

# Time-varying features as correlates of place of articulation in stop consonants

Diane Kewley-Port

Department of Psychology, Indiana University, Bloomington, Indiana 47405

(Received 1 July 1981; accepted for publication 12 July 1982)

Running spectral displays derived from linear prediction analysis were used to examine the initial 40 ms of stop-vowel CV syllables for possible acoustic correlates to place of articulation. Known spectral and temporal properties associated with the stop consonant release gesture were used to define a set of three-time-varying features observable in the visual displays. Judges identified place of articulation using these proposed features from running spectra of the syllables /b,d,g/ paired with eight vowels produced by three talkers. Average correct identification of place was 88%; identification was better for the male talkers (92%) than the one female talker (78%). *Post hoc* analyses suggested, however, that simple rules could be incorporated in the feature definitions to account for differences in vocal tract size. The nature of the information contained in linear prediction running spectra was analyzed further to take account of known properties of the peripheral auditory system. The three proposed time-varying features were shown to be displayed robustly in auditory filtered running spectra. The advantages of describing acoustic correlates for place from the dynamically varying temporal and spectral information in running spectra is discussed with regard to the static template matching approach advocated recently by Blumstein and Stevens [J. Acoust. Soc. Am. 66, 1001-1017 (1979)].

PACS numbers: 43.70.Dn, 43.70.Gr, 43.66.Fe

## INTRODUCTION

Recently a number of investigators have developed speech processing techniques which incorporate or model known spectral and temporal processing characteristics of the peripheral auditory system. Studies of auditory physiology by Kiang (1980), Kiang *et al.* (1979), Delgutte (1980, 1981), Sachs and Young (1979, 1980), and Young and Sachs (1979) have shown how the rapidly changing speech signal is represented in the neural signals output from the peripheral auditory system. Information gathered from this and other psychophysical research has been implemented in specific speech processing systems by several other investigators. Word recognition systems using auditory filters for processing the speech signal have been developed by Zwicker *et al.* (1979) and Klatt (1979). On a more limited scale, Bladon and Lindblom (1979) and Carlson and Granstrom (1980) have designed and experimentally evaluated auditory filters for classifying vowel spectra. Another application has included the development of speech vocoders which incorporate properties of auditory processing (Schroeder *et al.*, 1979; Flanagan, 1980; Flanagan and Christensen, 1980).

The present research was inspired directly by the recent work of Searle and his colleagues (1979, 1980). Searle *et al.* developed a frequency-by-amplitude analysis based on peripheral auditory filters approximated by analog 1/3-octave filters, updated at 1.6-ms intervals. Their approach emphasized not only the auditory transformation of the speech signal, but also the dynamic changes in the transformation. Their analyses permitted them to construct a three-dimensional representation, displaying the running spectra of a speech signal as it changed over time. In a preliminary study of the properties of stop consonants observed in these running spectra, Searle *et al.* were very successful in identifying

cues to voicing for stop consonants, but only partially successful in identifying cues to place. The idea of examining distinctive cues in running spectral displays provided the basis for developing a similar analysis technique in this investigation to study several long-standing questions concerning acoustic invariance in stop consonants.

In particular, the present research examined the problem of identifying invariant cues for place of articulation in initial stop consonants. Although this issue has been one of great interest in speech research, only recently have investigators actually claimed to have solved the problem. Stevens and Blumstein (1978, 1981; Blumstein and Stevens, 1979, 1980) in a series of articles have attempted to describe and experimentally verify the existence of static integrated acoustic cues for place of articulation in stops. They have argued that invariant acoustic properties for place can be found in the overall gross shape of the spectrum at the onset of the release burst. Stevens and Blumstein claim that a unique spectral shape can be found for each particular place of articulation. Furthermore, they have argued that these spectral shapes are correlates of the phonologically distinctive features that define place of articulation which can be observed across syllable position, consonant manner class, and talker (Jakobson *et al.*, 1952).

Stevens and Blumstein have developed specific templates of the gross spectral shapes for each place of articulation. These templates were derived from single 25.6-ms spectral sections taken at the onset of stop-vowel syllables and smoothed by linear prediction. Blumstein and Stevens (1979) have also carried out several experimental tests to assess the adequacy of these templates to identify place of articulation for syllable-initial and final stops produced in five vowel contexts by six talkers. Their results showed that the templates

were fairly successful in identifying the place of articulation of stops in syllable-initial position, but not of stops in syllable-final position.

Stevens and Blumstein also carried out two further studies to verify experimentally that the overall shape of the onset spectra contained important perceptual cues for identifying place of articulation. Both studies used synthetic CV syllables (Stevens and Blumstein, 1978; Blumstein and Stevens, 1980). The synthetic stimuli varied in overall duration from relatively long CV syllables to very brief truncated stimuli, although they were all constructed using the same synthesis principles. These principles involved preserving the natural details of the burst, VOT, and voiced formant transitions in their "full-cue" set. All these parameters were then manipulated to produce a stop consonant-vowel continuum from /b/ to /d/ to /g/ before three vowels. Subjects were able to identify some of these stimuli as /b/, /d/, or /g/ on 100% of the trials in a forced choice task, whereas other stimuli were identified ambiguously. Stevens and Blumstein then observed, in an informal way, that the onset spectra for the unambiguously identified stimuli were in agreement with the proposed gross spectral shapes. From these findings they argued that the gross shape of the spectrum at onset contains the perceptually distinctive acoustic information to identify place of articulation in stops.

Although many other experimental conditions were included in their two reports, the basic strategy for verifying the role of gross spectral properties in synthetic CV syllables remained basically the same. By their own admission, these data "do not constitute a strong test of the theory" (Stevens and Blumstein, 1978, p. 1367). Indeed, their experimental procedures appear to be unsatisfactory as a way of verifying whether the gross spectral shapes are sufficient perceptual cues for place of articulation, primarily because acoustic properties of the stimulus set were not manipulated in terms of the gross onset spectral properties themselves, but rather in terms of burst frequency, VOT, and formant transitions. Nevertheless, Stevens and Blumstein argued that invariant acoustic cues for specifying place of articulation can be located in the first 10–20 ms of a stop waveform. These particular claims motivated several aspects of the present investigation.

The research undertaken in this report examines the initial portion of the acoustic waveform, as Stevens and Blumstein have suggested, but used running spectral displays for the spectral representation. These running spectra differed considerably from Searle's running spectra since they were calculated digitally using linear prediction analysis (Markel and Gray, 1976). Linear prediction running spectra produced with a carefully chosen set of analysis parameters can in fact provide a good spectral representation for speech, with fine resolution of the formant structure in the frequency-by-amplitude domain and good temporal resolution by appropriately updating the analysis in the time domain. Continuous running spectral representations of speech are presumably a better model of the spectral information output from the peripheral auditory system than are Stevens' single integrated spectral sections. That is, the most basic property of auditory neural signals is that they vary directly with the time variations of the input acoustic signal. In fact, Schroeder *et al.* (1979) have modeled the spectral processing properties of the ear in a manner quite analogous to our calculation of a running spectrum. In their model, Schroeder *et al.* state: "The inner ear...performs a running short-time spectral analysis in which the frequency coordinate  $f$  is represented by a spatial coordinate  $x$  along the length of the basilar membrane. We approximate this process by short-time Fourier transformations over successive 20-ms time windows" (p. 1647).

Figure 1 shows six typical running spectral displays of stops in different vowel contexts. The first frame effectively displays 5 ms of the burst release. Subsequent frames are offset at 5-ms intervals. Further details of how these displays were generated are provided in Sec. IIA below.

Underlying the analysis of invariant cues to place of articulation used by Stevens and Blumstein is the assumption that the information specifying place is contained primarily in the first 10–20 ms of the stop-vowel waveform. While this assumption has enjoyed wide support over the years (cf. Jakobson *et al.*, 1952; Fant, 1960; Cole and Scott, 1974; Tekieli and Cullinan, 1979), competing accounts have also been proposed. Liberman and his colleagues at Haskins (Cooper *et al.*, 1952; Liberman *et al.*, 1954, 1967) have emphasized the importance of the voiced formant transitions in

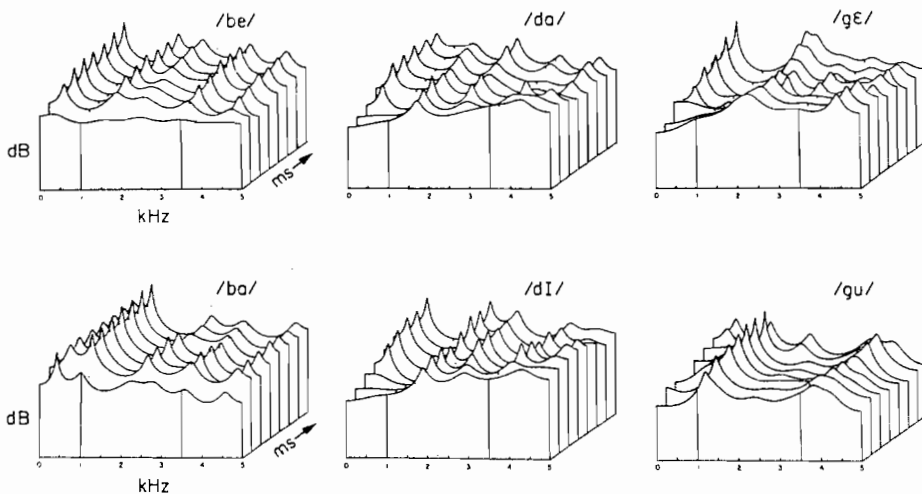


FIG. 1. Running spectral displays for the first eight frames of six stop-vowel syllables. Each frame is offset by 5 ms.

carrying place of articulation information. Liberman's argument has been that an essential property of speech is that it is *dynamic*. Formant transitions of stop-vowel syllables appeared to be the appropriate cues to place because they capture the important time-varying changes associated with both the underlying articulatory gestures and the resulting acoustic signal. Recent research, however, has cast doubt on the hypothesis that formant transitions are the most important cues to place of articulation in natural speech. Dorman *et al.* (1977) obtained poor identification of place of articulation from only the formant transition plus vowel segments edited from natural syllables. Furthermore, Kewley-Port (1980, 1982) measured numerous parameters of formant transitions in natural CV's and was unable to reliably distinguish place of articulation based on these measures across all eight vowel contexts studied.

Examination of the running spectral displays in the present investigation revealed that emphasis on properties of the voiced formant transitions was due, in part, to the representation of CV's in the spectrographic analysis typically used in earlier studies (Potter *et al.*, 1947; Liberman *et al.*, 1954). The running spectral displays of the present investigation showed continuous changes in the spectral prominences from the release burst into the voiced formant transitions. As can be seen in Fig. 1, the onset of voicing in these displays is indicated by the abrupt appearance of a well defined  $F_1$  peak at low frequencies, while the upper formants are more nearly continuous into the burst. The voiced formant transitions do not stand out visually as isolated distinctive segments of the speech signal in the running spectral displays.

Visual examination of numerous stop-vowel syllables from a male talker revealed several time-varying acoustic features or attributes that could potentially distinguish place of articulation in the running spectral displays. This observation led to pilot studies (Kewley-Port, 1979a, b) of running spectra which suggested that some of these features might specify place invariantly over different vowel contexts. The earlier feature definitions were carefully considered and re-defined for the larger and more formal experiment presented here. The features that we developed were similar to acoustic features derived from the acoustic theory of speech production as proposed by Fant and to the features used by Stevens (Stevens, 1975; Stevens and Blumstein, 1978) in his research. In fact, as will be seen below, the features have strong roots in the distinctive feature theory of Jakobson *et al.* (1952). The important aspect of these features, however, is that they are time varying and preserve the essential dynamic characteristics of the stop-vowel syllable.

### I. TIME-VARYING FEATURES OF PLACE OF ARTICULATION

The time-varying features proposed in this experiment are similar to the acoustic correlates for place described by Fant (1960, 1973) and Stevens (Stevens and Blumstein, 1978, 1981). For both Fant and Stevens the acoustic correlates of place directly reflect the underlying articulatory gestures according to the principles of the acoustic theory of speech production. Furthermore, since the articulatory gesture for a given consonant is assumed to be relatively fixed, the corresponding acoustic cues for place should be invariant. The

running spectral features proposed in the present study are similar to the relational, invariant features originally suggested by Fant and Stevens. A general description of the time-varying properties associated with each place of articulation follows below.

The first property observed in the running spectra was the tilt of the burst. In the running spectral displays, the spectrum of the first 5 ms of the burst was displayed in the first spectral section or frame. According to the predictions of the acoustic theory of speech production, both Fant (1960) and Stevens (1975; Stevens and Blumstein, 1978) have proposed that burst spectra for labials emphasize low frequencies with a general downward tilt of spectral energy from the low to the high frequencies. This can be seen in the running spectra in Fig. 1 as a flat or falling tilt of spectral energy in the burst frame for /be/ and /ba/. Alveolars, on the other hand, were predicted to have a generally rising tilt of energy from low to high frequencies which can be seen in the running spectra of /da/ and /d<sub>1</sub>/ in Fig. 1. Velar bursts have not been characterized as rising or falling because burst energy shifts from high to low as the vowel context moves from front to back (Halle *et al.*, 1957; Fant, 1960; Stevens, 1975).

Although the tilt of burst property discussed above is purely a *spectral* property, there is an implied temporal property necessary for *locating* the burst itself. In this experiment, as in most earlier studies of stop consonants, the burst frame was located visually in the waveform by the experimenter. In ongoing speech processing, however, the occurrence of the burst itself must also be detected. A time-varying acoustic feature for locating the burst in running speech has already been proposed by Stevens who suggested that the release burst can be detected when an abrupt change in energy occurs (Stevens, 1980; Stevens and Blumstein, 1981). This proposal corresponds well with physiological studies which have shown that the auditory system responds uniquely to abrupt changes at signal onset (Zhukov *et al.*, 1974; Delgutte, 1980, 1981). We suggest that an abrupt change in energy can be observed easily in running spectra which include the low energy frames occurring in the stop closure preceding the burst.

While velar place of articulation is not characterized by burst tilt, Fant and Stevens have proposed that the essential spectral property of velar bursts is the presence of a compact or prominent peak of energy in the mid-frequency region (see /ge/ and /gu/ in Fig. 1). Stevens and Blumstein (1978, 1981; Blumstein and Stevens, 1979) argue that this spectral property is sufficient to identify velar place when captured in a static onset spectrum. However, Fant has pointed out that the burst release has temporal properties as well (Fant, 1960, 1973). Labial and alveolar bursts are said to be between 5 and 10 ms in length which is shorter than the 20–30 ms velar bursts. Furthermore, velars are described as having a compact spectrum which lasts throughout the longer burst. This compact spectrum arises from a resonant pole produced in the cavity in front of the velar constriction. Fant (1973) pointed out that the velar release is relatively slow so that this resonance is sustained for approximately 30 ms. Therefore a distinctive property of velars is that only slow changes in spectral energy are observed over this interval compared

to the more rapid changes observed for labials and alveolars. The running spectral analysis as proposed here also captures these temporal properties of the release bursts. The 5-ms burst frame is quite prominent in this analysis and clearly displays the tilt of the bursts. As the linear prediction window slides along in succeeding frames for labials and alveolars, transient energy is encountered and a rapid change in spectra can be observed (see Fig. 1). For velars, however, successive spectra following the burst show little change in the prominent mid-frequency spectra.

One other feature appeared prominently in our visual examination of the running spectral displays. This correlate of place articulation was the delay in the onset of voicing relative to the burst. In the running spectra, the frame in which the  $F_1$  peak onsets relative to the burst is a direct measure of VOT (voice-onset-time). Vot has been shown to increase in length from labials to alveolars to velars (Lisker and Abramson, 1964; Zue, 1976). In running spectral displays, our earlier pilot work indicated that a delay of 20 ms (i.e., four frames) or more in the onset of voicing was strongly associated with velar place (see Fig. 1).

The observations of spectral and temporal features associated with labial, alveolar, and velar place were formalized for evaluation in an earlier pilot study (Kewley-Port, 1979a, b). Based on these results, three time-varying features having binary feature categories were defined for examination in this experiment. These definitions, in conjunction with an assignment matrix, were used by judges to determine place of articulation in running spectral displays as follows.

**Feature 1: Tilt of the spectrum at burst onset:** Tilt was estimated by visually fitting a straight line to the first frame between 0 and 3500 Hz. The feature categories were R = rising and F = flat or falling.

**Feature 2: Late onset of low-frequency energy.** Late onset was defined as the occurrence of high amplitude, low-frequency peaks ( $F_1$  peaks) starting in the fourth frame of the display or later. Feature categories were L = late onset and N = no late onset.

**Feature 3: Mid-frequency peaks extending over time.** This feature was defined as the presence of a single, prominent peak between 1000 and 3500 Hz occurring for three or more frames, although not necessarily consecutive frames. The feature categories were Y = yes, peaks exist and N = no, no such peaks are present.

After the feature categories were specified, place of articulation was assigned by the judges as /b/, /d/, or /g/ in accordance with the assignment matrix in Table I. An entry of "?" in the assignment matrix meant that either feature category could occur for that stop. The "\*" by the feature L

TABLE I. Matrix used by judges to assign place of articulation from specific feature categories.

Tilt of burst	Late onset	Mid-frequency peaks	Assigned consonant
F	N	N	b
R	?	N	d
?	L*	Y	g

indicated that in ambiguous cases, the presence of L was sufficient to assign the stop g.

Our earlier pilot study had examined stops before only the front vowels from one talker. In the present experiment we added more vowels and more speakers to the original data base. The purpose of this study was to determine if the features and new assignment matrix could adequately describe the invariant visual properties of running spectra for identifying place of articulation in stop consonants.

## II. EXPERIMENT

### A. Stimuli

Three talkers, two males (RP and TF) and one female (NL), produced the set of consonant-vowel syllables that were analyzed in this study. Syllables from talker RP, a phonetician, were a subset of those analyzed in a separate study of formant transitions (see Kewley-Port, 1980, 1982). Three repetitions each of the syllables /b,d,g/ paired with /i, ɪ, e, ε, æ, a, o, u/ were used in this study. The data base was then expanded by adding an additional male talker TF, and a female talker NL, both of whom were phonetically naive. They produced the syllables /b,d,g/ paired with /i, e, a, o, u/. All syllables were embedded in the carrier sentence, "Teddy said CV." Sentences were read from randomly ordered lists. Talkers were recorded in a sound attenuated IAC booth on an Ampex AG-500 tape recorder using an Electro-Voice D054 microphone. Three repetitions of each syllable were digitized for analysis from the middle of the ten lists recorded. Waveforms were first low-pass filtered at 4.9 kHz and then sampled at 10.0 kHz using a 12-bit A/D converter. The total number of syllables examined in this experiment was 162, 72 from speaker RP, and 45 each from speakers TF and NL.

All syllables were analyzed on a PDP 11/34 computer using the SPECTRUM program (Kewley-Port, 1979c) to produce running spectral displays such as those shown in Fig. 1. The waveforms were edited and first differenced (pre-emphasized). A Hamming window was then positioned so that the burst onset of the CV was located in the center of the window. Linear prediction coefficients were calculated for each window using the autocorrelation method of Markel and Gray (1976). Since this method requires at least two pitch pulses to fall within the analysis window, 20-ms windows were used for RP and NL, but TF, having a lower fundamental frequency, required a 25-ms window. Smoothed spectra were calculated by means of a discrete Fourier transform of the reflection coefficients with added zeros using the algorithm of Markel (1971; also see Markel, 1972). The resulting 256-point spectrum has a 19.5-Hz bandwidth. This narrow bandwidth reflects the potential accuracy of specifying formant frequency information when the linear prediction analysis parameters have been appropriately chosen. Monsen (1981) has recently shown that accuracy of measuring vowel formants was poorer than predicted at approximately 60 Hz.

A new spectral section or frame was calculated at 5-ms intervals. The 20- and 25-ms Hamming windows used in this analysis have effective durations of about 10 and 12.5 ms, respectively. Thus the 5-ms update interval produced some

spectral overlap between frames without severe oversampling as can be seen in Fig. 1. These temporal parameters were originally selected to preserve the onsets of formant transitions as accurately as possible within the limits of the linear prediction method (see Kewley-Port, 1982).

In these running spectral displays, the Hamming window was positioned visually by the experimenter in such a way that the first frame encompassing the burst had an effective duration of about 5 ms. Therefore the first frame can be said to display spectral energy from the release burst only. Very often the first several frames analyzed were voiceless, i.e., missing the *F*<sub>1</sub> peak. According to Markel and Gray (1976), fewer linear prediction coefficients are needed to specify the spectrum adequately in voiceless than in voiced frames. Spectral sections calculated with fewer coefficients had smoother peaks, closer to the underlying fricative spectrum, than did the rippled peaks often produced by extra coefficients. Therefore, in this analysis, four fewer coefficients were used in analyzing the voiceless frames. Fourteen coefficients were used to calculate voiced frames for the males with five formant peaks; 12 coefficients were used for the female with four formant peaks.

A running spectral display was then plotted for the first eight frames—or 40 ms—for each CV. Using a Tektronics hard-copy unit (Model 4631), an 8-1/2-by 11-in. display was produced for each of the 162 CV's. All the CV's were randomized, coded by number, and placed in looseleaf notebooks for examination by the judges in this experiment.

## B. Judges

Three members of the laboratory served as judges. Phonetically sophisticated judges were required for this task because the descriptions of both the displays and the phonetic features employed standard acoustic and phonetic terminology. The judges were not, however, familiar with running spectral displays or the specific nature of the present experiment. Two of the judges (ACW and TDC) were graduate students in Psychology; the third was a post-doctoral fellow (SEK) with a Ph.D. in Speech and Hearing Sciences.

## C. Procedure

The present experiment consisted of three parts: (1) training, (2) independent judging, and (3) collaborative judging. The training session was used to acquaint the judges with the feature definitions and the assignment matrix to be used to identify place of articulation from running spectral

displays. A typewritten page containing the feature definitions and assignment matrix for consonant identification as described earlier was given to each judge. A 20-min training session was conducted by the experimenter with 15 examples of running spectra, none of which were included in the 162 test displays. Figure 1 shows the six primary examples used in the training phase. Table II shows the correct feature responses and consonants on a facsimile of a response sheet.

During the training phase, judges learned to identify each feature category for a display independently and write corresponding letters on their response sheets. The judges were then asked to assign a consonant according to the entries in the assignment matrix. It was noted that the assignment matrix did not include all possible combinations of features. The judges were told, however, that combinations not represented in the matrix would probably occur infrequently. If they occurred, judges were instructed to assign a consonant in whatever way they could.

After training, the judges were asked to respond to each of the 162 displays *independently*. Each judge viewed half of the displays in the looseleaf notebooks in two separate 1-h sessions on different days. The response sheets from the independent sessions were scored for correct consonant identification only. One or more incorrect responses occurred on 46 of the 162 displays. To resolve errors which may have resulted from careless judgments, a collaborative judging session was arranged one week after the independent judging sessions were completed. The three judges met together as a group with the experimenter present and were given the feature definition sheets and additional instructions for rescoring the 46 displays in which errors occurred. Judges were instructed to write down on a separate response sheet whether they unanimously agreed or disagreed for a given display. When judges disagreed, they were asked to indicate in what ways the features or matrix were ambiguous. The displays were judged in a 1-1/2-h session which was taperecorded for later analysis.

## D. Results

In the collaborative judging, 20 of the 46 displays were unanimously assigned to the correct consonant, while ten of the displays were unanimously identified incorrectly. The remaining 16 were judged to be ambiguous. To obtain an overall score for correct identification, all unanimous assignments from the independent and the succeeding collaborative judgments were used in the final results. For the 16 ambiguous displays, the forced choice responses from the independent judging were used since the consonants had not

TABLE II. Correct responses for the training examples seen in Fig. 1.

Syllable	Tilt of burst	Features		Assigned consonant
		Late onset	Mid-frequency peaks	
/be/	F	N	N	b
/da/	R	N	N	d
/ge/	R	L	Y	g
/ba/	F	N	N	b
/di/	R	L	N	d
/gu/	R	L	Y	g

TABLE III. Results of assigning consonants to running spectral displays using three judges.

Talker	No. errors	(N)	% Correct	% Correct by sex
RP	14	(216)	94	Male = 92
TF	14	(135)	90	
NL	30	(135)	78	Female = 78
Total	58	(486)	88	

been assigned for ambiguous cases in the collaborative judging.

Table III shows the overall results for consonant identification. Consonants were correctly identified 88% of the time from the running spectral displays. The errors were not uniformly distributed by talker. Specifically there was a large difference in correct identification depending on the sex of the talker, 92% correct for male talkers, but only 78% correct for the female talker.

The distribution of errors by consonant and vowel is shown in Table IV. Overall, /d/ was identified most accurately, with performance at 93%. The consonant /b/ was identified in most contexts except in syllables containing the high vowels /i,e,u/. Most errors occurred for /b/ syllables produced by the female speaker. Analysis of the collaborative judging errors indicated that the tilt of the burst was ambiguous or slightly rising for bilabial stops before high vowels. The consonant /g/ was poorly identified in the syllable /gi/ with 56% correct. Most of these errors were also contributed by the female speaker whose mid-frequency peaks occurred above the 3500-Hz limit imposed in the original feature definitions. Additional /g/ errors occurred because the otherwise prominent mid-frequency peaks were not clearly "single" peaks in the display.

Table V displays the pattern of errors for the consonants identified in the running spectral displays. /b/ consonants were frequently mistaken as /d/'s, but not the converse. Bilabials and velars were rarely mistaken for one another. /g/'s were frequently mistaken for /d/'s.

The feature categories assigned for each CV were analyzed to determine whether the judges had reliably and consistently categorized the features. Judges were said to disagree on feature assignment when they had not unanimously assigned the same feature categories. The results showed that when the judges correctly identified the consonant in the independent judging, only 2% feature disagreement occurred. In the collaborative judging of the other 46 displays, only 8% feature disagreement was observed. Therefore an overall score of 5% feature disagreement was obtained from the three judges. These results indicate that it was relatively easy for the judges to categorize the features in the running spectral displays as specified by our feature definitions.

The specification of the original assignment matrix was based primarily on running spectra from a male speaker, RP.

TABLE IV. Percent correct identification of consonant by vowel for all talkers. Note, vowels /i, e, æ/ were produced by only one talker.

Vowel	(N)	b	d	g	Total
i	(81)	67	93	56	72
e	(81)	70	81	89	80
a	(81)	100	93	100	96
o	(81)	93	96	100	96
u	(81)	78	96	100	91
ɪ	(27)	89	100	100	96
ɛ	(27)	100	100	67	89
æ	(27)	100	100	67	89
Total	(486)	84	93	87	88

TABLE V. Percentage of response errors obtained for a given consonant in running spectral displays.

Displayed	Assigned consonant		
	b	d	g
b	...	14%	2%
d	2%	...	5%
g	0%	13%	...

To check the validity of the feature matrix for assigning stops for all three talkers, the percentage of feature assignments made by all three judges in the independent judging was calculated. These results are given in Table VI, where percentages are entered in terms of the categories as they appeared on the feature definition sheet.

It can be seen from the data in this table that the percent judgment of feature categories other than "?" was quite high, averaging 92%. The two entries of "?," which signified that either category might occur, were assigned equally to the categories. These results indicate that the original assignment matrix was appropriate for the features examined in this experiment. The slightly lower assignments of correct categories for Tilt of burst (85% and 88%) and the presence of Mid-frequency peaks for /g/ (83%) suggest that some improvement of these feature definitions may be needed in future studies.

Not all possible permutations of the feature categories were listed in the assignment matrix. As a consequence, there were possible combinations of feature categories which led to ambiguous consonant assignments. Most of these occurred for the 46 displays in which a consonant error occurred. These were dealt with in the analysis of the collaborative judging data. However, for the displays judged unanimously correct, only 1% of the responses resulted from ambiguous feature combinations. Thus possible ambiguities in the assignment matrix were not a problem in this study.

Only ten of the displays (6%) were unanimously assigned the incorrect place of articulation. The incorrect assignments for these displays provide some insights into problems with the current feature definitions. Four of the ten displays were /b/'s before front vowels. Because the burst tilt was rising, judges assigned d's to these displays. Three other displays were the female talker's /gi/'s each having a

TABLE VI. Percent of feature assignments obtained in independent judging. They are listed according to the feature categories appearing in the feature matrix used for consonant assignment.

Consonant	Tilt of burst	Late onset	Mid-frequency peaks
b	F = 85	N = 96	N = 98
d	R = 88	?N = 59 ?L = 41	N = 96
g	?F = 35 ?R = 65	L = 96	Y = 83

small mid-frequency compact peak at or above the 3500-Hz frequency limit used in the feature definitions. Judges also incorrectly assigned /d/'s to these displays. Analysis of the remaining three displays (2%), however, showed unusual compact peaks for these /g/ syllables.

Some special attention should be given to the feature Late onset of *F*1. All three features in this experiment were defined as binary. Since only two binary features are necessary to specify the three consonants, this feature system is redundant. However, the feature of Late onset was so prominent in the running spectral displays that we thought that it ought to play some role in distinguishing place of articulation among the stops, especially /g/ from /b,d/. In particular, since Mid-frequency peaks were sometimes difficult to identify for /g/, it was thought that in ambiguous cases, the additional feature of Late onset of *F*1 might facilitate the correct identification of the stop category. This was incorporated in the feature definition sheet using "L\*" in the matrix for /g/. *Post hoc* analysis revealed, however, that the Late onset feature was actually used by the judges to disambiguate /g/ in only three cases (0.6% of the judgments). Thus the current feature definition system did not adequately capture the intended usefulness of Late onset, even though Late onset was present in 96% of the /g/'s.

As a consequence, the running spectra were re-examined to determine if an alternative definition for Late onset could be developed. The frame in which the onset of *F*1 occurred was determined visually by the experimenter for the 162 displays. In doing this, we found that approximately 50% of the /g/ displays had *F*1 onset after 30 or more milliseconds, compared to only one /d/ display and no /b/ displays. This observation suggests a new definition for Late onset, namely, that the category L refers to onset of *F*1 peaks after 30 ms (six or more frames). An informal examination of the present running spectra indicated that the proposed change in the Late onset feature would produce only a small overall improvement in consonant identification if incorporated in the analysis. However, the proposed change represents a much better implementation of the concept which was intended for the Late onset of *F*1 feature, namely, that the presence of the Late onset of *F*1 (i.e., a long VOT) is strongly associated with the consonant /g/. Such a feature can be useful in disambiguating /g/ in running spectral displays where Mid-frequency peaks may not be clearly present.

### III. COMPARISON OF LINEAR PREDICTION AND AUDITORY FILTERED RUNNING SPECTRA

The results from this study have demonstrated that human observers can reliably identify place of articulation from visual features displayed in running spectra. Place of articulation was identified using time-varying features observed in the initial portions of the stop-vowel waveform which were independent of the following vowel context. While such findings were reliable and consistent across three judges for a large number of natural speech tokens, there still remains the question of the relation between linear prediction running spectra and spectral processing of speech by the human auditory system. Specifically, we were interested in

determining if the time-varying features used in the visual experiment would be as robust in auditory spectral representations as they were in linear prediction spectra.

Several investigators have constructed auditory processing models to produce spectral representations of speech signals. In some cases, the output of these models has been examined for the presence of spectral cues for speech recognition. In particular, as noted earlier, the research of Searle and his colleagues (Searle *et al.*, 1979, 1980) using running spectra to search for place and voicing cues in stops inspired the present study. Although Searle *et al.* explicitly advocated the development and use of auditory processing techniques for speech, analysis techniques currently implemented in speech research are simply not adequate models of known psychophysical properties of the human auditory system. In the case of Searle *et al.*'s study, a standard, commercial set of 1/3-octave filters was chosen as the basis of their speech processor. Design characteristics of 1/3-octave filters were developed to meet a set of engineering standards for commercial filters (American National Standards Institute ANSI S1.11-1966 class III). The popularity of 1/3-octave filters for speech processing (e.g., Klein *et al.*, 1970; Schouten and Pols, 1979) derives from their general availability and speed (i.e., analog processing). As we shall see below, however, 1/3-octave filter characteristics are only a gross approximation to the actual filtering properties of the human auditory system.

Before evaluating the success of several recent processing techniques, it will be useful to compare the properties of running spectral displays with several of the known properties of mammalian auditory systems as well as with other auditory processing models currently employed in speech analysis. In terms of the individual smoothed spectra, we will briefly discuss the characteristics of the analyzing filters (e.g., linear prediction, FFT, 1/3-octave filter banks) and the representation of the frequency dimension (e.g., linear, log, bark).

Differences in design characteristics of the filters for processing speech signals can produce quite different spectral displays. Based on psychophysical measures, the frequency resolution in the human auditory system has been described in terms of a set of critical bands (Scharf, 1970). A critical-band analysis corresponds roughly to the frequency analysis of a set of bandpass filters whose bandwidth is constant (about 100 Hz) below 500 Hz, and then becomes successively broader as frequency increases above 500 Hz. Bandwidth, however, is only one property of a bandpass filter. Two other properties less frequently discussed are the shape of the filter itself (e.g., rectangular versus Hamming) and the slopes of the skirts of the filter. While these two properties of auditory filters have generally been fitted by simple functions, Patterson and Nimmo-Smith (1980) have specified a unique, two-part filter shape to account for their data. The discussion that follows, however, emphasizes differences in bandwidth since it has received the most attention in the psychophysical literature. (See Klatt, 1976 and 1979, for a comprehensive list of appropriate design characteristics for critical-band filters for speech processing.)

Research has produced two sets of estimates for the

bandwidth of critical-band filters, one about one-half as wide as the other (see Sever and Small, 1979). This range varies approximately from 0.1 to 0.18 times the center frequency of the filter. Estimates of the well-known bark critical-band filters used by Zwicker (1961; Zwicker *et al.*, 1979) are compared with other filter bandwidths in Fig. 2. The hatched area includes bandwidths most often reported in the psychoacoustic literature. One-third-octave filters have bandwidths approximately 0.23 times the center frequency. Thus the bandwidths of 1/3-octave filters are considerably broader than estimates of the critical bandwidths derived from psychophysical data. This means that 1/3-octave filters provide poorer frequency resolution in the important mid-frequency range used in speech than does the human auditory system. As a consequence, 1/3-octave filtering of speech signals probably represents a lower limit on the poorest frequency resolution that the human auditory system might display. One-sixth-octave filters have been used in the speech analysis systems developed by Schroeder *et al.* (1979) and Flanagan and Christensen (1980). While 1/6-octave filter bandwidths are narrower than Zwicker's (1961) estimates as seen in Fig. 2, they correspond well with more recent

bandwidth estimates, especially those of Patterson (1976; Patterson and Nimmo-Smith, 1980). [See Flanagan and Christensen (1980) for a demonstration of mid-frequency differences between 1/3-octave filters and 1/6-octave filters.]

On the other hand, linear prediction spectra used in the present study are theoretically capable of providing a constant frequency resolution of 19.5 Hz (but see Monsen, 1981). As can be seen in Fig. 2, this resolution is considerably narrower than that of the auditory system, particularly at higher frequencies. We should note, however, that frequency resolution as measured in discharge patterns of auditory-nerve fibers in cats can be extremely accurate under optimal signal conditions. The analyses carried out recently by Delgutte (1980, 1981) and Young and Sachs (1979; Sachs and Young, 1979, 1980) have shown considerable accuracy in determining formant frequencies of synthetic vowels under certain conditions. It is not known if such precise formant frequency information is available under normal listening conditions for human speech recognition. However, linear prediction spectra when displayed on a log-frequency axis appear roughly similar in frequency resolution to many of the vowel-formant figures reported in both Sachs and Young (1980, cf. Fig. 13) and Delgutte (1981, Chap. 1). These figures represent vowel spectra in log frequency as a function of interval measures of the discharge patterns of auditory-nerve fibers. Thus linear prediction spectra may be considered as providing an upper limit on the best possible frequency resolution the human auditory system might display.

With spectrally analyzed speech, several possible representations of the frequency-by-amplitude dimensions can be chosen. Linear prediction spectra are typically represented on a linear frequency scale. In the auditory system, however, frequency on the basilar membrane is equally distributed in approximately bark intervals (Zwicker, 1961; Schroeder *et al.*, 1979), which is often approximated by a simple log-frequency scale. Thus, for research employing auditory filters, a bark frequency or modified log scale (technical Mel) is probably more appropriate for displaying frequency than is a linear scale.

Another property of running spectral displays is the representation of time. We know that the auditory system can closely track time variations in waveforms in terms of synchrony of discharge firings with the input signal (Kiang, 1980). Apparently, the important acoustic distinctions in speech vary much more slowly than the temporal processing capabilities of the ear. Therefore the limits of the representation of the time dimension for processing speech spectra should be set according to the observed rates of change in the speech signal. For speech, this limit would be placed somewhere between 1 and 20 ms. Searle *et al.* (1979) originally used a 1.6-ms time frame for running spectra. This time frame seemed to present too much detail, so they used averaged time frames of 8 ms in their feature analysis. The running spectra in the present study were 5 ms apart, while Klatt's (1979) spectra were 10 ms apart. Thus the time intervals between spectra currently employed by different investigators are in the 5- to 10-ms range.

The question originally posed in this section was

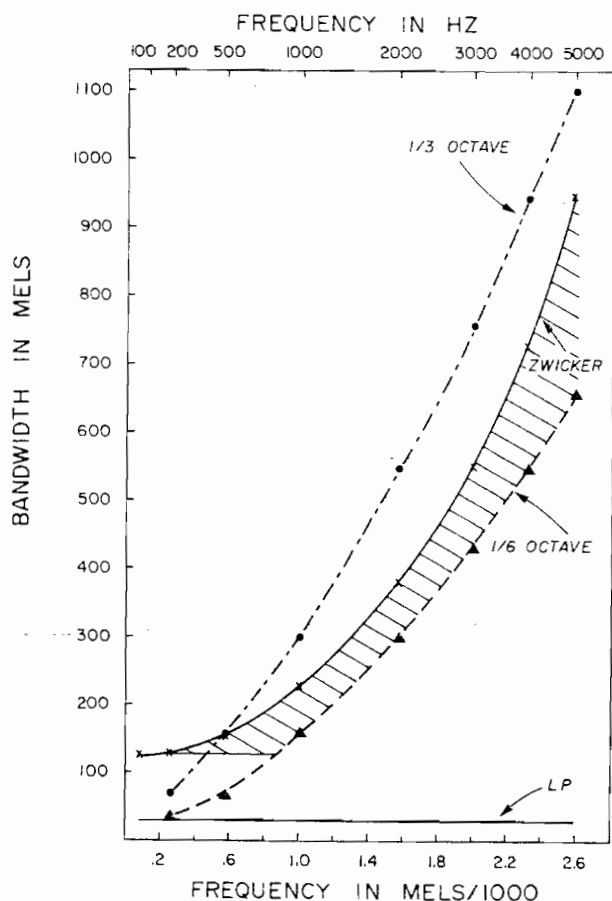


FIG. 2. Four different filter bandwidths as a function of frequency are displayed using the technical Mel scale. 1/3- and 1/6-octave filter bandwidths are shown according to ANSI standards. The function labeled Zwicker is derived from his definition of critical bands in terms of barks (Zwicker, 1961). The function LP is the constant 19.5-Hz bandwidth of our linear prediction spectra. The hatched area represents critical bandwidths typically reported in psychoacoustic research (see text).

whether running spectral features would be prominently displayed in auditory filter representations of running spectral displays. To explore this problem further, we decided to re-examine the stimuli in our data base with two auditory filter representations. SPECTRUM was modified so that a set of programmable, sliding filters could be applied to the previously computed spectral sections. Two properties of the filters were adjustable: bandwidth and the slopes of the skirts. Fixed properties included symmetry in a log-frequency space and a flat top giving an overall trapezoidal shape. These fixed filter properties were chosen to match 1/3-octave filter bank specifications. The filters were overlapped in close frequency intervals in order to produce smooth spectra. Pilot work showed that for the spectra used in this experiment, the convolution of the programmable filters with the smoothed linear prediction spectra produced filtered spectra almost identical to those produced by convolution with the equivalent 200-point FFT. Thus the smoothed linear prediction spectra were used as input to the programmable filters in this analysis.

The programmable filters were selected as follows. One-third-octave filters were chosen because they have frequently been used in speech processing, most recently by Searle *et al.* (1979, 1980), and they represent a frequency resolution which is probably poorer than that of the human auditory system. The 1/3-octave filters were digitally defined according to the ANSI S1.11-1966 class III standard, with a bandwidth constant of 0.23 times the center frequency, and skirts having a 50 dB/oct rolloff. The other filters, although similar to the auditory filters of Bladon and Lindblom (1979) and Klatt (1979), were patterned more closely

after the narrower filters proposed by Patterson (1976) and the 1/6-octave filters of Flanagan and Christensen (1980). The bandwidth constant was 0.13 and the skirts had a 75 dB/oct rolloff. Below 400 Hz, regardless of the bandwidth constant, the bandwidth was fixed at 95 Hz in keeping with standard critical-band measurements as shown in Fig. 2.

The running spectral display was also altered for spectra processed by the auditory filters by implementing a log-frequency scale. Amplitude was still displayed in decibels, and the time frame rate was kept at 5 ms. The 1/3-octave running spectra produced by this digital method are not directly comparable to 1/3-octave filter bank running spectra (see Searle *et al.*, 1979). In an analog filter bank, as well as in the ear, the time constant for high-frequency spectral components is shorter than that for low-frequency components. Thus rapid change in high-frequency information has a finer temporal resolution than low-frequency change. This relationship is not preserved in a digital analysis based on windowing where rapid changes in high frequencies are averaged over the window duration. This should not be a serious problem for our linear prediction analysis because our window size and the 5-ms frame rate appear to preserve the temporal variation observed the speech waveform as discussed earlier. Nonetheless, the 1/3-octave and auditory filter running spectra presented below are only approximations to true filter bank analyses.

With these programming changes, the running spectra that we previously examined could be redisplayed using 1/3-octave filters or auditory filters. Figure 3 compares three linear prediction running spectra with auditory filter displays. Figure 4 compares three different linear prediction

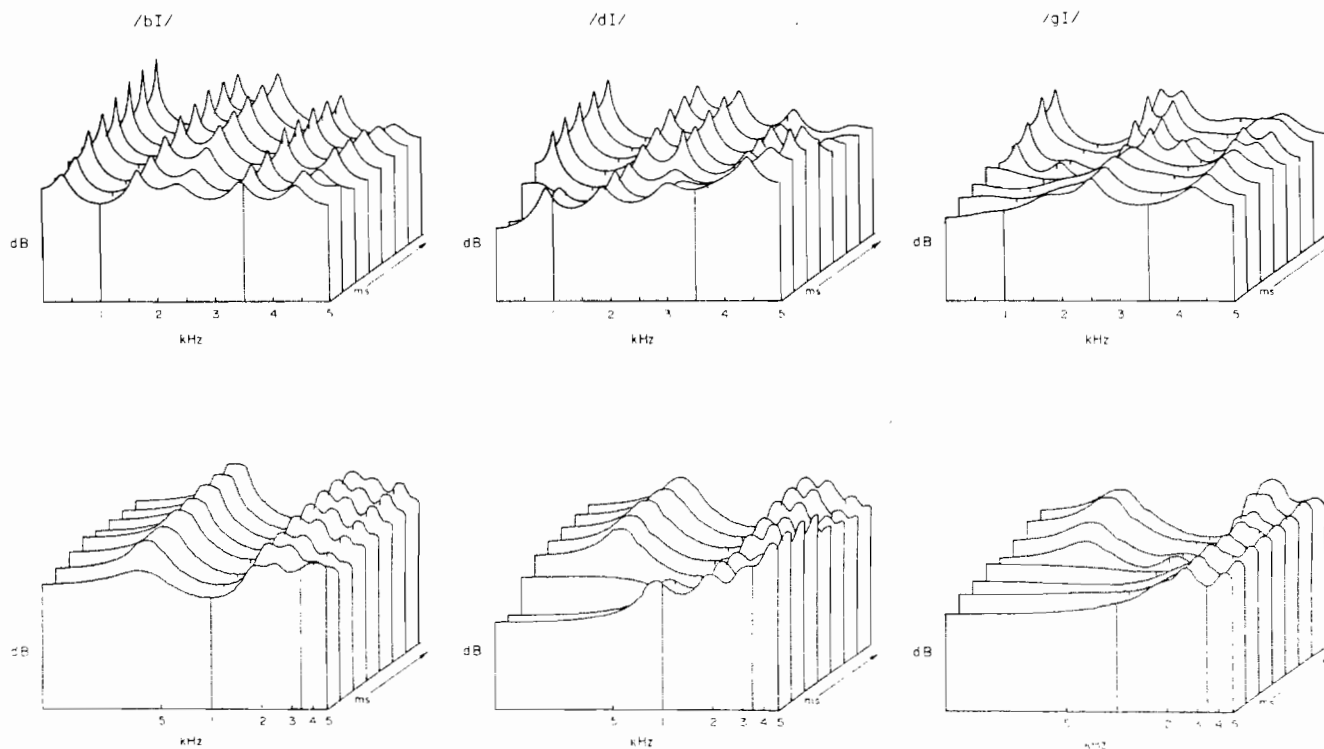


FIG. 3. Comparison of running spectral displays produced by either linear prediction analysis (top), or smoothed by auditory filtering of the Patterson type (bottom) for three stop-vowel syllables.

running spectra with 1/3-octave filter displays.

We re-examined the displays computed earlier using both types of filters to determine the extent to which the three visual features used to identify place were still present. In carrying out these analyses, we were interested in determining how such filtering would potentially alter our earlier feature descriptions. Approximately half of the 116 displays which had been correctly identified by all three judges in the independent judging were examined first. Then all 46 displays from the collaborative judging were redisplayed and examined visually by the experimenter.

An examination of Figs. 3 and 4 reveals that an auditory filter representation changes the frequency space in three important ways. First, the low-frequency region of  $F1$  is more prominently displayed. Second, filtering alters the spectral tilt of each spectral section. Because bandwidths are broader at higher frequencies, more energy is averaged into the high-frequency filters causing an upward spectral tilt. Thus both auditory filters and 1/3-octave filters result in a nonlinear transformation of spectral tilt which emphasizes high-frequency energy in comparison to the linear prediction spectra due to the imposed constant bandwidth of the filters at low frequencies (see Fig. 2). Finally, the spectral peaks move toward higher frequencies because the filters are symmetrical in log-frequency space, which means that they include more high-frequency energy than low-frequency energy in a linear frequency domain. Klatt (1979) has recently implemented Patterson's hypothesis (1974; Patterson and Nimmo-Smith, 1980) that auditory filters are symmetrical in linear frequency, but the log-frequency symmetry has been more commonly used. As a result of this shift, the feature

definitions referring to the 3500-Hz marks shown on the displays should probably be altered to approximately 4000 Hz in the following discussion.

The results of our examination of the auditory filter representations of running spectra may now be compared to the earlier feature definitions based on linear prediction spectra. The Tilt of burst category definitions are altered in a similar way using either 1/3-octave or auditory filters due to the high-frequency emphasis. The categories should now be *strongly rising* for /d/ versus *moderately rising* for /b/. Looking at running spectra on which an error occurred such that a /b/ had been misclassified as /d/, the filtering appeared to disambiguate these cases specifically because of the high-frequency emphasis. However, the current auditory filter representations did not incorporate a transformation of the present amplitude dimension from decibels to sones or some other measure of equal loudness. While the effects of implementing equal loudness contours on the Tilt of burst feature are yet to be explored, we note that this would alter *relative* amplitude. Since the Tilt feature is itself relative, we may presume that its specific definition might change, but not its effectiveness as a feature. To summarize our observations, both the 1/3-octave and auditory filter representations of the Tilt of burst categories appeared to be as successful as the earlier linear prediction displays.

The definition of the Late onset feature is not altered significantly by filter representations of running spectra. The  $F1$  peak, as shown in Figs. 3 and 4, is described as a broad, well-formed, low-frequency peak. Otherwise, no other aspects of this feature appeared to change for either 1/3-octave or auditory filters.

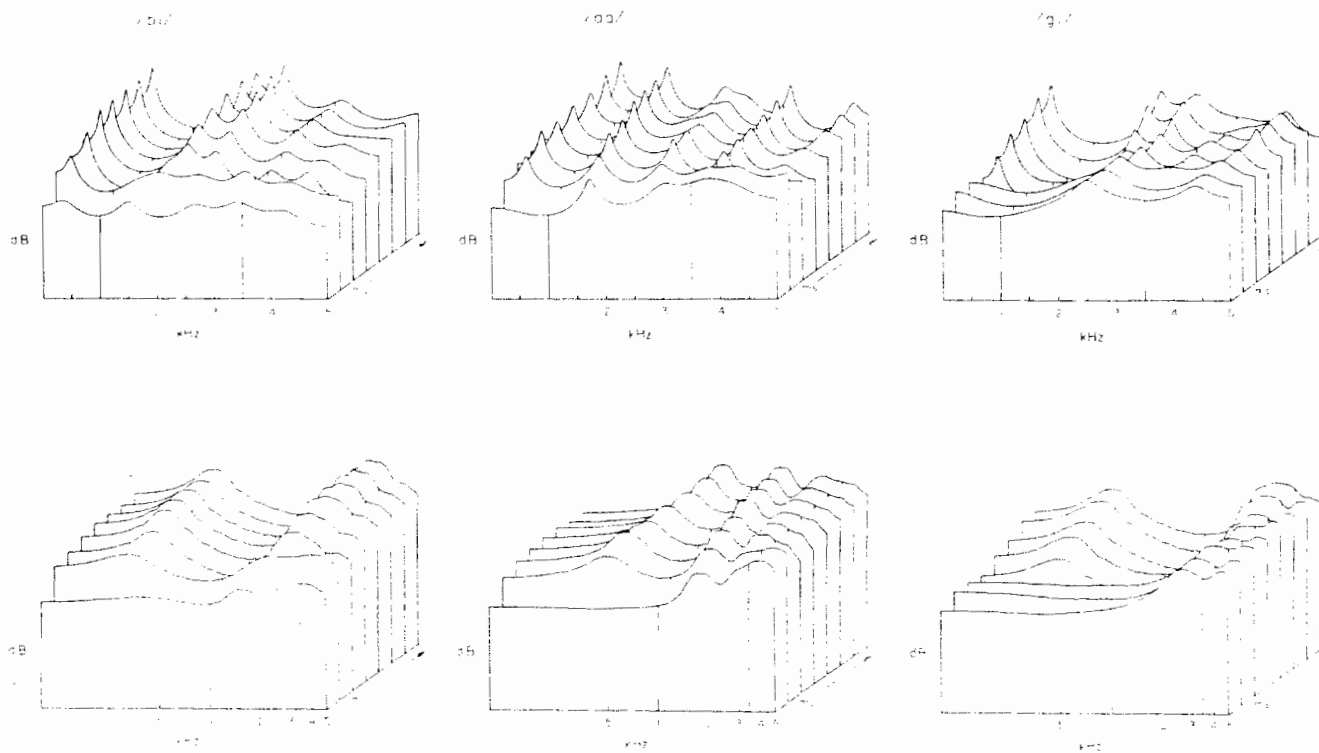


FIG. 4. Comparison of running spectral displays produced by either linear prediction analysis (top) or smoothed by 1/3-octave filters (bottom) for three stop-vowel syllables.

The visual display of the Mid-frequency peak feature changed quite a bit under the two types of filtering. Single prominent peaks were still readily observable for /g/'s, but they were narrower, had more "ripple," and appeared in more spectral sections (see Fig. 3). A significant problem occurred for the 1/3-octave filters which did not occur for the auditory filters. Many /b/'s analyzed by 1/3-octave filters acquired prominent mid-frequency peaks like those shown for /bi/ on Fig. 4. These peaks would cause them to be misclassified as /g/'s. For these same /b/ stimuli, auditory filtering did not produce mid-frequency peaks. Auditory filtering appeared to disambiguate many /g/ displays which had previously been misclassified either because two smaller peaks were superimposed on the prominent peak, or because the prominent peaks occurred on fewer than three spectra. Thus the results from this analysis indicate that mid-frequency peaks for /g/'s are more salient in running spectral displays using auditory filtering than in those using linear prediction filtering. However, the 1/3-octave filter representations would cause many more /b,g/ confusions.

In summary, the present study demonstrated that three acoustic features could be used to accurately identify place of articulation from linear prediction running spectra prepared for *visual* inspection. However, linear prediction spectra provide a much finer spectral resolution than the human auditory system carries out. Therefore two other spectral representations were used to construct running spectral displays. In the 1/3-octave filter representation, all features were preserved in the visual displays, but, unfortunately, too many mid-frequency peaks were erroneously produced for /b/'s. Thus it appears that 1/3-octave filtering may eliminate some spectral properties that are important for speech analysis. These filters generate spectra having a poorer frequency resolution than that of the human auditory system. The other filter representation was representative of a class of auditory filters currently used by other investigators. The original features proposed in this study were all displayed robustly in auditory filter running spectra, and, in some cases, appeared to display place of articulation more successfully than did the linear prediction running spectra. Thus it appears that the proposed time-varying features can be used to classify place of articulation in initial stop syllables displayed in two different spectral representations of speech, namely, linear prediction and auditory filtered running spectra.

#### IV. EXTENSIONS OF RUNNING SPECTRAL ANALYSIS FOR PLACE CUES

Based on the results of this study and on the preceding discussion, it is now possible to recommend specific changes in both the definitions of the features and the analysis procedures which can be used in future extensions of this research. Special attention will be given to the problem of vocal-tract normalization, i.e., accounting for the differences obtained between the male and female talkers in this study.

In terms of analysis techniques which compute the running spectra, further exploration of the auditory filter representations should be carried out. Since the human auditory

system is the best speech processing device currently available, implementation of analogous spectral processing capabilities may make the task of locating reliable acoustic cues for speech easier. This appeared to be the case for the auditory filters examined in this study.

Turning to the individual feature definitions, we begin with the feature of Mid-frequency peaks extending in time. The frequency range of 1000–3500 Hz used in the Mid-frequency peaks definition was determined from the earlier pilot study using male speaker RP. The results of the present investigation demonstrated that this range was unsuitable for the female talker because all potential peaks for /gi/ fell outside this range. Fant (1973) has already provided an explanation of this problem. Fant presented a table showing which formant is associated with the compact resonance peaks observed for /g/. For /i/ and /e/, the compact peak is associated with both  $F_3$  and  $F_4$ . That is, for high front vowels, the more palatal constriction for /g/ produces a short vocal-tract resonance cavity. The formant peak of this cavity is high and close to  $F_3$  and  $F_4$ . Thus, in order to capture a prominent peak for /g/, the mid-frequency range must be placed higher than  $F_4$  by approximately 500 Hz. The frequency range for  $F_4$  is, of course, dependent on a talker's vocal-tract size and that will have to be considered in future analyses.

Likewise, the lower value of the mid-frequency range was originally set at 1000 Hz based on talker RP's velar peaks. The lowest peaks for velars occur before the vowels /o/ and /u/ which are associated with the talker's  $F_2$  resonance (see /gu/ on Fig. 1). Although the prominent peaks for /o/ and /u/ are continuous with  $F_2$ , they fall from a higher frequency in the burst into the steady state  $F_2$  for the vowel. Based on these observations, it appears that a simple rule to account for vocal-tract size can be implemented in the definition of the frequency range for Mid-frequency peaks which can solve this problem. The lower frequency limit should be placed at the lower frequency of a talker's  $F_2$  for /o/ or /u/, and the upper limit should be placed 500 Hz higher than a talker's  $F_4$ . These values are for linear prediction spectra. If auditory filtering is used, the lower limit would not change, but the upper limit should be set about 1000 Hz higher than the talker's  $F_4$  because of the spectral averaging of the higher frequencies.

Next consider the Tilt of burst feature. We noted earlier that more /b/ errors occurred for the high vowels than the low vowels, and more errors occurred for the female than the male talkers. That is, rising spectral tilts were sometimes obtained in the 5-ms, burst-only spectra for /b/, apparently due to high-frequency energy associated with certain vowel contexts. Earlier in this discussion we suggested that this problem might be solved by changing the definition of rising tilt to be rising more prominently for /d/'s. We also note that with auditory filtering an emphasis of high frequencies was observed which had the beneficial effect of sorting more clearly /b/'s from /d/'s, particularly for the more ambiguous high vowel cases. We should note, however, that the Tilt of burst definition also includes an upper frequency limit, previously set to 3500 Hz. This limit was imposed because the burst-only spectra for /d/ are *not* in fact "diffuse rising"

up to 5000 Hz as Stevens and Blumstein (1978) have suggested. Rather, for /d/ before back vowels, a vowel dependent peak in the spectra occurs for males at about 3000 Hz so that the spectra fall above this frequency. Both Zue (1976) and Klatt (1980) have previously reported this spectral property. This peak is vowel dependent and therefore varies with vocal-tract size. Thus the upper frequency limit for the Mid-frequency peaks and Tilt of burst features must take into account vocal-tract size, and probably can be set to the same value using the previously suggested rule. The problems in the feature analysis discussed above arose from an interaction of vowel context effects and differences in vocal-tract size. The specific solution proposed here is that by properly accounting for differences in vocal-tract size in the feature definitions, the features will automatically specify place of articulation independent of the vowel context.

Finally, the feature of Late onset of  $F_1$ , as previously discussed, should be considered as a secondary feature for separating velars with long VOT's from bilabials and alveolars. Suggestions for a new definition of this feature and the resulting matrix were previously given at the end of Sec. IID. Treating delay in voicing onset as a secondary correlate to place of articulation has also been proposed by Fant (1973, p. 136). If Late onset can be used as a secondary cue to place, it might be worthwhile to determine if any empirical evidence could be found for the use of this feature in the perception of place of articulation. Several studies have already demonstrated that identification of place of articulation changes when the VOT of the stimuli was manipulated (Sawusch and Pisoni, 1974; Miller, 1977; Oden and Massaro, 1978). All three of these studies, however, used synthetic stop-vowel syllables without release bursts. Therefore a more careful synthesis study using velar syllables with bursts containing the important Mid-frequency peaks feature is needed to verify the specific role of the Late onset feature in perception of place of articulation.

The adequacy of these three features for specifying place of articulation was determined using human observers who examined visual displays of running spectra. The presence of the three features in the running spectra could, in principle, be determined algorithmically by computer. However, since this experiment was an initial study of these features, human observers were used to obtain feedback about possible ambiguities in the definitions or procedures in the collaborative judging session. In future experimentation with these features, a machine implementation of the feature definitions will certainly be included.

## V. DISCUSSION AND CONCLUSIONS

The results of this study demonstrate that invariant features for identifying place of articulation in initial stop consonants are readily observable in continuous running spectral displays of CV syllables. This experiment was an initial attempt to establish the adequacy of the analysis procedures and proposed feature descriptions for identifying place across a large number of vowel contexts and several talkers. Although this research has a theoretical framework very similar to that of Stevens and Blumstein (1978, 1981; Blum-

stein and Stevens, 1979), our analysis differs from theirs in a very important way: running spectral features incorporate the *time* dimension whereas Stevens and Blumstein onset spectra are basically static spectral "snapshots" of the continuously changing speech signal. Indeed, Blumstein and Stevens (1979, p. 1013) have acknowledged that the time dimension might have to be incorporated into their analysis. After our pilot work was completed (Kewley-Port, 1979a,b), Blumstein and Stevens (1979, p. 1013) suggested that an analysis procedure similar to the running spectra of Searle *et al.* (1979) might be an improvement over their single-spectrum static template procedure. The present findings demonstrate that this suggestion is correct.

Our running spectral analysis contrasts with the static template analysis proposed by Stevens and Blumstein in several ways. For example, by integrating energy over a 26-ms time window into a single onset spectrum, important spectral differences that are present for the labial and alveolar bursts are obscured. A 26-ms window will always include some transitional information about the vowel (aspirated or voiced) along with the energy in the burst. Furthermore, a single, fixed integration window cannot account for the differences in the *temporal* properties of the release burst as described previously by Fant (1960, 1973). The rapidly changing spectra following the burst for bilabials and alveolars cannot be observed in only a single 26-ms onset spectra. But more importantly, the velar property of a *slowly varying* compact spectrum extending in time cannot be represented in only a single onset spectrum having no temporal dimension. Blumstein and Stevens have stated in their recent publications that the compact spectra for velars must persist in time for listeners to identify velars correctly (1979, p. 1002; 1980, p. 652). Nevertheless, their single onset spectra cannot, in principle, represent this acoustic information adequately.

In fact, the velar template constructed by Blumstein and Stevens (1979) is essentially a peak detector. However, in carrying out their analysis, they discovered that simple integration of the first 26 ms of spectral energy produced numerous spectra containing prominent peaks which were not velars. To be precise, 27% of the alveolar consonants had spectral peaks near the  $F_2$  locus at 1800 Hz. This observation prompted Stevens and Blumstein to modify the original diffuse rising template so as to exclude peaks occurring around 1800 Hz (1979, p. 1005). Furthermore, alveolar peaks said to arise from subglottal resonances also occurred in the 800–1600-Hz region. These peaks were also arbitrarily excluded from the diffuse-rising template (1979, p. 1005). Moreover, a problem for the compact template was the presence of double peaks. Blumstein and Stevens (1979, p. 1006) apparently treated two spectral peaks separated by less than 500 Hz as a single peak. Thus it appears that attempts to locate the spectral feature compact as a simple peak in a single, 26-ms integrated spectrum have given rise to numerous exceptions and the postulation of *ad hoc* decision rules. Similar problems did not arise in our analysis of running spectral displays. We suspect that the problems were present for Stevens and Blumstein because the temporal dimension of the speech signal was eliminated in their static onset spectral analysis.

The present investigation was limited to the initial voiced stops /b,d,g/ and several other phonetic parameters still await investigation. Foremost is voicing. In particular, it is of some interest to determine if our analysis will correctly identify place of articulation for /ptk/ as well as /bdg/. Clearly, the two primary features, Tilt of burst and Mid-frequency peaks, are located mostly in the burst portion of the spectral displays. Since this portion of the spectra should be very similar for voiced and voiceless consonants, we may confidently predict that our proposed features will successfully identify place in /ptk/. The secondary feature of Late onset of  $F_1$ , on the other hand, will probably need some further modification. Only research with both the voiced and voiceless consonants will determine whether the Late onset feature can be defined in such a way that it can act as a reliable secondary cue for identifying /g/ in various vowel contexts.

The features used in this study were used to identify place in what has been called *initial* stops. Initial in the context of this experiment means *syllable initial* since, in fact, all CV's examined here were originally extracted from the carrier sentence "Teddy said CV." No claims or hypothesis are being proposed on the basis of this experiment for presence of these features in running spectra of syllable final stops, or in running spectra of segments from other manner classes such as nasals or fricatives. For example, the Tilt of burst feature, as discussed, contains as part of its definition the location of a burst following a closure interval. Since this sequence of acoustic events is not found in nasals, the feature would not apply to the class of nasal consonants. This conclusion is quite different from the proposals made by Stevens and Blumstein (1978, 1981). The invariant acoustic cues they propose in terms of onset spectra are linked to the general notion of distinctive features for place as proposed by Jakobson *et al.* (1952) and Chomsky and Halle (1968). Thus Stevens and Blumstein specifically claim that onset spectra can and should correctly specify place for final stops and nasals. However, their own research provides little convincing evidence to support this claim. For final stops in the Blumstein and Stevens' template study (1979), the average correct identification of place at closure was 53%, and identification of the final burst (which is not typically present in running fluent speech) was 76%. In the preliminary study of [n] versus [m], average place identification was 76%. However, [n]'s were accepted by both the labial and alveolar template 67% of the time, so that the unique identification of [n]'s was at best only about 33%. Therefore the combined results of uniquely identifying [n] versus [m] was near the 50% chance level for two choices. These results do not support the strong claims of Stevens and Blumstein that static onset-spectra templates can reliably capture the invariant properties for the distinctive feature of place in all environments or across several manner classes.

In conclusion, we have proposed three time-varying, relational acoustic features as a principled solution to the problem of invariant acoustic cues for place of articulation in initial stop consonants. These features are clearly observable in visual representations of running spectral displays of naturally produced CV syllables. The present study evaluated

this proposed analysis for voiced stops before a large number of vowels produced by three talkers. From our results, it appears that these features are invariant over vowel context, and with the addition of two simple rules, appear to be invariant over vocal-tract size as well. The running spectra used in this study appropriately model some of the peripheral processing characteristics of the human auditory system. The time-varying features used to specify place of articulation in this experiment were also shown to be robustly displayed in another visual representation which more closely approximated the filtering properties of the auditory system. Although the features examined in this experiment may be ultimately limited to identifying place in syllable-initial, voiced and voiceless stops, it is fully expected that examination of auditory spectral representations of speech signals in which the time dimension is properly preserved will also be successful in determining the acoustic correlates of other classes of speech sounds as well.

## ACKNOWLEDGMENTS

I am very grateful to David B. Pisoni who encouraged me to pursue this research as part of a doctoral dissertation submitted to the Graduate Center of The City University of New York. I am particularly indebted to him for the time and support he gave while conducting the research and in preparing the manuscript. In addition, I wish to thank Michael Studdert-Kennedy, Dennis H. Klatt, Katherine S. Harris, and Donald Robinson for carefully reading earlier versions of this manuscript. Some of the results of this research was presented before the Acoustical Society of America at the 97th meeting in Cambridge, MA and the 100th meeting in Los Angeles, CA. This research was supported by the National Institutes of Health, Research Grant NS-12179-05 and the National Institute of Mental Health, Research Grant MH-24027-06 to Indiana University in Bloomington.

- Bladon, R. A. W., and Lindblom, B. (1979). "Auditory modeling of vowels," in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, edited by J. J. Wolf and D. H. Klatt (Acoustical Society of America, New York), pp. 1-4.
- 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.
- Carlson, R., and Granstrom, B. (1980). "Model predictions of vowel dissimilarity," *Q. Prog. Status Rep. STL-QPRS 3-4*, Speech Transmission Laboratory, Stockholm, 84-104.
- Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English* (Harper and Row, New York).
- Cole, R. A., and Scott, B. (1974). "The phantom in the phoneme: Invariant cues for stop consonants," *Percept. Psychophys.* **15**, 101-107.
- 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.
- Delgutte, B. (1980). "Representations of speech-like sounds in the discharge patterns of auditory-nerve fibers," *J. Acoust. Soc. Am.* **68**, 843-857.
- Delgutte, B. (1981). "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," unpublished doctoral thesis, MIT, Cambridge, MA.

- Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues." *Percept. Psychophys.* **22**, 109-122.
- Fant, G. (1980). *Acoustical Theory of Speech Production* (Mouton, The Hague, The Netherlands).
- Fant, G. (1973). "Stops in CV-syllables," in *Speech Sounds and Features*, edited by G. Fant (MIT, Cambridge, MA), pp. 110-139.
- Flanagan, J. L. (1980). "Parametric coding of speech spectra," *J. Acoust. Soc. Am.* **68**, 412-419.
- Flanagan, J. L., and Christensen, S. W. (1980). "Computer studies on parametric coding of speech spectra," *J. Acoust. Soc. Am.* **68**, 420-430.
- Halle, M., Hughes, G. W., and Radley, J. P. A. (1957). "Acoustic properties of stop consonants," *J. Acoust. Soc. Am.* **29**, 107-116.
- Jakobson, R., Fant, C. G. M., and Halle, M. (1952). *Preliminaries to Speech Analysis* (MIT, Cambridge, MA).
- Kewley-Port, D. (1979a). "Continuous spectral change as acoustic cues to place of articulation," *Res. Speech Percept., Prog. Rep. No. 5*, Indiana University, 327-346.
- Kewley-Port, D. (1979b). "Spectral continuity of burst and formant transitions as cues to place of articulation in stop consonants," in *Speech Communication Papers Presented at the 97th Meeting of The Acoustical Society of America* (Acoustical Society of America, New York), pp. 175-178.
- Kewley-Port, D. (1979c). "Spectrum: A program for analyzing the spectral properties of speech," *Res. Speech Percept., Prog. Rep. No. 5*, Indiana University, 475-492.
- Kewley-Port, D. (1980). "Representations of spectral change as cues to place of articulation in stop consonants," *Res. Speech Percept., Tech. Rep. No. 3*, Indiana University.
- Kewley-Port, D. (1982). "Measurements of formant transitions in naturally produced stop consonant-vowel syllables," *J. Acoust. Soc. Am.* **72**, 379-389.
- Kiang, N. Y. S. (1980). "Processing of speech by the auditory nervous system," *J. Acoust. Soc. Am.* **68**, 830-835.
- Kiang, N. Y. S., Eddington, D. K., and Delgutte, B. (1979). "Fundamental considerations in designing auditory implants," *Acta Otolaryngol.* **87**, 204-218.
- Klatt, D. H. (1976). "A digital filter bank for spectral matching," in *Conference Record of the 1976 IEEE International Conference on Acoustics Speech and Signal Processing*, edited by C. Teacher (IEEE Catalog No. 76CH1067-8 ASSP, Philadelphia, PA), pp. 537-540.
- Klatt, D. H. (1979). "Speech perception: A model of acoustic-phonetic analysis and lexical access," *J. Phonet.* **7**, 279-312.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971-995.
- Klein, W., Plomp, R., and Pols, L. C. W. (1970). "Vowel spectra, vowel spaces, and vowel identification," *J. Acoust. Soc. Am.* **48**, 999-1009.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," *Psychol. Rev.* **74**, 431-461.
- Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," *Psychol. Monogr.* **68** (8, Whole No. 379), 1-13.
- Lisker, L., and Abramson, A. S. (1964). "A cross-language study of voicing in initial stops: Acoustical measurements," *Word* **20**, 384-422.
- Markel, J. D. (1971). "FFT pruning," *IEEE Trans. Audio Electroacoust.* **AU-19**, 305-311.
- Markel, J. D. (1972). "Digital inverse filtering—a new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.* **AU-20**, 129-137.
- Markel, J. D., and Gray, A. H. (1976). *Linear Prediction of Speech* (Springer-Verlag, New York).
- Miller, J. L. (1977). "The perception of voicing and place of articulation in initial stop consonants: Evidence for the nonindependence of feature processing," *J. Speech Hear. Res.* **20**, 519-528.
- Monsen, R. B. (1961). "Accuracy of formant frequency estimation by spectrograms and by linear prediction analysis," *J. Acoust. Soc. Am. Suppl.* **1** **69**, S17.
- Oden, G. C., and Massaro, D. W. (1978). "Integration of featural information in speech perception," *Psychol. Rev.* **85**, 179-191.
- Patterson, R. D. (1974). "Auditory filter shape," *J. Acoust. Soc. Am.* **55**, 802-809.
- Patterson, R. D. (1976). "Auditory filter shapes derived with noise stimuli," *J. Acoust. Soc. Am.* **59**, 640-654.
- Patterson, R. D., and Nimmo-Smith, I. (1980). "Off-frequency listening and auditory-filter asymmetry," *J. Acoust. Soc. Am.* **67**, 229-245.
- Potter, R. K., Kopp, G. A., and Green, H. C. (1947). *Visible Speech* (Van Nostrand, New York).
- Sachs, M. B., and Young, E. D. (1979). "Encoding of steady-state vowels in the auditory nerve: Representations in terms of discharge rate," *J. Acoust. Soc. Am.* **66**, 470-479.
- Sachs, M. B., and Young, E. D. (1980). "Effects of nonlinearities on speech encoding in the auditory nerve," *J. Acoust. Soc. Am.* **68**, 858-875.
- Sawusch, J. R., and Pisoni, D. B. (1974). "On the identification of place and voicing features in synthetic stop consonants," *J. Phon.* **2**, 181-194.
- Scharf, B. (1970). "Critical bands," in *Foundations of Modern Auditory Theory*, Vol. 1, edited by J. V. Tobias (Academic, New York), pp. 157-202.
- Schouten, M. E. H., and Pols, L. C. W. (1979). "CV- and VC-transitions: A spectral study of coarticulation—Part II," *J. Phonet.* **7**, 205-224.
- Schroeder, M. R., Atal, B. S., and Hall, J. L. (1979). "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.* **66**, 1647-1652.
- 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 Kimberly, B. P. (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.
- Sever, J. C., and Small, A. M. (1979). "Binaural critical masking bands," *J. Acoust. Soc. Am.* **66**, 1343-1350.
- Stevens, K. N. (1975). "The potential role of property detectors in the perception of consonants," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, New York), pp. 303-330.
- Stevens, K. N. (1980). "Acoustic correlates of some phonetic categories," *J. Acoust. Soc. Am.* **68**, 836-842.
- 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., 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. 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.
- Young, E. D., and Sachs, M. B. (1979). "Representations of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Am.* **66**, 1381-1403.
- Zhukov, S. Ya., Zhukova, M. G., and Chistovich, L. A. (1974). "Some new concepts in the auditory analysis of acoustic flow," *Sov. Phys. Acoust.* **20**, 237-240 [*Akust. Z.* **20**, 386-392 (1974)].
- Zue, V. W. (1976). "Acoustic characteristics of stop consonants: A controlled study," *Tech. Rep. 523*, Lincoln Laboratory, MIT, Cambridge, MA.
- Zwicker, E. (1961). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Am.* **33**, 248.
- Zwicker, E., Terhardt, E., and Paulus, E. (1979). "Automatic speech recognition using psychoacoustic models," *J. Acoust. Soc. Am.* **65**, 487-498.