Perception & Psychophysics
1984, 35 (4), 353-360

Time-varying features of initial stop consonants in auditory running spectra: A first report

DIANE KEWLEY-PORT and PAUL A. LUCE
Indiana University, Bloomington, Indiana

Kewley-Port (1983) recently demonstrated that place of articulation of initial voiced stops could be identified from time-varying features observed in visual displays of linear prediction smoothed spectra. The present study extends this method of analysis in several directions. First, both voiced and voiceless syllable-initial stops produced at three speaking rates—normal, fast, and slow—were examined. Second, a new rule for vocal tract size normalization was tested. Third, the earlier time-varying features were augmented in order to specify the burst and voicing as well as place of articulation. The four time-varying features were (1) an abrupt increase in energy at high frequencies, (2) the onset of a prominent low-frequency peak, (3) the relative tilt of voiceless energy at onset, and (4) the presence of extended midfrequency peaks. Finally, the visual displays were modified to incorporate filtering and other characteristics of processing of speech by the auditory system. Auditory running spectra were generated for stop consonant-vowel syllables read by two males and two females. Employing the four time-varying features, judges first located the burst and onset of voicing, and then identified place of articulation from the visual displays. Over all conditions, place of articulation was identified at an 86% level of accuracy. While these results constitute only a first step towards an automated analysis procedure, they nonetheless indicate that our new time-varying features are appropriate for identifying place of articulation across both voiced and voiceless stops produced by different speakers at different speaking rates.

Specifying the invariant acoustic correlates of place of articulation in initial stop consonants has been an extremely difficult problem in the field of speech research. Recently, however, Stevens and Blumstein (1978; Blumstein & Stevens, 1979) and Kewley-Port (1980, 1983) have demonstrated that invariant acoustic correlates of place of articulation can be found in the first 20 to 40 msec of the speech waveform following the onset of burst release. Both Stevens and Blumstein and Kewley-Port propose that these correlates, described in acoustic terms, are direct consequences of articulatory gestures as predicted by the acoustic theory of speech production (Fant, 1960; Stevens, 1975, 1980; Stevens & Blumstein, 1978). Stevens and Blumstein and Kewley-Port have hypothesized that these acoustic properties are relational invariants in the sense that they specify place of articulation for initial stop consonants irrespective of differences in vowel context, voicing, talker characteristics, and other sources of phonetic variation. However, these researchers disagree on whether the invariant acoustic correlates of place of articulation are best described as *static* or *dynamic*.[1]

Stevens and Blumstein have argued that static integrated acoustic cues found in the overall gross shape of the spectrum at the onset of the release burst specify place of articulation. Kewley-Port (1983), on the other hand, has argued that cues to place of articulation lie in the dynamic changes in spectral energy over time observed in the first 40 msec beginning with the release burst. A recent perceptual study by Kewley-Port, Pisoni, and Studdert-Kennedy (1983) compared these two approaches and found empirical support for the claim that cues to place of articulation were dynamic and not static. In addition, both Blumstein (see Lahiri & Blumstein, 1981) and Stevens (see Ohde & Stevens, 1983) have apparently modified their earlier views and have now adopted a description of the acoustic correlates that is more dynamic in nature.

The present study was designed to extend the earlier work of Kewley-Port (1980, 1983) in a number of ways to specify more precisely the nature of the acoustic correlates to place of articulation in initial stop consonants. First, we collected a new data base from

This is an expanded version of a paper presented at the 103rd meeting of the Acoustical Society of America, Chicago, April ,1982. This work was supported by NINCDS Research Grant NS-12179 to Indiana University in Bloomington. We would like to thank David B. Pisoni for his advice and assistance at various stages of this research and Thomas D. Carrell and Amanda C. Walley for serving as judges. We would also like to thank two anonymous reviewers for their insightful comments and helpful suggestions on an earlier version of this paper. Reprint requests may be sent to Diane Kewley-Port, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405.

Copyright 1984 Psychonomic Society, Inc.

two male and two female talkers that included both voiced and voiceless initial stop consonants. Second, in a preliminary attempt to generalize Kewley-Port's findings to connected speech, we collected data on consonant-vowel syllables spoken in a carrier phrase at three speaking rates: normal, fast, and slow. Third, we addressed the problem of defining the acoustic correlates that are invariant across talkers as well as vowel contexts by implementing a simple rule to account for vocal tract size variation. Fourth, we modified the analysis of the spectral characteristics of the waveform in order to more closely model the transformations of frequency and energy assumed to take place in the auditory system.

In a previous study, Kewley-Port (1983) used running spectral displays for consonant-vowel syllables that were calculated from linear prediction coefficients (Markel & Gray, 1976). In the present study, we constructed new displays that incorporated filtering characteristics of the auditory system derived from psychoacoustic research (see Patterson, 1976). The reason for changing our analysis to model these auditory properties was motivated by the belief that applying auditory processing models to the analysis of speech would improve our understanding of speech perception (cf. Carlson & Granstrom, 1982; Delgutte, 1980; Klatt, 1979). That is, employing a model of the auditory system to extract the frequency, time, and intensity parameters of speech might provide new insights into identifying the perceptually important properties of speech stimuli. Specifically, the present study examined the proposal made by Kewley-Port (1983) that many of the details of the description of the acoustic correlates of place of articulation would be enhanced with auditory filtered displays.

The purpose of the present investigation was to determine if new time-varying features modified for auditory running spectra would adequately specify both place of articulation and location of stop bursts from visual displays. Moreover, we were also interested in determining whether the underlying acoustic correlates were invariant across talkers, speaking rates, and vowel contexts. Because of the exploratory nature of this research, judging of the visual displays was performed by human observers in order to evaluate the potential success of this approach. The experiment consisted of two parts. In Part 1, we examined only the voiceless stop consonants /p/, /t/, /k/. In Part 2, we examined both voiced and voiceless stop consonants produced at three different speaking rates.

METHOD

Data Collection

The stimuli for both parts of the experiment were collected in a single session for each talker. The talkers were recorded while they read 30 stop consonant-vowel syllables embedded in the carrier phrase "Teddy said ______." The syllables consisted of all combinations of /b,d,g,p,t,k/ paired with the vowels /i,æ,ɑ,ɔ, u/. The 30 test sentences were randomized and presented one at a time on a CRT monitor under computer control. The talkers were recorded in a sound-treated booth (IAC Model 401A) using an Electro-Voice D054 microphone and an Ampex AG-500 tape recorder. Each talker read five blocks of the 30 sentences at a normal rate. In addition, one male and one female talker read five blocks of the same stimuli at both a fast rate and a slow rate. Because the normal rate condition was recorded first, no special instructions concerning speaking rate were given. For the fast rate, the subjects were instructed to speak as rapidly as possible without misarticulating. For the slow rate, the subjects were asked to slow their speaking rates by lengthening the words and not by inserting pauses.

Data analysis

Tokens from each talker from the third and fourth blocks of sentences were low-pass filtered at 4.8 kHz. The tokens were then digitized at a 10-kHz sampling rate via a 12-bit analog-to-digital converter using a PDP-11/34 computer. When a token was deemed unacceptable because of a mispronunciation or because of excessive noise in the signal, another token of the same utterance was taken from the fifth block of sentences. Mispronunciations were rare and occurred mostly at the slow speaking rate. All together, 288 tokens were selected for both parts of the experiment.

After digitizing, the stop consonant-vowel syllables were digitally spliced out of the carrier sentences for subsequent analysis. To produce the auditory running spectral displays, the following procedures were used to simulate the filtering by the auditory system. First, linear prediction analysis was carried out using the program SPECTRUM (Kewley-Port, 1979). Twenty-millisecond waveform segments were first-differenced to remove glottal source and lip impedance characteristics. Using a Hamming window, 14 autocorrelation coefficients were then calculated. The size of the window was 20 msec, except for a few low-pitch tokens from the second male talker for which a 25-msec window was used.[2] Smoothed spectral sections, or "frames," were calculated and updated every 5 msec. The first spectral section was positioned during the stop closure such that between one and six frames of closure preceded the burst frame. The number of closure frames was chosen randomly for each syllable. Altogether, 15 frames (70 msec) were analyzed for each syllable.

In an earlier study, one of us argued that representing energy in decibels and updating the spectral frames every 5 msec in the visual displays was appropriate for modeling the energy and temporal transformations of speech in the auditory system (see Kewley-Port, 1983). However, the linear prediction filters do not appropriately model the frequency analysis carried out by the auditory system. It is well known that the bandwidth of the frequency analysis performed by the ear is not equal across all frequencies, but instead is constant below 500 Hz and then increases as frequency increases (Scharf, 1970). Various estimates of auditory bandwidth, or critical bands, have been made in the past, and they range from at most 1/2 octave bandwidths to 1/10 octave bandwidths. To simulate auditory filtering in this study, 1/6 octave bandwidths were chosen on the basis of Patterson's results (Patterson, 1976; Patterson & Nimmo-Smith, 1980). The filter shape was trapezoidal and had a 75-dB/octave roll-off in the skirts. These filters were convolved with the linear prediction smoothed spectra to produce auditory filtered spectra. A sufficient number of overlapping filters were used to produce the smoothed spectra shown in Figure 1.

Finally, the frequency scale was changed from a linear scale in hertz to a technical mel scale to model more closely the frequency selectivity of the auditory system (see Fant, 1960, p. 241). Auditory filtered running spectra were produced in this way for each syllable.

Four time-varying features were formally defined for the auditory filtered running spectral displays. We will briefly summarize the features here and discuss them in detail in the following section. The first feature—the occurrence of an abrupt increase in energy at high frequencies—was defined to permit identification

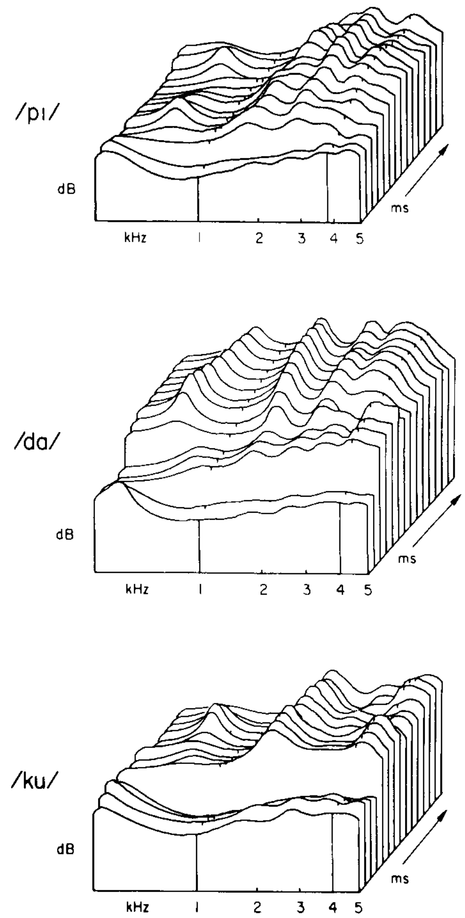

Figure 1. Auditory filtered running spectral displays for /pi/, /da/, and /ku/. Frequency in kilohertz is represented on the x axis; relative decibels are represented on the y axis. The spectral frames are offset along the z axis in 5-msec steps.

of the burst frame. Stevens initially proposed that this feature, which is based on dynamic changes in energy, is a correlate of an abrupt consonantal release (cf. Stevens, 1980; Stevens & Blumstein, 1981). This second feature—the onset of a prominent, narrow, low-frequency peak that is continuous with succeeding frames—was defined to permit identification of the onset of voicing.

The third and fourth features were defined in such a way as to reflect the proposed invariant properties that are correlates of place of articulation. These features were modifications of the features proposed in Kewley-Port's (1983) earlier study.[3] The third feature defined the relative tilt of the burst and voiceless frames for up to 40 msec (8 frames) of voiceless spectra. Relative tilt of the burst spectra has been proposed as a correlate of place of articulation; a rising tilt is correlated with alveolars and a flat or falling tilt is correlated with bilabials (Fant, 1960; Stevens & Blumstein, 1978). In an earlier study, which examined only voiced stops, tilt from low to high frequencies was estimated as a visual regression line through only the burst spectrum (see Kewley-Port, 1983). In this study, however, tilt was judged over a variable number of frames because pilot work had indicated that when many voiceless frames were present, more than a single spectral frame should be used to estimate tilt. Nevertheless, tilt was judged over *only* the first 40 msec of voiceless spectra on the basis of Kewley-Port et al.'s (1983) finding that accurate place identification was possible for stops truncated at 40 msec (see also Tekieli & Cullinan, 1979).

The fourth feature was the presence or absence of prominent midfrequency peaks extending in time, a definition taken directly from Kewley-Port (1983). The presence of "compact," midfrequency energy at the burst release has long been proposed as an acoustic correlate of velar place of articulation (see Fant, 1960; Jakobson, Fant, & Halle, 1952; Stevens & Blumstein, 1978).

Judging

To illustrate how these features were used in judging the displays, we will refer to Figure 1, where examples of the running spectral displays for the syllables /pi/, /d*a*/, and /ku/ are shown.

Two graduate students in psychology with formal training in phonetics served as judges in this experiment. Judges received about 15 min of supervised training in identifying the features from eight sample running spectra. For each display, the judges were instructed to locate the burst frame and the onset of voicing and then to decide on the spectral tilt and the presence or absence of midfrequency peaks. To locate the burst frame, the judges were instructed to find the first frame in which there was an obvious increase in high-frequency energy and to record the number of that frame on their response sheets. In Figure 1, the burst occurs in the third frame for /pi/ and /d*a*/ and in the fifth frame for /ku/.

Next, the judges indicated the number of the first frame in which a low-frequency peak became prominent and remained continuous for the remaining frames. This frame designated the onset of voicing. In Figure 1, the onset of voicing for /d*a*/ begins in the seventh frame. Although /pi/ and /ku/ show low-frequency peaks, the peaks do not continue in the remaining frames. Thus, according to our definition, no onset of voicing is present in these displays. In these cases, the judges recorded an "N" (for "not voiced"). When low-frequency peaks began in the closure frames and continued for the remaining frames, the judges recorded a "C" (for "continuous voicing").

In this manner, the burst frame and up to eight succeeding voiceless frames were identified. Judges then decided on the spectral tilt of these frames, or only the burst frame when voicing was continuous. They recorded an "F" if the spectral tilt of the burst and voiceless frames was flat, falling, or slightly rising toward higher frequencies and an "R" if the frames were distinctly rising. Spectral tilt was subsequently used to assign place to the bilabials and alveolars.

Finally, the judges were required to identify the presence or absence of a prominent midfrequency peak starting on the burst frame and extending for three or more frames. The ticks on the display demarcate the midfrequency region (see /ku/ in Figure 1). The ticks were positioned on the display according to a vocal tract normalization rule proposed in the previous study (Kewley-Port, 1983). This rule was derived from Fant's (1960, 1973) proposal that the spectral peak for velars is associated with F2 for low, back vowels and F3 or F4 for high, front vowels. Thus, the low-frequency tick in these displays was positioned at the talker's F2 for /u/ and the high-frequency tick was positioned at the F2 value of /i/ plus 500 Hz.

After judging the four time-varying features, the assignment matrix shown in Table 1 was used to determine place of articulation. If tilt was judged as relatively flat and no midfrequency peaks were observed, the display was labeled as a bilabial. If the tilt was distinctly rising and no midfrequency peaks were observed, the display was labeled as an alveolar. If midfrequency peaks were present, the display was labeled as a velar regardless of spectral tilt.

Table 1
Assignment Matrix Used in Judging Place of Articulation

Tilt	Midfrequency Peaks	Place
F	N	Bilabial
R	N	Alveolar
*	Y	Velar

*Note—F = falling; R = rising, Y = yes; N = no; and * = falling or rising.*

Judging of the visual displays for both parts of the experiment was carried out in two sessions. Written definitions of the four time-varying features were available to the judges during both sessions. The judges independently determined the location of the burst frame, the onset of voicing, and place of articulation for each of the 288 randomly ordered displays. Responses from the independent judging were then scored. If there was disagreement between the judges on the assignment of place of articulation, the displays on which the judges disagreed were presented again in a second session in which the judges collaborated. If the judges were unable to agree on the assignment of any of the features in the collaborative judging session, they were instructed to mark on their response sheets that the feature was ambiguous.

RESULTS

The results for identification of place of articulation for the individual judging session showed that the judges performed surprisingly poorly. Place of articulation was correctly identified in only 51% of the displays. To determine the source of the errors, the feature assignments for each judge from the individual judging session were compared, as shown in Table 2. The comparison of feature assignments revealed that the judges disagreed most often on assigning spectral tilt to the voiceless consonants. Upon reexamination of the displays for the voiceless consonants, we discovered that the spectral tilt of the voiceless stops often changed from flat in the burst frame to distinctly rising over the first eight frames (40 msec) for /t/. Because the judges had occasionally overlooked this information, in the second, collaborative judging session, they were reminded to examine the tilt over all eight voiceless frames. In addition, judges were instructed to weigh later frames more heavily in ambiguous cases (e.g., 4 flat and 4 rising frames).

Table 2
Results for Two Judges Assigning Features and Place of Articulation to 288 Running Spectral Displays

Category	N	Percent Correct
Burst Onset		
Agreed	227	79
One Frame Difference	36	12
Other Differences	35	9
Voicing Onset		
Agreed	173	68
One Frame Difference	37	14
Other Differences	46	18
Tilt of Burst		
Agreed	150	65
Midfrequency Peaks		
Agreed	235	88
Place of Articulation		
Correct in Individual Judging	148	51
Correct in Collaborative Judging	248	86

Note—N is the number of displays in each category. Percent correct is N/(288 − missing cases).

In the collaborative judging session, 122 of the 140 displays classified incorrectly in the individual judging session were presented again. The other 18 displays were not presented for judging a second time because the experimenters agreed with the judges that these displays clearly contained the inappropriate features for identifying place. These displays were thus scored as errors. Of the 122 displays presented in the collaborative judging session, 92% were identified correctly. In 6% of the displays, place was incorrectly identified, and 2% of the features were judged to be ambiguous. Collapsing across both judging sessions, correct place identification rose to 86%.[4]

Further analysis of these results is broken down in terms of the two parts of the experiment, the first for voiceless stops and the second for speaking rate differences. Table 3 shows the results for Part 1 of the experiment, which evaluated the new features for identifying place of articulation only for voiceless stops at the normal speaking rate. Two tokens from four talkers of all possible combinations of /p,t,k/ and /i,æ,ɔ,u/ resulted in 96 voiceless stop-vowel syllables. A three-way analysis of variance (talker × consonant × vowel) showed that the identification of place of articulation was not significantly different across talkers [$F(3,8) = 1.0$, $p > .4$], although the percent correct identification for the second male talker was lower than for the other three talkers. Place identification across consonants was essentially equivalent ($F < 1.0$). Vowel context, however, produced a significant difference [$F(3,8) = 5.67$, $p < .02$] because of the poor identification performance on the vowel /ɔ/. No interactions were significant. Overall, 89% of the voiceless stops were identified correctly.

The second part of this study was designed to examine some of the conditions found in fluent speech, namely the effects of speaking rate on the acoustic correlates of place of articulation. The experimental variables and results for Part 2 are shown in Table 4.

Running spectral displays were constructed from two tokens produced by one male and one female talker of all combinations of /p,t,k,b,d,g/ and /i,ɑ,u/ spoken at three different rates (fast, normal, and slow), resulting in 216 syllables. (Twenty-four of the normal-rate syllables were also included in the voiceless results shown in Table 3). A five-way analysis of variance (talker × consonant × vowel × rate × voice) showed that place of articulation was identified equally well across talkers ($F < 1.0$), consonants ($F < 1.0$), vowel context ($F < 1.0$), and rate ($F < 1.0$). Although the normal speaking rate produced better identification, it was not significantly different from the fast and slow rates.

Voiced stops were, however, identified more poorly than voiceless stops [$F(1,8) = 8.3$, $p < .01$] across all three speaking rates. On closer examination of the data, we found this result to be caused by signif-

Table 3
Percent Correct Place Identification for 96 Voiceless Stops in Part 1 of the Experiment Across Talkers, Consonants, and Vowels

Variable		Percent Correct Place Identification	Significance Level
Talker	Male 1	92	n.s.
	Male 2	79	
	Female 1	92	
	Female 2	92	
Consonant	/p/	91	n.s.
	/t/	87	
	/k/	87	
Vowel	/i/	96	$p < .02$
	/æ/	96	
	/ɔ/	71	
	/u/	92	
Total		89	

Table 4
Percent Correct Place Identification for 216 Stop Consonant-Vowel Syllables in Part 2 of the Experiment Across Talkers, Rate, Consonants, Voicing, and Vowels

Variable		Percent Correct Place Identification	Significance Level
Talker	Male 1	84	n.s.
	Female 1	88	
Rate	Normal	90	n.s.
	Fast	85	
	Slow	83	
Place	Bilabial	89	n.s.
	Alveolar	86	
	Velar	83	
Voicing	Voiced	80	$p < .01$
	Voiceless	93	
Vowel	/i/	90	n.s.
	/a/	85	
	/u/	83	
Total		86	

icantly poorer identification of the voiced stops at the fast and slow rates. Eighty-nine percent of the voiced stops were identified correctly at the normal rate, but only 75% of the voiced stops were identified correctly at the fast and slow rates. Rate had no effect, however, on the identification of place for the voiceless stops. Thus, the two-way interaction of rate by voicing was not significant. Overall, 86% of the displays were identified correctly.

Two of the higher order interactions were also significant: talker × consonant [$F(2,25) = 7.3$, $p < .01$] and consonant × vowel [$F(4,25) = 2.7$, $p < .03$]. Analysis of the data revealed that, for the male talker, identification of place of articulation was better for bilabials than for alveolars and better for alveolars than for velars. The reverse was true for the female talker. Analysis of the consonant × vowel interaction revealed lower identification performance for alveolars preceding */a/* than for the other consonant-vowel syllables. The underlying causes of these interactions is not apparent at this time.

The results from the present experiment may be compared with those of two similar studies by Blumstein and Stevens (1979) and Kewley-Port (1983) by considering only the results from stops produced at the *normal* speaking rate (see Table 4). Place-of-articulation identification was improved from 84% reported by Blumstein and Stevens and 88% reported by Kewley-Port to 90% reported here. These results indicate that the modifications of the procedures originally proposed by Kewley-Port (1983), including the use of auditory filtered running spectra, have successfully improved identification of place of articulation for syllable-initial stop consonants.

DISCUSSION

The present study extends Kewley-Port's (1983) earlier investigation of place of articulation in a number of ways. First, we examined time-varying features for place of articulation of voiceless syllable-initial stop consonants; only voiced stops were examined in the previous study. Identification of place of articulation for the voiceless stops was quite good (89%) and proved to be consistent across talkers and consonants. Identification was not consistent across vowel contexts, however, due to poor identification performance on voiceless stops preceding the vowel /ɔ/. Nevertheless, we believe we have successfully demonstrated the adequacy of the proposed time-varying features for place of articulation for voiceless stops.

In a second extension of Kewley-Port's (1983) earlier work, we examined the effects of speaking rate on the acoustic correlates of place of articulation. Both voiced and voiceless stops were produced at three speaking rates: normal, fast, and slow. Although identification of voiceless stops was good across all three speaking rates, identification of voiced stops dropped below that for voiceless stops at the fast and slow speaking rates. Overall, however, speaking rate did not have a significant effect on identification performance, thus demonstrating that the proposed time-varying features should generalize to conditions approximating the variation observed in more naturally produced speech.

In the present study, more appropriate modeling of the frequency characteristics of the auditory system was employed in generating the visual displays of the running spectra. Such changes, however, apparently made the identification of the tilt of the burst more difficult, particularly for the voiced consonants. Presumably, some of these difficulties were related to the method used to produce the auditory filtered spectra (see Method section). Further research will employ more direct methods of deriving the auditory filtered display, such as those proposed by Klatt (1976, 1979) and Flanagan and Christensen (1980). For example, direct filtering would result in a shortening of the analysis time window for high-frequency energy. The onset of the high-frequency energy for alveolar bursts, therefore, would occur earlier than the onset of low-frequency energy, making the distinction between bilabial and alveolar bursts potentially more salient in terms of the relative tilt of the burst spectra.

A new rule for locating a talker's midfrequency range on the displays was also implemented. The new rule was derived from Fant's (1960, 1973) proposal that the spectral peak for velars is associated with F2 for low, back vowels and F3 or F4 for high, front vowels. This rule resulted in improved performance as compared with the earlier study, permitting equally effective identification of place of articulation across male and female talkers. From the point of view of perceptual processing, this normalization rule does not require an active process ongoing during speech perception. Rather, the listener would determine a midfrequency range only once for a given talker. Thus, we may say that the correlate of the presence of midfrequency peaks is invariant over talkers in the sense that the proposed normalization rule merely defines the location of a talker's fixed midfrequency range.

If the time-varying features proposed here are actually *invariant perceptual* attributes of place of articulation, then the identification results from this visual identification study ought to compare favorably with results from human listeners identifying the same stop consonant-vowel syllables. To test this hypothesis, a short listening experiment was conducted. Twenty listeners identified place of articulation from randomized auditory presentations of each of the 288 stop consonant-vowel syllables. For the voiceless stop consonant-vowel syllables from Part 1 of this study, 99% were identified correctly in the listening experiment, compared with 89% in the visual judging experiment. For the stop consonant-vowell syllables produced at different rates from Part 2 of this study, 93% were identified correctly in the listening experiment, compared with 86% in the visual judging experiment. Thus, the overall difference in identification was about 8% between listeners and visual judges. Although these results indicate that there is room for improvement, they do appear to support the claim that the time-varying features are invariant as both acoustic correlates and perceptual cues over vowel context, voicing, talkers, and speaking rate. For example, listeners made more errors with the voiced stops at slow rates, similar to the pattern of errors observed with the visual judges. Perceptual studies directly investigating the perceptual salience of the time-varying features are currently underway.

Finally, the time-varying features employed in this study were used to locate two additional properties of stops, the burst and the onset of voicing. Judges agreed on the choice of the burst frame within one frame 91% of the time (see Table 2). The correlate of this feature, an abrupt change in high-frequency energy, was easy to observe in the visual displays and will hopefully serve to identify the classes of stop consonants from other phonetic classes in fluent speech. In fact, Mack and Blumstein (1983) have recently reported that a measure of relative energy change at consonant release served to distinguish stops from glides. While the details differ, both their approach and ours are based on discovering an invariant correlate for the class of stops in the dynamic changes in energy at the release burst.

The present study included a time-varying feature marking the onset of voicing. This feature served only to identify the voiceless frames for judging the tilt of burst. That is, the tilt-of-burst feature was not defined over a fixed temporal interval as in the Kewley-Port (1983) study, but rather varied according to the number of voiceless frames present. Delgutte (1980) has provided a good rationale for this approach. He suggested that the peripheral auditory system might process voiceless low-energy sounds differently from high-energy voiced sounds. In a study of auditory nerve-firing patterns, Delgutte hypothesized that information for voiceless sounds is derived from rate measures, whereas information for voiced sounds is derived from temporal measures. Furthermore, he suggested that the abrupt onset of either high-frequency energy or low-frequency voicing is a strongly marked event in the auditory system. Delgutte's proposal can be used to rationalize the features employed in this study as follows: An abrupt change in high-frequency energy signals the onset of a stop burst. The tilt of the spectral energy in succeeding voiceless frames is integrated over a variable length of time, not exceeding 30 to 40 msec. The onset of voicing terminates judgment of voiceless spectral tilt. If the burst and voicing onset occur simultaneously (as they normally would for /b/ and sometimes /d/), then the tilt of burst can be identified from the first 5 to 10 msec of energy. It is clear that the identification of place of articulation and voicing are interdependent in this analysis (see also Sawusch & Pisoni, 1974). While the judgment of the phonetic feature of voicing was not specifically made in the present study, it is obviously an important next step for our research.

In summary, we believe we have successfully extended Kewley-Port's (1983) original findings in a number of important ways. Our results also further validate the use of time-varying features derived from auditory running spectra to identify invariant acoustic correlates of place of articulation (cf. Kewley-Port et al., 1983). These features appear to be adequate for the identification of the feature of voicing as well. While this research is very encouraging with regard to defining a set of invariant time-varying features for identifying stop consonants in fluent speech, it can be considered only a first step. Future research should focus on developing computer algorithms for evaluating the proposed invariant features. Automating the feature identification process will allow us to fine-tune these definitions and to examine a considerably larger set of talkers and utterances over a wider range of conditions that produce variability in the acoustic-phonetic properties of speech. Nevertheless, the results of this study demonstrate that the overall rationale is correct. Time-varying correlates of place of articulation in stops can be identified from visual displays. Moreover, these correlates can be generalized to both voiced and voiceless stops produced by different talkers at different speaking rates.

REFERENCES

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, **66**, 1001-1017.

Carlson, R., & Granstrom, B. (1982). *The representation of speech in the peripheral auditory system*. New York: Elsevier Biomedical Press.

Delgutte, B. (1980). Representations of speech-like sounds in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America*, **68**, 843-857.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fant, G. (1973). Stops in CV-syllables. In G. Fant (Ed.), *Speech sounds and features* (pp. 110-139). Cambridge, MA: M.I.T. Press.

Flanagan, J. L., & Christensen, S. W. (1980). Computer studies on parametric coding of speech spectra. *Journal of the Acoustical Society of America*, **68**, 420-430.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, MA: M.I.T. Press.

Kewley-Port, D. (1979). *Spectrum: A program for analyzing the spectral properties of speech* (Research on Speech Perception: Progress Report No. 5, pp. 475-492). Bloomington: Indiana University, Department of Psychology.

Kewley-Port, D. (1980). *Representations of spectral change as cues to place of articulation in stop consonants* (Research on Speech Perception: Tech. Rep. No. 3). Bloomington: Indiana University, Department of Psychology.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **73**, 322-335.

Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, **73**, 1779-1793.

Klatt, D. H. (1976). A digital filter bank for spectral matching. In C. Teacher (Ed.), *Conference Record of the 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE Catalog No. 76CH1067-8 ASSP, pp. 537-540). Philadelphia: IEEE.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, **7**, 279-312.

Lahiri, A., & Blumstein, S. E. (1981). A reconsideration of acoustic invariance for place of articulation in stop consonants: Evidence from cross-language studies. *Journal of the Acoustical Society of America*, **70**, S39.

Mack, M., & Blumstein, S. E. (1983). Further evidence of acoustic invariance in speech production: The stop glide contrast. *Journal of the Acoustical Society of America*, **73**, 1739-1750.

Markel, S. D., & Gray, A. H. (1976). *Linear prediction of speech*. New York: Springer-Verlag.

Ohde, R. N., & Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America*, **74**, 706-714.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, **59**, 640-654.

Patterson, R. D., & Nimmo-Smith, I. (1980). Off frequency listening and auditory-filter symmetry. *Journal of the Acoustical Society of America*, **67**, 229-245.

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

Scharf, B. (1970). Critical bands. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (pp. 157-202). New York: Academic Press.

Stevens, K. N. (1975). The potential role of property detectors in the perception of consonants. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 303-330). New York: Academic Press.

Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, **68**, 836-842.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **64**, 1358-1368.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. Miller (Eds.), *Perspectives on the study of speech* (pp. 1-38). Hillsdale, NJ: Erlbaum.

Summerfield, A. Q. (1975). *Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables* (Speech Perception No. 4). Belfast: Queen's University of Belfast, Department of Psychology.

Tekieli, M. E., & Cullinan, W. L. (1979). The perception of temporally segmented vowels and consonant-vowel syllables. *Journal of Speech and Hearing Research*, **22**, 103-121.

NOTES

1. Our use of the terms "static" and "dynamic" follows Stevens and Blumstein (1978) and Kewley-Port et al. (1983). "Static" correlates refer to spectral energy integrated over a given temporal interval; in contrast, "dynamic" correlates refer to changes in spectral energy distributed over time.

2. The procedure used to calculate auditory spectra was necessitated by the limitations of the available computing facilities and may have introduced some undesirable frequency characteristics in the derived voiceless spectra. However, comparisons of the spectra generated by convolving the auditory filters with FFT spectra versus linear prediction spectra showed little difference for the purposes of the present time-varying feature analysis. On the other hand, the LPC analysis required a slightly longer analysis window for one male talker that need not have been used for FFT spectra. In future work, the auditory filtering will be improved by implementing a bank of digital filters applied directly to the speech waveform.

3. A feature used in the previous study was dropped from this analysis. This feature, called late onset of low-frequency energy, was essentially a measure of voice-onset time (VOT). Since three speaking rates were used in this study, we assumed that this feature would not be invariant across the utterances examined here (cf. Summerfield, 1975).

4. The results were collapsed across both sessions because the addition of the two new rules in the collaborative judging did not actually alter the judging procedure. The first rule allowed a feature to be identified as ambiguous; in such cases, identification of place was scored as incorrect. The second rule, which required the judges to weigh later frames more heavily for ambiguous cases, corrected an oversight in the original instructions for assigning tilt.

(Manuscript received August 19, 1983;
revision accepted for publication February 19, 1984.)