Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants

Sheila E. Blumstein

Department of Linguistics, Brown University, Providence, Rhode Island 02912

Kenneth N. Stevens

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 24 November 1978; accepted for publication 6 July 1979)

On the basis of theoretical considerations and the results of experiments with synthetic consonant-vowel syllables, it has been hypothesized that the short-time spectrum sampled at the onset of a stop consonant should exhibit gross properties that uniquely specify the consonantal place of articulation independent of the following vowel. The aim of this paper is to test this hypothesis by measuring the spectrum sampled at the onsets and offsets of a large number of consonant–vowel (CV) and vowel–consonant (VC) syllables containing both voiced and voiceless stops produced by several speakers. Templates were devised in an attempt to capture three classes of spectral shapes: diffuse-rising, diffuse-falling, and compact, corresponding to alveolar, labial, and velar consonants, respectively. Spectra were derived from the utterances by sampling at the consonantal release of CV syllables and at the implosion and burst release of VC syllables, and these spectra (smoothed by a linear prediction algorithm) were matched against the templates. It was found that about 85% of the spectra at initial consonant release and at final burst release were correctly classified by the templates, although there was some variability across vowel contexts. The spectra sampled at the implosion were not consistently classified. A preliminary examination of spectra sampled at the release of nasal consonants in CV syllables showed a somewhat lower accuracy of classification by the same templates. Overall, the results support an hypothesis that, in natural speech, the acoustic characteristics of stop consonants, specified in terms of the gross spectral shape sampled at the discontinuity in the acoustic signal, show invariant properties independent of the adjacent vowel or of the voicing characteristics of the consonant. The implication is that the auditory system is endowed with detectors that are sensitive to these kinds of gross spectral shapes, and that the existence of these detectors helps the infant to organize the sounds of speech into their natural classes.

PACS numbers: 43.70.Gr, 43.70.Dn, 43.70.Ve

INTRODUCTION

It is a known and well-accepted fact that the speech sounds occurring in natural language share a limited set of fundamental characteristics. Specifically, it has been hypothesized that man is born with a predisposition to perceive and produce sounds which are biologically significant. It has been suggested by Stevens (1972) that the defining properties of speech are determined by certain constraints imposed by the articulatory system on the one hand and by the perceptual system on the other. At one level, it is obvious that the articulatory and auditory systems have certain intrinsic limitations, i.e., a sound can not be produced if it is beyond the physiological capacity of the articulatory system, or be perceived if its acoustic characteristics are beyond the auditory capacity of the system. However, the set of speech sounds seems to be constrained in a more fundamental and theoretically interesting way. In particular, it has been shown that small changes in some articulatory configurations or states produce large changes in certain attributes of the resultant acoustic waveform. In contrast, similar changes in other articulatory configurations minimally affect the attributes of the acoustic waveform. Thus, only a limited set of articulatory configurations produce stable acoustic patterns. These acoustic patterns are not only relatively insensitive to perturbations in the articulatory configurations but are also distinctive in the sense that they have acoustic attributes that are different from the properties of the sounds produced by other configurations. It is this set of stable acoustic configurations with their distinctive properties which theoretically define the finite set of speech sounds used in natural language.

Similarly, there seem to be perceptual constraints on the human auditory system which also restrict the possible range of properties of sounds which can be used in natural language. These constraints are evidenced most clearly by the well-known phenomenon of categorical perception. (Liberman *et al.*, 1967, 1957; Cutting and Rosner, 1974; Pastore *et al.*, 1977). This phenomenon has been shown across modalities and for both speech and nonspeech stimuli. Suppose that a series of sounds is arranged along a continuum such that the sounds are equally spaced according to some physical measure. There are some acoustic continua for which the listener does not judge the items to be equidistant perceptually, but rather he has difficulty discriminating adjacent pairs of sounds in one part of the continuum and can easily discriminate adjacent pairs in another part of the continuum. The physical scale along which the sounds appear to differ in equal steps is not the same as the scale that is used by the auditory system to judge differences between the sounds. The scale

 0001-4966/79/101001-17$00.80

used by the auditory system is warped—in one part of the scale large physical differences are needed to produce small perceptual differences, whereas in another part of the scale, small physical differences are judged by a listener to be large. Sounds within the continuum can be discriminated only with difficulty, possibly requiring specially designed psychophysical procedures (Carney *et al.*, 1977), and thus can be regarded as having a common property. In the case of speech sounds, the pairs of stimuli which subjects cannot discriminate are identified as belonging to the same phonetic class, whereas the stimulus pairs that are discriminated belong to different phonetic classes. The perceptual system, then, seems to be restricted in the degree to which it can effectively perceive and ultimately use the various acoustic attributes in a given speech signal.

It is the complex interface of these articulatory and perceptual constraints which theoretically restrict the properties of speech sounds found in natural language. Nevertheless, given these constraints, it is far from clear what, in fact, constitutes the full range of properties of speech sounds used in both the perception and production of speech. It is the object of this paper to investigate the nature of the properties intrinsic to the place-of-articulation dimension for consonants. We will argue that there are different integrated acoustic properties which manifest acoustic invariance for different places of articulation. These properties reflect the configuration of acoustic events occurring at the release of the stop consonant, and reside in the short-term spectrum sampled at the moment of consonantal release. The existence of such invariance suggests that the perception of speech makes use of property-detecting mechanisms which can extract the necessary information for perceiving place of articulation directly from the acoustic signal. As a consequence, abstract theories of recoding (Liberman *et al.*, 1967; Studdert-Kennedy *et al.*, 1970) contextual dependencies, (Liberman *et al.*, 1967) and analysis by synthesis (Halle and Stevens, 1972; Stevens and Halle, 1967) are not required to account for the ability of the listener to categorize the sounds of speech.

I. EVIDENCE FOR INVARIANT PROPERTIES: SOME THEORETICAL CONSIDERATIONS

There is evidence from acoustic theory to support the notion that invariant properties that define place of articulation for stop consonants can be derived from analysis of the short-time spectrum sampled at consonantal release (Fant, 1960; Stevens and Blumstein, 1978). The theory predicts the characteristics of the burst of sound energy generated at the release, and the approximate values of the natural frequencies of the vocal tract in the vicinity of the consonantal release. As a consequence of these formant positions and of the burst characteristics, the spectrum sampled over the 10–20 ms following the consonantal release can be shown to have gross characteristics that are different for different places of articulation. The theoretical analysis shows that these gross characteristics are exhibited by the burst spectrum, produced with a source of vocal-tract excitation in the vicinity of the consonantal constriction, and by the spectrum following the burst, when the source of vocal-tract excitation is at the glottis. When both components of the onset are present, the spectral characteristics for a particular place of articulation are enhanced relative to the characteristics that exist for either one of the components separately. In the case of velar consonants, the theoretically predicted common attribute of the spectrum is a major spectral prominence in the midfrequency range; for alveolar consonants, the spectral energy is diffuse or distributed throughout the frequency range, but with greater spectral energy at higher frequencies; when the consonant constriction is made at the lips, the spectral energy is again diffuse, but the spectrum is weighted towards lower frequencies. The theory predicts that similar gross characteristics should be observed for spectra sampled at the implosion and at the burst release for syllable-final consonants.

The observation, based on acoustic theory, that short-time spectra sampled at consonantal release show distinctive gross characteristics for different places of articulation suggests that these properties are utilized by the human speech perception mechanism in order to extract information concerning place of articulation. This hypothesis has been tested by a series of perception experiments (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979) in which the acoustic characteristics of the bursts and the transitions in synthetic consonant-vowel syllables were systematically manipulated in accordance with the theoretical principles noted above, and the stimuli were presented to listeners for identification. These experiments showed that:

(1) For a continuum of stimuli for which the physical differences are systematically manipulated to produce a range of consonant responses from [b] to [d] to [g], the stimuli at the phoneme boundaries (where listener responses are equivocal) have physical characteristics that are intermediate between the theoretically distinct properties described above (Stevens and Blumstein, 1978).

(2) Appropriate consonantal responses to stimuli constructed in accordance with the theoretical principles noted above are obtained whether or not bursts are appended at the onsets of the stimuli, although responses are more consistent when bursts are present (Stevens and Blumstein, 1978).

(3) Information signalling place of articulation for stop consonants appears to reside in the initial 10–20 ms following the consonantal release (Blumstein and Stevens, 1979; Winitz *et al.*, 1972).

(4) There is some evidence that, if a stimulus with an abrupt onset, such as a stop consonant, is to give rise to an auditory representation containing a narrow spectral peak, the spectral peak must persist for a few tens of milliseconds following the onset (Blumstein and Stevens, 1979).

In summary, the theoretical considerations and the perceptual experiments suggest that, for each place of articulation for stop consonants, the short-time spec-

trum sampled at stimulus onset should show distinctive acoustic properties. The gross spectrum shape is diffuse-falling or flat for labials, diffuse-rising for alveolars, and compact for velars. These conclusions are, however, based on certain idealized models concerning the acoustic properties of the bursts and of the formant onsets as the place of articulation is manipulated. It is the object of the present study to investigate the validity of these assumptions by examining the short-time spectra measured at the consonantal release for a large number of natural speech utterances. In particular, the aim is to determine whether these spectra can be characterized in terms of the acoustic properties diffuse-rising, diffuse-falling or flat, and compact, and whether these properties for each place of articulation are found independent of the following vowel, in different phonetic segments (voiced and voiceless stop consonants and nasals) and phonetic contexts (initial and final), and across different vocal-tract sizes (speakers).

II. ACOUSTIC ANALYSIS OF NATURAL SPEECH: BACKGROUND AND METHODOLOGY

The possibility that spectra sampled at the release of stop consonants provide distinctive shapes for place of articulation has been noted in several investigations of natural speech. Halle, Hughes, and Radley (1957), and Zue (1976) have shown that spectral analyses of the burst in isolation give rise to three classes of patterns associated with the three places of articulation—labial, alveolar, and velar. Searle, Jacobson, and Kimberley (1979) have also shown that acoustic events derived from spectra sampled in the initial few tens of milliseconds following the consonantal release can be used to separate stop consonants into categories according to place of articulation. The work of Fant (1960) and Jakobson, Fant, and Halle (1963) in particular has attempted to characterize the distinctive patterns derived from short-time spectral analyses of stop-vowel utterances.

Examples of spectra for several naturally produced voiced and voiceless consonants in English are shown in Fig. 1. These are linear prediction spectra obtained by pre-emphasizing the higher frequencies and using a 26-ms time window beginning at the consonantal release (see also Stevens and Blumstein, 1978). Owing to the different burst length and voice-onset time for these stops, the portion of the consonantal onset actually measured varies across the different consonants. Note that for [b], the 26-ms time window includes both burst and some portion of voicing onset, whereas for [g] essentially only the burst is measured. In the case of voiceless stops, the window includes the initial frication burst and possibly a portion of the aspiration, but does not extend into the onset of voicing. Nevertheless, examination of the spectral shapes for these onset spectra reveals the distinctive patterns described by Fant (1960) and Jakobson, Fant, and Halle (1963), and predicted by the theoretical analysis.

In order to provide a more quantitative measure of the degree to which these gross spectral shapes do in fact correspond to each place of articulation, we developed a series of templates designed to reflect each of the spectral properties—diffuse-rising, diffuse-falling or flat, and compact. The configurations of these templates were determined in part from theoretical considerations and in part from an examination of a limited set of consonant-vowel utterances consisting of the initial consonants [b d g] in the environment of the vowels [i e ɑ o u] produced by two male speakers. The adequacy of these templates for classifying stop consonants according to place of articulation was then assessed with a larger set of CV and VC utterances produced by six different talkers. In all cases the tape-recorded syllables were low-pass filtered at 4800 Hz and sampled at 10 kHz. The first difference of the waveform was calculated (in effect pre-emphasizing the high frequencies), the waveform was multipled by a modified raised cosine time window, and a smoothed spectrum was calculated using a 14-pole linear prediction algorithm. The window shape, shown in Fig. 1, puts greater emphasis on the earlier portions of the signal for a spectrum calculated at an onset. The size of the window was determined informally by inspecting the spectra obtained with different-sized windows. A 26-ms time window was chosen because it seemed to produce spectral shapes that were optimally similar to the theoretically derived curves.[1]

The three templates were derived by examining a few hundred utterances of the two male speakers, through a process of minor adjustments of the templates and reanalysis of the data. These adjustments and refinements of each of the three different templates continued until we felt that (1) the template reflected the theoretical shapes discussed above, and (2) it accepted the majority of the CV utterances for which it was designed and rejected the majority of the other utterances.

III. THE TEMPLATES

The general form of the three templates is schematized in Fig. 2. The two diffuse templates are represented by two reference lines about 10 dB apart, and the requirement for diffuseness is that at least two spectral peaks must lie within the region between the reference lines and that these peaks be separated by at least 500 Hz. This requirement ensures that there is some spread of energy across a range of frequencies. In the case of the diffuse-rising template, the higher-frequency peaks must be higher in amplitude than the lower-frequency peaks, whereas the opposite is true for the diffuse-falling template. The template that tests for compactness contains a single spectral peak in the midfrequency range, and the requirement is that a midfrequency prominence in the spectrum must fit within this peak so that no other spectral peak protrudes through the reference line. The peak can lie anywhere within a specified midfrequency range; the figure shows an example of such a peak. In general, a spectrum that satisfies the compactness property would fail to fit within the diffuse templates, since only one peak would lie within the region defined by the two reference lines. We now describe in more detail each of these templates and the conditions for applying them.

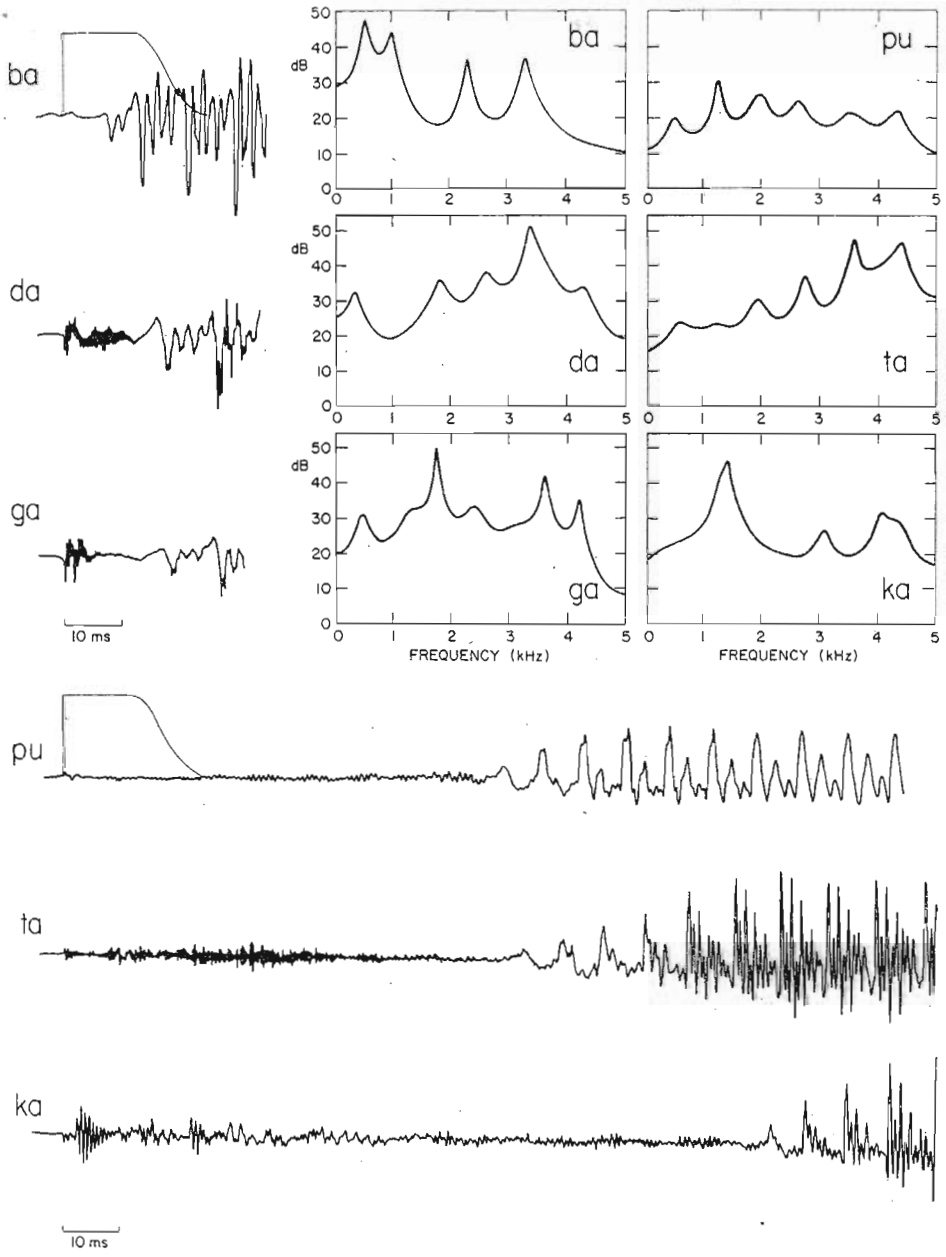

FIG. 1. Examples of waveforms and spectra sampled at the release of three voiced and three voiceless stop consonants as indicated. Superimposed on two of the waveforms is the time window (of width 26 ms) that is used for sampling the spectrum. Short-time spectra are determined for the first difference of the sampled waveform (sampled at 10 kHz) and are smoothed using a linear prediction algorithm, i.e., they represent all-pole spectra that provide a best fit to the calculated short-time spectra with pre-emphasis.

A. The diffuse-rising template

As indicated earlier, the property that describes the alveolar consonant can be characterized by a diffuse spread of energy in the spectrum sampled at the onset, with high-frequency peaks having greater amplitude than the lower-frequency peaks. This set of attributes was distilled into a template, as shown in the upper part of Fig. 3, and this template is interpreted in the following manner. We first identify a spectral peak above 2200 Hz such that this peak just touches an upward sloping reference line, and all other peaks above this frequency lie below the line. We next examine the distribution of peaks along the amplitude-frequency domain. For a spectrum to fit the template, at least two peaks must fall between the two reference lines, and a peak above 2200 Hz must be higher in amplitude than one other peak below it in frequency. Thus, the template describes a class of spectra with a diffuse spread of spectral energy characterized by a tilting slope upwards with no one peak dominating the entire spectrum. An example of a spectrum meeting the required characteristics is shown in the top half of Fig. 3, superimposed on the diffuse-rising template. Examples of spectra that do not have the diffuse-rising characteristics are shown in the bottom part of the figure. The spectrum of the [g] shows just one prominent peak, and thus does not satisfy the diffuseness requirement. Although the spectrum of [b] is diffuse, in that it contains energy spread over a range of frequencies, the spectral energy distribution is falling with increasing frequency, and thus does not fit within the template.

Analysis of a number of the natural CV utterances indicated that the spectra for the alveolar consonants sometimes contained two other characteristics not accounted for by the proposed diffuse-rising template. The first involved the occurrence of low-frequency peaks in the range 800–1600 Hz, which were found in many examples of the alveolar stop consonants. These

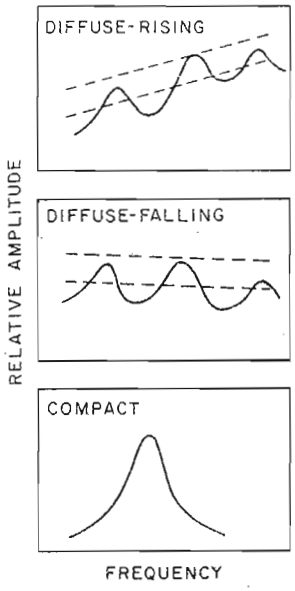

FIG. 2. Schematization of the diffuse-rising, diffuse-falling, and compact templates designed to capture the gross spectral shapes characteristic of alveolar, labial, and velar places of articulation, respectively. The diffuse templates require a spread of spectral peaks across a range of frequencies with increasing energy at higher frequencies for the diffuse-rising template and a falling or flat spread of energy for the diffuse-falling template. The compact template requires a single spectral peak in the midfrequency range.

peaks are probably a consequence of subglottal resonances that arise from an interaction between the subglottal system and the supraglottal system when there is a subglottal opening (Fant *et al.*, 1972). By convention, subglottal resonances, defined as peaks of energy occurring between 800 and 1600 Hz, that fall within 10–12 dB above the top reference line are ignored by the template. An example of such a spectrum superimposed over the diffuse-rising template can be seen in

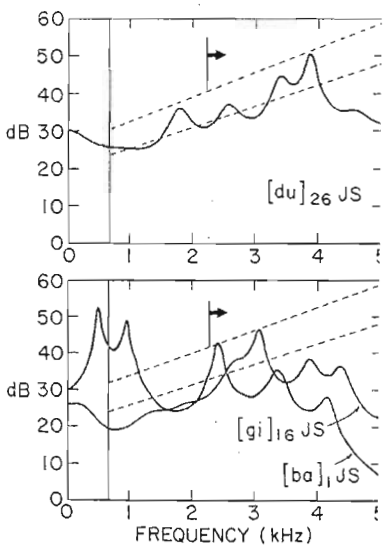

FIG. 3. Examples of short-time spectra which have been fit to the diffuse-rising template. The spectra are sampled at the consonantal releases of natural CV utterances. The top panel shows the onset spectrum of an alveolar consonant which meets the requirements of the template. The spectrum is adjusted vertically so that a spectral peak falling above 2200 Hz (indicated by the arrow) just touches the upper reference line. At least two peaks must fall within the reference lines, and a higher-frequency peak must be higher in amplitude than the lower-frequency peak. The bottom panel shows onset spectra for a velar and a labial consonant, neither of which meets the conditions of the diffuse-rising template. In the case of the labial consonant, only one peak of energy falls within the reference lines of the template, and a low-frequency peak juts above the top reference line. For the velar consonant, only one peak falls within the reference lines of the template.

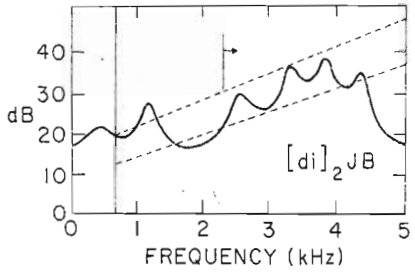

FIG. 4. Example of the onset spectrum for an alveolar consonant (superimposed on the diffuse-rising template) containing a subglottal resonance peak at about 1200 Hz which juts above the top reference line of the template. By convention, subglottal resonances are ignored by the template.

Fig. 4. Note that despite the fact that the spectrum contains low-frequency energy peaks that fall above the top reference line, the overall spectrum shape is still diffuse-rising, and two spectral peaks fall within the reference lines set by the template. These spectral peaks in the burst corresponding to subglottal resonances are usually not contiguous with peaks in the spectrum immediately following the burst, i.e., with vowel formant peaks. In effect, then, the convention to ignore these subglottal resonances amounts to ignoring a midfrequency spectral peak whose duration is rather brief (since it occurs only in the burst).

The second characteristic not accounted for by the diffuse-rising template concerned a set of consonants whose spectra contained a substantial peak of energy in the vicinity of 1800 Hz—a peak whose amplitude exceeded the amplitude allowed by the alveolar template. This peak of energy corresponds to the starting frequency of the second formant for the following vowel—the so-called hub or locus (Potter *et al.*, 1947; Delattre *et al.*, 1955). In constructing the template, we allowed a high-amplitude spectral peak in the vicinity of the locus to occur without violating the conditions required for fitting the diffuse-rising template. The only requirement is that two spectral peaks occur within the template, possibly including the peak arising from the second-formant locus, and that the peak that is higher in frequency (above 2200 Hz) also be higher in amplitude than the lower-frequency peak. Examples of such spectra which meet the requirements of the diffuse-rising template are given in Fig. 5. As the top part of the figure shows, despite the fact that the peak associated with the locus predominates in amplitude over the other peaks in the spectrum, there are still two peaks within the template reflecting the upward rising pattern. The bottom part of Fig. 5 gives an example of a spectrum which contains two peaks fitting within the template. One of them is the 1800-Hz peak which, although higher in amplitude than the upward sloping reference line, is lower in amplitude than the higher-frequency peak. Thus, the overall diffuse-rising spectral shape is still maintained. The peak in the onset spectrum corresponding to the second-formant locus is usually continuous with a second-formant spectral peak following the burst, as the formants undergo transitions toward the vowel. Thus this spectral peak cannot be ignored because of its brief duration, as in

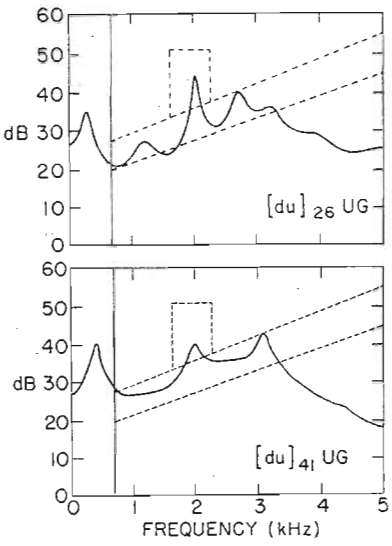

FIG. 5. Modification of the diffuse-rising template to allow a substantial energy peak in the vicinity of 1800 Hz corresponding to the so-called hub or locus. Two onset spectra are superimposed on the template. The top panel shows the onset spectrum of an alveolar consonant containing two peaks within the reference lines reflecting the upward rising pattern, in addition to a substantial energy peak at 1800 Hz. The bottom panel shows an example of a spectrum for an alveolar consonant containing two peaks within the template, one of them at 1800 Hz which, although higher in amplitude than the upward sloping reference line, is lower in amplitude than the higher-frequency peak.

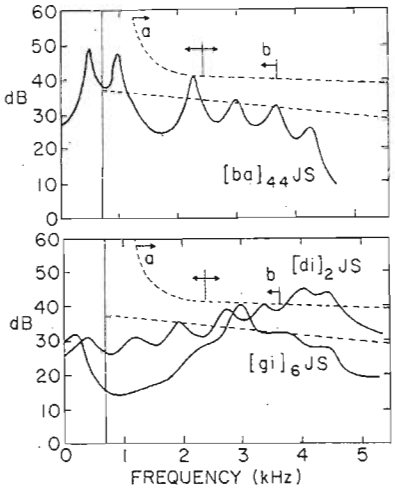

FIG. 6. Examples of short-time spectra which have been superimposed on the diffuse-falling template. A spectral peak between 1200 and 3600 Hz (between a and b) is fit to the top reference line. At least two peaks must fall within the reference lines, one peak falling below 2400 Hz (see arrow) and the other peak falling between 2400 and 3600 Hz. The top panel shows a labial spectrum which fits the diffuse-falling template. The bottom panel shows spectra for an alveolar and a velar consonant. In the former case, there is no energy peak falling below 2400 Hz; in the latter case, only one peak falls within the reference lines of the template.

the case of lower-frequency peaks that are thought to be a consequence of subglottal resonances.

B. The diffuse-falling template

The property characteristic of labial consonants is a diffuse spread of energy, with either a predominance of lower-frequency spectral peaks over high-frequency peaks or an equal distribution of energy among various peaks. The gross shape of the spectrum for labial consonants, then, is either diffuse-falling or diffuse-flat. The template designed to reflect these spectral shapes is shown in Fig. 6. In order to determine whether a spectrum fits the diffuse-falling template, a spectral peak between 1200 and 3500 Hz (between a and b on the figure) is fitted to the top reference line such that all other peaks in this frequency region lie on or below the line. The distribution of the spectral peaks is then examined in relation to the template. The condition required to fit the template is that at least two peaks must fall within the reference lines, one peak falling below 2400 Hz and the other peak falling in the range of 2400 and 3600 Hz. Note that there is no condition on the amplitudes of spectral peaks falling below 1200 Hz. Thus, to satisfy the conditions of the diffuse-falling template, the spectrum shape must be either falling or relatively flat, and spectral energy must occur at low to mid frequencies, although it may also occur at higher frequencies (above 3600 Hz). An example of a spectrum of a labial consonant containing the required characteristics is superimposed on the template in Fig. 6. Examples of spectra with shapes that do not fit these characteristics are shown in the bottom part of the figure. Although the [d] spectrum is diffuse, i.e., there are several peaks spread out in the frequency domain, the distribution of the spectrum is rising rather than either falling or flat; the [g] spectrum shows one prominent mid-frequency peak and thus is not diffuse.

C. The compact template

The property that describes a velar consonant is the presence of a prominent spectral peak in the mid-frequency range. Two spectral peaks that are separated by 500 Hz or less are treated for our purposes as comprising a single gross spectral peak. A peak is "prominent" if there are no other peaks nearby and if it is larger than adjacent peaks, so that the peak stands out, as it were, from the remainder of the spectrum. In this sense, the spectrum is compact. We have attempted to capture this property with the template shown in Fig. 7. This template shows an overlapping set of spectral peaks in the mid-frequency range (from 1200–3500 Hz). The widths of these peaks increase with increasing frequency. These widths were established on the basis of examining spectra for a number of velar consonants in different vowel environments. The fact that the widths increase with frequency is to be expected, on the basis of data on critical bands for the auditory system (Plomp, 1964). The widths (at the half-power points) were not, however, adjusted to be equal to critical bandwidths; they are, in fact, somewhat wider than critical bands, varying from about 300 Hz at 1200 Hz to about 800 Hz at 3500 Hz.

To determine whether a given spectrum fits the compact template, a spectral peak in the mid-frequency range is adjusted to touch a matching peak of approximately equal frequency on the template. The requirement for spectral compactness is that no other peak in

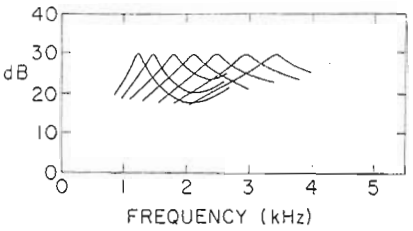

FIG. 7. A schematic of the compact template. This template shows an overlapping set of spectral peaks in the midfrequency range (from 1200–3500 Hz), the widths of these peaks increasing with increasing frequency. A single midfrequency peak in a measured spectrum must fit within one of these peaks to meet the conditions of the template (see Fig. 8).

the spectrum projects through the reference line, and further that there is no other peak of the same or greater magnitude occurring either below 1200 Hz or above 3500 Hz. (The first formant is ignored in these analysis procedures.) The bottom part of Fig. 8 shows an example of a spectrum that meets the requirements of the velar template. Note that the 1700-Hz peak of the spectrum is matched to a low-frequency peak on the template. Although there is an additional energy peak at 3500 Hz, it is lower in amplitude than the major spectral peak. The top part of Fig. 8 shows an example of a spectrum that does not fit the compact template. As the figure shows, a second peak (at 2500 Hz) juts through the major spectral peak of the template; further, there is an energy peak above 3500 Hz which is higher in amplitude than the major mid-frequency peak.

IV. TEMPLATE ANALYSIS

In order to determine the extent to which naturally produced stop consonants fit the hypothesized shapes reflected by the templates, we analyzed the spectra for

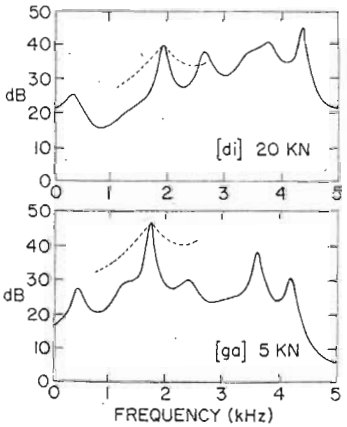

FIG. 8. Examples of short-time spectra which have been superimposed on the compact template. A spectral peak in the mid-frequency range is adjusted to touch a matching peak of approximately equal frequency on the template. The top panel shows an alveolar spectrum which is rejected by the compact template, as a second spectral peak projects through the matched spectral peak, and a third spectral peak above 3500 Hz is higher in amplitude than the mid-frequency spectral peak. The bottom panel shows a velar spectrum which is accepted by the compact template, as there is a single mid-frequency peak which fits within the reference lines of a corresponding peak in the template, and this peak is higher in amplitude than the spectral peak falling at 3500 Hz.

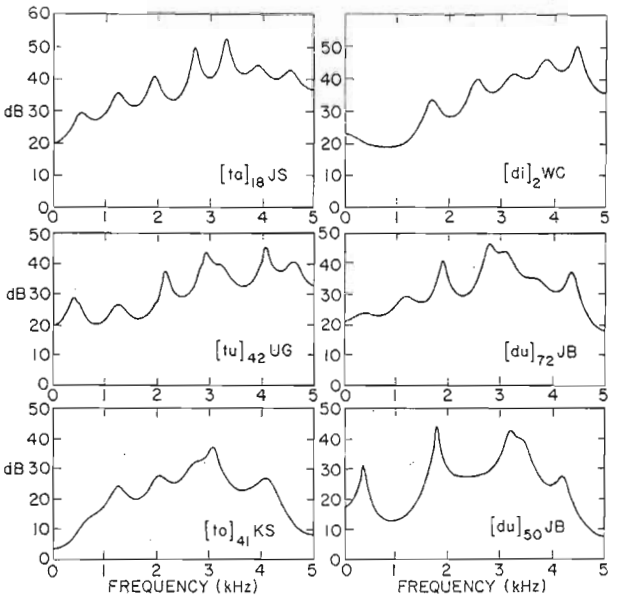

FIG. 9. Examples of short-time spectra of voiced and voiceless alveolar stop consonants followed by different vowels and spoken by different speakers. Except for the bottom right panel, these spectra show the diffuse-rising property characteristic of alveolar consonants.

a number of initial and final voiced and voiceless stop consonants produced by several speakers. Subjects were asked to read a listing of CV utterances containing five repetitions of each of the stop consonants [p t k b d g] in the context of the five vowels [i e ɑ o u], and a listing of VC utterances containing the same six stop consonants but preceded by the vowels [i ɛ ɑ ʌ u]. Six subjects (four male and two female) participated, producing a total of 1800 natural speech monosyllables, 900 with consonants in initial position and 900 in final position. These utterances were tape-recorded in a sound-treated room and subsequently analyzed using the procedures for spectral analysis described above.

As indicated earlier, for syllable-initial consonants the spectra were sampled at the point of consonantal release (see Fig. 1). Figure 9 shows some examples of onset spectra for voiced and voiceless alveolar stops produced by several of the speakers. The first five spectra fit the diffuse-rising template, and the lower right spectrum does not. Examples of spectra for voiced and voiceless labial consonants are given in Fig. 10. Five of these spectra are identified as labials by the diffuse-falling template; the lower right spectrum would be rejected. Figure 11 gives examples for the velars with the lower right spectrum again being rejected by the compact template.

For the syllable-final stop consonants, acoustic information with regard to place of articulation can reside at two points in the signal: at the point of consonantal closure at the end of the vowel, and in the burst that occurs at the consonantal release (if the final consonant is, in fact, released). In principle, the spectrum sampled at both of these points in time should contain the characteristics observed in the spectra sampled at onset. The procedures used to sample the spectra for final consonants are illustrated in Fig. 12. The figure

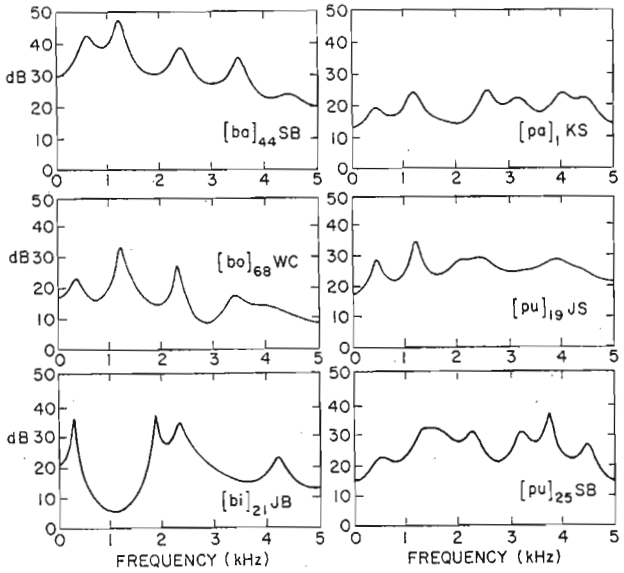

FIG. 10. Examples of short-time spectra of voiced and voiceless labial stop consonants followed by different vowels and spoken by different speakers. Except for the bottom right panel, the spectra show either the diffuse-falling or diffuse-flat property characteristic of labial consonants.

shows spectra sampled at closure and at burst release for the syllable [ɛk]. In the measurement of the closure portion of the syllable, the peak of the spectral window was placed at the point of closure, thus maximizing the weighting for the final tens of milliseconds of the closure. For the release burst, the spectral measurements were analogous to those used for CV onsets: measurements were made from the moment of burst release, and the spectral window emphasized the earlier portions of the

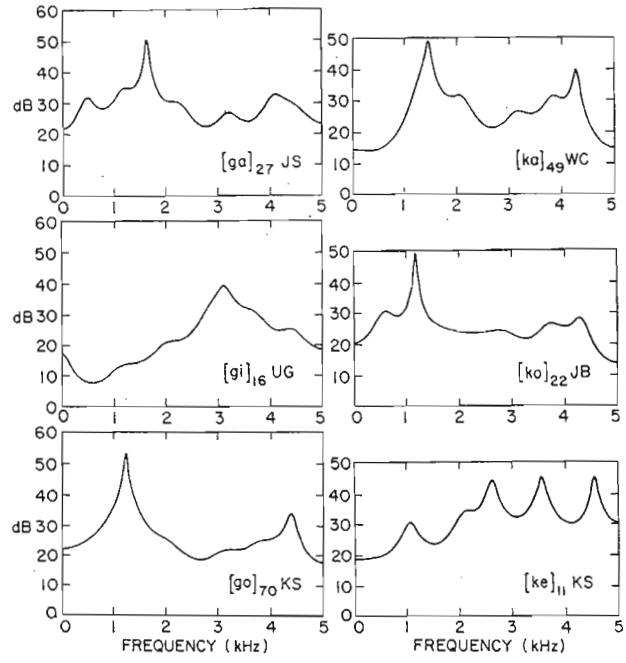

FIG. 11. Examples of short-time spectra of voiced and voiceless velar stop consonants followed by different vowels and spoken by different speakers. Except for the bottom right panel, the spectra show the compact midfrequency prominence characteristic of velar consonants.

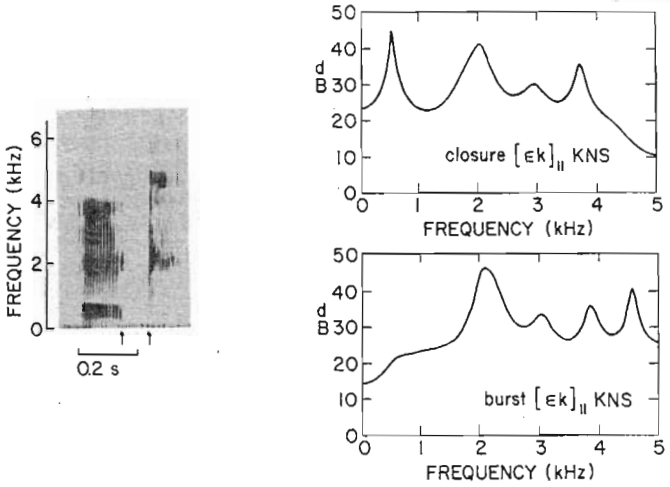

FIG. 12. The left side of the figure shows a spectrogram of the naturally produced VC utterance [ɛk]. The left arrow indicates the point of consonant closure and the right arrow indicates the point of release of the closure. The right side of the figure shows the short-time spectrum of this utterance sampled at these two points. The spectrum in the top panel is sampled at the point of closure, and the spectrum in the bottom panel is sampled at consonantal release. Both show the compact spectral shape characteristic of the velar consonants.

release. Both spectra shown in Fig. 12 fit the compact template.

The spectra of the 1800 natural CV and VC utterances were individually tested against each template. We adopted a conservative strategy for assessing whether the spectral shapes were accepted or rejected by the particular template. In order to fit the template, the spectrum had to meet all the conditions specified for the template. If it did not fit for any reason, e.g., the shape was clearly wrong or the shape was correct but a peak failed to fit within the reference lines, then the spectrum was rejected. A tabulation was made for each subject of the proportion of responses which fit and which were rejected by each of the templates.

Table I shows a summary of the results of applying the diffuse-rising, diffuse-falling, and compact templates to initial voiced and voiceless stop consonants

TABLE I. Template-matching results for initial voiced and voiceless stop consonants. The entries give the mean percentage of utterances of each consonant (based on 150 utterances of each consonant, occurring in five vowel environments, and obtained from six speakers) that were correctly accepted or rejected by the template.

Diffuse-rising template		
Correct acceptance	Correct rejection	
[d] 84.0	[b] 86.0	[g] 88.5
[t] 88.0	[p] 80.0	[k] 85.3
Diffuse-falling template		
Correct acceptance	Correct rejection	
[b] 82.5	[d] 80.7	[g] 90.0
[p] 80.0	[t] 95.3	[k] 94.7
Compact template		
Correct acceptance	Correct rejection	
[g] 86.7	[b] 91.3	[d] 82.7
[k] 84.7	[p] 86.7	[t] 88.0

TABLE II. Template-matching results for the closure and burst release of final voiced and voiceless stop consonants. The entries give the mean percentage of utterances of each consonant (based on 150 utterances of each consonant, occurring in five vowel environments, and obtained from six speakers) that were correctly accepted or rejected by the template.

Diffuse-rising template			
Correct acceptance		Correct rejection	
closure	burst	closure	burst
[d] 31.2	74.7	[b] 90.0	85.2
[t] 31.7	78.4	[p] 88.7	83.2
		[g] 85.8	79.7
		[k] 81.0	84.0
Diffuse-falling template			
Correct acceptance		Correct rejection	
closure	burst	closure	burst
[b] 77.3	78.7	[d] 42.5	82.0
[p] 76.5	75.0	[t] 40.0	91.2
		[g] 42.7	82.0
		[k] 40.0	83.2
Compact template			
Correct acceptance		Correct rejection	
closure	burst	closure	burst
[g] 57.8	70.0	[b] 71.3	86.8
[k] 46.0	80.8	[p] 75.2	89.6
		[d] 67.8	78.0
		[t] 73.2	87.2

across the six speakers. Overall, about 85% of the utterances are correctly accepted by these templates, and about the same percentage of utterances are correctly rejected. In other words, a given template (such as the diffuse-rising template) correctly accepts the voiced and voiceless stop consonants containing the appropriate place of articulation (e.g., alveolar consonants) and correctly rejects consonants having a different place of articulation (i.e., labial and velar consonants).

The results of the template-matching procedures applied to the final consonants are shown in Table II. Here, the spectra for the closure and release bursts are each matched to the three templates. One subject did not release final voiceless stops; he did, however, release voiced stops. It is quite clear that, overall, the closure and the burst do not equally reflect the spectral shapes diffuse-rising, diffuse-falling, and compact. The release burst is correctly accepted by the appropriate template approximately 76% of the time, and is correctly rejected by the two other templates approximately 84% of the time. The closure, on the other hand, fails to provide a reliable index for correctly accepting or rejecting a given place of articulation based on the gross shape of the short-time spectrum. The failure of the closure at the offset to show the spectral shapes predicted may, in part, be a result of the fact that final consonants are often devoiced prior to closure, and consequently the spectral analyses may not measure the actual closure but rather may characterize the utterance at a point several tens of msec before the closure occurs. The spectrum would reflect, then, the articulatory motions from the vowel to the target place of articulation rather than the closure itself. There are, however, differences in the scores depending on the place of articulation. The spectrum at closure for the labials is correctly accepted by the diffuse-falling template much more often than the other two places of articulation are accepted by their respective templates. Data from the perception of place of articulation in unreleased stop consonants indicate that identification is not as good as that obtained for final released stops, although the reported absolute level of identification varies across studies (Halle *et al.*, 1957; Malécot, 1958; Wang, 1959).

There seems to be little difference among the three templates in correctly accepting and rejecting the relevant place-of-articulation dimension, except for the offsets in VC syllables. Thus, overall, the results of the template-fitting procedures indicate that there are indeed unifying acoustic properties for place of articulation across phonetic contexts (i.e., different vowels) and syllable positions (i.e., initial and final), and among different voicing characteristics (i.e., voiced and voiceless stop consonants). These properties can be derived from the short-time spectrum at consonantal onset and offset and provide higher-order invariant acoustic evidence for place of articulation directly derivable from the acoustic signal. These properties represent "higher order" invariance in the sense that they reflect the relative shape of the spectrum for a particular consonantal configuration rather than an absolute measure (along the frequency-amplitude domain) of a particular attribute.

A. Role of F2 peak in diffuse-rising template

As described earlier, the diffuse-rising template was modified to reflect the fact that many alveolar consonants contained a fairly substantial energy peak at the hub or second-formant locus. In order to determine the extent to which the alveolar productions reflect this pattern, a tabulation was made across subjects of the percent of alveolar stop consonants which had a high-amplitude peak projecting above the upward reference line of the template. The results of this analysis are given in Table III. First, it is important to note that a fair proportion of the alveolar productions contain a significant spectral peak at the $F2$ locus; 27% of the total alveolar consonants contained such a peak. However, as the table shows, subjects do not seem to contribute equally to this pattern. Rather, some sub-

TABLE III. Summary of the percent of voiced and voiceless alveolar stop consonants which had a high-amplitude energy peak at about 1800 Hz, corresponding to the so-called locus. The entries give the mean percent occurrence for each of the six subjects and mean totals for initial consonants and the release burst of final consonants.

	Initial		Final		
Subject	[d]	[t]	[d]	[t]	Mean
1	36	9	10	5	15
2	33	16	55	5	28
3	15	0	0	5	5
4	16	20	16	a	17
5	63	50	54	68	59
6	53	36	42	20	38
Mean	36	22	30	12	27

[a]No released voiceless consonants.

jects have substantially more utterances with a large $F2$ peak in the spectrum than others. Subjects 5 and 6, who had the largest proportion of these kinds of productions, were the two female speakers. It is not clear at this time whether this is a function of sex differences in speech, whether it reflects differences in vocal-tract size, or whether the results are a consequence of sampling error. In any case, owing to the fairly large proportion of these utterances in this sample, it is not surprising that the locus provides a fairly strong perceptual cue for the alveolar place of articulation (Delattre *et al.*, 1955). Nevertheless, although a strong perceptual cue, it cannot be the primary cue to the alveolar place of articulation as it always occurs in conjunction with the diffuse-rising property, and further it constitutes only 27% of the total alveolar productions.

B. Properties for place of articulation: Vowel dependent or independent?

One of the major objectives of this research has been to determine whether invariant acoustic properties are found for place of articulation independent of the vowel context in which the consonant appears. The problem of vowel dependencies in speech perception is a familiar one (Liberman *et al.*, 1967), and, in fact, it was the failure to find a one-to-one correspondence between acoustic cue and phonetic percept independent of the following vowel which gave rise to the motor theory of speech (Liberman *et al.*, 1967; Studdert-Kennedy *et al.*, 1970) and the analysis-by-synthesis model (Halle and Stevens, 1972; Stevens and Halle, 1967). The summary of the results presented in Tables I and II did not focus specifically on the success or failure of the template-matching procedures in relation to the particular vowel context in which a consonant appeared. That is, although the diffuse-rising template, for example, accepted overall about 82% of alveolar consonants and rejected the same percentage of labial and velar consonants, it is of some interest to determine whether these results were equally distributed across the various vowel contexts or whether consonants in particular vowel contexts were systematically accepted or rejected.

Figure 13 shows the proportion of correct acceptance and rejection by each of the templates in relation to the vowel context of the particular consonant. To assess the accuracy of the templates in fitting stop consonants in final position, only the results for the burst (as opposed to closure) were used. It is clear that for all vowels across all consonants, the acceptance and rejection rate for the templates is above chance (50%). However, there seem to be some differences across the vowel contexts. In order to assess the reliability of these differences, a three-way (place of articulation X vowel X syllable position) repeated measures analysis of variance was conducted for each template. The Duncan test was used for all post-hoc comparisons.

The results of the analysis of variance on the scores for the diffuse-rising template indicate a significant main effect for vowels ($F_{(4,20)} = 5.62, p < 0.004$), and significant interactions for place X vowel ($F_{(8,40)} = 6.40, p < 0.001$), position X vowel ($F_{(4,20)} = 3.29, p < 0.04$)

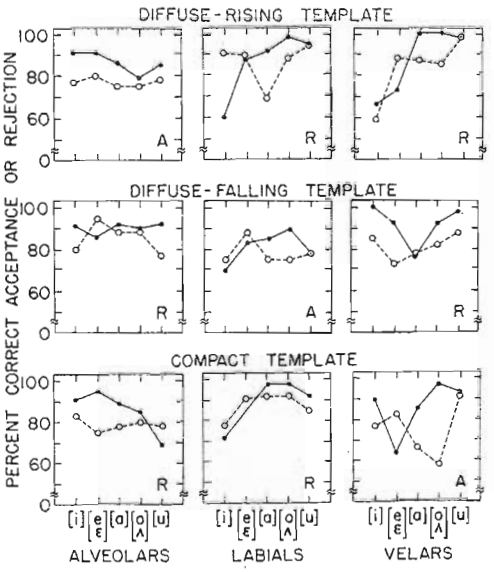

FIG. 13. The percent correct acceptance and rejection of each of the templates (as indicated) in relation to the vowel context for each of the stop consonants (alveolar, labial and velar). The filled circles represent initial consonants and the open circles final consonants. The letter R or A in the right corner of each panel corresponds to correct rejection or correct acceptance by the particular template respectively. For example, the top row shows the percent correct scores for the diffuse-rising template in accepting alveolar consonants and rejecting labial and velar consonants. For initial consonants, the vowel environments include [i e ɑ o u]; for final consonants, they include [iɛɑʌu]. Data for final consonants are for spectra sampled at the release of the burst.

and place X position X vowel ($F_{(8,40)} = 3.45, p < 0.005$). Post-hoc analyses revealed that these results were due to the following effects. Initial labial consonants (Fig. 13, upper middle panel) tended to be incorrectly interpreted as alveolar consonants in the context of the vowel [i], whereas final labial consonants were misclassified as alveolars in the context of the vowel [a]. Velar consonants (Fig. 13, upper right panel) tended to be misclassified as alveolar in the context of high front vowels, with initial velars tending to be classified as alveolar more frequently in the context of the vowels [i] and [e] and final velars classified as alveolars in the context of the vowel [i]. There were no significant vowel context effects for alveolar consonants (Fig. 13, upper left panel), i.e., the template correctly classified the alveolar consonants independent of the vowel context in which the consonant appeared.

Thus, for the diffuse-rising template, there seems to be a vowel context effect. The fact that both labial and velar consonants were misclassified as alveolars in the context of high front vowels ([i] and [e]) can be explained in terms of the relatively high-frequency onset of $F2$ and $F3$ for these consonants in the environment of these vowels. This characteristic of the onset can be attributed to the coarticulation of the tongue body in anticipation of the following front vowel. The higher frequencies for $F2$ and $F3$ would tend to raise the amplitude of the spectrum at higher frequencies, and this would tend to give a diffuse-rising shape to the spectrum. Further, the failure to correctly classify place

of articulation independent of vowel context may also reflect problems with the particular shape or slope of the template devised. To be sure, the templates themselves have not been optimized (see Sec. VI), and subtle changes in the templates could in fact modify the accuracy of the classifications.

Statistical analyses of vowel context effects for the diffuse-falling template (middle row of Fig. 13) revealed a significant main effect for position due to higher accuracy in fitting onset in contrast to offset spectra ($F_{(1,5)} = 15.46, p < 0.02$), and a significant place X position X vowel interaction ($F_{(8,40)} = 3.49, p < 0.001$). The results of the post-hoc tests showed some significant vowel effects, although they are less systematic and easy to interpret than those found for the diffuse-rising and compact templates. In particular, initial labial consonants in the context of [i] were less often correctly accepted by the diffuse-falling template in comparison to labials in the context of the vowels [ɑ] and [o]. Examination of Fig. 13 indicates that these incorrectly identified labial consonants were interpreted mostly as [d]'s, but they were also interpreted as [g]'s as well. In final position, the diffuse-falling template showed no vowel context effects. Thus, the percent of final labial consonants that were accepted correctly by the diffuse-falling template was about the same for all vowel contexts. For the alveolar consonants, there were no context effects in initial position. In final position, alveolar consonants in the context of the vowels [i] and [u] were incorrectly rejected by the diffuse-falling template (i.e., accepted as labial consonants) significantly more times than were alveolar consonants produced in the context of the vowel [ɛ]. It is not clear why this particular pattern of results emerged. Finally, initial velar consonants in the environment of [ɑ] were incorrectly accepted by the diffuse-falling template (i.e., called a labial) significantly more than were velars in the context of the other vowels [i ɛ o u]. Apparently, although there was a predominant mid-frequency peak, it was not sufficiently high in amplitude to dominate the spectrum relative to higher-frequency peaks. In the case of final consonants, velars in the context of [e] were interpreted as labials significantly more than were velars in the context of [u]. Again, no principled explanation can be provided. Vowel context effects for the diffuse-falling template, then, are less easily explainable in terms of the acoustic properties of the following vowels. Thus, although some differences were obtained for correctly assessing the labial place of articulation across vowel contexts, overall performance by the diffuse-falling template in correctly classifying place of articulation never fell below 70%.

Analysis of vowel context effects for the compact template (bottom row of Fig. 13) showed a significant place X vowel interaction ($F_{(8,40)} = 4.47, p < 0.001$). Post-hoc tests revealed that, as with the results for the diffuse-rising template, these effects can be accounted for in terms of the onset frequencies of the formants following the consonants in relation to the acoustic characteristics of the particular vowel environment. In particular, labial consonants (Fig. 13, lower middle panel) in the context of [i] were more likely misclassified as velars than were the other vowels. Presumably, the high-frequency onsets of $F2$ and $F3$ for the labial consonants in the environment of the vowel [i] enhance the spectral energy in the mid-frequency region relative to the other energy peaks, resulting in a misclassification of the consonant as compact. In contrast, alveolar consonants (Fig. 13, lower left panel) in the context of [u] were often misclassified as velars; here the low-frequency $F2$ and $F3$ onset presumably enhances energy in the lower-frequency regions more than the higher-frequency regions. Finally, the velar template (Fig. 13, lower right panel) was best at classifying velar consonants when they were in the context of the vowel [u]. As with the alveolar consonants, the low $F2$ enhances energy in this mid-frequency domain, the area to which the compactness property is most sensitive.

In sum, the results of the vowel context analyses indicate that the acoustic characteristics of the vowel environments do have an effect on the classification of the appropriate place of articulation. Nevertheless, this vowel dependency is in a much more limited sense than, for example, the vowel context effects found in the perception of various attributes of the acoustic signal, such as burst frequency and transition motions (Liberman *et al.*, 1967). Although some frequency-dependent effects were found, the templates correctly classified better than chance *all* consonants in the context of *all* of the vowel environments.

V. PRELIMINARY STUDY OF NASAL CONSONANTS

The spectra sampled at the onsets of stop consonants arise in part from the configuration of formants that result from glottal excitation of the vocal tract immediately following the release and in part from the burst of acoustic energy generated by vocal-tract excitation in the vicinity of the constriction. Theoretical analysis shows that both components of the release contribute to the gross spectral property that is characteristic of a particular place of articulation. In the case of nasal consonants there is no burst, but, for a given place of articulation, the configuration of formants at the consonantal release should be roughly the same as for stop consonants. Thus the spectrum sampled at the release of a nasal consonant should show the same properties as that for a stop consonant with the same place of articulation, although the property may be weaker due to the absence of the burst. A nasal consonant also differs from a stop consonant in that there is a nasal murmur preceding the release. Likewise, the configuration of formants at the implosion or point of closure of a syllable-final nasal consonant should be similar to that for a syllable-final stop consonant. Thus to the extent that the gross shape of the spectrum at offset characterizes the place of articulation for stop consonants it should also indicate place of articulation for nasals.

We have begun a study of the analysis of onset spectra for nasal consonants in syllable-initial position. In order to obtain spectra characteristic of the nasal release without being contaminated by the preceding or following nasal murmur, we have shortened the time window used for sampling the spectrum to 6 ms, i.e.,

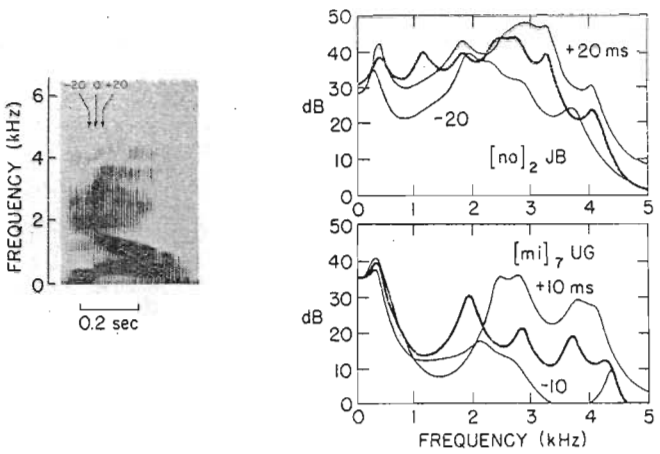

FIG. 14. Illustrating the method used to sample the spectrum at the release of a nasal consonant. A spectrogram of the syllable [no] is shown at the upper left. Three spectra are shown in the panel to the right of the spectrogram, sampled at the instant of release (heavy line), two glottal periods (−20 ms) preceding the release, and two glottal periods (+20 ms) following the release. These sampling times are indicated on the spectrogram. The lower right panel shows three spectra sampled at similar instants of time surrounding the release of the syllable [mi] produced by a female talker. The width of the time window (raised cosine) for sampling the spectra is 6 ms.

the order of one glottal period or less. Examples of spectra sampled in the vicinity of the release for two nasal consonants are shown in Fig. 14. Three spectra are displayed for each consonant: one spectrum is sampled two glottal periods preceding the release, and indicates the smoothed spectral shape for the nasal murmur; the second spectrum is sampled at the onset of the glottal pulse nearest to the release, as judged by a discontinuity in the waveform shape; the third spectrum is sampled two glottal periods following the release, and shows the spectrum shape during the transition to the vowel. As these examples demonstrate, the spectrum at release for the alveolar nasal has the diffuse-rising characteristic, and the labial nasal has the predicted diffuse-falling shape.

We have measured the onset spectra for 110 alveolar and labial consonants produced in consonant-vowel syllables by the same six speakers as those used to produce the stop consonants. Five different vowels were used as before. Results of this preliminary study are summarized in Table IV. We observe that alveolars and labials are correctly accepted by their respective templates about 76% of the time, with labials having slightly fewer errors. The templates incorrectly accept the wrong place of articulation for about 10% of labials, but the percentage of alveolar nasals incorrectly accepted by the diffuse-falling template is 67%. Apparently a large number of alveolars were accepted by both templates since there was not a sufficiently large rising slope on the alveolar spectra.

TABLE IV. Preliminary template-matching results for initial nasal consonants. The entries give the percentage of utterances of each consonant (about 55 utterances of each consonant occurring in five vowel environments and obtained from six speakers) that were correctly accepted or rejected by the templates.

	Diffuse-rising template	Diffuse-falling template
Correct acceptance	[n] 72	[m] 81
Correct rejection	[m] 90	[n] 33

Why should the error rate be substantially higher for the nasal consonants? A possible hypothesis is that the spectral representation in the auditory system for a signal with an abrupt onset that is preceded by silence is different from the representation when the onset is preceded by a nasal murmur. The presence of the nasal murmur for the 100-odd ms interval preceding the nasal release may serve to reduce the response of some neural units in the auditory system to the onset, particularly the response to the low-frequency components of the onset where the nasal murmur is most likely to produce this masking effect. If this were true, then the representation of the spectrum in the auditory system would show an attenuation of the low frequencies relative to the spectral representation if the onset were preceded by silence. In other words, if the measured spectra in Fig. 14 are to be compared with those at the onset of a stop consonant, the low frequencies in the nasal onset spectra should be attenuated. If this modification were made, then the onset spectra for the alveolar nasal consonants would tend to have a more sharply rising characteristic, and it is likely that the percent of these consonants accepted by the diffuse-rising template and rejected by the diffuse-falling template would increase.

There is some evidence to support this low-frequency masking hypothesis in data on the electrical response of single units in the auditory nerve, reported by Delgutte (1978). His results show that a masking stimulus that causes a continuous response of an auditory nerve fiber can reduce the response to an abrupt onset of a sinusoidal signal that follows the masker. This hypothesis is highly speculative, but it could help to explain how the speech-perception mechanism can classify stop and nasal consonants with the same place of articulation as having a common feature, even though there are differences in the shape of the spectrum sampled at the consonantal release. A corollary of this hypothesis is that the perception of labial to dental place-of-articulation continua for stops and nasal consonants would show different boundary values, with the nasal consonants showing an earlier cross-over point than the stop consonants (i.e., towards the labial category). This would presumably be due to the fact that the auditory representation of nasal consonants is attenuated at lower frequencies, and a relatively flat or slightly rising onset spectrum would be tilted up, resulting in a diffuse-rising pattern characteristic of the alveolar place of articulation. Results to this effect have been obtained in the identification of nasal and stop continua varying systematically along the labial and alveolar place-of-articulation dimensions (Miller, 1977; Miller and Eimas, 1977).

VI. GENERAL DISCUSSION

The acoustic analysis of natural speech utterances indicates that there are indeed invariant acoustic properties intrinsic to the place-of-articulation phonetic dimension. These properties reflect the gross shape of the short-time spectrum at consonantal release, and can be characterized in terms of the properties diffuse-rising, diffuse-falling, and compact for the alveolar, labial and velar places of articulation respectively. The fact that over 80% of the data for stop consonants were correctly classified using the template-matching procedures is interpreted as strong support for the theory of acoustic invariance. Further, these results represent a conservative estimate of the occurrence of these properties within the speech signal, as it may be possible to optimize further both the analysis procedures on the one hand and the templates on the other in order to capture more fully these invariant properties.

With regard to the analysis procedures, the particular window size and shape and the particular type of spectral analysis used in this study may not provide the best measures. For example, although pilot work was conducted to choose a window size which seemed to show the hypothesized spectral shape most effectively, we did not systematically compare different size windows and their effectiveness with a large body of data. In fact, it was found that for some stimuli which did not meet the conditions of the template, the spectral properties of the signal were either enhanced or emerged when the window size was considerably shortened. For example, the top part of Fig. 15 shows the onset spectrum for the syllable [to] using (a) a 26-ms time window and (b) a 3.2-ms time window. Although the spectrum shape using the longer window is diffuse-rising, the energy of the 3200-Hz peak is too high, and the spectrum meets the conditions of the compact rather than the diffuse-rising template. In contrast, the shorter time window reduces the high amplitude of the high-frequency peak and the spectrum shape is now appropriately diffuse-rising. The bottom part of the figure shows the spectrum for the syllable [ge] using both the long (26 ms) and short (3.2 ms) time window. The shape of the spectrum using the long window is not compact, as the two mid-frequency peaks are not close enough to be considered a single energy peak. This property emerges, however, with the shorter time window, as the two peaks become a single broad spectral peak. These examples suggest that it may not be desirable to postulate a single, fixed time window, but that the gross spectrum shape may be assessed on the basis of examining successive spectral samples extending over 10–20 ms, each computed using a relatively short time window. It is also possible that the time window should be a function of frequency, with the spectrum at low frequencies being sampled over a longer time interval than the spectrum at high frequencies. These differences in time resolution between low and high frequencies would occur automatically if the initial spectral analysis of the signal utilized bandwidths that reflected the properties of the peripheral auditory system (Searle, *et al.* 1979; Plomp, 1964). The narrower bandwidths at low frequencies would lead to a longer time window and the wider high-frequency bandwidths would result in a shorter time window.

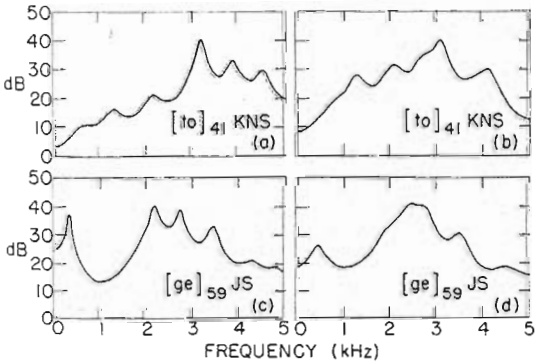

FIG. 15. The onset spectra for the natural CV utterances [to] and [ge] using a 26-ms time window (a) and (c) and a 3.2-ms time window (b) and (d). The gross spectral shapes for the longer time window of [to] and [ge] fail to fit the diffuse-rising and velar templates respectively. In contrast, using the shorter time window, the spectral properties corresponding to their appropriate templates emerge.

As described in Sec. II, the short-time spectra were derived by means of a linear prediction algorithm with high-frequency pre-emphasis. It would be instructive to reanalyze the natural speech data using alternative spectral analysis procedures, as for example, one that is more properly "tuned" to the auditory properties of the ear (Searle *et al.*, 1979; Klatt, 1976; Miller *et al.*, 1977; Houtgast, 1974). Such analyses may show an even greater division of the spectral shapes of these utterances into the three place-of-articulation categories, if it is the case that the natural properties inherent in speech reflect the interaction between articulatory and auditory properties of the system. Procedures for displaying a spectral analysis of a signal by simulating the filtering properties of the auditory system have been developed by Dennis Klatt (1976), and hence it is possible to compare spectral displays based on the linear prediction algorithm and on the simulated auditory filters. Examples of such a comparison for three different utterances are shown in Fig. 16. The gross spectral properties of diffuseness–compactness and rising-falling that we have observed for the pre-emphasized linear prediction spectra are also evident for the spectra based on simulation of the auditory filtering. In fact, the auditory simulation may tend to smooth out spectral fluctuations that are not relevant to categorization of the spectra. For example, the spectral peak associated with the second formant onset for alveolar consonants might be attenuated and broadened with this spectral representation, and it may be unnecessary to invoke a special condition on the application of the diffuse-rising template. With this spectral representation, then, the three basic properties of the onset spectra captured by the templates are evident to an even greater degree than the linear-prediction spectra.

As was pointed out in Sec. IV, the templates we have

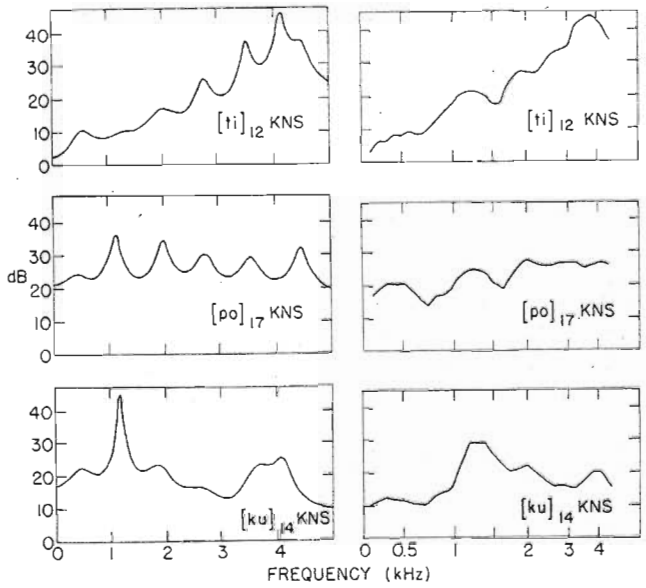

FIG. 16. The spectra at the left are linear-prediction spectra for three voiceless consonants, sampled in the manner described in connection with Fig. 1. The data at the right represent the outputs of a simulated filter bank (Klatt, 1976) consisting of 30 filters whose frequency characteristics approximate the filtering properties of the peripheral auditory system. The frequency scale is adjusted so that equal distances along the scale correspond roughly to equal frequency resolution of the peripheral auditory system.

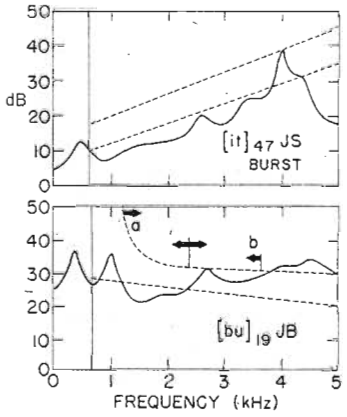

FIG. 17. Examples of spectra which failed to fit a particular template despite the fact that their gross shape reflected the appropriate property. The top panel shows the release burst of a final voiceless alveolar stop which, although showing the diffuse-rising property, fails to fit the diffuse-rising template because only one peak falls within the reference lines. The bottom panel shows an initial voiced labial stop which fails to fit the diffuse-falling template, as energy peaks project above the reference line of the template. Nevertheless, its gross shape is diffuse flat corresponding to the property characteristic of labial consonants.

devised represent only a first approximation of the characterization of the spectral properties intrinsic to the place-of-articulation dimension. The question of whether these templates can in fact be optimized rests on the nature of the misclassification of the templates. The failure of a particular spectrum to meet the conditions of a template was attributable to several different factors. First, there were some spectra whose shape did not reflect at all the hypothesized acoustic properties (see bottom right spectra in Figs. 9, 10, and 11). Here, it remains to be determined whether changing the analysis procedures as discussed above would modify the short-term spectrum in such a way that it would correspond to these properties. Second, and perhaps most important, certain spectra failed to fit a particular template although the spectrum shape was right from a descriptive point of view. Examples of two of these spectra are shown in Fig. 17. Some minor modifications of the actual template shape may be sufficient to increase the classification scores. For example, in the case of the spectra shown in the figure, changes in the shape of the template at higher frequencies for both the alveolar and labial templates would result in the correct acceptance of these particular spectra.

Some of the spectra were accepted by more than one of the templates, particularly if the property for which the template was devised was only weakly represented. The all-or-none procedure failed to distinguish between those spectra whose shape was clearly wrong and those whose shape was in fact correct but just missed fitting the template. A goodness-of-fit analysis procedure would reflect this important difference. For example, in our study, 15% of the initial stop consonant spectra fit not only the correct template but also one other. A goodness-of-fit analysis would presumably classify some of these spectra incorrectly thus reducing the score below the overall value of 85% given in Table I. On the other hand, 7% of the initial consonant spectra failed to fit any template. A goodness-of-fit analysis would classify some of these correctly, thus raising the overall performance scores. It is probable, then, that a template matching procedure that uses a measure of goodness-of-fit would yield scores similar to those given in Table I.

In summary, we can assert that acoustic invariance for place of articulation is directly derivable from the acoustic signal. This invariance for place of articulation has been shown in this study for stop consonants, and some preliminary data have suggested that, possibly with some modifications of the analysis procedures, there are also invariant properties for place of articulation for nasals. These properties reside in the short-term spectrum at consonantal release and are on the whole independent of vowel context effects, syllable position, and speakers. As has been discussed elsewhere (Stevens and Blumstein, 1978 and 1979), the notion that there is acoustic invariance directly derivable from the speech signal has important implications for the nature and characteristics of the speech processing mechanism. In particular, it has been hypothesized that the auditory system is endowed with innate property-detecting mechanisms which are specifically tuned to detect the invariant properties which reside in the short-term spectrum at onset and offset. Such a system requires that it be sensitive to abrupt changes in intensity at an onset or offset, and that it can assign certain simple properties to the resultant spectrum. There is, in fact, some preliminary evidence from auditory psychophysics and auditory physiology to sug-

gest that the auditory system processes discontinuities at abrupt onsets in a special manner (Zhukov *et al.*, 1974; Kiang *et al.*, 1965; Leshowitz and Cudahy, 1975). Presumably, the extraction of these properties is dependent upon the interaction of sets of individual neurons which together can extract particular features or attributes from a complex auditory pattern. Although there is no direct evidence from studies with humans, recent investigations in animal electrophysiology indicate that such a system is not out of the realm of possibility. In particular, studies using the cat (Nelson *et al.*, 1966), the monkey (Katsuki *et al.*, 1962), and the bat (Suga, 1972) have demonstrated that particular features or characteristics of complex auditory stimuli appear to be selectively detected or processed by limited sets of neuronal cells. These characteristics may be detected by fairly "abstract" pattern detectors that are not tied to particular frequencies. For example, it has been shown that there are units that are sensitive to rising or falling patterns across the time domain independent of the particular frequency range of the test stimuli (Whitfield, 1967).

It is instructive to consider the theoretical implications of such a system in terms of what we know about the behavior of the speech-processing mechanism. Recent investigations in speech perception have shown that infants at least as young as one month old can not only discriminate speech sounds but they do so in a manner similar to adults (Eimas *et al.*, 1971; Moffitt, 1971; Morse, 1972; Eimas, 1974). This discrimination ability and the resultant correspondences to adult perception suggest that these abilities could not have been acquired through experience and interaction with the linguistic environment, but rather they reflect the operation of innate biological mechanisms selectively tuned to the fundamental set of acoustic properties used in natural language. In this view, invariant, context-independent acoustic properties provide the first set of cues used by the infant for the perception of the phonetic dimensions of speech. Nevertheless, the human perceptual system in general and the infant perceptual system in particular does not function independent of its environment. What the linguistic environment provides are contextually dependent cues which always co-vary with the invariant acoustic properties. For example, the onset frequencies of the formants for given consonants change in relation to the particular vowel as do the formant motions. These context-dependent cues provide further information about the phonetic structure of speech, and presumably are learned by the infant and ultimately integrated as part of the speech processing mechanism.

There is little doubt that listeners can and do make use of these contextually determined cues both in the experimental situation and in the normal communication process. First, it has been shown that adult listeners can identify place of articulation for two-formant synthetic stimuli (Delattre *et al.*, 1955; Cooper *et al.*, 1952), which apparently do not possess the general properties discussed above. Presumably the listener invokes the context-dependent cues such as formant motions and the formant frequencies of the following vowel to make his phonetic decision. Second, the normal communication process is such that the perceptual system may need to rely on both contextually dependent and independent cues. Speech occurs in the context of a noisy channel where extraneous noise may mask out some of the acoustic information within the signal itself. A perceptual system that could not make use of the context-dependent information provided in the speech stream would most certainly be maladaptive. Thus, the normally processing speech system presumably depends upon both the *primary* acoustic properties and the *secondary* context-dependent cues for effective communication. The invariant properties are considered primary because they are invariant and can provide in the first instance the basis for phonetic categories. The context-dependent cues are considered secondary because they can only be defined with reference to adjacent segments and must be learned through interaction with the linguistic environment.

The results presented in this study for place of articulation far from complete an analysis of the inventory of the place-of-articulation dimension for stop consonants used in natural language. Of particular interest are the acoustic properties characteristic of the so-called coronal consonants including the dental, alveolar, palato-alveolar, and retroflex places of articulation (Stevens and Blumstein, 1975; Chomsky and Halle, 1968; Ladefoged, 1971). Presumably, all of these phonetic categories would show the diffuse-rising acoustic property in the form presented in this study or in a slightly modified form. Differences among these categories would reside in other attributes of the sound, probably in terms of the direction of the formant transitions or the frequencies of particular formants at onset. Thus, in the case of those languages which have phonologically more than one coronal consonant (e.g., Hindi, Malayalam), other acoustic properties would need to be invoked to distinguish within the coronal class. The nature of these properties and the degree to which they are invariant requires further study.

That speech requires a complex interface between the perceptual and production systems is obvious. This paper has attempted to begin to characterize the nature of this interface. In particular, it has been shown that the speaker provides the listener with invariant acoustic cues, cues which can be directly derived from the speech signal itself. Thus, the interface between the perceptual and production systems resides in the acoustic signal where the properties of speech can be uniquely and invariantly specified. It remains to be determined in what ways the production system codes and ultimately programs these invariant properties and in what way the perceptual system extracts and categorizes them. However, further investigation of the basic properties of speech and a delineation of their structure should provide us with some important clues as to the nature and operating principles of these systems and ultimately lead us to a further understanding of the speech process.

ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health, Grant NS-04332. Dr. Blumstein's participation was supported in part by the John Simon Guggenheim Foundation and the Radcliffe Institute. We thank Dennis Klatt and Victor Zue for their positive influence and helpful comments, and Roman Jakobson for his longstanding interest in these issues and many helpful discussions. Measurement and display of the linear prediction spectra were accomplished using a program written by Dennis Klatt.

[1]Another approach to the selection of a suitable time window and a suitable procedure for calculating the spectrum would be to base the selection on certain known properties of the peripheral auditory system (Searle *et al.*, 1979). Until we have more detailed information concerning the characteristics of this system, particularly in response to abrupt onsets, we have chosen to take a more empirical approach to the selection of spectrum matching procedures.

Blumstein, S. E., and Stevens, K. N. (1979). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. (in press).

Carney, A. E., Widin, G. P., and Viemeister, N. F. (1977). "Noncategorical perception of stop consonants differing in VOT," J. Acoust. Soc. Am. 62, 961–970.

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English* (Harper and Row, New York).

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.

Cutting, J., and Rosner, B. (1974). "Categories and boundaries in speech and music," Percept. Psychophys. 16, 564–570.

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

Delgutte, B. (1978). "Codage des changements rapides d' intensité dans nerf auditive: expériences avec des sons pures," Neuvième journées d'étude sur la parole, Lannion 31 May–2 June.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (1971). "Speech perception in infants," Science 171, 303–306.

Eimas, P. D. (1974). "Auditory and linguistic processing of cues for place of articulation by infants," Percept. Psychophys. 16, 513–521.

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

Fant, G., Ishizaka, K., Lindquist, J., and Sundberg, J. (1972). "Subglottal formants," Speech Transmission Laboratory QPSR, Royal Institute of Technology, Stockholm, No. 1, pp. 13–17.

Halle, M., Hughes, G. W., and Radley, J.-P. A. (1957). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. 29, 107–116.

Halle, M., and Stevens K. N. (1972). Speech recognition: a model and a program for research, in *The Structure of Language*, edited by J. A. Fodor and J. J. Katz (Prentice-Hall, New Jersey), pp. 604–612.

Houtgast, T. (1974). "Auditory analysis of vowel-like sounds," Acustica 31, 320–324.

Jakobson, R., Fant, G., and Halle, M. (1963). *Preliminaries to Speech Analysis* (MIT Press, Cambridge, MA).

Katsuki, Y. N., Suga, N., and Kanno, Y. (1962). "Neural mechanisms of the peripheral and central auditory systems in monkeys," J. Acoust. Soc. Am. 34, 1396–1410.

Kiang, N. Y.-S., Watanabe, T., Thomas, E. C., and Clark, L. F. (1965). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve* (M.I.T. Press, Cambridge).

Klatt, D. H. (1976). "A digital filter bank for spectral matching," Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, IEEE Catalogue No. 76, CH 1067-8 ASSP, 537–540.

Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics* (University of Chicago Press, Chicago).

Leshowitz, B., and Cudahy, E. (1975). "Masking patterns for continuous and gated sinusoids," J. Acoust. Soc. Am. 58, 235–242.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). "The discrimination of speech events within and across phonetic boundaries," J. Exptl. Psychol. 54, 358–368.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. 74, 431–461.

Malécot, A. (1958). "The role of releases in the identification of released final stops," Language 34, 370–380.

Miller, J. D., Engebretson , A. M., Spenner, B. F., and Cox, J. R. (1977). "Preliminary analyses of speech sounds with a digital model of the ear," J. Acoust. Soc. Am. 62, 513.

Miller, J. L. (1977). "Nonindependence of feature processing in initial consonants," J. Speech Hear. Res. 20, 519–528.

Miller, J. L., and Eimas, P. D. (1977). "Studies in the perception of place and manner of articulation: a comparison of the labial-alveolar and nasal-stop distinctions," J. Acoust. Soc. Am. 61, 835–845.

Moffitt, A. R. (1971). "Consonant cue perception by twenty- to twenty-four-week-old infants," Child Develop. 42, 717–731.

Morse, P. A. (1972). "The discrimination of speech and non-speech stimuli in early infancy," J. Exptl. Child Psychol. 14, 477–492.

Nelson, P. G., Erulkar, S. D., and Bryan, J. S. (1966). "Responses of units of the inferior colliculus to time-varying acoustic stimuli," J. Neurophysiol. 29, 834–860.

Pastore, R. E., Ahroon, W. A., Baffuto, K. J., Friedman, C., Puleo, J. S., and Fink, E. A. (1977). "Common-factor model of categorical perception," J. Exptl. Psych.: Hum. Perc. Perf. 3, 686–696.

Plomp, R. (1964). "The ear as a frequency analyzer," J. Acoust. Soc. Am. 36, 1628–1636.

Potter, R. K., Kopp, G. A., and Green, H. C. (1947). *Visible Speech* (Van Nostrand, New York).

Searle, C. L., Jacobson, J. Z., and Kimberly, B. (1979). 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) (in press).

Stevens, K. N., and Halle, M. (1967). Remarks on analysis by synthesis and distinctive features, in *Models for the Perception of Speech and Visual Form*, edited by W. Wathen-Dunn (M.I.T. Press, Cambridge), pp. 88–102.

Stevens, K. N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data, in *Human Communication, A Unified View*, edited by P. B. Denes and E. E. Davids (McGraw-Hill, New York), pp. 51–66.

Stevens, K. N., and Blumstein, S. E. (1975). "Quantal aspects of consonant production and perception: a study of retroflex consonants," J. Phonet. 3, 215–234.

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. (1979). 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 Assoc., New Jersey) (in press).

Studdert-Kennedy, M., Liberman, A. M., Cooper, F. S., and Harris, K. S. (1970). "Motor theory of speech perception: a reply to Lane's critical review," Psychol. Rev. 77, 234–249.

Suga, N. (1972). "Analysis of information bearing elements in complex sounds by auditory neurones of bats," Audiol. 11, 58–72.

Wang, W. S.-Y. (1959). "Transition and release as perceptual

cues for final plosives," J. Speech Hear. Res. 3, 66–73.
Whitfield, I. (1967). *The Auditory Pathway* (Edward Arnold, London).
Winitz, H., Scheib, M. E., and Reeds, J. A. (1972). "Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech," J. Acoust. Soc. Am. 51, 1309–1317.
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. Zh. 20, 386–392 (1974)].
Zue, V. (1976). "Acoustic characteristics of stop consonants: a controlled study," Ph.D. thesis, MIT (unpublished).