Statistical analysis of word-initial voiceless obstruents: Preliminary data

Karen Forrest
Speech Motor Control Laboratories, Waisman Center, University of Wisconsin, Madison, Wisconsin 53705-2280

Gary Weismer
Speech Motor Control Laboratories, Waisman Center and Department of Communicative Disorders, Goodnight Hall, University of Wisconsin, Madison, Wisconsin 53706

Paul Milenkovic
Speech Motor Control Laboratories, Waisman Center and Department of Computer and Electrical Engineering, University of Wisconsin, Madison, Wisconsin 53706

Ronald N. Dougall
Speech Motor Control Laboratories, Waisman Center and Department of Communicative Disorders, University of Wisconsin, Madison, Wisconsin 53706

(Received 16 October 1987; accepted for publication 31 March 1988)

A statistical procedure for classifying word-initial voiceless obstruents is described. The data set to which the analysis was applied consisted of monosyllabic words starting with a voiceless obstruent. Each word was repeated six times in the carrier phrase "I can say ____, again" by each of ten speakers. Fast Fourier transforms (FFTs), using a 20-ms Hamming window, were calculated every 10 ms from the onset of the obstruent through the third cycle of the following vowel. Each FFT was treated as a random probability distribution from which the first four moments (mean, variance, skewness, and kurtosis) were computed. Moments were calculated from linear and Bark transformed spectra. Data were pooled across vowel contexts for speakers of a given gender and input to a discriminant analysis. Using the moments calculated from the linear spectra, 92% of the voiceless stops were classified correctly when dynamic aspects of the stop were included. Even more important, the model constructed from the males' data correctly classified about 94% of the voiceless stops produced by the female speakers. Classification of the voiceless fricatives when all places of articulation were included in the analysis did not exceed 80% correct when the moments from either the linear or Bark transformed scales were used. However, classification of only the voiceless sibilants was 98% correct when the moments from the Bark transformed spectra were used. As with the stops, the classification model held across gender.

PACS numbers: 43.70.Fq

INTRODUCTION

The purpose of the present article is to describe a quantitative approach to classifying obstruent spectra. The notion of classifying obstruent spectra according to certain articulatory dimensions—such as place of articulation—has held the interest of researchers for over 3 decades (see summary in Fant, 1973, pp. 160–170). Specifically, researchers have been interested in the nature of variability in consonant spectra, especially as it may be conditioned by vowel contexts. This interest stems, in part, from a desire to determine how perceptual constancy is derived from acoustically distinct signals. Recently, Blumstein and Stevens (1979), Kewley-Port (1983), and Kewley-Port and Luce (1984) have shown how qualitative analysis of spectral shape for stops, either confined to a single interval following the burst or spanning a sequence of such intervals, provides a means to reveal unique spectral characteristics of the three places of stop articulation in American English. The appeal of these results is that the spectral uniqueness is maintained even in the face of variation in (1) following vowel context, (2) voicing status of the stop, and, to a lesser degree, (3) speaking rate.

The analysis employed by Blumstein and Stevens (1979) involved the construction of graphic templates of spectral features associated with the three places of articulation in English. These templates, which were based on a linear predictive code (LPC) analysis of a 25.6-ms interval beginning at the stop burst (half-Hamming window, 0- to 5-kHz bandwidth), were developed by careful examination of multiple stop spectra and trial and error adjustment of the template features. Classification was dependent on the shape of the spectrum; bilabials had either a flat or falling spectrum, alveolars were characterized by an upward spectral tilt, and velars were defined by a compact, central spectral peak. Blumstein and Stevens (1979) used the final version of the templates to classify correctly an average of 84% of stops produced in the syllable-initial position by six talkers.

Kewley-Port (1983) and Kewley-Port and Luce (1984) argued that the *static* spectral analyses of Blumstein and Stevens could miss important, time-varying spectral fea-

0001-4966/88/070115-09$00.80

tures in the interval following the stop burst, and implemented a *running spectral* analysis [LPC, Hamming window (20 ms), 0- to 5-kHz bandwidth] of stops to classify place of articulation. In Kewley-Port's experiments, the classification of spectra was based on experimenter-defined features that were applied by trained judges to the running-spectra displays. Kewley-Port's judges identified correctly an average of 88% of voiced stops from absolute frequency displays (Kewley-Port, 1983) and 90% of voiced and voiceless stops from spectral displays modified to reflect the spectral processing characteristics of the human auditory system. Based on these results, Kewley-Port has argued that time-varying spectra are preferable to static spectra in the classification of stop place of articulation.

In addition to the obvious application of these findings to theories of speech perception (see, for example, Stevens and Blumstein, 1978; Kewley-Port *et al.*, 1983), the categorization of obstruent spectra is also interesting for the understanding of disordered speech production, specifically concerning the relationship of a speaker's intended phonological units to his/her vocal tract output. In our own work with misarticulating children and adults, for example, we have been perplexed by obstruent productions that seem on perceptual analysis to be unclassifiable in the phonemic system of English. Some of these sounds appear to be "between" two phoneme categories (e.g., not a /d/ or /g/, but having both /d/- and /g/-like characteristics), or simply unlike any phonemes.

A rare example of the application of a spectral categorization strategy to the analysis of disordered speech is found in Shinn and Blumstein (1983), who used the Blumstein and Stevens (1979) templates to classify production errors of aphasic speakers as phonetic or phonemic. Shinn and Blumstein (1983) found that, when the templates were used to classify those stops that were *perceived consistently* as the intended target, classification scores were very similar to those reported for normal speakers by Blumstein and Stevens (1979). Such a restriction on the material to be classified, however, makes the success of the study unsurprising and supports Ziegler's (1984) contention that the Blumstein and Stevens template system is too coarse to be of much success in understanding obstruent errors in aphasia. The system described by Kewley-Port and Luce (1984) would also not be likely to produce useful results with disordered speakers, as these investigators had to impose several *ad hoc* adjustments of their system to get acceptable scores for normally produced stops. A highly desirable approach to spectral analysis of obstruents produced by disordered speakers would be one that was strictly objective and could specify spectral distances between an aberrant token and some model of the target obstruent. Such distance measures would be especially useful for obstruents that are unclassifiable on the basis of spectral template or auditory analysis.

A quantitative approach to the classification of voiceless stops described by Kobatake and Ohtani (1987) provides the desired analysis objectivity. Using principal components analysis of the onset spectra, they found unique patterns that correctly classified 90% of their voiceless stop tokens. However, the classification data were the same set used to construct the patterns, thereby making high classification accuracy unsurprising. The use of a model derived from one data set to classify a new data set presents difficulties for a linear model, such as principal components analysis. A simple frequency shift in the spectrum peak can result in a complex variation of the linear parameters that may be difficult to normalize across speakers.

In the present study, we selected the moments of a probability density function, which can be normalized for shifts in center frequency, as numerical indices of spectral shape and center of gravity. If successful in classifying obstruent spectra, such indices would be preferable to the classification schemes employed by Blumstein and Stevens (1979) and Kewley-Port and Luce (1984) which require human judgments that are subject to error even under optimal conditions (i.e., when the judges are sophisticated). These indices would also be preferable to Kobatake and Ohtani's (1987) procedure in that they are easily normalized for frequency shifts and could provide a metric to determine the relationship between correct and aberrant productions. The particular statistical measures that we chose to investigate were the mean, skewness, and kurtosis of the computed FFTs in the region of the stop burst and onset of frication noise. These measures were chosen because of their ability to summarize the concentration, tilt, and peakedness of the energy distributions, the three spectral characteristics prominent in the categorical classification systems described previously (Blumstein and Stevens, 1979; Kewley-Port and Luce, 1984). Furthermore, the nature of the relationship between the mean and higher moments would leave shape changes, as indexed by skewness and kurtosis, unaffected by scalar frequency shifts. We decided to investigate the classification utility of moments for both linear and Bark representations of stop and fricative spectra, because previous work (Kewley-Port and Luce, 1984; Bladon and Seitz, 1986) has suggested that "auditory" spectra provide the most meaningful profile of obstruent acoustics. Other investigators have used moments to index the spectra associated with impulse noises (Erdreich, 1986), spontaneous neuronal activity (Lansky and Radil, 1987), tidal ventilation (Butler and Mohler, 1979), and EMG tests of endurance (Hary *et al.*, 1982). Furthermore, moment analysis may be used as a general approach to evaluating systems (Bendat and Piersol, 1980).

I. METHOD

A. Subjects

Five males and five females ranging in age from 18–31 yr (mean age = 23.6, s.d. = 3.0) served as subjects. All subjects reported normal hearing and had no history of speech or hearing problems.

B. Speech sample

Table I presents a list of the test words that were analyzed in this study. These words were a part of a larger speech sample (31 words) that was collected to serve as a baseline for data collected from phonologically disordered

TABLE I. Speech sample used in analysis.

paid	pop	pay
keen	cot	key
tea	tot	two
she	see	fat
fought	thought	

children. This larger speech sample was developed to sample errors commonly made by phonologically disordered speakers. Towards that end, the larger speech sample included series of words that could test for consonant-related phonological processes, such as consonant deletion, cluster reduction, etc. The speech sample, then, was not geared toward testing vowel context effects. For that reason, the words that were analyzed in this study were not balanced for vowel context. However, there is more than one vowel context for most of the obstruents discussed in this report (/k/, /t/, /p/, /f). The vowels that are paired with these obstruents are distinct from one another and occupy rather different formant spaces, thereby affecting the obstruents differently (Stevens *et al.*, 1966). This claim is supported further by Cohn's (1987) data that showed that templates constructed to classify stop + /a/ sequences did a poor job of classifying stop + /u/, /i/, or /e/. Again, this would suggest that the vowels used in this study, though limited, would be expected to affect the obstruents rather differently.

C. Procedures

The subject's task was to repeat each stimulus word after it was presented via a loudspeaker. This elicitation procedure was used to make the present task comparable to that used with the disordered speakers with which we, eventually, want to compare the present data. The eliciting tape was prepared from a single recitation of each stimulus word provided by a male speaker. The words were then low-pass filtered (f_c = 5kHz) and digitized at 10 kHz on a Harris 800 computer. After the peak intensity of the words was equated, each word was reproduced six times. The resulting list of 186 words (31 words × 6 repetitions) was randomized and output to an audio tape with 5 s between items.

Each subject was seated in a sound-treated room. The eliciting tape was played on a Tandberg 440 cassette recorder/reproducer, passed to a power amplifier (Amber 50 A), and transduced by a loudspeaker (Polk Audio 7c). The stimulus words were presented to the subject at a comfortable listening level. Upon hearing the stimulus word, the subject repeated the word in the carrier sentence "I can say ____, again." The subject's speech was transduced by a Shure (SM10A) microphone placed about 6 cm from his/her lips and recorded on a Tandberg 420 tape recorder/reproducer.

All data processing was accomplished on an IBM PC AT desktop computer. The tape recording of the acoustic speech signal was low-pass filtered at 10 kHz with an eight-pole Butterworth filter (model 901F1, Frequency Devices, Haverhill, MA) and subsequently sampled at 20 kHz using an analog-to-digital converter with 12 bits of numeric resolution (Labmaster, Scientific Solutions, Solon, OH). Following sampling, a two-pole digital high-pass filter with a 70-Hz cutoff frequency (Milenkovic, 1986) was applied to the speech waveform. The high-pass filter served to reduce oscillations resulting from room vibration as well as to suppress the microphone air blast artifact associated with plosive speech productions.

The speech waveform was subsequently displayed on a computer CRT screen using CSpeech, a speech waveform analysis program developed by the third author, and cursors were manually placed to designate measurement points. In the case of stops, the initial cursor was placed at the onset of the burst. With fricatives, the initial cursor was placed at the onset of the frication noise waveform. In all cases, the final cursor was positioned at the end of the third pitch period cycle of the following vowel.

For each waveform token, a sequence of spectra was computed. A 20-ms analysis window was used to compute each spectrum in the sequence. The initial spectrum in the sequence was obtained by centering the analysis interval at the initial cursor position. Subsequent spectra in the sequence were obtained by moving the analysis interval forward in time by 10-ms increments, resulting in 50% overlap of adjacent analysis intervals. The number of analysis intervals varied with the phone analyzed, ranging from a minimum of 7 intervals for /p/ to 25 intervals for /s/.

The Fourier spectrum was computed for each analysis interval using the following procedure. The speech signal was preemphasized by first differencing. A 400-point Hamming window was applied to the preemphasized speech signal within each analysis interval, the 400-point data sequence was extended with zeros, and a 512-point fast Fourier transform (FFT) was computed.

The linear frequency scale spectral moments were derived from the Fourier spectrum by computing the power spectrum $P(k) = X_{\rm re}(k)^2 + X_{\rm im}(k)^2$, where $X_{\rm re}$ and $X_{\rm im}$ are the real and imaginary components of the Fourier spectrum, and k denotes the Fourier frequency sample. For frequency samples k in the range $1 \leqslant k \leqslant 256$, the normalized power spectrum $p(k)$ is computed as $p(k) = P(k)/[P(0) + \cdots + P(256)]$. The frequency sample at dc, $k = 0$, is left out of the calculation because the acoustic recording system does not give meaningful data at dc. Frequency samples above $k = 256$ are disregarded, because for real valued signals $x(n)$, $P(k) = P(512 - k)$.

The spectral moments were obtained by treating $p(k)$ for $1 \leqslant k \leqslant 256$ as the probability values for a discrete random variable k, which assumes integer values over the range (1,256). Because of the manner in which $p(k)$ was computed, we can be assured that $p(1) + \cdots + p(256) = 1$ and $p(k) \leqslant 1$ and $p(k) \geqslant 0$, giving p the properties of a probability that allow us to compute moments of a discrete probability distribution. The linear frequency scale moments are given by

$$\begin{aligned}
L_1 &= f_1 p(1) + \cdots + f_{256} p(256), \\
L_2 &= (f_1 - L_1)^2 p(1) + \cdots + (f_{256} - L_1)^2 p(256), \\
L_3 &= (f_1 - L_1)^3 p(1) + \cdots + (f_{256} - L_1)^3 p(256), \\
L_4 &= (f_1 - L_1)^4 p(1) + \cdots + (f_{256} - L_1)^4 p(256),
\end{aligned}$$

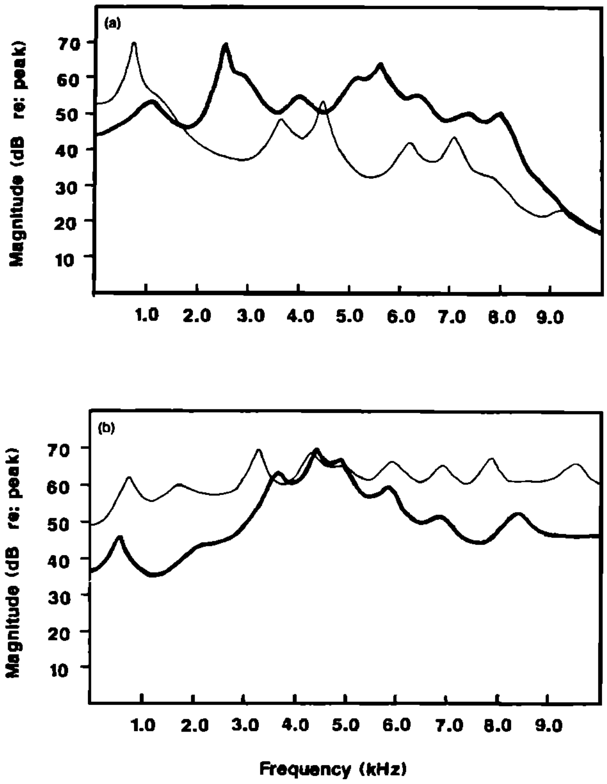

FIG. 1. (a) Two spectra that differ in mean and skewness. The thin-lined spectrum has a mean of 1.0 kHz and a positive skewness of 4.0. By contrast, the thick-lined spectrum has a mean of 4.5 kHz and a slight negative skewness of -1.2. (b) Two spectra with similar means and skewness, but different kurtoses. The thin-lined spectrum has a kurtosis of -0.7, which is reflective of its diffuse peaks, while the thick-lined spectrum has a kurtosis of 6.7. All spectra were LPC smoothed for easier viewing.

where the frequency f_k of the kth frequency sample is given by $f_k = f_s k/512$, where f_s is the sampling frequency at which the speech waveform is acquired. The moment L_i has the units of frequency to the ith power. Dimensionless versions of the third and fourth moments are computed according to $l_3 = L_3/(L_2)^{3/2}$, where l_3 is the coefficient of skewness and $l_4 = [L_4/(L_2)^2] - 3$, where l_4 is the coefficient of kurtosis (see Newell and Hancock, 1984). These dimensionless versions of the third and fourth moments are normalized with respect to shifts in center frequency and frequency scale that can occur between subjects producing the same sound. Examples of spectra with different values for each of these moments are presented in Fig. 1.

Bark transform frequency scale moments were computed from the power spectrum using a somewhat different procedure from the one used for the linear frequency scale moments. The Bark transform is a nonlinear warping of the frequency scale performed according to formulas stated by Syrdal and Gopal (1986). This FFT algorithm computes the Fourier spectrum at the uniformly spaced frequency samples f_k. If we apply the Bark transform to the frequency scale, the corresponding Bark samples b_k will be nonuniformly spaced. The Bark moments were computed using the procedure outlined for the linear moments, where we replaced the power spectrum $P(k)$ with the weighted power spectrum $(b_k - b_{k-1})P(k)$ and we replaced the frequency samples f_k with their corresponding Bark samples b_k. In this manner, nonuniform spacing of spectrum values on the Bark scale gave more weight over the region where the Bark scale was more sparsely sampled.

The moments data from each subject were grouped across all repetitions of each target obstruent, independent of the vowel context. Moments calculated for each obstruent were then grouped according to the time interval from which they were derived. For example, all moments calculated from the first 10-ms interval for all repetitions of one subject's production of /k/, were grouped; a similar grouping was made for the next 10-ms interval, etc. The data from all subjects of a given gender were then combined.

Though the first four moments were calculated, it was determined that the second moment, the variance, did not add to the discriminability of the different obstruents, so it was not used as a discrete variable in our analyses. The efficacy of the remaining moments in differentiating place of articulation was evaluated by means of graphic representation and discriminant analyses.

Stepwise discriminant analyses (BMDP7M) were performed separately on the stops and fricatives. Basically, the discrimant analysis is a linear combination of "n" variables such that the resulting functions, or canonical variates, provide maximum distance between members of different categories while minimizing the distance between members of like categories. The number of canonical variates is dependent on the number of categories to which the data are to be assigned; there is one less canonical variate than the number of categories. Discriminant analysis is very sensitive to deviations of the input variables from normality. For this reason, a Shapiro–Wilk (Shapiro and Wilk, 1965) test of normality was performed on all input-variable distributions. All variables were normally distributed ($p > 0.05$).

D. Results

The moments data were used to construct three-dimensional graphs for each target place of articulation. The time interval (*re*: obstruent onset) from which the moments were calculated was used as a parameter. In this way, four relevant dimensions (mean, skewness, kurtosis, time) could be displayed in three-dimensional space. Figure 2 presents representative graphs from one subject for the linear moments calculated from the first 10 ms (burst to burst + 10 ms) of each voiceless stop. The data are collapsed across all vowels, as are all data presented in this article.

It can be seen in Fig. 2 that the simultaneous evaluation of mean, skewness, and kurtosis differentiates the place of stop articulation. For example, /p/ and /t/ differ consistently in skewness and mean but not in kurtosis; /k/ is similar to /p/ in mean and skewness but differs from the other places of articulation in kurtosis. The graphic representation of the linear moments suggests that labial and alveolar stops are distinct in terms of mean and skewness, while the velar stops are distinguished by their kurtosis.

Figure 3 presents the Bark moments for the same time interval and subject as the linear moments presented above. As with the linear moments, differences can be seen in the three-dimensional space occupied by each place of articula-

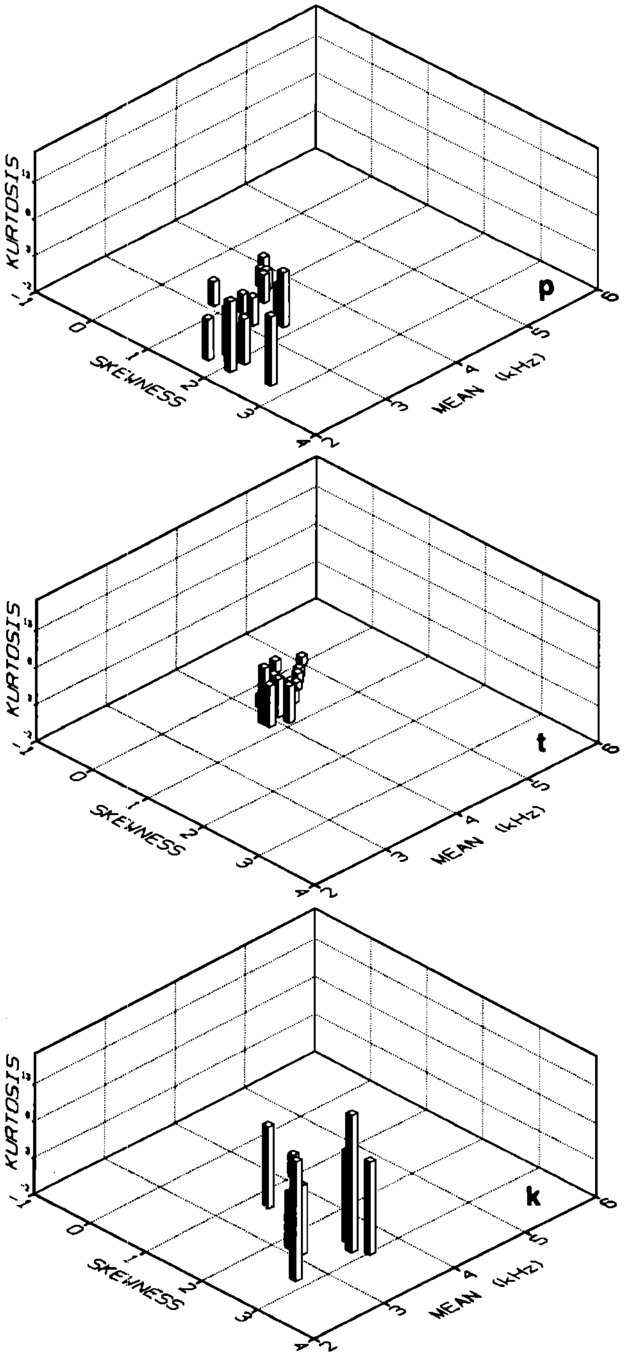

FIG. 2. The mean frequency, skewness, and kurtosis are displayed for each voiceless stop for the burst interval. The moments plotted on this graph were calculated from linear spectra.

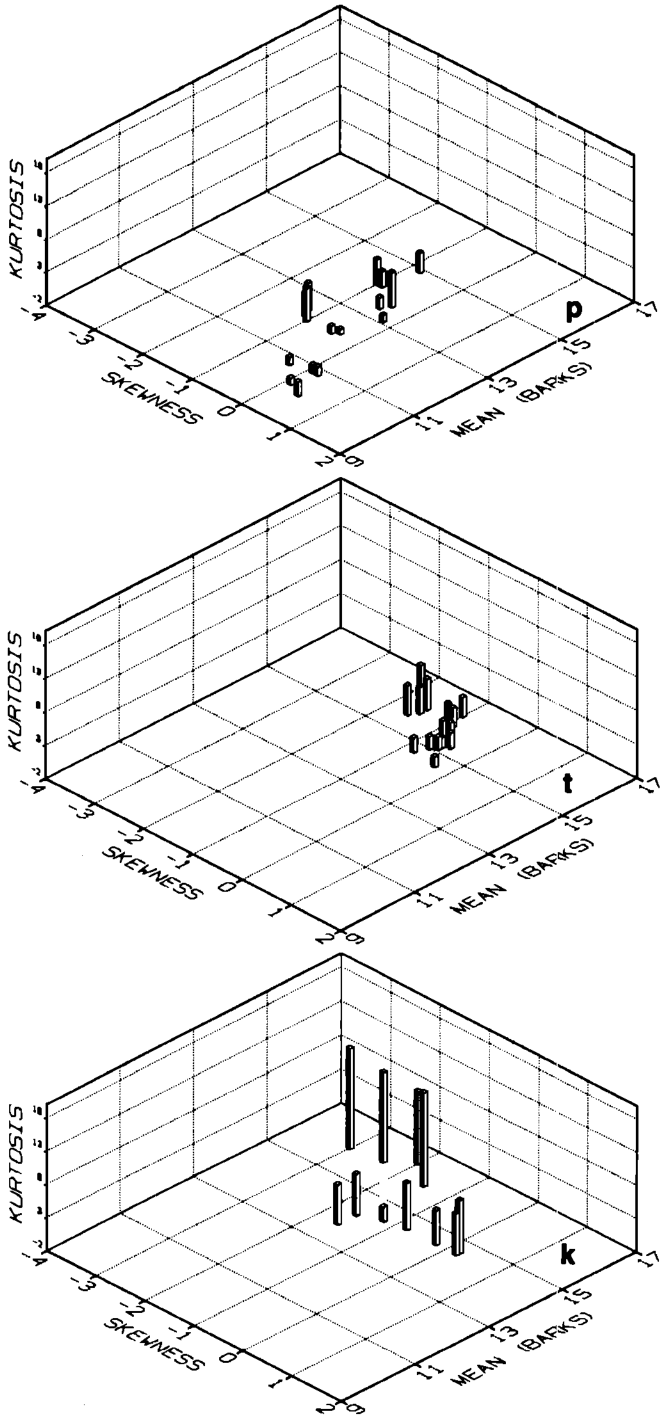

FIG. 3. Mean frequency, skewness, and kurtosis are plotted for the Bark transformed spectra. Data for each voiceless place of articulation are plotted.

tion, but perhaps on different dimensions. For example, skewness is not as useful in distinguishing /p/ from /t/ on the Bark scale, although means remain distinct for these two places. Kurtosis, however, is still the primary feature that distinguishes the velar stops from the other places of articulation.

Graphic representation of the moments is useful if discrete areas in the three-dimensional space can be ascribed to distinct obstruents. For example, if a unique space could be defined for each stop, phonemic accuracy could be determined. Productions that fell outside of the target space would indicate phonemic errors. Unfortunately, we were unable to create unique areas with the moments data since only three variables could be viewed at one time, whereas earlier studies (Kewley-Port and Luce, 1984) would suggest that stop consonants can best be viewed in multidimensional space.

Perhaps a better way to determine whether word-initial obstruents can be differentiated by the linear and/or Bark moments is to perform a discriminant analysis. As stated

TABLE II. Classification of voiceless stops from moments calculated from first 10 ms, *re*: burst.

			Males			Females			
		Percent	Number of cases			Percent	Number of cases		
	Phoneme	correct	p	t	k	correct	p	t	k
Linear	p	84.0	74	10	4	88.7	79	2	8
	t	77.7	11	63	7	88.6	7	28	3
	k	78.0	11	7	64	83.9	6	7	68
Bark	p	86.4	76	12	0	82.0	73	11	5
	t	90.1	4	73	4	83.1	10	74	5
	k	67.1	9	18	55	70.3	9	15	57

earlier, investigators have shown that dynamic aspects of the onset spectra are important in the description and perception of word-initial stops (Kewley-Port *et al.*, 1983). If the moments analysis is to be useful in the quantification of the spectral properties of these obstruents, classification performance should improve as moments from sequential temporal slices are added. We, therefore, performed discriminant analyses on the moments calculated in the interval surrounding the burst (burst to burst + 10 ms) and then performed additional analyses as we added the moments from spectral cross sections of successive intervals.

In our first analysis, the input variables were the mean, skewness, and kurtosis from one spectral cross section, covering the burst to burst + 10-ms interval for voiceless stops produced by the five male subjects. Table II presents the results of this analysis. On average, the linear moments correctly classified 79.9% of the voiceless stops, whereas 81.2% of the voiceless stops were categorized correctly using the Bark moments.

The data from the female subjects for this time interval were used to validate the classification functions derived from the male data. This procedure provides an empirical test of the validity of the discrimination of the classifying variables, in this case the voiceless stops. On the linear scale, the male categories correctly classified 87.1%, on average, of voiceless stops produced by the female subjects. The male categories derived from the Bark moments correctly classified 78.5% of the voiceless stops produced by the females. This validation procedure suggests that the linear moments provide a better model for the classification of voiceless stops.

The addition of a second cross section, thereby including spectral information from the first 20 ms of the stop, improved the classification accuracy for males based on the linear moments but did not affect the overall classification from the Bark moments. On the linear scale, 88.8% of the voiceless stops were, on average, correctly classified, while the Bark moments correctly classified 82.6% of the voiceless stops.

The validity of the classification of voiceless stops from the linear moments from the first 20 ms (*re*: burst) was tested with the data from the female subjects. On average, the classification functions derived from the male speakers' linear moments correctly categorized 91.2% of the females voiceless stops. While there was no improvement in the classification of the stops using the Bark moments from the first 20 ms of the VOT interval, the validity of the classification functions improved, compared to the model obtained with the first 10 ms of the stop. The classification functions derived from the male Bark moments from the first 20 ms of the stop correctly classified 85.7% of the female stops. The percent correct classification of each stop is presented in Table III.

In an effort to improve classification, we added two more spectral cross sections and repeated the discriminant analysis of the males' data. Based on the linear moments from the first 40 ms, 95.4% of the /p/'s, 88% of the /t/'s, and 92.6% of /k/'s were classified correctly. Classification of the females' voiceless stop consonants based on the males' discriminant functions was correct for 90.5% of the /p/'s, 96.5% of the /t/'s, and 93.6% of the /k/'s. The /p/'s and /t/'s were misclassified as /k/ (i.e., no overlap between /p/ and /t/ categories), whereas /k/ was misclassified as either /p/ or /t/.

TABLE III. Classification of voiceless stops from moments calculated from first 20 ms, *re*: burst.

			Males			Females			
		Percent	Number of cases			Percent	Number of cases		
	Phoneme	correct	p	t	k	correct	p	t	k
Linear	p	89.7	79	3	6	87.6	78	3	7
	t	91.3	3	74	4	100	0	89	0
	k	85.3	7	5	70	86.1	6	5	70
Bark	p	94.3	83	5	0	89.8	80	7	2
	t	81.5	10	66	5	92.1	2	82	5
	k	72.0	3	20	59	75.3	2	18	61

There was little improvement in the classification functions based on the Bark transformed moments when additional spectral cross sections were included. Classification of the stops based on the Bark moments from the first 40 ms of the VOT interval yielded, on average, 85.7% correct classification. When the female data were classified by these functions, classification improved to approximately 89% correct. On the Bark scale, classification errors caused /p/ and /t/ to be confused, while /k/ was consistently misclassified as /t/.

From these analyses, it appears that the voiceless stops can be discriminated well by the mean, skewness, and kurtosis calculated from linear FFTs over the first 40 ms of the VOT interval. Group means of the discriminant functions calculated from the linear moments for the first 40 ms, and the distance of all points from those means, are plotted for the first two canonical variates in Fig. 3. *Post hoc* inspection of the canonical weights indicates that the first canonical variate relates to the mean frequency of the second and later spectral cross sections. This would include the first 40 ms of the VOT interval exclusive of the burst itself. The second canonical variate, as determined by the second highest canonical weights, is an index of the skewness and kurtosis of all spectral cross sections investigated. This variate includes the burst through the first 40 ms of the stop.

Classification of the fricatives when all voiceless fricatives were included in the analysis was not as successful as the classification of the voiceless stops. When the linear moments from only the first 10-ms spectral cross section of the males' fricatives were used, correct classification ranged from 41.4% for /s/ to 71.4% for /f/. The addition of a second spectral cross section did not improve classification (41.4% for /s/ to 74.5% for /f/). The Bark moments from the first 20 ms of the fricatives yielded slightly better classification than the linear moments, but was still a rather poor estimate of the fricative categories (58.3% for /θ/ to 75.4% correct classification of /f/). The addition of two more spectral cross sections, thereby attempting classification on the basis of the first 40 ms of the fricatives, did not improve overall performance for either the linear or Bark moments. The Bark moments did, however, provide a better classification than the linear moments (average correct classification of 77.7% for the Bark moments versus 74.5% correct for the linear moments).

Since /θ/ and /f/ may be discriminated by information in the transition region rather than by the noise itself (Harris, 1958), we attempted a discriminant analysis on the Bark moments from the first 40 ms plus the moments from a 20-ms interval of the transition. While the addition of this transition information improved categorization to an average of 80.4% correct, only 61% of the /θ/'s were correctly categorized. The remaining 39% were classified as /f/. Validation of this model with the female data was not attempted since the model derived from the male data performed so poorly.

Figure 4 provides a representation of the overlap of the phoneme groups based on the first two canonical variates calculated from this last discriminant analysis. It can be seen that there is little overlap between the two sibilants, while the classes of the less intense fricatives almost completely overlap one another. When either sibilant was misclassified, the error was most often in classifying the sibilant as either /f/ or /θ/. This result suggests that all relevant variables used to discriminate the sibilants from the other fricatives may not have been represented in our analysis. For example, if we included fricative intensity in our analysis, the classification errors of the sibilants might disappear.

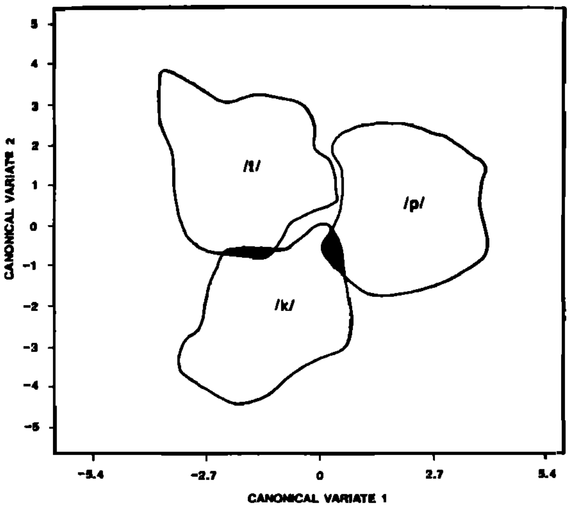

FIG. 4. Cluster centers, marked by the appropriate phoneme, and boundaries enclosing all voiceless stops are plotted for the first two canonical variates. Shading represents areas of classification ambiguity.

Given the poor discrimination of /f/ and /θ/ combined with the type of classification errors of the sibilants (i.e., sibilants classified as nonsibilants), discriminant analyses were performed on the sibilants alone. When the moments from the first 20 ms (i.e., two spectral cross sections) of the males' sibilants were included in the analysis, average discrimination improved to 82.7% when the linear moments were used and 98.3% corect for the Bark moments. Further, the function derived from the Bark moments correctly classified 95% of the sibilants produced by the female subjects. Contingency tables from these analyses are presented in Table IV.

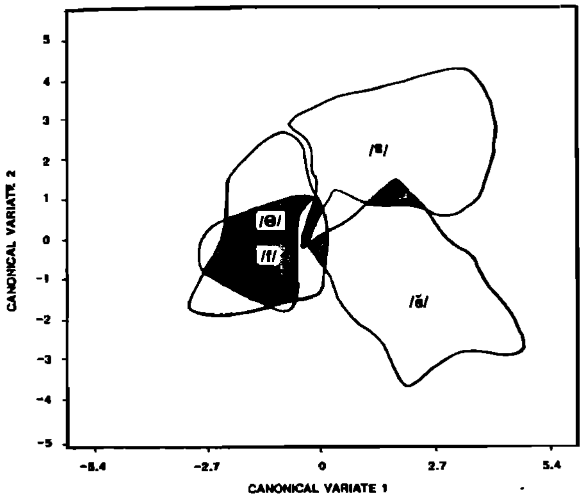

FIG. 5. Cluster centers, marked by the appropriate phoneme, and boundaries enclosing all voiceless fricatives are plotted for the first two canonical variates. Shaded regions indicate cluster overlap.

TABLE IV. Classification of voiceless sibilants from moments calculated from first 20 ms, *re*: frication onset.

		Males			Females		
		Percent	Number of cases		Percent	Number of cases	
	Phoneme	correct	s	š	correct	s	š
Linear	s	70.0	21	9	100	30	0
	š	90.0	3	27	100	0	30
Bark	s	96.6	29	1	93.3	28	2
	š	100	0	30	96.7	1	29

Post hoc inspection of the weightings of the variables used in the discriminant functions indicates that the sibilants were discriminated primarily on the basis of skewness. Further, this discrimination improved when the Bark moments were used.

In summary, the voiceless stops could be classified with an average of 92% accuracy from the linear moments derived from the first 40 ms of the voiceless stops. Further, the classification functions can be generalized across gender. Correct classification of 98% of the sibilants resulted from the discriminant analysis of the Bark moments obtained from the first 20 ms of the consonants. Again, these effects hold across gender.

II. DISCUSSION

The results of the present experiment demonstrate that a quantitative procedure can be applied to the classification of word-initial voiceless obstruents. Compared to qualitative descriptions of obstruent features (Kewley-Port, 1983; Blumstein and Stevens, 1979, 1980), the moments analysis described in this article provided greater classification accuracy. For example, Kewley-Port and Luce (1984) found that sophisticated judges could classify voiceless stops with 89% accuracy. In the present experiment, 92% of the voiceless stops were correctly categorized using linear moments from the first 40 ms of the VOT interval.

Performance on the Bark scale was similar to the accuracy noted by Kewley-Port and Luce (1984) but inferior to the linear moments in the classification of the voiceless stops. This is in contrast to Kewley-Port's (1983) findings that the "auditory filter representation" of LPC spectra for voiced stops "appeared to display place of articulation more successfully" (p. 332) than the linear spectra. It may be that the filter bandwidths used in the present investigation were sufficiently different from Kewley-Port's to account for the different results. The Bark scale is essentially a log-frequency scale, thereby providing a wider analysis band in the high-frequency compared to the low-frequency region. This causes a greater weighting of energy in the higher frequencies and an upwards shift in the spectral peaks. The effect was to distort the indices of specral shape, namely, skewness and kurtosis, since the Bark spectra for all voiceless stops have a strong negative slope and little dispersion around the mean frequency. It is possible that the narrower bandwidths used by Kewley-Port did not cause as much distortion as was seen in the Bark scale used in the present work. Also, Kewley-Port used LPC spectra as the basis of analysis. The LPC filtering prior to "auditory filtering" may have reduced the excessive high-frequency weighting that distorted shape indices.

On the linear scale, the important variables for the classification of stops were the mean frequency of the frication interval after the burst and the shape of the spectra from all temporal intervals including the burst. It seems reasonable, on psychoacoustic grounds, that the mean frequency of the burst would not be relevant to the discrimination of voiceless stops. The burst for stop consonants is of such short duration that frequency discrimination would be extremely poor, if not impossible. For this reason, the burst is probably best described as noise with little frequency specificity.

A high rate of classification accuracy was also found for the voiceless sibilants. Unlike the stops, however, the sibilants were discriminated better on the Bark scale than on the linear scale. The superiority of the Bark over the linear scale in the classification of sibilants has been demonstrated by Bladon and Seitz (1986). They found that voiceless sibilants were discriminated best by the slope of the Bark transformed spectra. Our results, which demonstrated that skewness was the most important moment used to discriminate /s/ from /š/, are consistent with Bladon and Seitz's data.

Neither the linear nor the Bark moments provided accurate classification of /f/ and /θ/. In fact, nearly half of the /θ/ tokens were misclassified as /f/. One explanation for this poor classification is that /f/ and /θ/ are, preceptually, among the most confusable consonants (Miller and Nicely, 1955). One might, therefore, infer that the acoustic cues for these two phonemes are not highly distinctive. It is possible that classification accuracy of these phonemes could be improved with different analysis parameters. For example, distinctive spectral information may reside in the later portion of these less intense fricatives. Since our analysis concentrated on phoneme onsets (the first 40 ms), we may have failed to capture salient spectral differences. The addition of limited information from the transition region (i.e., 20 ms prior to the vowel onset), may have been inadequate to improve classification. Alternatively, a different window size may have yielded better spectral differentiation of /f/ and /θ/. We chose to use a 20-ms Hamming window for all of our analyses. Since fricatives are characterized by a relatively stationary articulatory configuration, a larger window could have been used. This may have provided greater frequency specificity, which may have aided in differentiating /f/ from

$/\theta/$. The effect of these changes on classification of the fricatives is being explored.

One of the most striking features of the present data is that the classification functions of the voiceless obstruents could be generalized across gender. On average, categorization of females' obstruents was accomplished with 96% accuracy using the same model established from the males' moments data. To the best of our knowledge, this is the first demonstration of a high rate of cross-gender classification in the absence of any additional normalization procedure. It suggests that a quantitative procedure may provide scale-independent shape information upon which spectral classification can be made.

As we have indicated, these results are preliminary in the sense that the speech sample was rather limited. It is possible that the high degree of classification accuracy may diminish when additional sources of variability (e.g., more vowel contexts, changes in rate) are introduced. However, the data of Forrest *et al.* (1987) suggest that this is not that case, at least for male speakers. The improvement in classification accuracy in the present study compared to the Kewley-Port's and Blumstein and Stevens' work may also be attributable to differences in sampling rate; that is, both Kewley-Port and Blumstein and Stevens sampled the speech signal at 8 kHz. In the present investigation, a sampling rate of 20 kHz was employed. It is possible that there is information in the higher frequencies that aids classification of voiceless obstruents. These possibilities are being explored in ongoing research.

Finally, the application of discriminant function analysis to the continuous distributions of the spectral moments may hold promise as a technique for indexing the distance between a segmental error and a target phoneme. The output of the discriminant function analysis includes a metric (Mahalanobis D^2) that describes the distance of each item (phone) from the mean of the cluster. This metric may be particularly useful when error sounds are not easily placed into a phoneme category. If the appropriate acoustic-perceptual studies show that these ambiguous sounds can be described meaningfully by Mahalanobis D^2 values, we may have a means for application of spectral classification systems to the description of disordered speech, which is, in fact, our goal.

ACKNOWLEDGMENTS

The authors would like to thank Deborah Hoffman for her careful editing and processing of this article. We also acknowledge Patricia Zander and Shirley Hunsacker for their assistance with the figures. This project was supported, in part, by NINCDS Grants NS-13274 and NS-21516.

Bendat, J. S., and Piersol, A. G. (**1980**). *Engineering Applications of Correlation and Spectral Analysis* (Wiley, New York).

Bladon, A., and Seitz, F. (**1986**). "Spectral edge orientation as a discriminator of fricatives," J. Acoust. Soc. Am. Suppl. 1 **80**, S18–S19.

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.

Blumstein, S. E., and Stevens, K. N. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

Butler, J. P., and Mohler, J. G. (**1979**). "Estimating a distribution's central moments: A specific tidal ventilation application," J. Appl. Physiol. **46**, 47–52.

Cohn, A. C. (**1987**). "Quantitative characterization of degree of coarticulation in CV tokens," J. Acoust. Soc. Am. Suppl. 1 **82**, S115.

Erdreich, J. (**1986**). "A distribution based definition of impulse noise," J. Acoust. Soc. Am. **79**, 990–998.

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

Forrest, K., Weismer, G., and Milenkovic, P. (**1987**). "Statistical representation or word-initial obstruents: Further data," J. Acoust. Soc. Am. Suppl. 1 **82**, S84.

Harris, K. S. (**1958**). "Cues for the discrimination of American English fricatives in spoken syllables," Lang. Speech **1**, 1–7.

Hary, D., Belman, M. J., Propst, J., and Lewis, S. (**1982**). "A statistical analysis of the spectral moments used in EMG tests of endurance," J. Appl. Physiol. **53**, 779–783.

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

Kewley-Port, D., and Luce, P. A. (**1984**). "Time-varying features of initial stop consonants in auditory running spectra: A first report," Percept. Psychophys. **35**, 353–360.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (**1983**). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **73**, 1779–1793.

Kobatake, H., and Ohtani, S. (**1987**). "Speech transition dynamics of voiceless stop consonants," J. Acoust. Soc. Am. **81**, 1146–1151.

Lansky, P., and Radil, T. (**1987**), "Statistical inference on spontaneous neuronal discharge patterns," Biol. Cybernet. **55**, 299–311.

Milenkovic, P. M. (**1986**). "Glottal inverse filtering by joint estimation of an AR system with linear input," IEEE Trans. Acoust. Speech Signal Process. **34**, 28–42.

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Newell, K. M., and Hancock, P. A. (**1984**). "Forgotten moments: A note on skewness and kurtosis as influential factors in inferences extrapolated from response disributions," J. Motor Behav. **16**, 320–335.

Shapiro, S. S., and Wilk, M. B. (**1965**). "An analysis of variance test for normality (complete samples)," Biometrika **52**, 591.

Shinn, P., and Blumstein, S. E. (**1983**). "Phonetic disintegration in aphasia: Acoustic analysis of spectral characteristics for place of articulation," Brain Lang. **20**, 90–114.

Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.

Stevens, K. N., House, A. S., and Paul, A. P. (**1966**). "Acoustical description of syllabic nuclei: an interpretation in terms of a dynamic model of articulation," J. Acoust. Soc. Am. **40**, 123–132.

Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.

Ziegler, W. (**1984**). "What can the spectral characteristics of stop consonants tell us about the realization of place of articulation in Broca's aphasia? A reply to Shinn and Blumstein," Brain Lang. **23**, 167–170.