A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study

Aditi Lahiri,[a] Letitia Gewirth,[b] and Sheila E. Blumstein
Brown University, Department of Linguistics, Providence, Rhode Island 02912

(Received 6 July 1983; accepted for publication 23 April 1984)

This study explored the claim that invariant acoustic properties corresponding to phonetic features generalize across languages. Experiment I examined whether the same invariant properties can characterize diffuse stop consonants in Malayalam, French, and English. Results showed that, contrary to theoretical predictions, we could not distinguish labials from dentals, nor could we classify dentals and alveolars together in terms of the same invariant properties. We developed an alternative metric based on the change in the distribution of spectral energy from the burst onset to the onset of voicing. This metric classified over 91% of the stops in Malayalam, French, and English. In experiment II, we investigated whether the invariant properties defined by the metric are used by English-speaking listeners in making phonetic decisions for place of articulation. Prototype CV syllables—[b d] in the context of [i e a o u]—were synthesized. The gross shape of the spectrum was manipulated first at the burst onset, then at the onset of voicing, such that the stimulus configuration had the spectral properties prescribed by our metric for labial and dental consonants, while the formant frequencies and transitions were appropriate to the contrasting place of articulation. Results of identification tests showed that listeners were able to perceive place of articulation as a function of the relative distribution of spectral energy specified by the metric.

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

INTRODUCTION

In recent years, a theory of acoustic invariance has been elaborated which makes two principal claims. The first is that there are invariant acoustic patterns in the speech signal which correspond to phonetic features and which remain invariant across speakers, phonetic contexts, and languages. The second claim is that human perceivers use these properties to provide the phonetic framework for natural language and to process the sounds of speech in ongoing perception.

A number of studies have examined the invariance hypothesis for place of articulation in English stop consonants. While there has been some disagreement about the particular form of invariance, acoustic patterns have been defined which have remained invariant in the sense that the same acoustic pattern corresponds to a particular place of articulation across vowel contexts and speakers (Searle *et al.*, 1979; Blumstein and Stevens, 1979; Kewley-Port, 1983), manner of articulation (Searle *et al.*, 1979; Blumstein and Stevens, 1979), and syllable position (Blumstein and Stevens, 1979).

The theory of acoustic invariance claims that a particular phonetic dimension should be realized by the same invariant property across all languages (Stevens and Blumstein, 1978). For example, the phonetic feature corresponding to the labial place of articulation should show the same invariant property irrespective of the language in which this phonetic feature occurs. The present investigation explored this claim by focusing on invariant properties for diffuse stop consonants—including the labial, dental, and alveolar places of articulation—in Malayalam, French, and English. Experiment I examines whether the same invariant acoustic properties can characterize place of articulation in these different languages. Experiment II investigates whether these invariant properties are used by the listener in making phonetic decisions about place of articulation.

I. EXPERIMENT I

Fant's acoustic theory of speech production predicts that different vocal tract configurations for place of articulation will result in distinct spectral patterns (Fant, 1956, 1960). Based on Fant's theory, Stevens and Blumstein (1978) described three distinct patterns corresponding to labial, alveolar, and velar stop consonants in terms of the shape of the spectrum in the vicinity of the stop release. For labial consonants, there were a number of peaks in the spectrum which were fairly spread out or diffuse, and the amplitudes of the peaks either had more energy in the low frequencies than the high frequencies (diffuse-falling pattern) or they were evenly distributed throughout the spectrum (diffuse-flat pattern). For alveolar consonants, there was also a diffuse spread of peaks of energy, but the amplitudes of these peaks were greater in the high frequencies (diffuse-rising pattern). Finally, for velar consonants, there was one prominent spectral peak, usually occurring in the midfrequency region, which dominated the entire spectrum (compact pattern).

It is important to note that there are two invariant properties that characterize the class of labial and alveolar consonants. Both places of articulation *share* the acoustic property of diffuseness (cf. also Jakobson *et al.*, 1963), and they are *distinguished* by the shape of the spectral energy distribution, with labials showing a flat or falling spectrum and alveolar consonants showing a rising spectrum. The first ques-

[a] Now at Department of Linguistics, University of California, Santa Cruz, CA 90054.
[b] Now at Department of Psychology, University of Pennsylvania, Philadelphia, PA 19174.

 0001-4966/84/080391-14$00.80

tion we wanted to address was whether the invariant acoustic property for labial stop consonants, i.e., diffuse-flat or diffuse-falling, will uniquely distinguish the class of labial stops from the class of nonlabial diffuse stops in other languages.

The second question focuses on nonlabial diffuse stop consonants. These consonants include at least two different places of articulation—alveolar and dental. The work of both Stevens and Blumstein (1978) and Kewley-Port (1983) indicated that alveolar consonants in English exhibited the diffuse-rising property. Fant investigated the spectral characteristics of dental consonants in Swedish. Results from both theoretical considerations and analyses of measured spectra indicated that dental consonants are also diffuse with a predominance of high-frequency energy (Fant, 1960). While dental and alveolar consonants do not share the exact same place of articulation, it has been suggested that the spectral characteristics of dental and alveolar stop consonants should be similar, characterized by a diffuse spread of energy with a predominance of high-frequency energy (Blumstein and Stevens, 1979; Halle and Stevens, 1979). In fact, phonological theory classifies both dental and alveolar consonants in terms of the same phonetic feature—[coronal] (Chomsky and Halle, 1968; Halle and Stevens, 1979). Therefore, we wanted to explore whether alveolar and dental stop consonants do in fact share the same invariant acoustic property, i.e., diffuse-rising, when the same measurement procedures are applied across a number of languages.

To this end, we focused on three languages: Malayalam, French, and English. Malayalam uses labial, dental, and alveolar stop consonants contrastively. However, these contrasts occur only intervocalically in voiceless geminate stop consonants.[1] We also studied French and English where labial consonants contrast with either dental (French) or alveolar (English) stop consonants in similar phonetic environments.

A. Pilot studies

Two pilot studies were conducted. In the first (Lahiri, 1980), three subjects (two males and one female) were asked to read a list of five repetitions of five real words in Malayalam containing the stops [t t̪] followed by the vowels [i e a ə u]. The 150 utterances were analyzed using the template-fitting procedures of Blumstein and Stevens (1979). In particular, LPC spectra[2] were derived at the burst release using a 25.6-ms half-Hamming window, and the obtained spectra were individually tested against the Blumstein and Stevens diffuse-rising template. We also analyzed a total of 25 labial stops spoken by one of the male speakers in the same phonetic context. The obtained spectra for the labial stops were individually tested against the Blumstein and Stevens diffuse-falling template. (The diffuse-falling template and the criteria used to fit it accept both diffuse-falling and diffuse-flat onset spectra.)

Results showed that the labial consonants were correctly accepted by the diffuse-falling template 88% of the time. The remaining 12% of the spectra for labial stops had a diffuse-rising shape. With regard to the alveolar and dental stop consonants, 71% of the alveolar stops and only 57% of the dental stops were correctly accepted by the diffuse-rising template. Although the dental consonants were diffuse, their spectral shape was similar to labials in that the distribution of energy was predominantly flat. Thus the shape of the spectrum could not distinguish labial from dental consonants in Malayalam. Figure 1 shows the onset spectra of a typical dental and a typical alveolar stop consonant in Malayalam. In both cases, the overall spectrum has a diffuse-flat shape, rather than the predicted diffuse-rising shape.

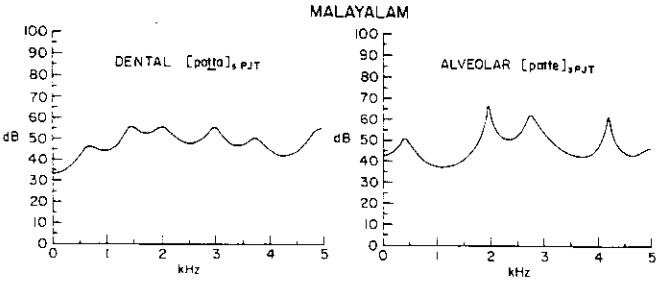

FIG. 1. Examples of spectra sampled at the burst release (sampled at 10 kHz) of dental and alveolar intervocalic stop consonants in Malayalam. LPC spectra (with pre-emphasis) were derived using a 25.6-ms half-Hamming window. The initials to the right of the utterance indicate the speaker and the number indicates the particular utterance measured.

In the second pilot study (Lahiri and Blumstein, 1981), 100 utterances of a male French speaker were analyzed. These included 50 voiced and 50 voiceless labial and dental stops produced in initial position in the environment of the vowels [i e a o u]. The spectra of these utterances were analyzed again following the procedure of Blumstein and Stevens (1979). The spectra of the labial consonants in French were found to be either diffuse-flat or -falling, similar to the labial stops in English and Malayalam. However, on examining the spectra for the French dentals, it was clear that they were very similar to the Malayalam dentals, in that they were more diffuse-flat than rising. To quantify these observations, we tested the obtained spectra for the dental stops against the Blumstein and Stevens diffuse-rising template. Only 40% were correctly accepted by the template. Figure 2 shows examples of the onset spectra for a typical French labial and dental stop consonant. The gross shape of the spectra for both consonants is diffuse-flat.

Thus the results of the two pilot studies indicated that, contrary to predictions from phonological and acoustic the-

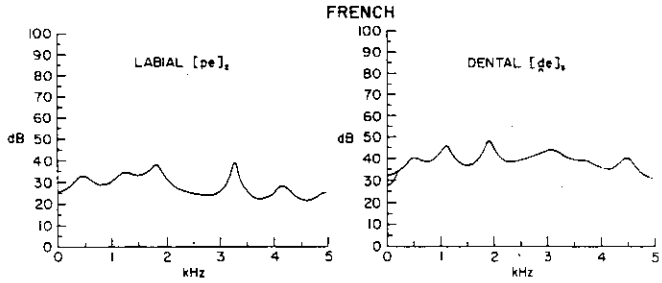

FIG. 2. Examples of spectra sampled at burst onset of labial and dental stop consonants in French. A 25.6-ms half-Hamming window was used to derive the spectra. The number to the right of the utterance indicates the particular utterance measured.

ory (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979; Halle and Stevens, 1979), the gross shape of the spectrum (1) fails to distinguish the labial from the dental stop consonants, and (2) fails to group the dental with the alveolar stop consonants.

B. Reanalysis

Given these results, it was necessary to reconsider the question of acoustic invariance for place of articulation in diffuse stop consonants. The main issues of concern were whether the acoustic invariance theory was incorrect in its claims that invariant properties corresponding to phonetic features would indeed generalize across languages, and whether invariant properties could be found that uniquely characterize labial versus nonlabial diffuse stop consonants.

Since the gross shape of the onset spectrum failed to appropriately classify stop consonants, we considered whether other properties inherent in the burst could be invoked. One possible property is that of the amplitude of the burst. It has been shown that the amplitude of the burst in alveolar stop consonants in English is generally higher than that of labials. In the case of alveolar stops, the amplitude of spectral peaks in the burst may be somewhat greater than the amplitude of the formants at the onset of voicing, particularly at the higher frequencies (Zue, 1976). We explored whether the amplitude of the burst was greater for dental and alveolar consonants compared to labial consonants by analyzing the stimuli of the pilot experiments, as well as 100 English voiced and voiceless labial and alveolar stops produced by a male speaker in the environment [i e a o u]. LPC spectra were made of each utterance at the burst using a 10-ms full-Hamming window and at the onset of voicing. We determined by inspection whether the amplitude of the burst was higher than the onset of voicing throughout the spectrum, only at high frequencies, or only at low frequencies. Results showed that, as expected, labial stops have weak bursts. None of the Malayalam labial stops and only 4% of the English and 14% of the French labial stops showed greater energy in the burst compared to voicing onset. This greater energy occurred in the low-frequency region. Although 87% of the alveolar stops in English and 75% of the Malayalam alveolar stops had bursts with greater energy than the vowel onset, only 54% of the Malayalam and French dental stops showed greater energy in the burst. For those alveolar and dental stops with greater energy in the burst, this energy occurred primarily in the high frequencies, with 93% of English alveolar stops, 87% of French dental stops, 75% of Malayalam alveolar stops, and 94% of Malayalam dental stops showing this pattern. The remaining stimuli with greater energy in the burst than in voicing onset had greater energy in both low and high frequencies. Nevertheless, since approximately one-half of the Malayalam and French dental stops (i.e., 46%) had relatively weak bursts similar to those of labials, and these bursts showed no predominance of energy in any particular frequency region, this measure of the burst amplitude could not distinguish dental stops from labial stops, nor could it reliably group dental stops with alveolar stops.

On the basis of this pilot work, it was clear that static properties as measured by the gross shape of the onset spectrum or by the amplitude of the burst failed to adequately classify labial, dental, and alveolar stops. Kewley-Port (1983) examined invariant acoustic properties for place of articulation in English stop consonants by focusing on time-varying properties of stop consonants from the release burst into the vowel portions of CV syllables. One of the properties of her metric was the spectral tilt of the burst. Since our pilot work showed that the gross shape (i.e., tilt) of the spectrum at stimulus onset could not appropriately classify dentals and alveolars, it was clear that her metric would not work either. However, an advantage of Kewley-Port's analysis over that of Blumstein and Stevens (1979) was that it explored spectral changes over time. Consequently, we attempted to determine whether the sought-for invariant might lie in dynamic spectral changes over time, rather than in the gross shape of the spectrum at a static point in time (corresponding to the burst release). We focused on changes in both the spectral and amplitude characteristics of the waveform over time, since these characteristics were critical in earlier investigations of the acoustic properties for place of articulation in stop consonants (Fant, 1960; Stevens and Blumstein, 1979; Zue, 1976).

Three-dimensional plots of a series of LPC spectra were made of a large number of utterances using a full-Hamming window of 10 ms with a window movement of 5 ms (Mertus, 1979). We were thus able to observe changes in spectral characteristics from burst onset well into the vowel (cf. Searle *et al.*, 1979, 1980; Kewley-Port, 1983). Rather than looking at the absolute *shape* or tilt of the spectrum, as Kewley-Port (1983) did, we focused on the relative changes in the distribution of energy from the burst release to the onset of voicing. The top half of Fig 3 shows an example of two three-dimensional displays, one for the syllable [do] (left panel) and the other for the syllable [bo] (right panel) spoken by a French speaker. Inspection of these three-dimensional plots revealed that the changes in *distribution* of energy from burst release to the onset of voicing were distinctively different for these two classes of stops.

For labial stop consonants, either the difference in energy between the burst release and the onset of voicing was less at low frequencies than at high frequencies, or the difference in energy was about the same at low and high frequencies. These patterns were obtained whether the shape of the spectrum at stimulus onset was diffuse-falling (as in Fig. 3) or diffuse-flat. This smaller difference in energy change in the low as compared to the high frequencies could be interpreted as indicating greater low-frequency energy in the spectrum, since there was less acoustic change over time in this frequency region. An equal difference in energy at low and high frequencies could be interpreted as a flat distribution of energy in the spectrum, since similar changes occurred in both the low and high frequencies.

For dental and alveolar consonants, the differences in energy between the stimulus onset and onset of voicing was less at high frequencies than at low frequencies, whether the gross shape of the onset spectrum was diffuse-flat (as in Fig. 3) or diffuse-rising. This smaller difference in energy change in the high as compared to low frequencies could be inter-

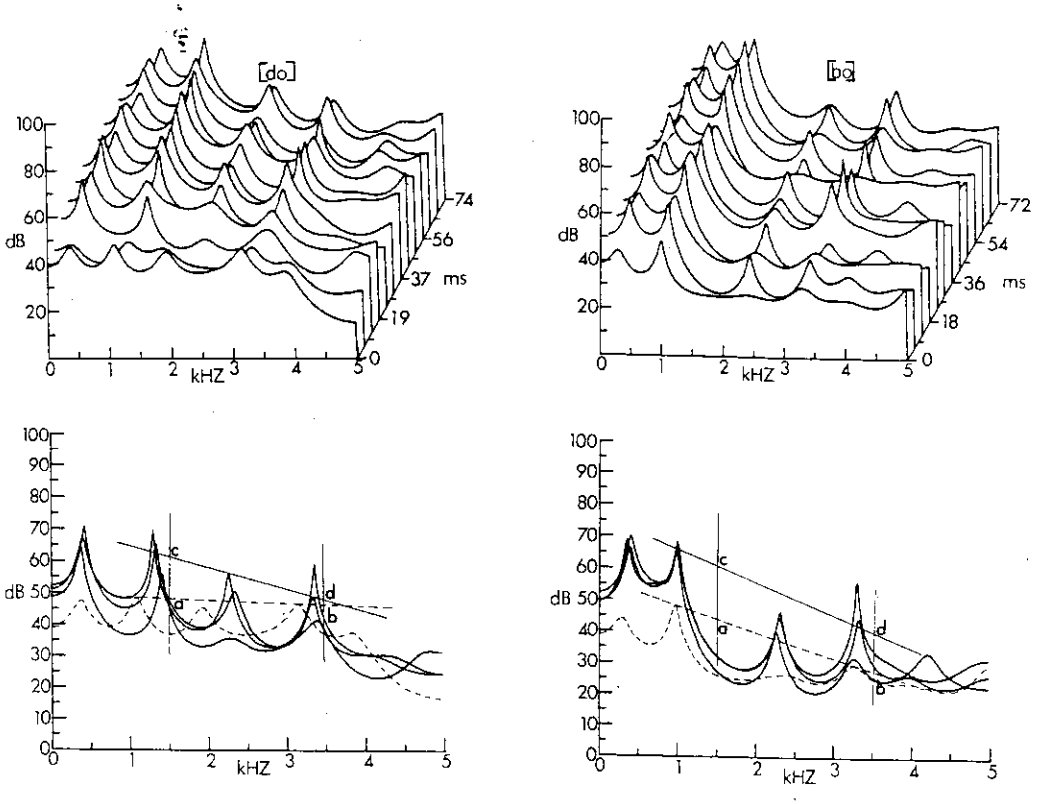

FIG. 3. The top half of the figure shows examples of two three-dimensional plots of LPC spectra of French voiced labial and dental stop consonants in the environment of the vowel [o]. The spectra were derived using a full-Hamming window of 10 ms with a window movement of 5 ms. The bottom half of the figure shows the changes in the spectral characteristics of the same syllables. The spectrum at the burst is indicated by the dotted line and the spectra of the first three pulses at the onset of voicing are indicated by the solid lines. The window was placed at the beginning of the burst, and then pitch synchronously at the first three glottal pulses at the onset of voicing. Line ab shows the derived slope of the burst joining the $F2$ and $F4$ peaks. Line cd indicates the derived slope of the onset of voicing taking the average values of the $F2$ and $F4$ peaks. The vertical lines at 1500 and 3500 Hz represent the low- and high-frequency markers.

preted as greater high-frequency energy in the spectrum, since there was less acoustic change in this frequency region over the time domain.

The bottom panel of Fig. 3 shows an example of the changes in the spectral characteristics for a labial and dental consonant from the burst release compared to the onset of voicing. The dotted line corresponds to the burst release and the solid lines correspond to the energy of the first three pulses of voicing. Note that for the labial, despite the diffuse-falling onset spectrum, there was little change in the relative distribution of energy at high and low frequencies from the burst release to the onset of voicing. In contrast, for the dental consonant, with a diffuse-flat onset spectrum, there was a small change of energy at the high frequencies but a large change at low frequencies.

These observations suggested that the hypothesis of acoustic invariance for place of articulation might still be maintained if a measure based on the changes in energy distribution were used instead of a measure based on the shape of the spectrum in the vicinity of the stop release. It was our goal to define measurement procedures that would appropriately classify labial versus dental and alveolar stop consonants across French, Malayalam, and English.

C. The metric

Since it was the change in energy from the release burst to the onset of voicing which seemed critical in characterizing these diffuse consonants, we focused on these two portions of the waveform in establishing the metric. As a result, the segmentation boundaries for the analysis of both voiced and voiceless stop consonants were very straightforward and easily determined by visual inspection of the waveform. However, by choosing these segmentation procedures, the measurements of the burst and the onset of voicing encompass somewhat different portions of the waveform for voiced and voiceless consonants. Nevertheless, on the basis of the original inspection of the three-dimensional spectra, the portions of the waveform defined in relation to the emergence of an $F1$ peak (i.e., the onset of voicing) seemed to provide a consistent pattern within each place of articulation, regardless of the voicing characteristics of the consonant.

Using a 10-ms full-Hamming window, we took spectral samples at the burst and a pitch-synchronous sample at each of the first three glottal pulses at the onset of voicing. We found that sampling the first three glottal pulses at the onset of voicing, rather than only the first glottal pulse, provided a more consistent measure of the location of the spectral peaks corresponding to the individual formant frequencies (see below). For the voiced consonants which contained prevoicing, the window was placed at the burst release, excluding the prevoicing portions of the waveform. The onset of voicing measure was made at the first three glottal pulses starting after the burst. For the voiceless consonants, the burst portion of the stimulus was often longer than 10 ms. In that case, we took multiple spectra of the burst, sampling at every 10 ms from the burst onset up to (but not including) the onset of voicing. By sampling the full burst, we avoided the problems of segmenting the burst into separable components including the burst, aspiration, and frication.

To compare the changes in energy distribution from the burst onset to the onset of voicing, we took the values of the $F2$ and $F4$ peaks of the spectra for the burst onset. In the case where multiple spectra were sampled at the burst, we took the average value of these spectra. We also took the average values of the $F2$ and $F4$ peaks for the three glottal pulses at the onset of voicing. The slopes of the spectra for the burst and the onset of voicing were determined by drawing straight lines through the two points marking these average

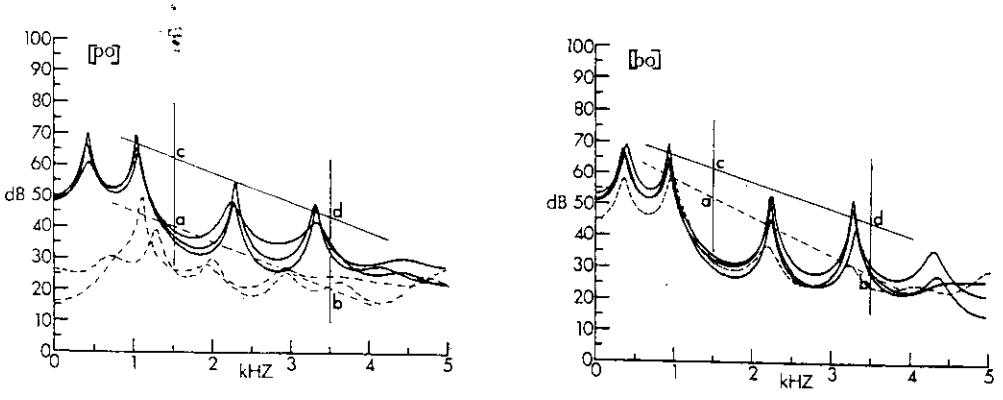

FIG. 4. Examples of spectra of French labial voiceless and voiced consonants. The dotted lines correspond to the burst release and the solid line indicates the spectra of the first three glottal pulses at the onset of voicing. For the voiceless consonant [po], multiple spectra taken every 10 ms were required to sample the burst. Line ab shows the derived slope of the onset taking the average values of the $F2$ and $F4$ peaks. Line cd indicates the derived slope of the onset of voicing taking the average values of the $F2$ and $F4$ peaks. See also legend for Fig. 3.

values. In the case of the voiced consonants, where only one burst spectrum was obtained, a straight line was drawn between the $F2$ and $F4$ peaks. The bottom panel of Fig. 3 shows the derived slopes of the burst (line ab) and of the onset of voicing (line cd) for the syllables [do] and [bo]. Figure 4 shows the derived slopes for the burst (line ab) and the onset of voicing (line cd) for a voiceless and a voiced labial consonant.

We next compared the differences in energy at high and low frequencies between the burst and the onset of voicing. Comparing the difference in energy at the $F2$ and $F4$ peaks was not appropriate, as the frequency values would differ not only between the burst and onset of voicing within a given utterance, but also across different utterances. Consequently, we arbitrarily defined 1500 and 3500 Hz as representative of low- and high-frequency energy, respectively, for all the spectra, and computed the ratio of the differences in energy between the burst and the onset of voicing at high frequencies in relation to low frequencies. In this way, changes in energy were computed at exactly the same frequencies in all spectra. We computed the *ratio* of the difference in energy at high and low frequencies, i.e., (d − b)/(c − a), rather than the *absolute difference* between the differences at high and low frequencies, i.e., (d − b) − (c − a), because computing the ratios provided a measure of relative change which absolute differences did not. For example, if (d − b) = 20 and (c − a) = 22, the absolute difference would be 2. The same value would be obtained if (d − b) = 2 and (c − a) = 4. Notice, however, that in the first case, the ratio of change is 0.91, indicating very little change in energy at the high relative to low frequencies. However, in the second example, the ratio value is 0.5 indicating that there is half as much energy change at high relative to low frequencies.

The bottom panels of Figs. 3 and 4 show the procedure involved in establishing the metric. Line ab corresponds to the slope of the burst, line cd corresponds to the slope of the onset of voicing, and the vertical lines at 1500 (ca) and 3500 Hz (db) represent the low- and high-frequency markers, respectively. The metric computed the ratio of the differences of energy between the onset of voicing and the burst at high frequencies (d − b) and at low frequencies (c − a), i.e., (d − b)/(c − a).

The different ratio values were used to classify dentals and alveolars versus labial consonants. The particular critical values chosen were derived from analysis of a set of pilot data, in which we determined what values and criteria provided the best means of dividing the data into the appropriate categories. A positive ratio of <0.5 or a negative ratio with the numerator being negative characterized dental and alveolar consonants. For labial consonants, the ratio was either >0.5 or it was negative, with the denominator having a negative value. Because the ratio values are arbitrary insofar as they were determined by our choice of 1500 and 3500 Hz as the defined low and high frequencies, and the ratio values were chosen because they divided the pilot data systematically into two classes, we are not making any particular theoretical claim about the particular ratio *value* used as the cutoff between dental and alveolar versus labial consonants.

The value of <0.5 for dentals and alveolars indicates that there is at most half as much energy change from the burst to the onset of voicing at high frequencies compared to low frequencies. In this sense, there is less energy change at high frequencies relative to low frequencies. The top panel of Fig. 5 shows an example of a dental consonant with a ratio of <0.5. The actual computed ratio for this consonant (d − b)/(c − a) was 0.14.

In the second pattern for dental and alveolar consonants, there was greater energy at the burst than at the onset of voicing. In most cases, this held only at high frequencies, such that the numerator (d − b) was negative, giving a negative ratio. There were some dental and alveolar consonants, however, where the energy at the burst was greater in all frequency regions compared to the energy at the onset of voicing. In this case, both numerator (d − b) and denominator (c − a) were negative. However, the ratio of two negative numbers was *not* taken to be a positive number. Instead, a negative numerator was automatically interpreted as a dental or alveolar stop consonant, regardless of the positive or negative value of the denominator. The middle and bottom panels of Fig. 5 show these two types of patterns. In the middle panel, the slope of the burst (ab) shows greater energy only at the high frequencies relative to the slope at the onset of voicing (line cd), whereas in the bottom panel, the slope of the burst (ab) shows greater energy at both high and low frequencies relative to the onset of voicing (line cd).

All of the patterns described by the metric for dental and alveolar stop consonants are consistent with earlier views of the acoustic properties for dental and alveolar stop consonants. In particular, both places of articulation display

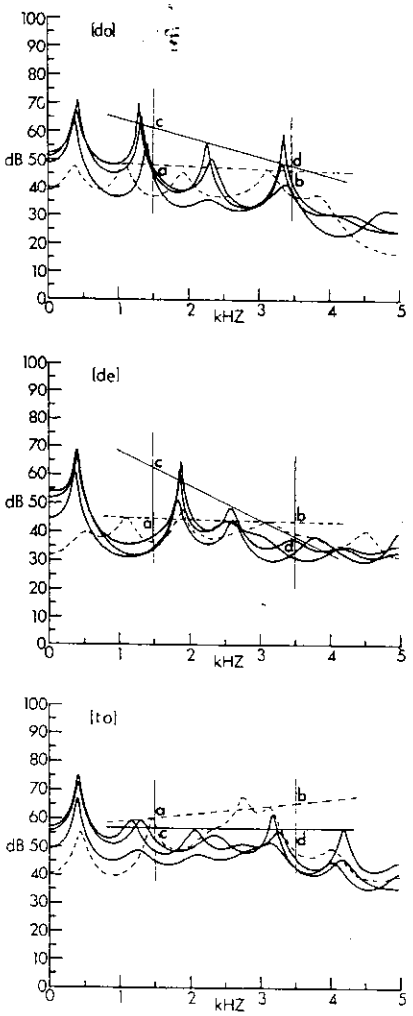

FIG. 5. Examples of spectra of French dental consonants at the burst release (indicated by the dotted line) and onset of voicing (indicated by the solid lines). The lines ab and cd correspond to the derived slopes of the burst onset and onset of voicing as described in Fig. 3. The formula for computing the ratio was d − b/c − a. The top panel is an example of a dental consonant with a ratio of 0.14. In the middle panel, the ratio is negative, the slope of the burst having greater energy at higher frequencies. The actual computed ratio is − 0.27. The computed ratio of the bottom panel is 3 (− 9/ − 3) indicating greater energy at both high and low frequencies for the burst release (note: the ratio of two negative numbers was *not* taken to be a positive number).

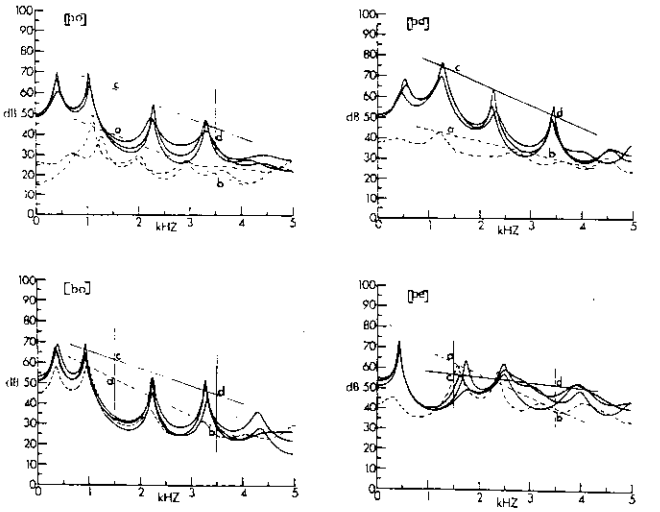

FIG. 6. Examples of spectra of French labial consonants at burst release (shown by the dotted lines) and at the onset of voicing (indicated by the solid lines). The lines ab and cd and the vertical markers at 1500 and 3500 Hz correspond to the description given in Fig. 3. The top left panel shows an example of a voiceless labial stop consonant with a ratio of 1.02 indicating almost no change in the distribution of energy from the burst onset to the onset of voicing. The top right panel gives an example of a labial consonant with a ratio of 0.69. The computed ratio of the bottom left panel is 1.88 indicating sustained low-frequency energy. The bottom right panel provides an example of a labial stop with a negative ratio of − 1.9 showing a predominance of energy at low frequencies.

greater spectral energy in the high frequencies compared to low frequencies or a burst amplitude which is larger than the amplitude of the formants at the onset of voicing.

The patterns obtained for labial consonants either showed a smaller change in energy at low frequencies compared to high frequencies or no real change in energy distribution at high and low frequencies. These patterns are also consistent with earlier claims that labial consonants are characterized by greater energy in the low frequencies or by a relatively flat distribution of energy throughout the spectrum. As mentioned earlier, if the ratio were > 0.5, the consonant was classified as a labial. The various patterns subsumed under the metric value of > 0.5 are shown in Fig. 6. The top right panel of Fig. 6 shows the spectra of the syllable [pa] containing a ratio of > 0.5 [the actual computed ratio (d − b)/(c − a) is 0.69]. The top left panel of Fig. 6 shows the spectra for [po] with a ratio of close to 1, indicating that there was no change in the distribution of energy at high and low frequencies. The bottom left panel of Fig. 6 shows an example of a labial consonant displaying a ratio of > 1, where there is less energy change at low frequencies compared to high frequencies. The actual computed ratio for this consonant (d − b)/(c − a) was 1.88. The bottom right panel of Fig. 6 shows an example of the labial consonant [pe] with more energy at the burst at low frequencies than at the onset of voicing.

D. Analysis and results

The metric was originally developed on the basis of the analysis of 100 utterances spoken by one French male speaker (ten occurrences of each of initial voiced and voiceless unaspirated labial and dental stops produced in the context of [i e a o u]). These were the same utterances used in the second pilot study discussed earlier. In order to determine the extent to which the metric could appropriately categorize labial, dental, and alveolar consonants across different speakers within the same language, we recorded another 200 utterances spoken by two male French speakers (only 198 of these tokens were analyzed as two had to be discarded because of their poor audio quality).

To assess the generality of the metric across different languages, we also applied it to a set of utterances from Malayalam and English. The Malayalam utterances taken from the first pilot study included 70 voiceless unaspirated dental and alveolar stop consonants and 25 unaspirated voiceless labial stop consonants. Fifty of the Malayalam dentals and alveolars and all the labials were spoken by one male speaker and the remaining 20 dentals and alveolars were spoken by

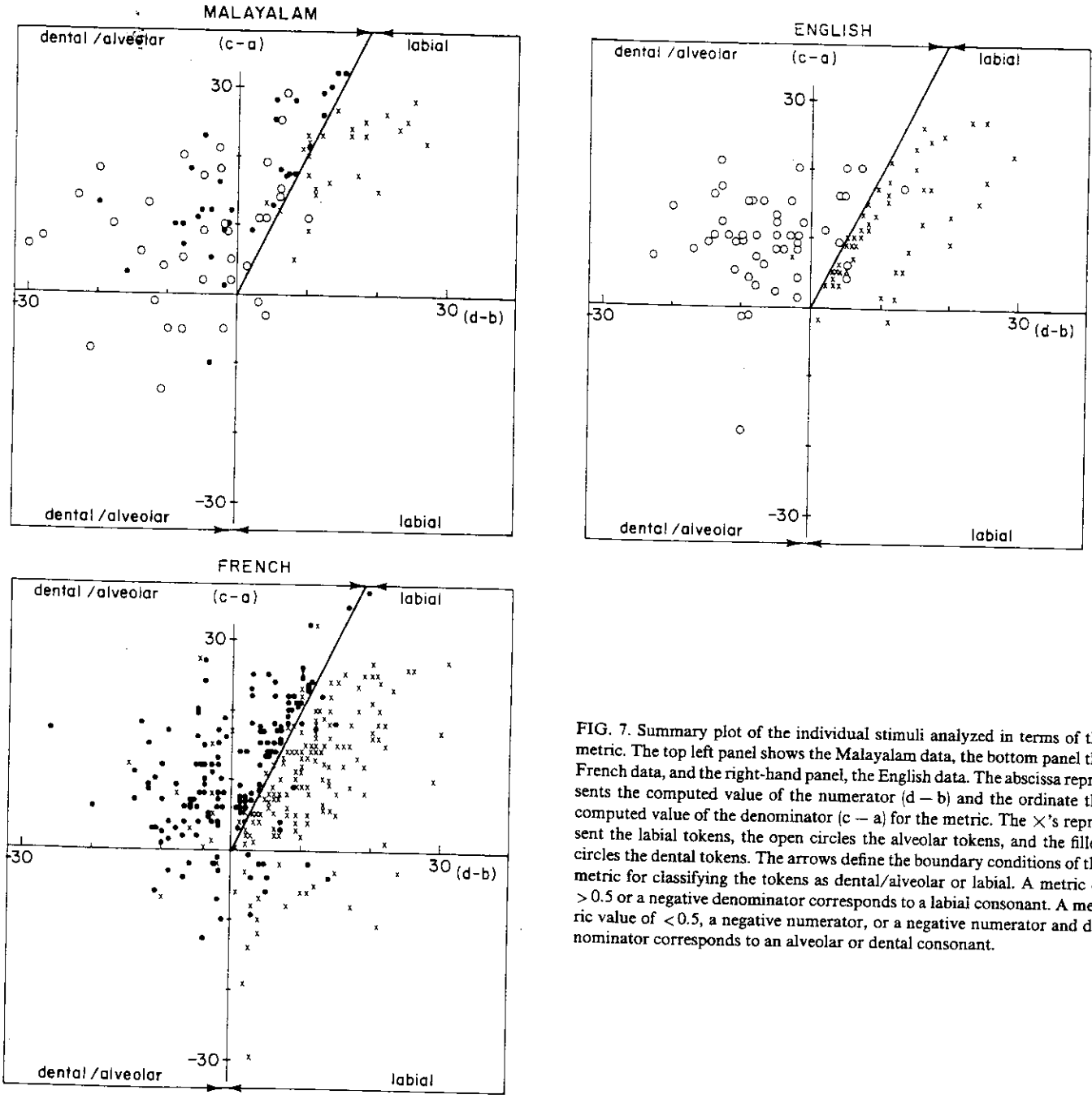

FIG. 7. Summary plot of the individual stimuli analyzed in terms of the metric. The top left panel shows the Malayalam data, the bottom panel the French data, and the right-hand panel, the English data. The abscissa represents the computed value of the numerator (d − b) and the ordinate the computed value of the denominator (c − a) for the metric. The ×'s represent the labial tokens, the open circles the alveolar tokens, and the filled circles the dental tokens. The arrows define the boundary conditions of the metric for classifying the tokens as dental/alveolar or labial. A metric of > 0.5 or a negative denominator corresponds to a labial consonant. A metric value of < 0.5, a negative numerator, or a negative numerator and denominator corresponds to an alveolar or dental consonant.

another male speaker. These utterances were all bisyllabic real words with the stop consonants occurring in intervocalic position. Finally, we analyzed 100 English voiced and voiceless aspirated labial and alveolar stop consonants. The English utterances, also used in the pilot study, consisted of initial voiced and voiceless labial and alveolar stops followed by the vowels [i e a o u] produced by one male speaker, who was one of the speakers used by Blumstein and Stevens (1979) in their original template analysis.

Figure 7 shows the distribution of all of the stimuli plotted as a function of the difference values obtained for the numerator (d − b) and the denominator (c − a) in the computation of the metric. Table I provides an overall summary of the data. Over 91% of the stops were correctly classified with a range from 84% (for Malayalam labial consonants) to 100% (for Malayalam dental consonants). There is little question then that the metric reliably classified labial, alveolar, and dental stop consonants in Malayalam, French, and English.

Looking at the distribution of the stop consonants in

TABLE I. Percent correct classification of the metric. Numbers in parentheses indicate the total number of utterances on which the percentages are based. × indicates that the phonetic dimension does not exist in the language.

	Bilabial	Dental	Alveolar	Total
French (298)	86	88	×	87.0
Malayalam (95)	84	100	91	91.6
English (100)	98	×	94	96.0
Total (493)	89.3	94.0	92.5	91.5

Fig. 7, there was a fairly even distribution of positive and negative indices for dental consonants in both Malayalam and French (46% vs 54% for both languages). The alveolar consonants, however, were characterized predominantly by greater energy at the burst at high frequencies, with Malayalam having 85% negative ratios and English having 87% negative ratios. In contrast, only a small proportion of labial consonants show greater energy at the burst than at voicing onset, and in those cases that do, the burst has, as expected, more energy at low as compared to high frequencies.

E. Discussion

This study was motivated by the fact that the invariant properties for diffuse stop consonants based on the gross shape of the spectrum in the vicinity of the consonant release could not adequately characterize place of articulation in stop consonants occurring in languages other than English. In particular, the class of labial stop consonants could not be uniquely distinguished from the class of dental stop consonants, and the class of alveolar and dental consonants did not share the same invariant property.

Nevertheless, a reanalysis of the acoustic data in terms of dynamic properties suggests that, although the original form of the invariant property was incorrect, an alternative metric, based on relative changes in distribution of energy from the burst release to the onset of voicing, is consistent with earlier theoretical claims. This metric appropriately classified over 91% of the stops of six different speakers from three languages belonging to different language families.

Although the metric is similar to those of Searle *et al.* (1979, 1980) and Kewley-Port (1983) in relying on time-varying spectral properties, it differs from both, and particularly from Kewley-Port's, in ignoring the absolute tilt (or shape) of the spectrum in determining place of articulation. Kewley-Port (1983) and Kewley-Port *et al.* (1983) claimed that invariant acoustic cues for place of articulation in stop consonants are to be found in the changes in spectral distribution of energy over time, rather than in the gross shape of the onset spectrum as proposed earlier by Stevens and Blumstein (1978) and Blumstein and Stevens (1979, 1980). Based on running spectral displays of voiced English stops, Kewley-Port (1983) developed three visual features to characterize acoustic invariance for place of articulation—tilt of burst spectrum, late onset of low-frequency energy (i.e., occurrence of high amplitude $F1$ peaks in the fourth frame of the display or later), and occurrence of midfrequency peaks extending over time. The late onset feature was considered a measure of VOT and classified velars against alveolars and labials. The third feature also was mainly relevant for distinguishing velars from the other places of articulations. Therefore, the main distinguishing factor for the labial and alveolar consonants was the "tilt of the burst"—flat or falling for labials and rising for alveolars. This "tilt" feature is strongly reminiscent of the "gross shape of the onset spectrum" feature proposed by Blumstein and Stevens.

Although previous research has shown that the shape or tilt of the spectrum in the vicinity of the consonant release can correctly classify labial and alveolar stops in English (Blumstein and Stevens, 1979; Kewley-Port, 1983), the present study showed that the shape or tilt of the spectrum in the vicinity of the consonant release cannot correctly classify labial versus dental and alveolar stop consonants in a number of different languages including English, French, and Malayalam. As a result, a stop consonant with a diffuse-flat spectral shape at the release burst might still be classified as a dental consonant, depending on the relative changes in the distribution of energy occurring later.

Although dental and alveolar consonants share a common property, they too should theoretically be distinguished by some invariant property. Recent research measuring the *total* energy in the burst compared to the onset of voicing indicates that dental and alveolar stop consonants can be distinguished with the burst of alveolar stop consonants having an overall greater amplitude than that of dental consonants (Jongman *et al.*, 1984).

The fact that invariant properties can be derived for phonetic features corresponding to place of articulation dimensions across a number of languages provides strong support for a theory of acoustic invariance in speech. However, such evidence only partially fulfills the requirements of such a theory. In particular, it is necessary to demonstrate that invariant properties derived from the acoustic analysis of natural speech have perceptual relevance. It is the object of experiment II to investigate whether listeners are indeed sensitive to the particular form of invariance captured by the metric established in experiment I in making phonetic categorizations for place of articulation in diffuse stop consonants.

II. EXPERIMENT II

A number of studies have been conducted to explore the perceptual significance of invariant properties for place of articulation in stop consonants (Cole and Scott, 1974a,b; Stevens and Blumstein, 1978; Blumstein and Stevens, 1980; Blumstein *et al.*, 1982; Kewley-Port *et al.*, 1983; Walley and Carrell, 1983). Some studies seemed to support the view that invariant perceptual cues for place of articulation reside in the vicinity of the stop release. In a tape-splicing experiment, Cole and Scott (1974a,b) showed that a burst excised from its original vowel context and transposed onto a different vowel context was still identified accurately for place of articulation. These results were obtained for labial and alveolar consonants, although velar consonants were not identified consistently across such transformations. Stevens and Blumstein (1978) showed that the place of articulation categories to which subjects assigned synthetic speech stimuli seemed to be based on the spectral shape of the first 26 ms of the CV stimulus. That is, despite the fact that all of the stimuli varied in the frequencies of the formant transitions and occurred in three vowel environments, the consonants classified as labials shared the diffuse-falling shape, the consonants classified as alveolars shared the diffuse-rising shape, and the consonants classified as velars shared the compact shape.

Blumstein and Stevens (1980) and Kewley-Port *et al.* (1983) showed that listeners were able to identify the appropriate place of articulation in synthetic stimuli containing only the initial portions of a CV stimulus, i.e., the burst and

the first few ms of formant transitions. These results were obtained even when the higher formant transitions, normally moving to the frequencies for the steady-state vowel, were straightened. Thus neither the full complement of formant transitions nor the steady-state vowel normally present in a CV syllable is necessary for the perception of place of articulation. These results suggest that the acoustic information in the vicinity of the stop release is sufficient for perception of place of articulation across different vowel environments.

Nevertheless, in a more detailed tape-splicing experiment than that of Cole and Scott (1974a,b), Dorman *et al.* (1977) found that, although the same burst spliced onto different vowel steady-states was always associated with a particular place of articulation, and thus was "functionally" invariant, it did not provide *sufficient* cues to place of articulation, since the transposed burst often did not maintain a high level of performance in place of articulation identification.

Similar conclusions were drawn from a series of studies designed specifically to test whether it is the *shape* of the spectrum in the vicinity of the stop release which in fact cues the perception of place of articulation (Blumstein *et al.*, 1982; Walley and Carrell, 1983). In the synthetic speech experiments reported above (Stevens and Blumstein, 1979; Blumstein and Stevens, 1980; Kewley-Port *et al.*, 1983), a particular shape of the spectrum corresponding to a particular place of articulation was always associated with the appropriate formant frequency characteristics for that place of articulation. Thus it is not clear whether the subjects' identification of place of articulation was based upon the invariant properties corresponding to the shape of the spectrum or upon the context-dependent cues corresponding to the formant transitions. Theoretically, if a CV stimulus contains a particular spectral shape, it should be identified as having the corresponding place of articulation, *even if* the formant frequencies are not appropriate to that place of articulation. For example, if a CV stimulus is synthesized with formant frequencies corresponding to a [da] but with a spectral shape corresponding to a [ba], the stimulus should be identified as a [ba].

A series of experiments was designed to explore this issue. Blumstein *et al.* (1982) attempted to eliminate context-dependent information present in the formant transitions and following vowel by synthesizing 40-ms CV stimuli containing no formant motions in the second and higher formants. They then manipulated the spectral shape of stimuli with onset formant frequencies appropriate for [ba bi bu da di du]. The spectra of the labial consonants were changed from diffuse-falling to diffuse-rising and the spectra of the alveolar consonants were changed from diffuse-rising to diffuse-falling. In a similar experiment, Walley and Carrell (1983) also manipulated the shape of the consonant spectrum. However, rather than using shortened stimuli with no formant transitions, they synthesized 255-ms CV syllables containing formant motions and steady-states appropriate to the consonants [b d g] in the vowel environments [a u].

The results of both experiments showed that, while shape did contribute somewhat to the phonetic decision of the listener, it did not *direct* perception. That is, subjects did not identify place of articulation as a function of the gross shape of the spectrum, but rather they made such identifications as a function of the *formant frequencies* of the stimuli.

The implications of these results are that the hypothesized invariant property, shape of the onset spectrum, does not seem to provide the listener with sufficient information for perception of place of articulation in stop consonants. The question we intend to pursue in experiment II is whether the relative change in the distribution of energy from the moment of consonantal release to the onset of voicing has perceptual consequences. If the invariant properties for place of articulation reside in such relative changes in spectral energy, then manipulating the spectral characteristics of the speech stimuli dynamically, i.e., the spectral distribution of the burst relative to voicing onset, or vice versa, should result in perceptual shifts in phonetic categorization for place of articulation.

A. Method

1. Stimuli

The stimuli were generated using a computer simulation of a terminal analog parallel formant synthesizer, so that individual control for formant amplitudes could be effected (Klatt, 1980). The synthesizer output was sampled at 10 kHz and was low-pass filtered with a cutoff frequency of about 4800 Hz. The formant frequency, duration, and amplitude values of the stimuli were originally derived from LPC and formant track analysis of ten CV tokens, [pi pe pa po pu di de da do du], spoken by a native French speaker. The values used were from French rather than from English because our metric was initially worked out with reference to the French data, and the English data were only used to corroborate it. Moreover, given that the same invariant properties were found across languages, it should theoretically make no difference for perception from what language the parameter values were taken. Voiceless rather than voiced labials were chosen as the exemplar stimuli because we wanted all of the synthesized stimuli to have the short lag VOT characteristic of voiced consonants in English. The obtained formant and amplitude tracks for the ten CV natural speech tokens were then smoothed to eliminate abrupt changes, sometimes shown by the LPC analyses. Using these values, ten prototype CV syllables were synthesized. For all stimuli, the burst was of 15-ms duration and was followed by the onset of voicing. The fundamental frequency for each stimulus was 130 Hz for the first 100 ms and then fell linearly to 105 Hz at the end of the stimulus. The amplitude of voicing remained constant at between 55 and 65 dB, depending on the token, until the last 35 ms of the stimulus, where it fell to 0 in a piecewise linear fashion. The length of each stimulus, as well as the formant frequency values and formant transition durations, varied according to the natural speech measurements. The amplitude and bandwidth values were then adjusted until the spectral properties at the burst onset and onset of voicing were consistent with the metric described in experiment I. The top panel of Figs. 8 and 9 show the spectral properties of the prototype [bo] and [do] stimuli.

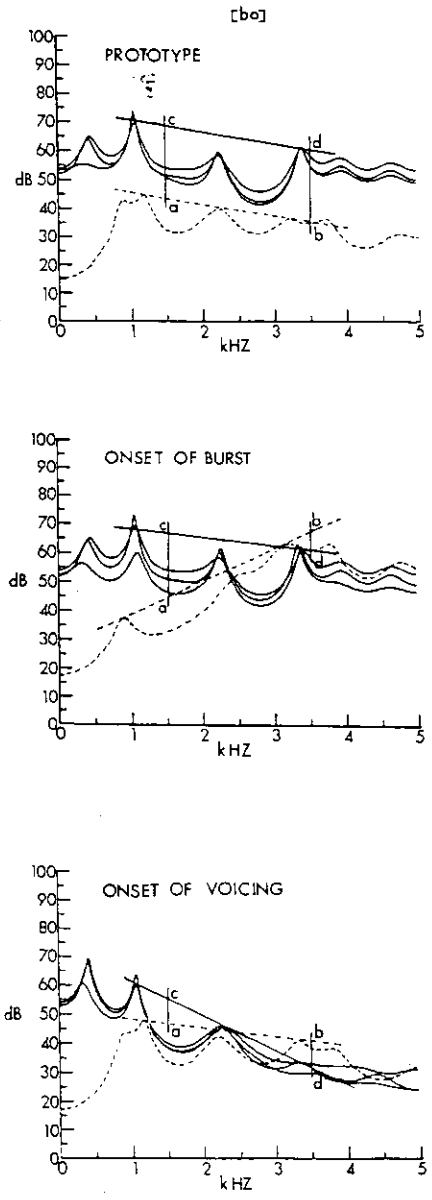

FIG. 8. Spectra for initial burst and first three glottal pulses of synthesized [bo] prototype and stimuli with the onset of burst and onset of voicing spectral manipulations. The top panel shows the spectrum of the prototype stimulus. In the middle panel, only the onset of burst spectrum has been manipulated, and in the bottom panel, the three glottal pulses at the onset of voicing have been manipulated. The dotted line represents the burst spectrum, and the solid line represents the spectra for the first three glottal pulses of the onset of voicing.

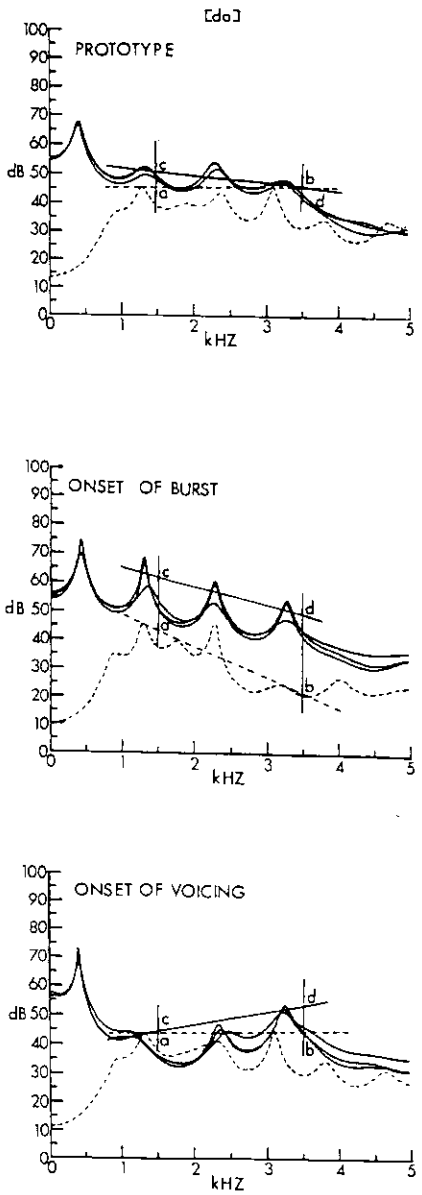

FIG. 9. Spectra for initial burst and first three glottal pulses of synthesized [do] prototype and stimuli with the onset of burst and onset of voicing manipulations. The top panel shows the spectrum of the prototype stimulus. In the middle panel, the onset of burst spectrum has been manipulated, and in the bottom panel, the three glottal pulses at the onset of voicing have been manipulated. The dotted line represents the burst spectrum, and the solid line represents the spectra for the first three glottal pulses of the onset of voicing.

In the [bo] stimulus, the distribution of energy remains unchanged from the burst to the onset of voicing, a pattern typical of labial stop consonants. In the [do] stimulus, there is a smaller change in energy in the high frequencies relative to the low frequencies, a pattern typical of dental or alveolar stop consonants. The metric values for these two prototype stimuli can be found in Table II. Recall that a metric value of <0.5, or one with a negative numerator, indicates a dental consonant, while a value of >0.5, or one with a negative denominator, indicates a labial consonant.

Once the prototype stimuli were constructed, we then synthesized two new sets of stimuli. Using the same parameters as the prototype stimuli, we first manipulated the spectral shape of the burst onset leaving the rest of the stimulus intact, such that the change of energy from the burst to the onset of voicing would be that of the alternative category. For example, for a labial prototype, we increased the amplitude of the burst at higher frequencies so that the stimulus as a whole had the relative energy distribution of a dental consonant. For the second set, the burst was left untouched while we manipulated the spectral characteristics at the onset of voicing by changing the amplitude and bandwidth of the second and higher formants. The gain of the synthesized stimuli was then adjusted to insure that there was no peak clipping. Figures 8 and 9 illustrate the results of the spectral manipulations for these two types of stimuli. Table II shows the metric values for each stimulus.

TABLE II. Metric values for synthesized stimuli. P indicates prototype; O indicates stimulus manipulated at the onset of burst; and V indicates stimulus manipulated at the onset of voicing.

		Vowel context				
[b]		[i]	[e]	[a]	[o]	[u]
	P	1.39	1.58	0.58	1.0	0.61
	O	0.21	0.13	−0.25[a]	−0.28[a]	−0.42[a]
	V	−0.08[a]	−0.61[a]	−1.33[a]	−1.0[a]	−1.0[a]
[d]						
	P	−0.47[a]	0.11	−0.5[a]	0.33	0.42
	O	0.68	1.23	17.0	1.52	1.05
	V	2.87	−19.0[b]	−1.77[b]	4.0	4.0

[a] Indicates value with a negative numerator.
[b] Indicates value with a negative denominator.

While, in principle, the methodology for creating the stimuli was fairly straightforward, in fact it was not. In making the adjustments, it became clear that because there was overlap in the skirts of the filters, a change in the amplitude of one formant peak often resulted in a change of spectral shape for other formant peaks. In addition, changes in amplitude parameter values of 1 dB, particularly of $F4$, often produced changes in the amplitude of the synthesized stimulus (as measured by LPC) corresponding to 6 dB or more. This is evident in reviewing the synthesis of the prototype [do] stimulus (cf. the top panel of Fig. 9). The spectrum of the burst is represented by the dotted line. The intent in modifying the spectrum at the onset of voicing was to raise $A2$ (the amplitude of $F2$) and lower $A4$ (the amplitude of $F4$) relative to the energy distribution of the burst. However, raising only $A2$ above a certain value resulted in the elevation of the *entire* onset of voicing spectrum, and lowering $A4$ resulted in the disappearance of the fourth formant peak in the onset of voicing spectrum. As a result, it required, in the end, adjustment of $A3$ and the bandwidth of $F4$ to effectively lower $A4$, and yet preserve some semblance of an $F4$ peak. In applying the metric to those spectra for which $F4$ was very weak, or failed to appear in spite of our efforts, we took the amplitude value at the frequency point corresponding to the synthesized parameter of $F4$. The actual loss of a high-frequency peak occurred in only three of the 30 stimuli.

In all, 30 CV stimuli were synthesized. (The actual parameter values used in the synthesis of the stimuli are available upon request.) There were ten prototype tokens, five of which were dental stops in the environment of [i e a o u], and five of which were labial stops in the same vowel environments. The 20 permuted tokens each had the formant frequency and duration characteristics of one place of articulation and spectral shapes characteristic of the opposing place of articulation.

Two test tapes were constructed, the first containing only the ten prototype stimuli, and the second the 20 manipulated stimuli. At the beginning of each tape, one occurrence of each of the tokens for that tape was presented in order to familiarize the subjects with the test stimuli. The rest of the tape consisted of a randomized sequence of ten occurrences of each of the stimuli. There was a 4-s delay between tokens and an 8-s delay between each ten tokens.

2. Subjects and procedure

As indicated above, the synthesis parameter values were based on the analysis of natural speech tokens of a French speaker. As a result, the values used for the [d] stimuli were appropriate to dental stop consonants, and, therefore, we chose subjects who would be familiar with the dental place of articulation. There was a total of 30 subjects, all native speakers of English, but with moderate to advanced knowledge of French or Russian. Twenty-eight of the subjects lived at the French House at Brown University, and their facility in French ranged from moderately good to bilingual. The remaining two subjects had moderate to excellent facility in Russian. All subjects were paid for their participation.

The tape containing the prototype stimuli was always presented first, then the tape containing the manipulated stimuli was presented. The subjects were told that for each syllable they heard, to write the letter *b* if they heard a syllable beginning with a [b] or a [p], and the letter *d* if they heard a syllable beginning with [d] or [t]. They heard the test syllables through headphones, and were tested in groups of one, two, or three in one session.

B. Results

To ensure that the subjects could reliably identify the prototype stimuli, we required that subjects correctly identify the prototype stimuli with a minimum score of 70% for *each* of three of the five vowel contexts in both the [b] and [d] phonetic categories. Eighteen subjects met this criterion. For the 12 subjects who failed, overall performance level was 68% with a range of 52% to 84%.[3]

Figure 10 shows the results. Figure 10(a) displays the percentage of [b] responses for stimuli containing [b] formant frequencies and either the appropriate spectral change from the onset and into the first three glottal pulses of the vowel [prototype (P)], or the modified spectrum appropriate for dental stops, achieved by changing the onset of the burst (O) or the onset of voicing (V). Figure 10(b) shows the percentage of [d] responses for stimuli containing [d] formant frequencies in the three stimulus conditions (P,O,V). As the figure shows, the perception of place of articulation shifted according to the predictions of the metric, regardless of the formant frequencies. This was particularly striking for the [b] stimuli, although less so for the [d] stimuli. With the exception of only one manipulation the onset for [da], all manipulated pairs showed a significant difference (by t tests) in perception compared to the prototype stimuli.

Nevertheless, although significant changes in perception occurred as a result of the spectral manipulations, these changes could have reflected an increase in stimulus ambiguity, rather than a change in perception of phonetic categories for place of articulation. Consequently, we attempted to apply more stringent criteria to evaluate whether reliable perceptual changes had in fact occurred. To this end, we operationally defined a reliable shift in phonetic perception as a change in performance level from at least 70% identification in the prototype condition to at least 70% identification in the *alternate* phonetic category in the manipulated conditions. We reanalyzed the data according to these criteria.

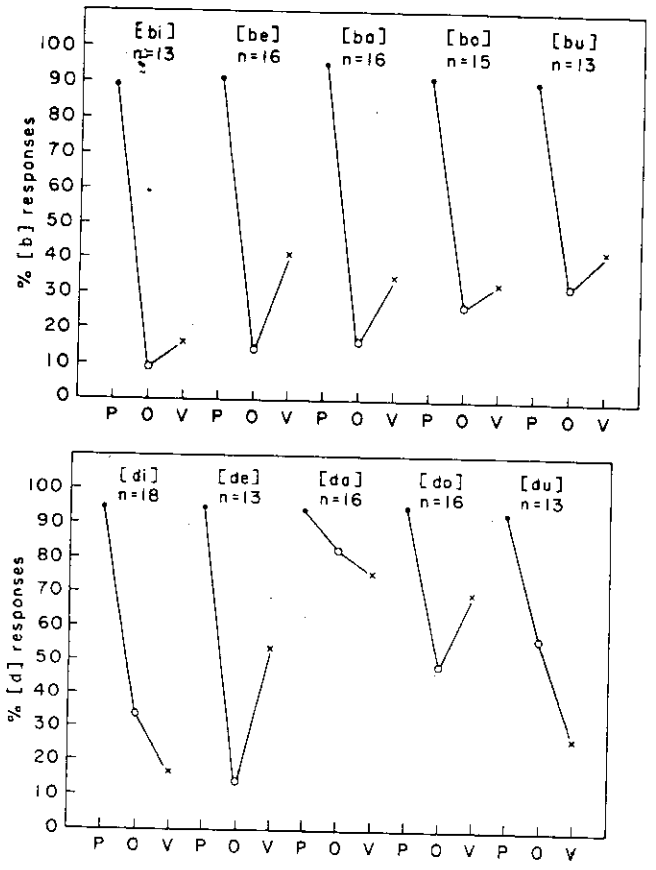

FIG. 10. (a) Percentage of [b] responses for stimuli containing [b] frequencies and either the spectrum with the appropriate relation between the onset and first three glottal pulses of the vowel for labial stops [prototype (P)], or the spectrum with the appropriate relation for dental stops, achieved by shifting the spectrum at the onset of the burst (O) or at the onset of voicing (V). (b) Percentage of [d] responses for stimuli containing [d] frequencies and either the spectrum with the appropriate relation between the onset and first three glottal pulses of the vowel for dental stops (P), or the spectrum with the appropriate relation for labial stops, achieved by shifting the spectrum at the onset of burst (O) or onset of voicing (V). See text for elaboration.

Results showed that for the [b] stimuli [Fig. 10(a)], reliable perceptual shifts to [d] occurred in the context of the vowels [i e a o] when the burst onset was manipulated, and they occurred in the context of the vowel [i] when the onset of voicing was manipulated. While not as consistent, phonetic category shifts were obtained with the [d] stimuli as well [Fig. 10(b)]. Reliable shifts to the [b] phonetic category occurred for [di] with both types of manipulations, for [de] with the manipulation of the burst, and for [du] with the manipulation of the onset of voicing. Subjects did not shift for the [da] manipulations, and they did minimally so for [do].

Comparing whether the burst onset manipulations or the onset of voicing manipulations had a bigger effect, we found that, although the [b] stimuli showed slightly larger shifts in perception when the burst spectrum was changed than when the onset of voicing spectrum was changed, this was only significant for the [bi] and [be] tokens ($t = 2.52$ and 3.10, respectively, $p < 0.05$). The [d] stimuli showed no systematic trends. These results reaffirm that it is not an invariant specifically in either the burst or the onset of voicing spectrum which is perceptually salient, but rather it is the dynamic change from the one to the other which provides the invariant acoustic property for place of articulation.

C. Discussion

The results of this study show that the relative change in spectral shape was able to override both formant frequency and transition motion cues to place of articulation. That it is the relative change in distribution of energy and not the gross shape of the onset spectrum alone or the gross shape of the onset spectrum relative to what follows is supported by the fact that manipulating *either* the burst *or* the onset of voicing produced perceptual shifts in the phonetic category.

Ohde and Stevens (1983) did achieve shifts in the perception of place of articulation in labial and alveolar stop consonants by raising or lowering the *entire* burst spectrum without altering its gross shape; however, the shifts in our study were elicited by changes in the *relative distribution* of energy across the frequency spectrum and across time, either in the burst or in the onset of voicing. In addition, the shifts in the Ohde and Stevens study occurred only at the phonetic boundary; the endpoints of their labial-alveolar continuum were little affected by the movement of the entire burst spectrum. This seems to indicate that changes in the amplitude of the gross shape of the spectrum are not sufficient to direct perception from one phonetic category to the other, at least for the end-point stimuli, although they have perceptual consequences for boundary value stimuli.

The results of experiment II contrast with those obtained by Blumstein *et al.* (1982) and Walley and Carrell (1983), who found that the spectral tilt was not sufficient to change the perceived place of articulation in synthetic stimuli containing formant frequencies appropriate for the alternate place of articulation. An explanation for this difference lies in the fact that, in both studies, it was only the spectral tilt at onset that was manipulated, not the distribution of spectral energy at onset relative to what followed. Our results provide support for the view that it is the dynamic spectral change from burst release into the beginning portions of the transitions which is important for determining the perception of place of articulation for stop consonants.

The fact that phonetic category shifts occurred in all vowel contexts for the [b] stimuli suggests that the perceptual effects of the invariant property for place of articulation are context-independent. The results for the dental consonants, however, are less clear-cut. The perceptual shifts were greater in the case of the front vowels than back vowels, although a reliable category shift was obtained for [du] when the onset of voicing was manipulated. It has been suggested that the perceptual saliency of spectral and transition cues for place of articulation will vary depending upon the vowel context (Fischer-Jørgensen, 1972; Dorman *et al.*, 1977). In particular, spectral properties of the burst may be particulary salient in the environment of front vowels where transition motions are minimal, whereas they may play a lesser role in the environment of back vowels where formant motions are greater. Formant transitions were fairly flat in nearly all vowel contexts for the [b] stimuli, and in the context of [i] and [e] in the [d] stimuli. The largest formant transition motions occurred in the [da] stimulus, the stimulus that showed negligible perceptual shifts. It could be that, in this case, the changes in the distribution of spectral properties from the burst release to the onset of voicing could not over-

ride the context-dependent formant frequencies and transition motions. Nevertheless, we are hesitant to attribute the failure of some of these stimuli to show the expected perceptual shifts to vowel context effects. Rather, we believe that the difficulty in demonstrating perceptual shifts was, at least in part, a function of the synthesized stimuli used, for, as reported in Sec. II A, we had particular difficulty in moving the formant amplitudes of the [d] stimuli.

III. GENERAL DISCUSSION

In this study, we have explored two claims of a theory of acoustic invariance in speech. The first claim is that there is acoustic invariance in the speech signal corresponding to the phonetic features of language. To test this hypothesis, we investigated the degree to which a particular invariant property would generalize to other languages sharing the same phonetic feature. Results indicated that indeed such invariance could be derived for diffuse stop consonants in Malayalam, French, and English, and could account for over 91% of the stop consonants analyzed. The second claim of a theory of acoustic invariance is that the perceptual system is sensitive to these invariant properties. Thus we investigated whether listeners perceived place of articulation in diffuse stop consonants according to the metric derived in the production study. Results showed that listeners were sensitive to the invariant properties in making place of articulation categorizations, even in the presence of formant frequency and transition cues for the alternative place of articulation category. Thus, although the invariant properties did not override the context-dependent cues for place of articulation 100% of the time, the listener, nonetheless, could and did make use of the invariant properties to make a phonetic decision concerning place of articulation. We take these results to provide strong support for a theory of acoustic invariance in speech.

Nevertheless, it is worth considering the fact that in both the production and perception study, 100% of the data could not be accounted for. Does a theory of acoustic invariance make such a requirement? We think not. On the one hand, it may be that invariant properties are present in the signal all the time, and the listener uses these properties all the time. In that case, it is possible that we have not focused on the optimal form of the metric, nor have we synthesized stimuli that contain all of the necessary characteristics of the invariant properties. Nevertheless, the fact that 91% of the production data could be accounted for, and listeners showed reliable perceptual changes, indicates that, at the very least, the speech signal does contain *context-independent* information corresponding to at least one phonetic feature of natural language in both production and perception. On the other hand, it may be the case that invariant properties are *not* present in the signal all the time. After all, many variables contribute to the speech production process, and these properties may not appear under certain conditions. This possibility does not seem to us to argue against the view that, in general, there are stable acoustic patterns in the speech signal corresponding to phonetic features, and these patterns can occur *independent* of local contextual cues. If one of the functions of the invariant properties is to provide the listener with the phonetic framework of natural language, then the presence of these invariances at all times would theoretically not be necessary. In fact, in perception, the listener most likely makes use of all acoustic information, both context-dependent and context-independent, in making phonetic decisions. It would be highly unlikely for the speech system to have evolved using only *one* of many cues present in the signal contributing to the phonetic percept.

The form of invariance which we have characterized for place of articulation in diffuse stop consonants is dynamic relative invariance. The properties are *dynamic* in the sense that they are determined by comparing the spectral properties of the signal across the time domain (cf. also Searle *et al.*, 1979, 1980; Kewley-Port, 1983). They are *relative* in the sense that the invariant patterns are derived on the basis of relative changes in the spectral characteristics of the signal in regions of high information (cf. also Stevens, 1975; Blumstein and Stevens, 1981; Stevens and Blumstein, 1981; Ohde and Stevens, 1983). The invariant properties corresponding to place of articulation are based on the relative changes in the distribution of energy at high and low frequencies from the release of the stop consonant to the beginning of the voiced formant transitions.

While the metric that we have developed to capture these properties accounted for most of the natural speech utterances, we are not claiming that this is the only possible way to characterize the invariant properties for place of articulation, *nor* are we claiming that the perceptual system must compute ratios to determine place of articulation dimensions. Although we do not know how such invariances as we have described are realized by the perceptual system, there is some recent research by Goldhor (1983) which suggests that the properties of the peripheral auditory system are such that they may transform the acoustic signal into complex combinations of acoustic properties similar to those described here for place of articulation. In particular, Goldhor has developed a model of the peripheral auditory system which transforms the acoustic signal in terms of frequency, amplitude, and temporal dimensions. The temporal dimension models adaptation of the auditory neurons to sustained energy in a particular frequency band. Goldhor has shown distinctive response patterns of the model to the first 40 ms of labial and alveolar stop consonants. For labial consonants, there is more energy at low frequencies, as evidenced by a greater saturation of channels in the lower-frequency bands over the first 40 ms of the stimulus, whereas for alveolars, there is more energy in the high frequencies as evidenced by a greater saturation of channels in the higher-frequency bands. Thus, using a model of the peripheral auditory system, Goldhor has shown that simple properties of auditory response patterns produce complex acoustic properties—acoustic properties which correspond in a fairly direct way to those we have shown for place of articulation in diffuse stop consonants.

ACKNOWLEDGMENTS

Many thanks to Kathleen Kurowski for her help in the synthesis of the stimuli and to Mark Naigles and Kenneth Stevens for their comments on an earlier draft of this paper.

Part of this research was conducted while S. Blumstein was a Visiting Luce Professor at Wellesley College. Their support is gratefully acknowledged. This research was supported in part by Grant NS15123.

[1]Labial and dental stops contrast in a number of other syllabic positions. However, alveolar stop consonants only occur as intervocalic geminates (i.e., long consonants).

[2]The LPC analysis calculates the first difference of the waveform, in effect pre-emphasizing the high frequencies at 6 dB per octave, multiplies the waveform by a Hamming window, and smooths the spectrum using a 14-pole linear prediction algorithm.

[3]It is problematic that 12 out of the 30 subjects were unable to reliably identify three of the five prototype stimuli in each phonetic category, especially since other experiments using synthetic speech (Stevens and Blumstein, 1978) had found quite good identification. The difference may have occurred because the values for our stimuli were taken from natural speech. In comparing the formant frequency values of our stimuli with those of Stevens and Blumstein, large discrepancies were noted particularly for the labial consonants. In the Stevens and Blumstein stimuli, the first three formants had large formant frequency excursions, ranging from 150–520 Hz. In the stimuli for our study, few such large frequency excursions were observed. In fact, in comparing the first three formants across the vowel contexts [a i u], only $F1$ and $F2$ of [a] and $F3$ of [i] and [u] fell within this range in the natural speech stimuli. This suggests that, in the case of synthetic stimuli, larger frequency excursions are perceptually important for perceiving the labial place of articulation. However, given the fact that we wanted to follow the natural formant frequencies, we had to compromise the better perceptual quality of stimuli containing sharp frequency transitions for those stimuli which followed the natural values. Of the 12 subjects who were eliminated from experiment II, nine were unable to perceive labials in a reliable manner.

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.

Blumstein, S. E., and Stevens, K. N. (**1981**). "Phonetic features and acoustic invariance in speech," Cognition **10**, 25–32.

Blumstein, S. E., Isaacs, E., and Mertus, J. (**1982**). "The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **72**, 43–50.

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

Cole, R., and Scott, B. (**1974a**). "The phantom in the phoneme: invariant cues for stop consonants," Percept. Psychophys. **15**, 101–107.

Cole, R. A., and Scott, B. (**1974b**). "Towards a theory of speech perception," Psychol. Rev. **81**, 348–374.

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. (**1956**). "On the predictability of formant levels and spectrum envelope from formant frequencies," in *For Roman Jakobson*, edited by M. Halle (Mouton, The Hague).

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

Fischer-Jørgensen, E. (**1972**). "Perceptual studies of Danish stop consonants," Ann. Rep. Inst. Phonet. Univ. Copenhagen **6**, 75–168.

Goldhor, R. (**1983**). "The representation of speech signals in a model of the peripheral auditory system," J. Acoust. Soc. Am. Suppl. 1 **73**, S4.

Halle, M., and Stevens, K. N. (**1979**). "Some reflections on the theoretical bases of phonetics," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Ohman (Academic, London).

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

Jongman, A., Blumstein, S. E., and Lahiri, A. (**1984**). "Invariant acoustic properties distinguishing alveolar and dental stop consonants," in preparation.

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

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

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Lahiri, A. (**1980**). "Coronal stops in Malayalam," Brown University Working Papers **4**, 81–95.

Lahiri, A., and Blumstein, S. E. (**1981**). "A reconsideration of acoustic invariance in stop consonants: Evidence from cross-language studies," J. Acoust. Soc. Am. Suppl. 1 **70**, S39.

Mertus, J. (**1979**). "The waveform editing program and manual," unpublished manuscript.

Ohde, R. N., and Stevens, K. N. (**1983**). "Effect of burst amplitude on the perception of place of articulation for stops," J. Acoust. Soc. Am. **74**, 706–714.

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 patterns in the 3-space of time and frequency," in *Perception and Production of Fluent Speech*, edited by R. A. Cole (Erlbaum, Hillsdale).

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).

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. L. Miller (Erlbaum, Hillsdale).

Walley, A. C., and Carrell, T. D. (**1983**). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 1011–1022.

Zue, V. (**1976**). "Acoustic characteristics of stop consonants: A controlled study," Sc. D. thesis, MIT (unpublished).