Method for the location of burst-onset spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants

Allard Jongman[a] and James D. Miller
Central Institute for the Deaf, St. Louis, Missouri 63110

(Received 13 December 1989; accepted for publication 6 September 1990)

A method for distinguishing burst onsets of voiceless stop consonants in terms of place of articulation is described. Four speakers produced the voiceless stops in word-initial position in six vowel contexts. A metric was devised to extract the characteristic burst-friction components at burst onset. The burst-friction components, derived from the metric as sensory formants, were then transformed into log frequency ratios and plotted as points in an auditory-perceptual space (APS). In the APS, each place of articulation was seen to be associated with a distinct region, or target zone. The metric was then applied to a test set of words with voiceless stops preceding ten different vowel contexts as produced by eight new speakers. The present method of analyzing voiceless stops in English enabled us to distinguish place of articulation in these new stimuli with 70% accuracy.

PACS numbers: 43.71.An, 43.71.Cq, 43.71.Es, 43.66.Ba

INTRODUCTION

Over the past 40 years, much research has been devoted to the question of whether distinct spectral patterns that correspond to phonetic dimensions, such as place and manner of articulation, can be derived from the acoustic waveform. While early research failed to find any consistent mapping between acoustic properties and phonetic features (e.g., Cooper *et al.*, 1952; Schatz, 1954; Delattre *et al.*, 1955), more recent research suggests that stable properties corresponding to phonetic features may indeed be found in the speech signal, provided that the signal is analyzed in the appropriate way.

Much of this research has focused on the search for properties distinguishing place of articulation in American English stop consonants. The acoustic theory of speech production (Fant, 1960) predicts that such properties can be derived from the short-time spectrum sampled at consonantal release. In this view, the acoustic information residing in the burst and approximately 20 ms of formant transitions combine into a single integrated property for place of articulation. For example, Blumstein and Stevens (1979) found that the gross spectral shape derived from the first 25 ms of a stop consonant provided unique and invariant information about its place of articulation. Their research was thus based on static properties of the speech signal in that a spectral shape was derived over a single 25-ms window. However, dynamic time-varying properties of the speech signal seem to provide a more reliable cue to place of articulation (Searle *et al.*, 1979, 1980; Kewley-Port, 1983; Lahiri *et al.*, 1984). Instead of deriving one spectral shape encompassing all information present within a single 25-ms window, changes in spectral properties over time were postulated as invariant cues to place of articulation in stop consonants.

The present study attempts to derive stable properties corresponding to place of articulation for stop consonants within the auditory-perceptual theory (APT) as motivated and described in detail by Miller (1987, 1989). In this approach, the spectral shape is assumed to be highly correlated with the locations of the "significant spectral prominences." Furthermore, a method is suggested for using these locations to define an auditory-perceptual space (APS) of three dimensions based on log frequency ratios in which both consonants and vowels can be mapped.

Briefly, these dimensions are defined by the short-time spectrum of speech sounds. Two types of spectra are distinguished: glottal-source spectra and burst-friction spectra. In general, the positions of the first three prominences for glottal-source sounds, and the position of two selected prominences for burst-friction sounds, in addition to their relation to a reference low-frequency component, define the location of each speech sound within the three-dimensional auditory-perceptual space. The following equations define this space for glottal-source spectra:

$$x = \log(\mathrm{SF3}/\mathrm{SF2}), \quad (1)$$

$$y = \log(\mathrm{SF1}/\mathrm{SR}), \quad (2)$$

$$z = \log(\mathrm{SF2}/\mathrm{SF1}), \quad (3)$$

where SF1, SF2, and SF3 represent the frequency locations of the first three significant prominences of the short-term spectral envelope of the acoustic waveform. SR is a reference frequency which is shifted slightly by the current speaker's average pitch and by significant pitch modulations (see Miller, 1989).

In the present paper, only the burst-friction spectra of

[a] Also at Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. Address all correspondence to Allard Jongman, Dept. of Linguistics, UCLA, Los Angeles, CA 90024.

 0001-4966/91/020867-07$00.80

syllable-initial voiceless stops of American English are considered. The center frequencies of the sensory formants of burst-friction spectra are signified as burst-friction sensory formants BF1, BF2, and BF3. Since voiceless burst-friction spectra are characterized by the absence of significant low-frequency energy, BF1 is arbitrarily set equal to SR, and such spectra are thus described by the values of BF2 and BF3. For the present study, the sensory reference was set equal to 168 Hz, the average pitch of male and female speakers, as reported by Peterson and Barney (1952).[1] The spectra are located in the auditory-perceptual space by the following equations:

$$x = \log(\mathrm{BF3}/\mathrm{BF2}), \tag{4}$$

$$y = \log(\mathrm{SR}/\mathrm{SR}), \tag{5}$$

$$z = \log(\mathrm{BF2}/\mathrm{SR}). \tag{6}$$

Since in the case of burst-friction spectra the absent first formant is set equal to SR, $z = \log(\mathrm{BF2}/\mathrm{SR})$ and the y component is zero. Therefore, burst-friction spectra are located in the xz plane of the three-dimensional space.

The application of the auditory-perceptual theory to stop consonants involves the following steps: First, a burst spectrum is derived using LPC analysis. Then, two spectral peaks (BF2 and BF3) are selected and converted into x and z coordinates in the APS on the basis of log frequency ratios by means of Eqs. (4)–(6). It is hypothesized that the values for x and z for a given burst-spectrum spoken by different speakers will uniquely define a two-dimensional enclosed region in this space, which is called a target zone.

A necessary condition for the success of the auditory-perceptual interpretation of the stop consonants is that the burst spectra of the stops be separable into distinct regions or target zones. That is, the regions occupied by the burst-spectra of labial, alveolar, and velar stop consonants must be mutually exclusive. This paper is an attempt to characterize the burst onsets of English voiceless stops as occupying distinct regions in the auditory-perceptual space.

I. EXPERIMENT I

The first experiment was conducted to devise a metric that would, for each place of articulation, reliably pick two characteristic burst-friction peaks from LPC spectra, and to establish preliminary target zones to distinguish voiceless stop consonants in terms of place of articulation in the APS.

A. Methods

1. Subjects

Four students, two males and two females, served as speakers. All were native speakers of a general Midwestern dialect of American English, with no known history of either speech or hearing disorders.

2. Stimuli

Stimuli consisted of CVC syllables with each of the three voiceless stop consonants [p, t, k] in initial position, followed by each of the six vowels [i, ɪ, ɛ, a, ʊ, u] and a final consonant.[2] All stimuli were existing English words. Two repetitions of each stimulus were read in random order by the speakers at a normal speaking rate. Speakers first familiarized themselves with the stimuli and then read all stimuli without interruption. Due to equipment malfunction, one token of [pit], produced by a female speaker, was lost. Thus the data analyzed consisted of 143 stimuli (four speakers $\times$ three consonants $\times$ six vowels $\times$ two repetitions).

3. Recording

Speakers were recorded in an anechoic chamber, using a special low-noise microphone/preamplifier combination (Bruel and Kjaer, model 4179/2660). The microphone was placed at a height equal to 0.5 m in front of the speaker's mouth (0° angle of incidence). The microphone output was fed directly to a digital audio recorder in a 16 bit mode (Sony, model PCM-501ES) with a video cassette recorder (JVC, model 720) serving as the storage medium.

The speakers initially read the tokens while the experimenter set the recording levels. Once an appropriate level had been determined, a calibration tone was recorded directly onto the tape. Recording levels were not varied after this time. A reading timer device, designed and built in-house, was used to regulate speakers' speed for recitation of the tokens.

4. Analysis

The recordings were digitized at 20 kHz with 16 bit precision and stored as files to be processed by the commercial software package ILS (Interactive Laboratory System). The sampled data files were then high-pass filtered, using a second-order 50-Hz Butterworth Filter to remove any incidental low-frequency noise.

For our analyses, the initial aperiodic portion defined as the interval from release burst up to onset of voicing (i.e., burst, frication, and aspiration noise) of each stimulus was located using a graphics display terminal. A burst spectrum was then derived using LPC analysis. In our LPC analyses, we used a 24-ms full Hamming window, a preemphasis factor of 0.98, and 24 poles. The number of spectral peaks to be extracted was set equal to 5.

Our analysis focused on the spectral characteristics of the onset of the burst (see Kobatake and Ohtani, 1987). That is, the 24-ms Hamming window was centered over burst onset, with the left tail of the window positioned 12 ms prior to burst onset. Frequency and amplitude values were then obtained.

5. Selection of burst-friction formants

Within the framework of APT, voiceless burst-friction sounds are characterized by the location of two burst-friction prominences. Inspection of the spectral displays did show a general tendency for spectral peaks to occur at frequency locations expected on the basis of the acoustic theory of speech production (Fant, 1960). However, these peaks did not always dominate the spectrum. In other words, only when the place of articulation of a particular stop consonant was known could the appropriate spectral peaks be picked. The problem was to devise a metric that would pick those

peaks without any prior information about the identity of the consonant.

In order to develop such a metric, we first handpicked those peaks from the spectral displays that we felt were potential candidates for consistently distinguishing among bilabial, alveolar, and velar stop consonants. The tokens were grouped in terms of place of articulation and analyzed by hand. For example, both authors analyzed all alveolar tokens, and picked for each token those two peaks that seemed characteristic for the alveolar place of articulation. We then devised a metric that would pick these handpicked peaks as often as possible.

The following metric was devised. First, the spectral peak with maximum amplitude below 6 kHz was located and labeled $P(\max)$. Then, moving from 60 Hz–6000 Hz, the first two peaks within 10 dB of $P(\max)$ were picked as BF2 and BF3. Thus, in those cases where there were two peaks within 10 dB of, and to the left of, $P(\max)$, the maximum peak itself would not be picked as a burst-friction component, as shown in Fig. 1 [panel (A)]. Furthermore, in those cases where BF2 had been picked and BF3 was separated from BF2 by 2500 Hz or more (a pattern that often occurred for the velars), the frequency value for BF2 was also used as that for BF3, as shown in Fig. 1 [panel (b)]. If, however, there were no peaks within 10 dB of $P(\max)$, the frequency value for $P(\max)$ was used for both BF2 and BF3, as shown in Fig. 1 [panel (c)]. These burst-friction prominences were then converted into x and z coordinates on the basis of log frequency ratios by means of Eqs. (4)–(6) and plotted in the auditory-perceptual space.

II. RESULTS

This metric was then systematically applied to the initial voiceless stop consonants of four speakers, without prior knowledge of their place of articulation. The metric was implemented as a computer program such that it was applied automatically to each token. As shown in Fig. 2, the different places of articulation generally cluster into distinct regions. In an attempt to minimize the amount of overlap, preliminary target zones were drawn by hand, using a computer-aided method with a resolution of 0.005 log units.

The locations of the zones are somewhat reminiscent of those devised by Fant (1973) for the Swedish voiceless stops [p, t, k], using a linear $F2$ by $F3$ space. Fant obtained high identification scores, but these were based on only 27 tokens produced by one male speaker. The present target zones are based on 143 tokens produced by four speakers. As can be seen, the present target zones are highly irregular in shape. Due to these irregular contours, many vowel-dependent effects can be accommodated. These regions enable us to distinguish the voiceless stops in terms of place of articulation with 98% accuracy. Table I presents the number of tokens of each stop consonant enclosed by each target zone.

The traditional characterization of bilabial, alveolar, and velar stops in terms of a diffuse-flat, diffuse-rising, and compact spectral shape (cf. Jakobson *et al.*, 1963; Blumstein and Stevens, 1979), respectively, is represented in the APS as follows: The low-frequency component of the bilabials is apparent from the relatively small values along the z axis

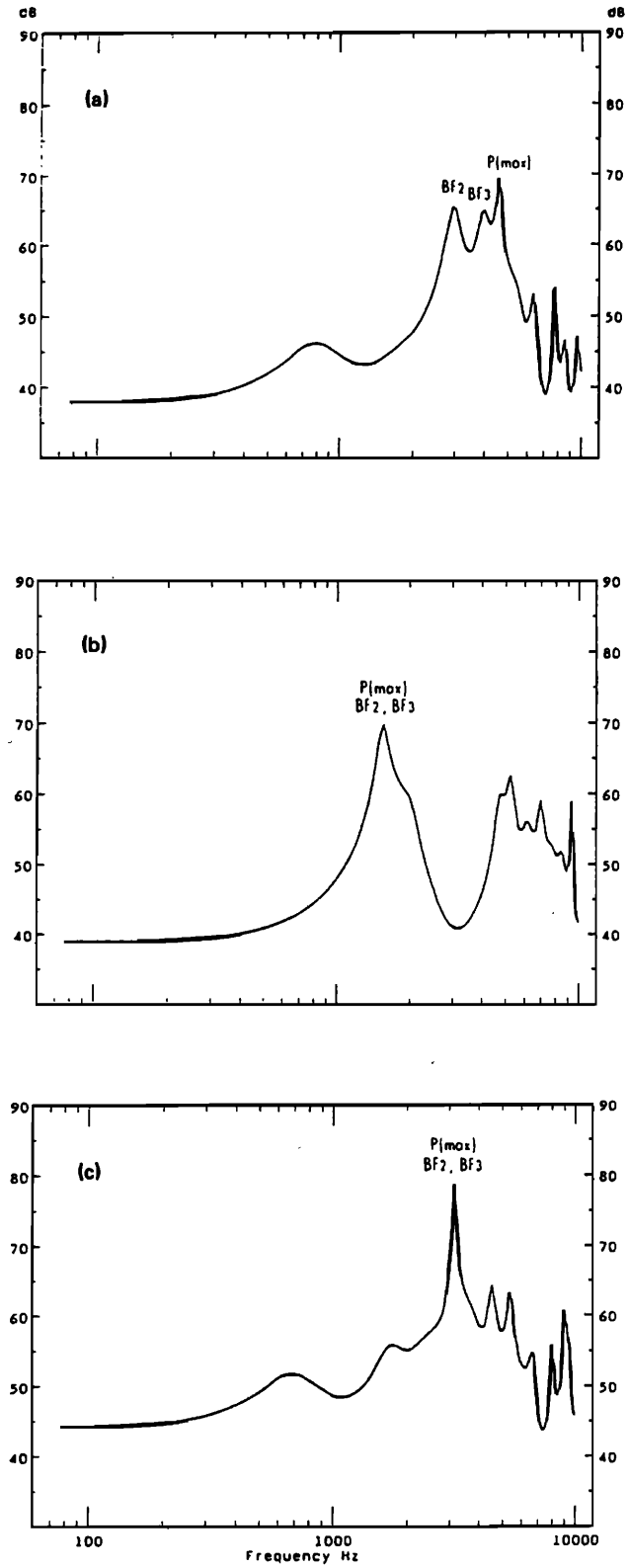

FIG. 1. Panel (a): burst spectrum of [t] as produced in the word "teen" by a male speaker; $P(\max)$ is located at 4523 Hz, with an amplitude of 69 dB. The method picks, going from left to right, BF2 at 2924 Hz (65 dB), and BF3 at 3905 Hz (65 dB). Panel (b): burst spectrum of [k] as produced in the word "could" by a female speaker; $P(\max)$ is located at 1458 Hz (67 dB). Going from left to right, the method picks this maximum peak as both BF2 and BF3, since the next peak (5160 Hz, 62 dB) is separated from BF2 by more than 2500 Hz. Panel (c): Burst spectrum of [k] as produced in the word "Ken" by a female speaker; $P(\max)$ is located at 3112 Hz (74 dB). Since there are no peaks within 10 dB of $P(\max)$, $P(\max)$ will be picked as both BF2 and BF3.

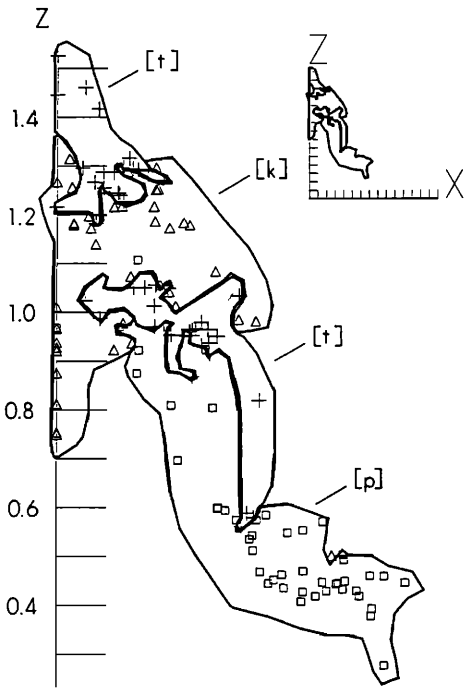

FIG. 2. Plot of the burst spectra in the *xz* plane of the auditory-perceptual space (APS) of 143 voiceless stops produced by four speakers. Preliminary target zones have been drawn. Bilabial stops [p] are represented by squares, alveolar stops [t] by crosshatches, and velar stops [k] by triangles. The figure shows a closeup of the *xz* plane. Note its relation to the inset, showing the location of the zones with respect to the point of origin. The *y* axis is perpendicular to the *xz* plane. The *x* and *z* axes are in 0.1 log units, and the point of origin is (0, 0).

{the lower the frequency value of BF2, the smaller *z* [see Eq. (6)]}; the high-frequency component of the alveolars is apparent from the relatively high *z* values; and the merger of BF2 and BF3 for many velars is represented by the very small *x* values {the closer BF2 and BF3, the smaller *x* [see Eq. (4)]}.

Moreover, in English, the velar stops are considered to have two distinct allophones, one before front vowels and the other before back vowels (Ohman, 1966; Zue, 1976; Sereno and Lieberman, 1987). This distinction does show up, in that all velar tokens with a small *z* value are followed by back vowels, whereas all other velars (for which *z* is relatively large) are followed by front vowels. In this way, for those velars that were correctly classified by the algorithm (98% of all velar tokens), front and back allophones of [k] could be distinguished with 96% accuracy.

As is apparent from Fig. 2, there seem to be two distinct regions for the alveolars as well. The alveolars were represented by two different spectral shapes. In the one case, as shown in Fig. 3, the method would pick a mid- and a mid- to high-frequency peak, a pattern corresponding to the "classic" alveolar pattern, with the center frequency of BF2 close to the alveolar *F*2-locus around 1800 Hz, as reported by Delattre *et al.* (1955). In the other case, as shown in Fig. 1 [panel (a)], the algorithm would pick two high-frequency peaks with frequency values close to those often found for [s]. However, it should be noted that these two distinct areas are not due to vowel context effects.

TABLE I. Number of tokens of each stop consonant in the training set enclosed by each target zone. The column on the far right indicates the percentage of stop consonants that were correctly classified in terms of place of articulation.

	Target zone			
Stimulus	P	T	K	Correctly classified
[p]	45	···	2	96%
[t]	···	48	···	100%
[k]	1	···	47	98%
			total correct	98%

III. EXPERIMENT II

The peak-picking algorithm and the preliminary target zones enabled us to distinguish the voiceless stops in the training set with 98% accuracy with respect to place of articulation. A second experiment was conducted to verify the algorithm and the location of the preliminary target zones using a set of voiceless stops, which were produced by a new set of speakers in a greater number of phonetic contexts.

A. Methods

1. Subjects

Eight students, four males and four females, served as speakers. All were native speakers of a general Midwestern dialect of American English, with no known history of either speech or hearing disorders. None of the speakers had participated in the previous experiment.

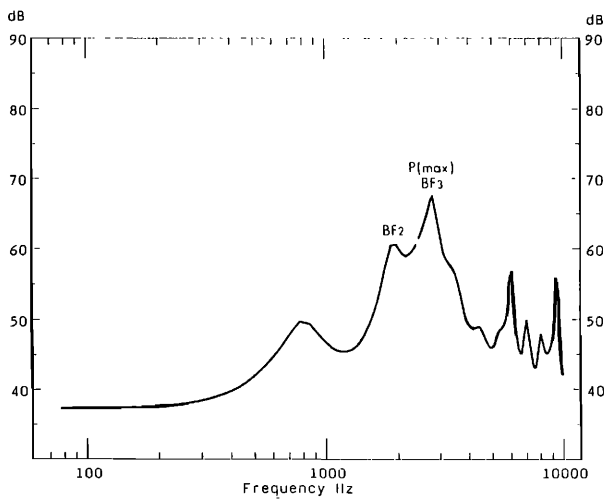

FIG. 3. Burst spectrum of [t] as produced in the word "tomb" by a male speaker. Our method picked the two "classic" alveolar peaks around 1900 and 2700 Hz.

TABLE II. Number of tokens of each stop consonant in the test set enclosed by each target zone. UNCL stands for unclaimed region. The column on the far right indicates the percentage of stop consonants that were correctly classified in terms of place of articulation.

	Target zone				
Stimulus	P	T	K	UNCL	Correctly classified
[p]	52	12	8	8	65%
[t]	8	50	18	3	63%
[k]	8	15	53	4	65%
				total correct	65%

2. Stimuli

Stimuli consisted of CVC syllables, with each of the three voiceless stop consonants [p, t, k] in initial position, followed by each of the ten vowels [i, ɪ, ɛ, æ, ʌ, ɚ, a, ɔ, ʊ, u] and a final consonant.[3] All stimuli were existing English words and included those that had been used in experiment I. All stimuli were read in random order at a normal speaking rate. The data analyzed consisted of 240 stimuli (eight speakers × three consonants × ten vowels).

3. Recording and analysis

The methods used for the recording and analysis of the tokens were those described for experiment I. As in experiment I, after LPC spectra had been obtained, the peak-picking algorithm was applied, and the peaks were converted into x and z values and plotted in the APS.

4. Results

The 240 tokens were plotted in APS and scored in terms of whether they fell in the appropriate preliminary target zones that had been established in experiment I. These target zones enabled us to distinguish the new set of voiceless stops in terms of place of articulation with 65% accuracy. Table II presents the number of tokens for each stop consonant enclosed by each target zone. The addition of four new vowel contexts did not substantially change the classification score: The overall classification score for [p, t, k] when followed by the six vowels used in experiment I was 60%. There are a number of tokens that fall into regions not claimed by any target zone. A slight modification of our preliminary target zones to include these points would bring the total correct classification to 70%.

It must be noted that, for the present study, we have simply drawn the target zones such that they enclose as many appropriate data points as possible. However, with this relatively small sample, it remains an open question whether such zones will prove to have any explanatory value. In order to address this question, perceptual verification experiments must be conducted to establish the "psychological reality" of the target zone boundaries. However, until this issue has been settled, we believe that there is no *a priori* reason to prefer smooth zones (e.g., circles or ellipses) over the present irregular zones.

IV. DISCUSSION AND CONCLUSIONS

The present study involved the acoustic analysis of voiceless stop consonants. Our method of analysis focused on the spectrum at burst onset. A metric was devised to extract the relevant burst-friction components from that spectrum. These components were then converted into log frequency ratios and represented as points in an auditory-perceptual space (APS). In APS, each place of articulation was seen to be associated with a distinct region, or target zone. Application of the metric to a training set of stimuli, which allowed for the initial drawing of the target zones, resulted in a 98% correct classification of the voiceless stop consonants. A new test set of voiceless stops, which included more speakers and additional vowel contexts, was then analyzed. The emphasis on the burst onset, the peak-picking metric, and the conversion of BF2 and BF3 into points in the APS subsequently enabled us to distinguish this new test set of voiceless stop consonants in terms of place of articulation with 70% accuracy.

The decrease in performance from training set to test set raises several questions. First, one may wonder whether burst spectra can be appropriately described in terms of the locations of spectral peaks. Although a few studies have documented the location of frequency peaks in burst spectra, these peaks have been shown to vary considerably as a function of the following vowel (e.g., Zue, 1976; Repp and Lin, 1989). One obvious reason for the decrease in performance is, therefore, the increase in spectral variability resulting from the use of different phonetic contexts and additional speakers.

Current research emphasizes the characterization of burst spectra in terms of global spectral shape, rather than the absolute frequency locations of spectral peaks, to reduce this variability (e.g., Stevens and Blumstein, 1981). Within the framework of the auditory-perceptual theory, the location of spectral peaks is assumed to reflect gross spectral shape. The present peak-picking metric does take into account spectral shape and spectral tilt in that it is sensitive to amplitude relations between spectral peaks in addition to frequency location. In the auditory-perceptual theory, extraction of spectral peaks is necessary in order to define the location of each speech sound in the auditory-perceptual space. While this approach has been shown to be successful for the representation of vowels (Miller, 1989; Jongman *et al.*, 1989), the present study forms only a first step toward the representation of (stop) consonants in the auditory-perceptual space. If successful, this approach will provide a unified framework in which vowels and consonants can be represented in the same perceptual space.

Second, one may ask whether the burst portion of voiceless stop consonants contains sufficient information for classification of place of articulation. In a perceptual study, Tekieli and Cullinan (1979) have shown that listeners, when presented with the first 10 ms of the signal, are able to distinguish voiceless stop consonants in terms of place of articulation significantly above chance. Tekieli and Cullinan (1979) presented subjects with the first 10 ms of the aperiodic portion of CV syllables, consisting of [p, t, k] followed by eight different vowels. Subjects were to identify both the conso-

nant and the vowel. For the consonants, the response set consisted of [p, b, t, d, k, g, tʃ, dʒ], as well as the option to report no consonant at all. Listeners in this experiment identified place of articulation for [p, t, k] with 54% accuracy. In addition, recent statistical approaches have been quite successful at classifying word-initial obstruents on the basis of the first 10 ms of the speech signal. In one study (Forrest *et al.*, 1988), a series of fast Fourier transforms (FFTs) was calculated every 10 ms from the onset of the obstruent. Each FFT was treated as a random probability distribution from which the first four moments were computed, reflecting spectral tilt and shape as well as concentration of energy. These moments were then input into a discriminant analysis for classification in terms of place of articulation. Performance was quite high based on one analysis window over the first 10 ms of the speech signal. Specifically, classification of voiceless stop consonants reached a level of approximately 85% accuracy. However, it is yet unclear how performance would be affected by changes in phonetic context. Forrest *et al.* (1988) used a very restricted set of vocalic contexts. Moreover, this set was not balanced across consonantal contexts (i.e., only four different vowels were used, and only one of these was paired with all three voiceless stops), and the test set consisted of the same limited set of vowels as the training set. Nevertheless, these studies suggest that the initial portion of voiceless stops does indeed contain sufficient information for correct classification of place of articulation.

While burst onsets contain important cues to place of articulation, it is generally recognized that classification of stop consonants improves when taking into account larger stretches of the speech signal or changes of the spectrum over time (e.g., Blumstein and Stevens, 1979; Searle *et al.*, 1979; Kewley-Port, 1983; Lahiri *et al.*, 1984). The auditory-perceptual approach to stop consonant representation can be modified to incorporate the use of dynamic information. The present study was based on only a single analysis frame. However, instead of representing the burst spectrum as a single point in APS, it is intended that a point be calculated for every millisecond of the speech signal. Thus, over time, a sequence of points (or "path") through APS is generated. Such a dynamic representation would allow for the continuous tracing of consonants and vowels and would enable us to take spectral changes over time into account within the auditory-perceptual theory.

Another issue of interest involves the perceptual verification of the hypothesized target zones. Further research is planned to explore whether the extracted spectral peaks and irregular target-zone boundaries have any perceptual significance. This will be accomplished by obtaining listeners' responses to stimuli that are synthesized from x and z values taken from APS. For example, some of the peaks selected by the algorithm as representative of [p] are rather low in frequency and might, in fact, arise from subglottal resonances. As such, these peaks would not contribute to perception of the labial place of articulation. Perception experiments will therefore help to refine the peak-picking algorithm.

As a preliminary attempt to determine whether the irregular target-zone boundaries have any perceptual significance, we used a synthetic continuum from the literature in

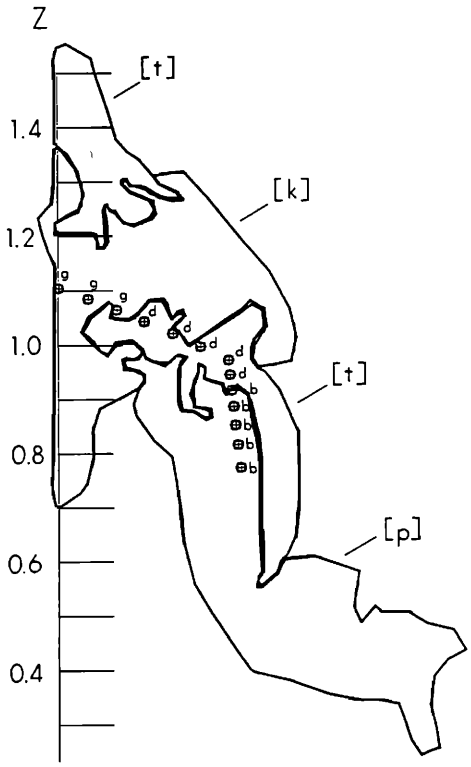

FIG. 4. The $F2$ and $F3$ values for a [b–d–g] continuum (Stevens *et al.*, 1969) converted into x, z coordinates and plotted in APS. Subjects' labeling is indicated next to each of the 13 continuum members (see Fig. 2 for axes).

which place of articulation was signaled by the locations of $F2$ and $F3$ (Stevens *et al.*, 1969). Stevens *et al.* synthesized a voiced stop continuum without bursts. It should be noted that the peak-picking metric and location of target zones developed in the present study were based on the analysis of the burst onset of natural voiceless stop consonants. Unfortunately, we were unable to find in the literature any synthetic continua for voiceless stops with bursts. Despite the differences, however, the Stevens *et al.* synthetic continuum provides a first step to a perceptual evaluation of the proposed target zones. For the evaluation of our approach, $F1$ values from Stevens *et al.* were not used, to reflect voiceless stops. Since $F2$ and $F3$ values from the synthetic continuum would be extracted by the metric as BF2 and BF3, $F2$ and $F3$ values (as BF2 and BF3) were converted into APS coordinates using Eqs. (4)–(6), with SR set equal to 168 Hz (i.e., the same SR value that was used to establish the target zones). The coordinates were then plotted in the auditory-perceptual space. As can be seen in Fig. 4, subjects' identifications obtained by Stevens *et al.* map very well onto our preliminary target zones. Most continuum members fall into the appropriate target zones, with the phoneme boundaries almost coinciding with the target-zone boundaries. These preliminary results from the Stevens *et al.* (1969) synthetic stimuli are consistent with the hypothesis that listeners may

be sensitive to fine acoustic differences represented by the irregular boundaries of the target zones. Of course, more detailed perceptual experiments using synthesized voiceless stop consonants with bursts are needed to systematically evaluate the perceptual validity of the target zones.

In sum, the present approach, based on a single analysis frame at burst onset, yielded promising results for the classification of place of articulation in voiceless stop consonants within the auditory-perceptual theory. Additional research, involving the inclusion of dynamic spectral information and the use of perception studies to evaluate the peak-picking metric and to verify the location of the target zones, is needed to further assess the auditory-perceptual approach to the representation of stop consonants.

ACKNOWLEDGMENTS

We thank Elizabeth B. Ottenberg, Melissa S. Piasecki, Jason Taff, and LaDeana F. Weigelt for assistance in analyzing the data, Steven J. Sadoff for the computer-implementation of the metric, and Joan A. Sereno and an anonymous reviewer for helpful comments and suggestions. This work was supported by NIH Grant NS 21994 and AFOSR Grant 86-0335 to Central Institute for the Deaf.

[1] In running speech, SR is shifted from 168 Hz to a value appropriate to the current speaker's vocal characteristics by incorporating that speaker's mean pitch. This mean pitch is calculated from the $F0$ in the adjacent vocalic context (Miller, 1989). However, since our analysis focused on the onset of the aperiodic portion of word-initial voiceless stop consonants and did not include any information about adjacent vowels, SR was fixed at 168 Hz for the present study.

[2] Test words used in experiment I: peat, teen, keel; pit, tin, kit; pet, ten, Ken; pot, top, cot; put, took, could; pool, tomb, cool.

[3] Additional test words used in Experiment II: pat, tan, cat; putt, ton, cut; perk, term, curt; paw, taugt, caught.

Blumstein, S. E., and Stevens, K. N. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fant, G. (**1973**). *Speech Sounds and Features* (MIT, Cambridge, MA).

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (**1988**). "Statistical analysis of word-initial voiceless obstruents: Preliminary data," J. Acoust. Soc. Am. **84**, 115–124.

Jakobson, R., Fant, G., and Halle, M. (**1963**). Preliminaries to speech analysis (MIT, Cambridge, MA).

Jongman, A., Fourakis, M., and Sereno, J. A. (**1989**). "The acoustic vowel space of Modern Greek and German," Lang. Speech. **32**, 221–248.

Kewley Port, D. (**1983**). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 322–335.

Kobatake, H., and Ohtani, S. (**1987**). "Spectral transition dynamics of voiceless stop consonants," J. Acoust. Soc. Am. **81**, 1146–1152.

Lahiri, A., Gewirth, L., and Blumstein, S. E. (**1984**). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. **76**, 391–404.

Miller, J. D. (**1987**). "Auditory-perceptual processing of speech waveforms," in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Erlbaum, Hillsdale, NJ), pp. 257–266.

Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**, 2114–2134.

Ohman, S. (**1966**). "Coarticulation in VCV utterances: Spectographic measurements," J. Acoust. Soc. Am. **39**, 151–168.

Peterson, G., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Repp, B. H., and Lin, H. B. (**1989**). "Acoustic properties and perception of stop consonant release transients," J. Acoust. Soc. Am. **85**, 379–397.

Schatz, C. D. (**1954**). "The role of context in the perception of stops," Language **30**, 47–56.

Searle, C. L., Jacobson, J. Z., and Rayment, S. G. (**1979**). "Stop consonant discrimination based on human audition," J. Acoust. Soc. Am. **65**, 799–809.

Searle, C. L., Jacobson, J. Z., and Kimberley, E. (**1980**). "Speech as patterns in the 3-space of time and frequency," in *Perception and Production of Fluent Speech*, edited by R. A. Cole (Erlbaum, Hillsdale, NJ), pp. 73–102.

Sereno, J. A., and Lieberman, P. (**1987**). "Developmental aspects of lingual coarticulation," J. Phon. **15**, 247–257.

Stevens, K. N., Liberman, A. M., Studdert-Kennedy, and Ohman, S. E. (**1969**). "Crosslanguage study of vowel perception," Lang. Speech **12**, 1–23.

Stevens, K. N., and Blumstein, S. E. (**1981**). "The search for invariant acoustic correlates of phonetic features," in *Perspectives on the study of speech*, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, NJ), pp. 1–38.

Tekieli, M. E., and Cullinan, W. L. (**1979**). "The perception of temporally segmented vowels and consonant-vowel syllables," J. Speech. Hear. Res. **22**, 103–121.

Zue, V. W. (**1976**). "Acoustic characteristics of stop consonants: A controlled study," Tech. Rep. 523, Lincoln Lab.