13 were found in the vocalic portion. Thus, there should be little difference when the cues in the vocalic portion were eliminated to model the frication-only condition. This was confirmed as this model performed at 70.1% on the probabilistic rule and 78.7% on the discrete rule (compared with 74.5% and 83.3%). Given the small difference between models, and the fact that both models were in the range of the listeners (M=76.3%), we do not report more on this model.

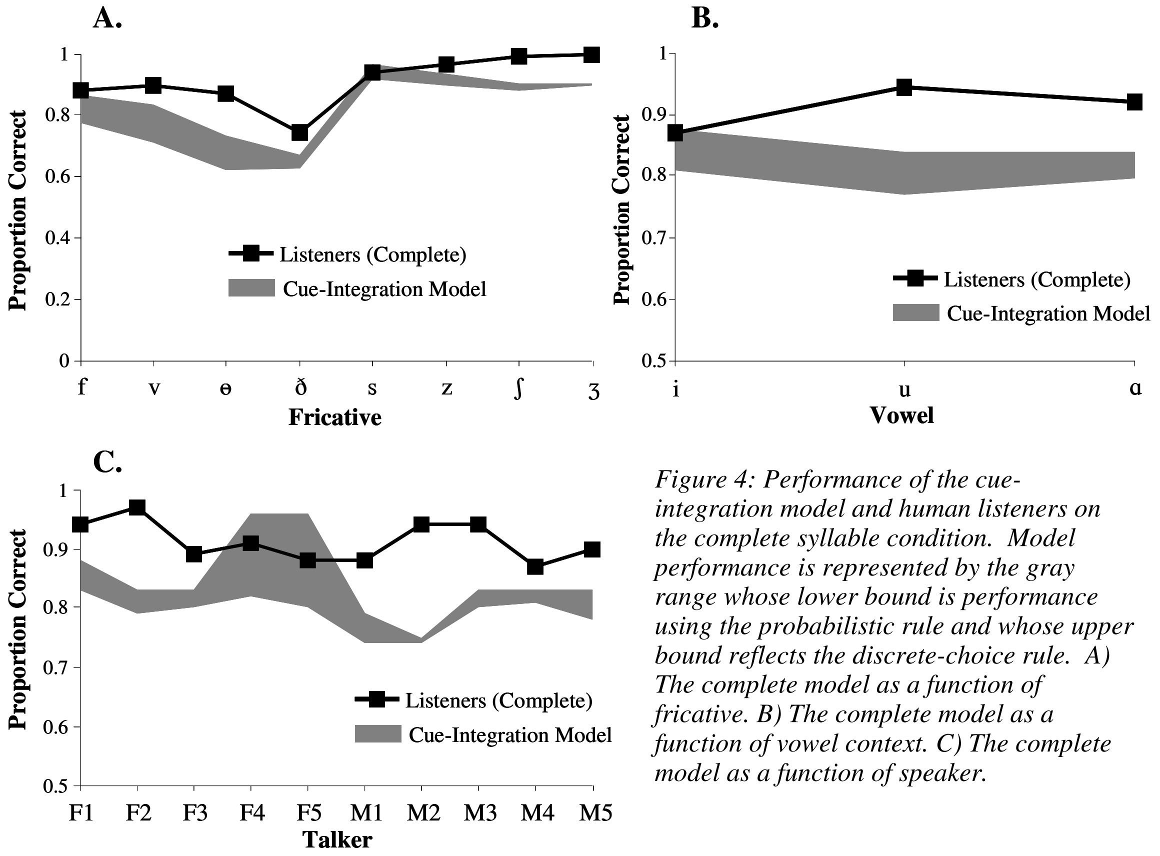

Figure 4: Performance of the cue-integration model and human listeners on the complete syllable condition. Model performance is represented by the gray range whose lower bound is performance using the probabilistic rule and whose upper bound reflects the discrete-choice rule. A) The complete model as a function of fricative. B) The complete model as a function of vowel context. C) The complete model as a function of speaker.

4.2 Cue-integration.

4.2.1 Complete Syllables

When all cues were used, model fit improved markedly. The intercept-only model had a BIC of 11034, and the full model showed a substantial decrease to 3381—lower than the naïve invariance model even when penalized for additional cues. The χ^2 analysis of fit was highly significant ($\chi^2(168)=8977$, $p<.0001$), however, the model did not take advantage of all the cues. Likelihood ratio tests showed that five were not used: F2 ($\chi^2(7)=8.7$, $p=.27$), F3AMP$_v$ ($\chi^2(7)=12.7$, $p=.08$), F5AMP$_v$ ($\chi^2(7)=7.1$, $p=.41$), M3$_{trans}$ ($\chi^2(7)=9.7$, $p=.21$) and M4$_{trans}$ ($\chi^2(7)=12.7$, $p=.08$)[8]. Interestingly, these were the cues that were most affected by vowel context. All other cues were highly significant (all $\chi^2(7)>15$, $p<.03$).

The model performed at 85.0% with the discrete-choice rule, and 79.2% with the probabilistic rule, an increase over the naïve invariance model (2.5%, 5% respectively). The new cues also allowed the model to better approximate listener performance. As Figure 4A shows, the model now exhibits differential performance within the non-sibilants: the interdentals are now worse than the labiodentals. In other ways, accuracy did not reflect listeners. The model performs better on alveolars than postalveolars, better for /i/ than the other two vowels (Figure 4B), and its performance across talkers is not correlated with listeners (R=-.01). Thus, this model improves over the naïve invariance model in accuracy and match to listeners, but it does not fully capture the pattern of errors and context effects.

4.2.2 Frication Only

Next, we examined whether the model could account for performance on the frication noise alone by training a new model on just the 10 cues in the frication (Table 2).

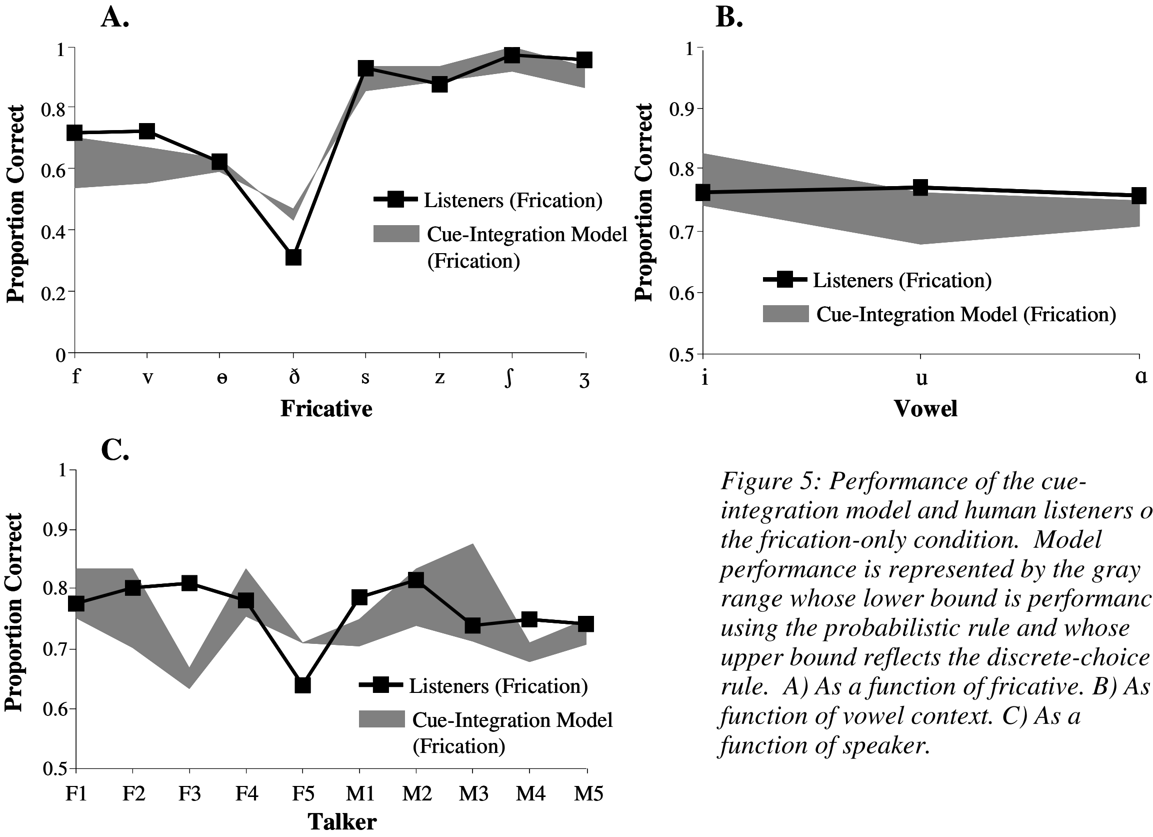

Figure 5: Performance of the cue-integration model and human listeners on the frication-only condition. Model performance is represented by the gray range whose lower bound is performance using the probabilistic rule and whose upper bound reflects the discrete-choice rule. A) As a function of fricative. B) As a function of vowel context. C) As a function of speaker.

This model fit the training data well, with a BIC of 11034 for the intercept-only version and 3431 for the final model ($\chi^2(70)=8155$, p<.0001), and all 10 cues significantly contributed to performance (all $\chi^2(7)>31$, p<.001). Performance was worse than the full model, mimicking the effect of eliminating the vocalic portion, averaging 77.9% for the discrete choice rule and 70.9% for the probabilistic one. This was quite close to listeners (75.4%). This model also offers a close fit to the listeners' accuracy across fricatives (Figure 5A). While it outperforms them on /ð/[9], it correctly captures differences between the other seven, particularly the sibilants. Its accuracy across talkers and vowels was more variable. Listeners showed little differences across vowel while the model was again best with /i/, although its range of performance mostly included the listener data. The effect of talker, however, was different between listeners and the model (R=-.04), though, as with vowels, listener performance is largely in the model's range.

Thus, the frication-only version of the cue-integration model offers a better fit to the corresponding empirical data than the complete-syllable version, particularly in overall accuracy. It is imperfect for some of the context effects, but many of the broad patterns are there. If anything, the complete-syllable version needs improvement to account for listener performance.

4.3 Compensation / C-CuRE.

The compensation model showed the best fit of all ($\chi^2(168)=10381$, p<.0001) with BIC reducing from 11977 to 2990. In contrast to the cue-integration model, likelihood ratio tests showed that all 24 cues affected performance (all $\chi^2(7)>21$, p<.004), suggesting that compensating for contextual variance helps the model gain access to new information sources.

Accuracy was excellent. The model was 92.9% correct using the discrete decision rule and 87.0% correct using the probabilistic rule (Figure 6A). This is a large improvement (over 7%) over the cue-integration model that puts performance in the range of human listeners. As

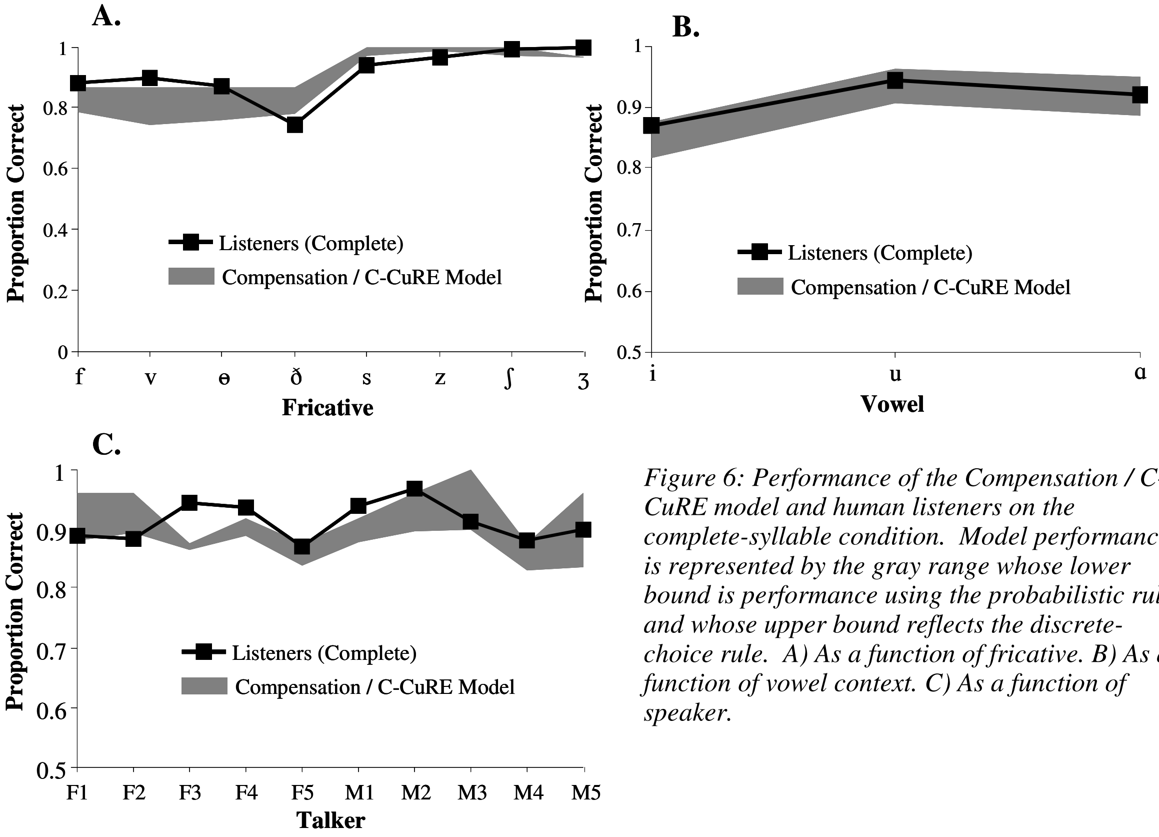

Figure 6: Performance of the Compensation / C-CuRE model and human listeners on the complete-syllable condition. Model performance is represented by the gray range whose lower bound is performance using the probabilistic rule and whose upper bound reflects the discrete-choice rule. A) As a function of fricative. B) As a function of vowel context. C) As a function of speaker.

Figure 6A shows, the model performed equivalently to listeners for sibilants and /θ/, although it slightly undershot them for /f/ and /v/ and overshot them for /ð/. Perhaps most impressively, the effect of vowel context has completely reversed from prior models and now fits the human data (Figure 6B), and the talker differences are also well correlated (R=.52) (Figure 6C). At a statistical level, this is surprising – we have partialed the effects of talker and vowel *out* of the raw cues, and yet, we are now seeing the correct effects in performance. This suggests that listeners' differences across vowels may be due to differences in compensation, not differences in the raw information available.

4.4 Model Comparison.

The results thus far (Table 5) indicate that compensated cues using C-CuRE offer the closest match to listeners in the complete-syllable condition, while the cue-integration approach works well in the frication-only condition. Our goodness of fit measure, however, compared each model's fit to the training data (the intended productions), showing only that the C-CuRE model is the better classifier of this data. While the qualitative differences between models in predicting vowel and talker effects suggest the C-CuRE model is a better fit to listeners, we have not reported goodness of fit to the perceptual data. Here we develop the tools to do so.

We focus on comparing the three models in the complete-syllable condition. The naïve invariance and cue-integration models were similar for the frication-only data, and differed only by a single cue (M4). Second, applying the C-CuRE model to the frication-only data makes little sense theoretically—without the vocalic portion it would be difficult to identify the talker or vowel to parse their effects from the cues in the frication.

The perception data takes the form of a frequency distribution: for each token, the number of times each category was chosen. The output of logistic regression is analogous: the probability of each category given the input. From these probabilities and frequencies, we can use the multinomial distribution to compute the likelihood of getting a particular distribution of responses (the listener data) given the probabilities computed by the model:

$$L = \frac{N!}{X_1!X_2!\ldots X_8!} p_1^{X_1} p_2^{X_2} \ldots p_8^{X_8} \qquad (6)$$

Here, N is the total number of responses; X_i is the number of times category i was selected; and p_i is the probability of that category from the model. Multiplying this across each token in the dataset (with $p_1 \ldots p_8$ for each computed from the model based on that token's cues) gives the total likelihood of the entire perceptual dataset given the model.

This allows us to compare any two models using odds-ratios (the ratio of the two likelihoods) to determine how much more likely one model is over the other. Generally, to compute the odds-ratio, we divided the likelihoods by 240 to compute the average likelihood of each token, and used this to compute the average odds ratio across tokens. We can also compute BIC from the log-likelihoods, to compute a BIC value relative to the observed perceptual data.

Thus, we first used each of the three models to compute the probabilities for each response for each token in the dataset. We next computed the likelihood of obtaining the distribution of responses observed in perceptual data for each token. These were logged and summed to obtain the total log-likelihood of the data given each model.

Consistent with the accuracy data, the cue-integration model fit the listener data better than the naïve invariance model. Its log-likelihood was larger (-3823 vs. -4740), and it was 45.7 times more likely to give rise to the responses for any perceptual token than the naïve invariance model. Even when penalized for its cues, its BIC was still lower (8605.3 vs. 10018.2).

Next, we compared the cue-integration and compensation / C-CuRE models. Surprisingly, the C-CuRE model (LL=-7142.7; BIC=15245) was a worse fit than the cue-integration model (LL=-3823.1; BIC=8605.3). This was unexpected given the compensation model's better overall performance and its closer match to the perceptual data.

Examining the log-likelihoods for individual tokens, we noticed that while most were in the 0 to -50 range, the C-CuRE model had a handful of very unlikely tokens: 16 (out of 240) had log-likelihoods less than -100, and one was less than -1000. This was due to the fact that it was extremely confident in categorizing some tokens, outputting probabilities near zero (1e-50 or less) for dispreferred fricatives. If subjects responded even once for these near-zero probability events (e.g. if they were guessing), this dramatically decreased the likelihood of the model (since this tiny probability, multiplied by all the others, squashes the more likely probabilities from other trials). In contrast, the cue-integration model was less confident in its decision so the probabilities for dispreferred fricatives were orders of magnitude larger. As a result, guessing was not nearly as deleterious—the model expected to be wrong occasionally.

Table 5: Summary of Performance of the Models and Human Listeners.

Condition	**Model**	**% Correct**		**Model Fit**	
		Discrete	**Prob.**	**BIC**	**Cues**
Complete	***Listeners***	**91.2**		-	-
	Invariance	83.3	74.5	3399	13
	Cue Integration	85.0	79.2	3381	24
	Compensation / C-CuRE	92.9	87.0	2990	24 (parsed)
Frication Only	***Listeners***	**76.3**		-	-
	Invariance	78.8	70.1	3537	9
	Cue Integration	79.2	69.7	3431	10

Thus, we modified the logistic model to include a chance of guessing, by constructing a mixture model in which the probability of a category was a mixture of logistic and guess trials.

$$p_{category} = p_{\log istic}\left(1 - p_{guess}\right) + p_{guess}\left(\frac{1}{8}\right) \tag{7}$$

Here, $p_{logistic}$ is the probability of a given fricative from the logistic regression, p_{guess} is the likelihood of guessing, and there is a 1/8 chance of selecting any fricative on guess trials.

It was not clear how to estimate p_{guess}, as we had no independent data on listeners' guessing. Ideally, one would estimate this parameter with the rest of the parameters in the logistic model. However, the logistic model was fit to intended productions, not perceptual data, which left no way to estimate a property of the listener from an unambiguous training signal. Thus, we examined a roughly logarithmic range of guess-rates from 0 (the previously reported results) to 20%, to allow a comparison of each model at the same p_{guess}. We also estimated the optimal p_{guess} for each model and compared models at their optimal guess rates.

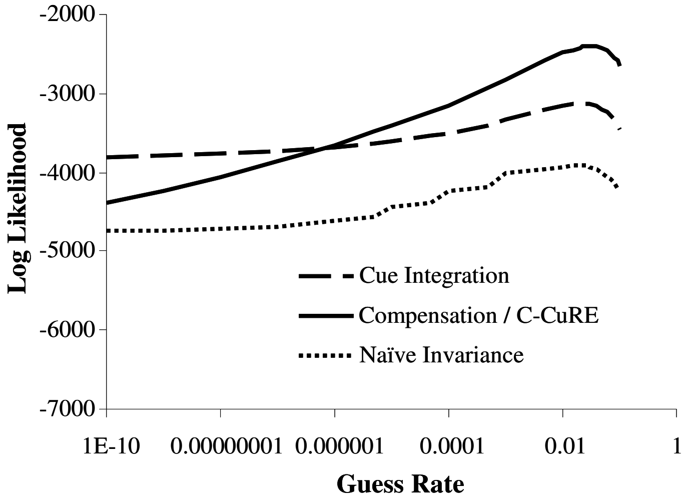

Figure 7: Log-likelihood for each of the models as a function of the guess rate.

Figure 7 shows the results. At very low guess rates, ($0 \leq p_{guess} \leq$ 1e-6), the cue-integration model was more likely. However, once p_{guess} exceeded 1e-6, the C-CuRE model was better at all values. At 1e-6, the average odds-ratio (C-CuRE / cue-integration) for individual tokens was 1.2, and this increased to 13.5 by .005, and stayed above 20 at when p_{guess} was greater than .0225. Thus even assuming extremely low rates of guessing, the C-CuRE model was more likely to generate the perceptual data than the cue-integration model. Our analysis at the optimal guess rates confirmed this. For the cue-integration model, the optimal p_{guess} was .02122 with a log-likelihood of -3131.2; for the C-CuRE model the optimal p_{guess} was slightly higher at .03094 but with a much higher log-likelihood of -2397.6, and the average odds ratio was 21.26 in favor of C-CuRE. Thus, comparing models at their optimal p_{guess} still favored the C-CuRE model.

4.5 Further Issues.

There are a few important caveats to the claim that the C-CuRE model offers the best fit. First, we made the simplifying assumption that listeners can identify talkers and vowels perfectly. This is unreasonable, of course, so these data should be interpreted as an upper bound of performance given this type of compensation. We probed the limit of this (Supplementary Note #4) by allowing the model to mis-identify the talker/vowel on some percentage of trials (thus compensating for the *wrong* talker or vowel). We found that until the mis-identification rate reaches 30-35% (for both talker and vowel simultaneously) C-CuRE still outperforms the cue-integration model (details in Supplementary note #4). Moreover, the variation and accuracy across talkers and vowels that only the C-CuRE model displayed was seen at every level of mis-parsing tested.

Similarly, we also examined the assumption that one must identify individual talkers and vowels for successful compensation (Online Supplement, Note #5). We simplified the C-CuRE model to categorize talkers only by gender, and to categorize vowels only in terms of height and backness (independently). These are easier to identify than individual talkers and phonemes, mitigating our assumption of perfect performance. This also reduces the number of parameters in each regression used for compensation to 4 (from 25), yielding a simpler model. This did not substantially affect performance with accuracy between 83.5% and 90.8%.

Finally, the C-CuRE framework is just one approach to compensation, and contrasts from classic approaches that posit purely bottom-up combinations of cues. We examined this in a more limited model that compare specific compound cues that have been proposed in the literature (e.g. locus equations, duration ratios, etc) to the raw versions of these cues with and without C-CuRE (Online Supplement, Note #6). The C-CuRE model outperformed the relative cue model substantially. Given its broader generality (compensation via the same mechanisms can be used with any cue), and the fact that it preserves (rather than discards) fine-grained detail, C-CuRE may be the better approach to compensation.

5.0 General discussion.

5.1 Summary

Our primary question concerned the information in the speech signal that is necessary to support categorization. We collected a corpus of productions that was intended to capture as complete a description as possible for a large sample of fricatives. We measured every cue that had been proposed, and discovered some new ones (LowF, F4 and F5, Supplementary Note #1). Acoustic analysis showed that these cues are heavily context- dependent, but also that there is substantial information for categorization: every cue had some correlation with fricative identity.

This database of measurements is the information available in the signal. We manipulated its quantity and format in the context of a common categorization model, and compared that to listener performance on the same tokens, testing three sets of informational assumptions: 1) those of invariance models, that a small number of raw cues is sufficient; 2) those of exemplar and cue-integration models: that a large number of uncompensated cues is sufficient; and 3) those of compensation models, using cues after effects of context have been compensated for. Compensation was instantiated in the C-CuRE framework, a mechanism that preserves fine-grained acoustic detail, and posits categorization as a basis of compensation.

Only the model using compensated cues yielded listeners' accuracy level and pattern of errors with complete-syllables, and no other model showed the right effects of vowel or talker. This was not due to our simplifying assumptions: the C-CuRE model can cope with mis-identified talkers and vowels, factor out variance with a reduced feature set, and is superior to complex relative cues (Online Notes #4-6). Minimally, this argues for some form of compensation, and it more specifically suggests that C-CuRE is a useful way to implement it. However, given the lack of formal models of compensation, it is possible that other approaches to compensation will offer a similar benefit.

C-CuRE, however, offers a unique description of the difference between the frication-only and complete-syllable conditions. If each portion of the signal subserves decisions about multiple properties (segmental and talker), listeners' differences between these conditions will be due in part to first-order cues in the vocalic portion that directly cue fricatives, but also to their abilities to use this portion to identify the talker and vowel as the basis of compensation of cues in the fricative. This is underscored by the good fit of the cue-integration to the frication-only condition, where the information necessary for compensation with C-CuRE may not be available. It can also explain the beneficial performance for /u/ (relative to the other values) – the coarticulatory effect of /u/ on the frication due to lip-rounding is strong, and identifying this vowel will offer a unique compensatory benefit for cues in the frication that is not seen with the other vowel (in the complete-syllable condition), or when compensation is not possible (via C-CuRE), in the frication-only condition.

5.2 Could uncompensated cues have worked?

There are a number of reasons the cue-integration and invariance models may have failed that do not bear on their informational assumptions. First, did we measure enough cues, or the right ones? Could the cue-integration model have succeeded with better information? We think not. We examined every cue reported by Jongman et al (2000), the most thorough fricative examination to date, and added 10 new ones (some of which were quite useful). We also tried cue combinations that should have been more invariant, with little benefit (Supplementary Note #6). Thus, our corpus did not lack information (although there may still be undiscovered cues for /ɵ, f, v/—even the C-CuRE model underperformed slightly on these).

Second, it is possible that the cues were not scaled properly. The auditory system represents some information nonlinearly (e.g., log scales for duration). We were confronted with many such choices during model development: how to scale the cues (e.g. bark vs. Hz frequency; log vs. linear duration); whether to include polynomial terms; and how to compute residuals (standardized, unstandardized, studentized). We explored many of these options and none affected model performance by more than 1%.

While there may be some yet to-be-discovered cue or transformation that will offer a magic bullet, we doubt it is forthcoming. Rather, the simulations suggest that once we include many redundant sources of information, and particularly when information is coded relative to expectations driven by other categories (e.g. vowel, talker), the details of which specific cues and how they are scaled matter less. It is the redundancy, the context sensitivity, and the statistical structure in the input that does the work, not the details of measuring and coding cues.

Third, perhaps this finding is unique to the statistics of fricatives. It is possible that the statistics of cues associated with other phonemic contrasts may better support categorization. Cole et al (2010)'s, analysis of vowels, suggests that for at least one other class of phonemes C-CuRE offers a decided advantage. However, there is a need for these sorts of comparative informational analyses on other phonological feature, a fruitful (though laborious) undertaking.

Fourth, we didn't model lexical or statistical factors that contribute to perception—perhaps with such things, uncompensated cues would be sufficient. But such factors could not have helped our listeners either: the stimuli were mostly nonwords, they uniformly spanned the space of possible CVs, and each was spoken by each talker. These statistics were thus uniform in our experiment, yet listeners performed better than the cue-integration model predicted.

Finally, perhaps the problem is our categorization model. For example, the mechanisms of categorization proposed by exemplar theory differ substantially from logistic regression, and are potentially more powerful. We cannot rule this out. However, we have experimented with three-layer neural networks which are capable of learning non linearly-separable distributions, and should offer better categorization. This network performed at 87.5% on the discrete-choice rule and 83.5% on the probabilistic rule, better than the cue integration model (85.0% / 78.2%) but less accurate than listeners. Uncompensated cues in the dataset may not support accurate categorization under any categorization model. In fact, we found even better performance when the same network was trained on cues parsed with C-CuRE (90.4% / 88.5%).

Thus the failure of the cue-integration models was not likely due to these simplifications and we are left to conclude that simply using as much information as possible in its raw form may be insufficient to account for listener performance. Interestingly, when this is the only route available to listeners (in the frication-only condition), the cue-integration model succeeds.

5.3 Is this result obvious?

Superficially, these results seem obvious. Of course, when we use many cues, performance improves. Of course, compensation improves categorization. However, this misses several important points. First, our criterion wasn't simply perfect categorization; it was match to listeners. Listeners were not at ceiling, averaging 90% correct—if invariant cues were sufficient, for example, the cue-integration or C-CuRE models could have overshot performance.

Second, our match to listener data was not based on accuracy alone, the effect of context (talker and vowel) was equally important. This was not built into the models (they were not trained on listener data) and there was no *a priori* reason any of them would give rise to such effects. Indeed, as we added cues, moving from the invariance to the cue-integration model, there was little improvement in this regard; it was only when we added compensation that we saw such performance. While one might expect that adding cues or compensation could increase the fit of the model to its training data, there was no reason to expect it to fit better to a completely independent dataset reflecting idiosyncratic performance across context. Thus, our close fit in this regard suggests that these differences across speakers and vowels are not so much a function of the statistical distribution of speech cues within speakers and vowels, but rather of the differential sensitivity of these distributions to compensation.

Third, C-CuRE was not optimized to the goal of identifying the fricative. This component of the model was trained independently of fricative categorization, simply recoding the input as distance from the expected cue-values for that talker and/or vowel. There was no guarantee that this would yield a cue-space better suited to fricative categorization, nor that it would result in the pattern of context effects we observed. In fact, it is surprising that we only see the effects of talker and vowel in the categorization model after we have parsed their effects *out* of the cue-set.

Given these factors, success of the C-CuRE model was by no means a foregone conclusion and its findings should not be dismissed easily.

5.4 Implications for Theories of Speech Perception

Fundamentally, the problem of lack of invariance is a question of information. On this issue, our acoustic analyses confirm for fricatives what most researchers have concluded in general: there is no invariance in the signal (e.g., Ohala, 1996; Lindblom, 1996). Even measuring 24 cues and assessing the effects of context on each, we found no truly invariant cues, and even the best of what we had were not sufficient to match listener categorization performance.

More importantly, however, the lack of invariance is not problematic, and one does not need to go to extremes to surmount it. Motor theory (Liberman & Mattingly, 1985; Liberman & Whalen, 2000) explicitly argues that the only solution to the lack of invariance is to code speech in terms of articulatory gestures. Exemplar theory (e.g. Pierrehumbert, 2001; Hawkins, 2003) makes a similar point: if listeners retain every exemplar they hear in fine-grained detail this can be overcome without compensation. While there are other reasons to argue for these theories, the lack of invariance does not require any specific approach to representation—categorization built on prototypes and acoustic cues can yield listener-like performance as long as many information sources are used with simple compensation schemes. Lack of invariance does not equal lack of information, and does not require a particular solution.

As our emphasis was information, we did not examine the categorization process, and any theory of speech categorization must describe both. Thus, here, we discuss the implications of these findings for a number of theories of speech perception where they are directly relevant.

First, our cue-integration model shares the informational assumptions of ***exemplar theory****:* use every bit of the signal, but without compensation. Clearly, the redundancy in a large cue-set offers advantages in performance and when compensation was not available (lacking the vocalic portion) listeners behave in a way that is consistent with these informational assumptions. However, C-CuRE offered substantially better accuracy (7-8%), and uniquely fit the context effects on performance. This would seem to disfavor exemplar models.

One concern with this is that in exemplar models, categorization may do more of the work. Storing complete exemplars may capture contextual dependencies, making compensation less necessary and enabling better processing based on raw cues. Testing this will require a formal implementation, which raises several issues. First, categorization decisions in exemplar models are made by comparing the incoming input to clouds of stored exemplars. But how many exemplars take part in this? If the input is compared to every exemplar, the model will act like a prototype model as the entire distribution is relevant to categorization and will perform similarly to logistic regression. On the other hand, if the input is only compared to the closest matching exemplar (or a handful of nearby ones), a model would harness more exemplar-like processing to achieve a better decision, if there is a close match. However, when one was not available (e.g. a new talker), it may perform worse. Second, depending on the scope of the exemplars, all types of contextual dependencies may not be captured, for example, coarticulation that crosses a word boundary (Cole et al, 2010). Thus, evaluating exemplar models may require concrete decisions about the categorization rule and exemplar scope.

Second, without oversimplifying the differences, our use of logistic regression closely overlaps with a range of models we termed ***cue integration models,*** models like FLMP (Oden & Massaro, 1978), NAPP (Nearey, 1990, 1997), and HICAT (Smits, 2001a,b). FLMP and NAPP are not strongly committed to any particular form of input (other than cues being continuous and independent), and we see no reason that input parsed with C-CuRE could not be used (though this would introduce feedback which may be incompatible with FLMP: Massaro, 1989, 2000).

Of the cue-integration models, HICAT (Smits, 2001a,b) is closest to our approach, in that a cue's interpretation is conditioned on other decisions (e.g. the vowel). In HICAT, this is embedded (and optimized) as part of the categorization problem, using interaction terms (e.g. F1 x Speaker) in the categorization model. This potentially creates a problem of generalization, as the influence of categories on cues is encoded within a single categorization decision. For example, one would have to learn the influence of a vowel and speaker on $F1_{onset}$ for /s/ decisions separately from the same influences on F1 for /ʃ/ decisions. This may lead to an explosion of such interaction terms. In C-CuRE, on the other hand, context effects are independent of specific categories: rather than conditionalizing the interpretation of cues on context, cues are recoded relative to expectations derived from context, making them available for many processes. This accounts for findings that listeners hear the signal as compensated: for example, hearing a nasalized vowel as more oral if nasalization can be attributed to coarticulation from a nasal consonant (Fowler & Brown, 2000; see also Pardo & Fowler, 1997; Beddor, Harnsberger & Lindemann, 2002). Of course, it is an open question whether cues are *only* encoded in compensated form, and the frication-only models suggest that both raw and compensated cues may need to be available to listeners. Either way, however, by recoding cues the parameters needed for compensation are independent of those needed for categorization (unlike HICAT), which makes model estimation much more tractable. Crucially, our simulations suggest this simpler approach is sufficient to account for listener performance in this corpus.

Cue-integration models like NAPP and HICAT have framed debates over *cue-sharing*, situations like the one studied here where a single cue is affected by multiple factors

(Mermelstein, 1978; Whalen, 1989, 1992; Nearey, 1990, 1992; Smits, 2001a) and model fit to complex perceptual datasets has been an important tool for comparing hypotheses. Nearey (1990) has argued that compensation effects on fricative identity can be accounted for without category→cue relationships by assuming listeners are simply biased toward particular pairs of phonemes, while Whalen (1992) and Smits (2001a) argue that fricative categorization is dependent on how the vowel is categorized. Our analysis supports the latter view, but using stimuli that capture the natural statistical distribution of cue-values (clustered), and a richer information source. The generality of the C-CuRE compensation mechanism, however, extends this by suggesting that phoneme categorization may also be contingent talker identity (cf., Strand, 1999; Nygaard et al, 1994) and thus offers a more unified account.

Finally, in the last few years, a number of **statistical learning** accounts of speech perception have emerged (de Boer & Kuhl, 1997; McMurray, et al, 2009; Vallabha et al, 2007; Toscano & McMurray, 2010; Feldman, Griffiths & Morgan, 2009). Our logistic model was not meant to advocate for any particular categorization framework, nor do we make strong claims about learning. However, it shares with statistical approaches the intuition that the statistical structure of the input is fundamental to categorization and it represents a powerful proof of concept by demonstrating that speech perception may in principle be learnable from the input, and that fairly complex variation in listener performance (e.g. the effect across talkers and vowels) can be derived largely from the information in input.

Ultimately, however, statistics (and hence, information) will not be sufficient to fully describe perception, we must also consider processing. Herein lies a limitation of our implementation of the ideas proposed here: in this specific domain, how do listeners identify the vowel to compensate during fricative perception, when vowel identification may also benefit from knowing the fricative? We suggest that listeners must simultaneously and interactively identify the talker, vowel and fricative. While these factors are identified in parallel, the cues for each may be available at different times, meaning that at some points in processing (e.g. before the vowel arrives) listeners may rely on an approach closer to our cue-integration approach, while once these contextual sources of information are available, they may be able to revise their initial decisions. This favors an approach more akin to interactive activation (e.g. Elman & McClelland, 1986), where a partial decision about talker or vowel could be used to parse cues to the fricative, increasing confidence in the fricative decision; while simultaneously, partial decisions about the fricative can be used to parse cues to the vowel (see also, Smits' [2001b], fuzzy parallel version of HICAT). Over time, the system gradually settles on a complete parse of all three factors without ever making a discrete decision about any one. Ultimately, though, understanding such mechanisms will require detailed analyses of the timecourse of processing (e.g. McMurray, Clayards, Tanenhaus & Aslin, 2008b) and more dynamic models of perception.

5.5 Computing Cues Relative to Expectations

The C-CuRE approach builds on parsing approaches which have historically been associated with gestural or acoustic accounts (Fowler, 1984; Gow, 2003). In contrast, our work shows that such operations do not require a particular representational form (cf., Ohala, 1981; McMurray et al, in press). C-CuRE also builds on auditory contrast accounts (Lotto & Kluender, 1998, Holt, 2006; Kluender, Coady & Kiefte, 2003) by proposing that cues are interpreted relative to expectations, though these expectations can be driven by categories (perhaps in addition to lower-level expectations). This generality allows a simple implementation using linear regression to partial out both articulatory (vowel) and non-articulatory (talker) factors as the difference between expected and actual cue values.

When compared against other ways of relativizing cues, C-CuRE has several advantages. By relying on remembered prototype values it avoids having to wait to accumulate information. For example, relativizing frication duration on the basis of vowel duration means that listeners must wait until the end of the vowel to identify the fricative. In fact, recent work (McMurray, et al, 2008b) on asynchronous cues to voicing suggests listeners do not do this. C-CuRE also does not require a lifetime of phonetic studies to determine relativizing relationships for each cue, and it can be used equally well with any cue. It also is consistent with prototype and statistical accounts of phonetic categories since in order to parse out the effect of a category on a cue you must know its mean and variance (McMurray & Farris-Trimble, in press).

Finally, like exemplar accounts, C-CuRE stresses the importance of fine-grained, continuous detail including indexical information, and the fact that it must be retained and used for multiple decisions during perception. Thus, it is not vulnerable to critiques leveled at normalization models (e.g. Pisoni, 1997). Finally, also like exemplar accounts, we stress the importance of indexical cues, while positing a different role for them. Rather than simply lumping indexical information in with phonetic cues, indexical cues are used to identify talkers, and that in turn is used to interpret cues signaling phonetic contrasts.

5.6 Conclusions

Speech categorization fundamentally requires massive cue-integration, but categorization must be performed at the same time as compensatory mechanisms that cope with contextual influences. When we approach categorization in this richer framework, many problems appear easier. While studies of small numbers of cues are valuable for exploring which cues are used (e.g. Summerfield, 1981; Massaro & Cohen, 1976) and answering theoretical questions (e.g. Pisoni & Tash, 1974; Miller & Volaitis, 1989), they may also oversimplify issues and exaggerate problems (e.g., Shinn, Blumstein & Jongman 1985).

Massively redundant information is the norm in speech categorization, but at the same time, cue-sharing happens everywhere and compensation using information from other types of categories is needed to cope with it. That is, categorization and compensation mechanisms may be deeply intertwined, challenging the conception that compensation occurs autonomously and pre-categorically. This has not been extensively explored outside of speech but may be crucial for understanding domains in which the information that supports categorization is variable and context-dependent, domains like face perception, color perception and even abstract category systems like syntactic categories (e.g. Monaghan, Christensen & Chater, 2005).

C-CuRE suggests important interactions between categorization and the encoding of perceptual cues. However, it is not the only such interaction that has been proposed. Categorical perception (Liberman et al, 1957), for example, implies that cue-encoding is accomplished in terms of categories. Categorical perception has not held up to empirical scrutiny in speech (Massaro & Cohen, 1983; Schouten, et al., 2003; Toscano et al, 2010), largely due to evidence that fine-grained detail is retained. C-CuRE suggests a more interesting way in which categories may affect the encoding of continuous cues, one that preserves continuous detail by recoding cues relative to expectations derived from categories. Thus, understanding compensation in perception may require us to understand higher-level processes like categorization, object recognition, and scene organization, and vice versa. More importantly, the generality of mechanisms like C-CuRE suggests that debates over representation may be of less importance in understanding categorization than debates over process: when the information is right, the framework for categorization may matter less than the content it works on.

Acknowledgements

The authors would like to thank Yue Wang for help with the perceptual experiment, Dan McEchron for assistance coding the additional cues and data management, and especially Jennifer Cole in helping to lay the groundwork for the parsing / C-CuRE approach. We'd also like to thank Terry Nearey, Bruce Hayes, Ben Munson, Gregg Oden, and Kristian Markon for helpful feedback on some of the ideas presented here. This research is supported by NIH DC-008089 to BM and DC-02537 to AJ.

References

Andruski, J.E., Blumstein, S.E. & Burton, M.W. (1994) The effect of subphonetic differences on lexical access. *Cognition, 52,* 163-187.

Ashby, G., & Perrin, N. (1988) Toward a unified theory of similarity and recognition. *Psychological Review, 95(1),* 124-150.

Balise, R.R., & Diehl, R.L. (1994) Some distributional facts about fricatives and a perceptual explanation. *Phonetica, 51,* 99-110.

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,* 1073–1077.

Beale, J.M. and Keil, F.C. (1995) Categorical effects in the perception of faces. *Cognition, 57(3),* 217-239.

Beddor, P.S., Harnsberger & Lindemann, S. (2002) Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics, 30,* 591-627.

Behrens, S. J., & Blumstein, S. E. (1988) Acoustic characteristics of English voiceless fricatives: A descriptive analysis. *Journal of Phonetics, 16,* 295–298.

Biederman, I. (1995). Visual object recognition. In S. F. Kosslyn and D. N. Osherson (Eds.). *An Invitation to Cognitive Science, 2nd edition, Volume 2, Visual Cognition.* MIT Press. Chapter 4, 121-165.

Blumstein, S.E. and Stevens, K.N. (1980) Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67(2), 648*-662.

Blumstein, S.E. and Stevens, K.N. (1981) Phonetic features and acoustic invariance in speech. *Cognition, 10,* 25-32.

Boersma, P. & Weenink, D. (2009): Praat: doing phonetics by computer (Version 5.1.38) [Computer program]. Retrieved June 1, 2009, from http://www.praat.org/

Bornstein, M.H. and Korda, N.O. (1984) Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research, 46(3),* 207-222.

Bradlow, A.R., Nygaard, L.C. and Pisoni, D.B. (1999) Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61(2),* 206-219.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Cole, J.S., Linebaugh, G., Munson, C., and McMurray, B. (2010) Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics, 38(2), 167*-184.

Connine, C. (2004) It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin & Review. 11(6)*, 1084-1089.

Connine, C., Ranbom, L., & Patterson, D. (2008) Processing variant forms in spoken word recognition: The role of variant frequency. *Perception & Psychophysics, 70(3)*, 403-411.

Creel, S. Aslin, R.N., & Tanenhaus, M.K. (2008) Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106(2),* 633-664.

Cronbach, L. J. (1987). Statistical tests for moderator variables: flaws in analyses recently proposed. *Psychological Bulletin, 102,* 414–417.

Crystal, T. & House, A. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America,* 83, 1553-1573.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16,* 507–534.

de Boer, B. & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustic Research Letters Online, 4,* 129-134.

Delattre, P., Liberman, A.M. & Cooper, F.S. (1955) Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27,* 769-773.

Elman, J. & McClelland, J. (1986) Exploiting Lawful Variability in the Speech Wave. In J. S. Perkell & D. Klatt (Eds.*) Invariance and Variability in Speech Processes* (pp. 360-380). Hillsdale, NJ: Erlbaum

Ernst, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429-433.

Feldman, N.H., Griffiths, T.L., & Morgan, J.L. (2009) The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116(4),* 752-782.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America, 84,* 115–124.

Fowler, C. (1984) Segmentation of coarticulated speech in perception. *Perception & Psychophysics, 36,* 359-368.

Fowler, C. A. (1994) Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics, 55,* 597–611.

Fowler, C. (1996) Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99(3),* 1730-1741.

Fowler, C., & Brown, J. (2000) Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception & Psychophysics, 62(1),* 21-32.

Fowler, C., & Smith, M. (1986) Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell, & D. Klatt (eds.) *Invariance and variability in speech processes* (pp. 123-136). Hillsdale, NJ: Erlbaum.

Ganong, W.F. (1980) Phonetic Categorization in Auditory Word Recognition. *Journal of Experimental Psychology: Human Perception and Performance, 6(1),* 110-125.

Gibson, J.J. (1966) *The Senses Considered as Perceptual Systems.* Boston, MA: Houghton Mifflin.

Goldinger, S.D. (1998) Echoes of Echos? An episodic theory of lexical access. *Psychological Review, 105,* 251-279.

Goldstone, R.L. (1995). The effects of categorization on color perception. *Psychological Science, 6(5),* 298-304.

Goldstone, R. L., & Kersten, A. (2003). Concepts and Categories. In A. F. Healy & R. W. Proctor (Eds.) *Comprehensive handbook of psychology, Volume 4: Experimental psychology (pp. 591-621).* New York: Wiley.

Goldstone, R.L., Lippa, Y. & Shiffrin, R.M. (2001) Altering object representations through category learning. *Cognition, 78,* 27-43.

Gow, D. (2001) Assimilation and Anticipation in continuous spoken word recognition. *Journal of Memory and Language, 45,* 133-139.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics,* 65, 575-590

Hansen, T., Olkkonen, M., Walter, S. & Gegenfurtner, K.R. (2006) Memory modulates color appearance. *Nature Neuroscience, 9(11),* 1367-1368.

Hawkins, S. (2003) Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31,* 373-405.

Hillenbrand, J.M., & Houde, R.A. (2003). A narrow band pattern-matching model of vowel perception, *Journal of the Acoustical Society of America, 113,* 1044-1055.

Hintzman, D. L. (1986). Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411-428.

Holt, L. (2006). The mean matters: Effects of statistically-defined non-speech spectral distributions on speech categorization. *Journal of the Acoustical Society of America*, *120*, 2801-2817.

Holt, L. & Lotto, A. (2010) Speech perception as categorization. *Attention, Perception & Psychophysics, 72,* 1218-1227.

Homa, D, & Cultice, J. (1984) Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(1),* 83-94

Honorof, D. & Whalen, D. (2005) Perception of pitch location within a speaker's F0 range. *Journal of the Acoustical Society of America, 117(4),* 2193-2200.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression, 2nd Edition.* New York: John Wiley & Sons, Inc.

Huette, S., & McMurray, B. (2010) Continuous dynamics of color categorization. *Psychonomic Bulletin & Review, 17(3),* 348-354.

Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants, *Journal of the Acoustical Society of America, 28,* 303–310.

Jacobs, R.A. (2002). What determines visual cue reliability? *Trends in Cognitive Science, 6,* 345-350.

Johnson, K. (1997) Speech perception without speaker normalization: An exemplar model. K Johnson - Talker variability in speech processing. In K. Johnson and J.W. Mullenix (Eds) *Talker Variability in Speech Perception (pp. 145-166).* New York: Academic Press.

Johnson, K. & Mullenix, J.(1997) *Talker Variability in Speech Perception.* New York: Academic Press.

Johnson K.C., Strand E.A. & D'Imperio M. (1999) Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *24*, 359-384.

Jongman, A. (1989). Duration of fricative noise required for identification of English fricatives. *Journal of the Acoustical Society of America, 85*, 1718-1725.

Jongman, A., Wang, Y., & Sereno, J.A. (2000). Acoustic and perceptual properties of English fricatives. *Proceedings of the 6th International Conference on Spoken Language Processing* , *I*, 536-539.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives, *Journal of the Acoustical Society of America, 106*, 1252-1263.

Kewley-Port, D. & Luce, P.A. (1984) Time-varying features of initial stop consonants in auditory running spectra - a 1st report. *Perception & Psychophysics, 35(4),* 353-360.

Kluender, K., Coady, J., & Kiefte, M. (2003) Sensitivity to change in perception of speech. *Speech Communication,* 41(1), 59-69

Kulikowski, J.J. & Vaitkevicius, H. (1997) Colour constancy as a function of hue. *Acta Psychologica, 97,* 25-35

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, 391-404.

LaRiviere, C., Winitz, H., & Herriman, E. (1975). The distribution of perceptual cues in English prevocalic fricatives. *Journal of Speech and Hearing Research, 18*, 613-622.

Lee, C.Y**.**, Dutton, L., & Ram, G. (2010). The role of speaker gender identification in F0 height estimation from multispeaker, brief speech segments. *Journal of the Acoustical Society of America, 128,* 384-388,

Liberman, A.M., Harris, K.S., Hoffman, H.S. and Griffith, B.C. (1957) The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54(5),* 358-368.

Liberman, A.M. & I.G. Mattingly (1985) The motor theory of speech perception revised. *Cognition*, 21, 1-36.

Liberman, A.M., & Whalen, D.H. (2000) On the relation of speech to language. *Trends in Cognitive Sciences, 4,* 187-196.

Lindblom, B. (1996) Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, 99(3), 1683-1692.

Lipsitz, S., Kim, K., & Zhao, L. (2006). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine, 13,* 1149-1163.

Lotto, A.J., & Kluender, K.R. (1998) General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60,* 602-619.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Maddox, W.T., Molis, M. & Diehl, R. (2002) Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception & Psychophysics, 64(4), 584-597.*

Maniwa, K., Jongman, A. & Wade, T. (2008). Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *Journal of the Acoustical Society of America, 123,* 1114-1125.

Maniwa, K., Jongman, A. & Wade, T. (2009) Acoustic characteristics of clearly spoken English fricatives. *Journal of the Acoustical Society of America, 126(6),* 3962-3973.

Mann, V.A. and Repp, B. (1981) Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69(2),* 548-558.

Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York: Freeman.

Marslen-Wilson, W. D. & Warren, P. (1994) Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101,* 653–75.

Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America, 69(2),* 559-567.

Martin, J. G., & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance, 8(3),* 473-488.

Massaro, D.W. (1989) Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21(3),* 398-421.

Massaro, D.W. (2000) The horse race to language understanding: FLMP was first out of the gate, and has yet to be overtaken. *Behavioral and Brain Sciences, 23(3),* 338-339.

Massaro D. W. & Cohen, M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/- /si/ distinction. *Journal of the Acoustical Society of America*, *60*, 704-717.

Massaro, D.W. and Cohen, M.M. (1983) Categorical or continuous speech perception: a new test. *Speech Communication, 2,* 15-35.

Maye, J., Werker, J.F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111

McCann, J.J., McKee, S.P., & Taylor, T.H. (1976). Quantitative studies in retinex theory. A comparison between theoretical predictions and observer responses to the "color Mondrian" experiments. *Vision Research, 16(5),* 445-458.

McLennan, C. and Luce, P.A. (2005) Examining the Time Course of Indexical Specificity Effects in Spoken Word Recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31(2),* 306-321.

McMurray, B., Aslin, R., Tanenhaus, M., Spivey, M., & Subik, D. (2008a). Gradient sensitivity to within-category variation in speech: Implications for categorical perception. *Journal of Experimental Psychology, Human Perception and Performance, 34(6),* 1609-1631

McMurray, B., Aslin, R.N., and Toscano, J. (2009a) Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science, 12(3),* 369-379.

McMurray, B., Clayards, M., Tanenhaus, M., & Aslin, R. (2008b) Tracking the timecourse of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin and Review, 15(6),* 1064-1071.

McMurray, B., Cole, J.S., & Munson, C. (in press) Features as an emergent product of computing perceptual cues relative to expectations. In R.Ridouane and N. Clement (Eds) *Where do Features Come From?*

McMurray, B., Dennhardt, J., and Struck-Marcell, A. (2008c) Context effects on musical chord categorization: Different forms of top-down feedback in speech and music? *Cognitive Science, 32(5),* 893 – 920.

McMurray, B., Tanenhaus, M., and Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access, *Cognition, 86(2),* B33-B42.

McMurray, B., Tanenhaus, M.K., and Aslin, R.N. (2009b) Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60(1),* 65-91.

McMurray, B., & Farris-Trimble, A. (in press) Emergent, yet unintended, coupling of perception and production: General processing principles and statistical learning. Invited submission to A. Cohn, C. Fougeron, and M. Huffman (Eds) *The Oxford Handbook of Laboratory Phonology,* Oxford, UK: Oxford University Press.

Medin, D. L., Lynch. E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology, 51,* 121–147.

Medin, D. L., & Schaffer, M. M. (1978) Context theory of classification learning. *Psychological Review, 85*, 207-238.

Mermelstein, P. (1978) On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception & Psychophysics, 23,* 331-336.

Mertus, J. (1989) *BLISS Manual*. Brown University, Providence, R.I..

Miller, J.L. (1997) Internal structure of phonetic categories. *Language and Cognitive Processes, 12,* 865-869.

Miller, J.L., & Volaitis, L.E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*, 505-512.

Mitterer, H. & de Ruiter, J. P. (2008). Recalibrating color categories using world knowledge *Psychological Science, 19,* 629-634

Munson, B. (2007). The acoustic correlates of perceived sexual orientation, perceived masculinity, and perceived femininity. *Language and Speech*, 50, 125-142.

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006) The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics, 34,* 202-240.

Nearey, T.M. (1978). *Phonetic feature systems for vowels.* Bloomington: Indiana University Linguistics Club.

Nearey, T.M. (1990). The segment as a unit of speech perception. *Journal of Phonetics, 18,* 347–373.

Nearey, T.M. (1992) Context effects in a double-weak theory of speech perception. *Language and Speech, 35(1-2),* 153-171.

Nearey, T.M. (1997) Speech perception as pattern recognition. *Journal of the Acoustical Society of America, 101(6),* 3241-3254.

Nearey, T. M. & Hogan, J. (1986). Phonological contrast in experimental phonetics: relating distributions of measurements in production data to perceptual categorization curves. In J. Ohala, & J. Jaeger (Eds.), *Experimental Phonology,* Academic Press: New York.

Nissen, S.L., & Fox, R.A. (2005) Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective, *Journal of the Acoustical Society of America, 118,* 2570-2578.

Nosofsky, R. (1986) Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115(1),* 39-57.

Nygaard, L., Sommers, M. & Pisoni, D. (1994) Speech perception as a talker contingent process. *Psychological Science, 5,* 42-46.

Oden. G. (1978) Integration of Place and Voicing Information in the Identification of Synthetic Stop Consonants, *Journal of Phonetics, 6(2),* 82-93.

Oden, G. & Massaro, D.W. (1978) Integration of featural information in speech perception. Psychological Review, 85(3), 172-191.

Ohala, J.J. (1981) The listener as a source of sound change. In: C. S. Masek, R. A. Hendrick, & M. F. Miller (eds.), *Papers from the Parasession on Language and Behavior* (pp. 178 - 203) Chicago: Chicago Linguistics Society.

Ohala, J. (1996) Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America, 99(3),* 1718-1725.

Öhman , S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements, *Journal of the Acoustical Society of America, 39,* 151–168.

Oliva, A., & Schyns, P. (1997) Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34(1),* 72-107.

Palmeri, T., Goldinger, S., and Pisoni, D. (1993) Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology: Learning Memory and Cognition, 19(2),* 309-328.

Pardo, J.S., & Fowler, Carol A. (1997) Perceiving the causes of coarticulatory acoustic variation: consonant voicing and vowel pitch. *Perception & Psychophysics, 59(7),* 1141-1152.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure.* Amsterdam: John Benjamins.

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46,* 115-154.

Pind, J. (1995). Speaking rate, VOT and quantity: The search for higher-order invariants for two Icelandic speech cues. *Perception & Psychophysics, 57,* 291–304.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson and J.W. Mullennix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press, 1997, pp. 9-32.

Pisoni, D.B. & Sawusch, J.R. (1974) On the identification of place and voicing features in synthetic stop consonants. *Journal of Phonetics, 2,* 181-194.

Pisoni, D.B. & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*, 285–290.

Port R.F. (2007) How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology, 25,* 143–170

Port, R.F. and Dalby, J. (1982) Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics, 32(2),* 141-152.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Reed, S. (1972) Pattern recognition and categorization. *Cognitive Psychology, 3(3),* 382-40.

Roberson, D. & Davidoff, J. (2000) The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory and Cognition, 28(6),* 977-986.

Roberson, D., Hanley, J. R., & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition, 112(3),* 482-487.

Rosner, B. & Pickering, J. (1994) *Vowel perception and production.* Oxford, UK: University Press.

Salverda, A.P., Dahan, D. and McQueen, J. (2003) The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90(1),* 51-89.

Schouten, M.E.H., Gerrits, E. and Van Hessen, A.J. (2003) The end of categorical perception as we know it. *Speech Communication, 41(1),* 71-80.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Schyns, P.G. & Rodet, L. (1997) Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23(3), 681-696.*

Shinn, P.C., Blumstein, S.E. & Jongman, A. (1985). Limitations of context-conditioned effects on the perception of [b] and [w]. *Perception & Psychophysics, 38***,** 397-407.

Smits, R. (2001a). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 1145-1162.

Smits, R. (2001b). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception & Psychophysics, 63,* 1109-1139.

Smits, R., Jongman, A., & Sereno, J. (2006) Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance, 32(3),* 733-754.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative–vowel coarticulation, *Journal of the Acoustical Society of America, 70,* 976–984.

Soto, F., and Wasserman, E. (2010) Error-driven learning in visual categorization and object recognition: A common-elements model. *Psychological Review, 117(2),* 349-381.

Stevens, K.N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111,* 1872-1891.

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.

Stevens, K.N., Blumstein, S.E., Glicksman, L., Burton, M. & Kurowski, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America, 91,* 2979-3000.

Stoel-Gammon, C., Williams, K., & Buder, E. H. (1994). Cross-language differences in phonological acquisition: Swedish and American /t/. *Phonetica, 51,* 146–158.

Stevens, K.N. & Keyser, S.J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics, 38(1),* 10-19.

Strand, E. (1999) Uncovering the Role of Gender Stereotypes in Speech Perception. *Journal of Language and Social Psychology, 18,* 86-100.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech, 3,* 32–49.

Summerfield, Q. (1981) Articulatory rate and perceptual constancy in phonetic perception. *Journal of the Acoustical Society of America, 7(5),* 1074-1095.

Sussman, H., Fruchter, D., Hilbert, J. & Sirosh, J. (1998) Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Science, 21,* 241-299.

Sussman, H. & Shore, J. (1996) Locus equations as phonetic descriptors of consonantal place of articulation. *Perception & Psychophysics, 58(8),* 936-946.

Tomiak, G. (1990) *An acoustic and perceptual analysis of the spectral moments invariant with voiceless fricative obstruents*. Doctoral dissertation, SUNY Buffalo.

Toscano, J. & McMurray, B. (2010) Cue Integration with Categories: A Statistical Approach to Cue Weighting and Combination in Speech Perception. *Cognitive Science, 34(3),* 436-464.

Toscano, J., McMurray, B., Dennhardt, J., & Luck, S. (2010) Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science, 21(10),* 1532-1540.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin and Review, 16,* 74-79.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147-162.

Whalen, D. (1981). Effects of vocalic formant transitions and vowel quality on the English /s-ʃ/ boundary. *Journal of the Acoustical Society of America, 69*, 275-282.

Whalen, D. (1989) Vowel and consonant judgments are not independent when cued by the same information. *Perception & Psychophysics, 46,* 284-292.

Whalen, D. (1992) Perception of overlapping segments: Thoughts on Nearey's model. *Journal of Phonetics, 20,* 493-496.

You, H-Y. (1979) *An acoustic and perceptual study of English fricatives*. M.Sc. thesis, University of Alberta.

Appendix A: Measurement and Data Processing of the Corpus of Cue Values

This appendix describes each of the individual cues that were measured in the corpus of fricatives and how they were measured. JWW refers to the original Jongman et al (2000) study.

Peak frequency was taken from JWW and measured from a 40 ms window at the center of the frication noise. It is the frequency of the highest-amplitude peak of the FFT spectrum.

Frication Duration and *Vowel Duration* were also obtained from JWW and measured from zero-crossings. Fricative onset was the first point at which high-frequency energy appeared in the spectrogram. Fricative offset/vowel onset was marked at the intensity minimum prior to the onset of periodic voicing energy for voiceless fricatives; and as the earliest period at which the waveform changed substantially (with respect to the frication) for voiced fricatives. Vowel offset was identified as the onset of the closure portion of the /p/.

Frication RMS amplitude and *Vowel RMS amplitude* were obtained from JWW, and measured by computing the RMS amplitude in dB for the entire frication as well as three consecutive pitch periods at the point of maximum vowel amplitude, respectively.

Spectral mean, variance, skewness, and *kurtosis* were JWW measurements and computed from spectra obtained from three 40 ms Hamming windows centered at the onset, midpoint and end of the frication. Spectra were based on a linear frequency scale as Jongman et al. (2000) reported little difference when values were derived from bark-scaled frequencies.

Transition moments were also derived in the same way from a window that included the last 20 ms of the frication and the first 20 ms of the vowel.

All spectral moment data were taken directly from the JWW database, with three modifications. First, Jongman et al. (2000) reported the second moment as spectral *variance* which can have very high values. We converted the second moment to standard deviations by taking the square root of each value (see Stoel-Gammon et al., 1994). Second, Window-3 (the last 40 ms of the frication) was removed from the analysis because it overlapped with Window-4 (which included the final 20 ms of the frication and the first 20 ms of the vowel) and would hence violate the independence assumptions of most statistical tests. Third, the moments in Windows-1 and -2 were highly correlated, particularly for the first two moments (M1: R=.85; M2: R=.82; M3: .63; M4: .43). This made it difficult for some of the models to converge. Therefore, moments in these two windows were averaged to create estimates of spectral mean, variance, skew and kurtosis that spanned the two windows.

We also measured the following new cues, using a combination of the Praat speech analysis software (Boersma and Weenink, 2009) and several custom Matlab scripts.

Low Frequency Energy was included as a potential measure of voicing during the frication. A spectrum over the entire frication noise was computed, and the average amplitude of the components below 500 Hz was measured.

Formant Frequencies. The frequency of the first five formants over the first 23.3 ms of vowel onset was measured in two stages. First, frequencies were automatically extracted for all files using the Burg algorithm method with two different parameter-sets (one selected for men and one for women). Next, a trained phonetic coder viewed plots of both formant tracks on top of the corresponding spectrograms and determined if either of the automatically coded tracks was correct. If not, formant frequency values were entered by hand from the spectrogram.

Fundamental frequency (F0) was computed for the first 46.6 ms of each vowel.

Narrow-band amplitude. This is a modification of the relative amplitude measure reported by JWW. In their paper, relative amplitude was computed in two stages. First, the amplitude of F3 at vowel onset for sibilants (/s, z, ʃ, ʒ/), and of F5 for non-sibilants (/f, v, ɵ, ð/) was measured using a discrete Fourier transform over a 23.3 ms window. Second, a spectrum was derived over the middle 23.3 ms of the fricative and the amplitude of the frequency component closest to the F3 or F5 values was obtained. Relative amplitude was then the difference between fricative amplitude and vowel amplitude.

While this is an excellent cue to place of articulation, we were concerned that this cue was measured differently depending on sibilance. In the vocalic portion, the amplitude in the F3 region (used for sibilants) is almost certainly greater than in the F5 region (used for non-sibilants). Thus this cue could artificially distinguish sibilants from non-sibilants. To avoid this, we measured both F3 and F5 amplitude for all fricatives and treated them as two separate cues.

Jongman et al. (2000) relativized this measure by subtracting the amplitude in the fricative from that of the vowel. We chose not to do this for two reasons. First, several of the analyses were intended to examine the cues in the frication noise alone and it was unclear whether such cues should be counted as frication cues or vowel cues – there is clearly amplitude information in the fricative alone even if it cannot be relativized against the vowel. Second, and more importantly, we wanted our models to use first-order cues (e.g. without normalization). Thus, it made sense to treat these as four independent cues: the amplitude in the F3 and F5 regions for the frication and for the vocalic portion. We refer to these measurements as *narrow-band amplitudes*.

Finally, some of the cues (spectral mean and variance, in particular) were large in real valued terms (spectral mean averaged 5879 Hz, and standard deviation averaged 2121 Hz in window 1. Including these with values in the 0-1 range (e.g. duration in seconds) posed a problem for fitting the logistic models used in Sections 4-8. Thus, prior to analysis, all variables were converted to Z-scores (relative to the overall mean and standard deviation), a form of centering that is common in regression and other generalized linear models (Cronbach, 1987).

Notes

[1] There is a long history of empirical work showing striking commonalities between speech perception and other domains of perceptual categorization. Most famously, categorical perception is seen in the perception of speech (Liberman, Harris, Hoffman & Griffith, 1957), color (Bornstein & Korda, 1984), and faces (Beale & Keil, 1995) (to name a few); and the later refutations of categorical perception (e.g. Schouten, Gerrits & Van Hessen, 2003; Toscano, McMurray, Dennhardt & Luck, 2010) have also been observed in color (Roberson, Hanley & Pak, 2009; Roberson & Davidoff, 2000) and faces (Roberson & Davidoff, 2000). Prototype effects are seen in both dot patterns (Posner & Keele, 1968) and speech categories (Miller & 1997) (among many other domains). Effects of top-down expectations can be observed in color categorization (Mitterer & de Ruiter, 2008), and speech (Ganong, 1980), though they may work differently in music (McMurray, Dennhardt & Struck-Marcel, 2008c). Finally, evidence for a graded competition between categories can be seen in both speech (McMurray, Aslin, Tanenhaus, Spivey & Subik, 2008a) and color perception (Huette & McMurray, 2009). Perhaps most importantly, many of the models of categorization in speech rely on similar principles to categorization in non-speech, principles like exemplar encoding (Goldinger, 1998), or prototypes (Miller, 1997), and many of the same models have been applied to both speech and non-speech problems (e.g. Oden & Massaro, 1978; Goldinger, 1998).

[2] Such accounts did not imply that such factors were eliminated from every level of encoding; rather that they are stripped away from encodings used during phonetic categorization. Indexical variation, for example, would be posited to be eliminated in the representations that support phonological categorization while still being available to support speaker identification.

[3] The exception to this is the second formant frequency (F2) which was remeasured in order to ensure consistency between it and the other four formants in terms of the procedure and the measurer.

[4] Dummy coding is a standard technique in regression (Cohen & Cohen, 1983) in which an independent factor with multiple levels (e.g. talker) is recoded into several variables. Each of the N-1 levels of the factor is given a single variable which is coded 1.0 if the current data-point has that level and 0 otherwise. These variables are then entered as a group into the regression.

[5] Multinomial logistic regression generalizes the binary form of logistic regression by implementing a series of binary comparisons between each category and the reference category—so a common reference category is required. The choice of which outcome serves as the reference category is arbitrary and will make no difference in the resulting probability predictions (Hosmer & Lemshow, 2000).

[6] This is how classification tables in statistics packages like SPSS are constructed.

[7] Parsing regressions were run entering speaker codes first and then vowel. However, this choice does not affect the residuals as the ultimate regression equation (using all of the terms) is the same regardless of order of entry. However, during online perception this may matter as certain factors may not be available at every time point – for example, the presence of a carrier sentence prior to the fricative may make speaker available for parsing before the vowel (see McMurray, et al., in press, for a discussion).

[8] We were initially surprised at the failure of F2 to participate in the categorization model given the wealth of studies positing a role for either F2 at onset or F2 locus equations. We can think of two reasons for this. First,

many of these studies have only examined sibilants (and voiceless sibilants at that)—the utility of this cue for sibilants may not have been sufficient to reach significance given the other meaningful contrasts and the redundant information in the signal like the other formants (F3 in particular). Second, Nearey (personal communication) suggested that F2 should show a quadratic relationship to place of articulation. We thus ran a second model including both linear and quadratic terms for each formant. Model performance was increased by .8% (85.8% vs. 85.0%), but not enough to outweigh the penalty of the additional parameters ($\text{BIC}_{\text{quadratics}}$: 3453 vs. $\text{BIC}_{\text{linear:}}$ 3381 without). However in the quadratic model, both the linear and quadratic terms for F2 were now significant (L: $\chi^2(7)=25.6$, p=.01; Q: $\chi^2(7)=120.1$, p<.0001), and significant quadratic effects were also observed for F4 ($\chi^2(7)=18.4$, p<.0001) and F5 ($\chi^2(7)=15.1$ p=.035). Given the higher BIC, however, and the complexities of using quadratic terms in the parsing model, we used only the linear term in this and subsequent models.

[9] This may have been due to cognitive factors that were not modeled. The fact that /ð/ generally only appears word-initially in function words and has the same orthographic representation as /ɵ/ may have led listeners to select /ð/ less than other responses.