Perceptual invariance and onset spectra for stop consonants in different vowel environments

Sheila E. Blumstein
Department of Linguistics, Brown University, Providence, Rhode Island 02912

Kenneth N. Stevens
Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 3 July 1979; accepted for publication 12 October 1979)

A series of listening tests with brief synthetic consonant-vowel syllables was carried out to determine whether the initial part of a syllable can provide cues to place of articulation for voiced stop consonants independent of the remainder of the syllable. The data show that stimuli as short as 10-20 ms sampled from the onset of a consonant-vowel syllable, can be reliably identified for consonantal place of articulation, whether the second and higher formants contain moving or straight transitions and whether or not an initial burst is present. In most instances, these brief stimuli also contain sufficient information for vowel identification. Stimulus continua in which formant transitions ranged from values appropriate to [b], [d], [g] in various vowel environments, and in which stimulus durations were 20 and 46 ms, yielded categorical labeling functions with a few exceptions. These results are consistent with a theory of speech perception in which consonantal place of articulation is cued by invariant properties derived from the spectrum sampled in a 10-20 ms time window adjacent to consonantal onset or offset.

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

INTRODUCTION

Theories of speech perception developed in the past 20 years have been based on the notion that there is a lack of one-to-one correspondence between attributes of the acoustic signal and the phonetic percept. This lack of invariance in the speech signal has resulted in a theory of speech perception that hypothesizes an active perceptual system in which the acoustic signal is interpreted at some higher level in terms of abstract features, possibly based on the speech-production system (Liberman et al., 1967; Studdert-Kennedy et al., 1970; Pisoni, 1978). Evidence for this theoretical viewpoint is based upon a series of studies investigating the role of individual components of the speech signal as cues to the place-of-articulation phonetic dimension. The results of these studies indicate that neither the spectrum of the initial burst nor the onset frequencies or directions of formant transitions provide invariant cues to place of articulation in stop consonants (Cooper, et al., 1952; Liberman et al., 1954; Schatz, 1954; Dorman et al., 1977). Rather, these dimensions seem to be context dependent, determined primarily by the following vowel. Further studies have investigated the "trading" relation between bursts and formant transitions in providing cues to stop-consonant perception and have found again that the perceptual importance of these components varies as a function of the following vowel (Dorman et al., 1977; Fischer-Jorgensen, 1972). For example, transitions seem to play a secondary role to the burst in the context of [i] and [u], whereas they play a primary role relative to the burst in the context of [a].

All of this research has focused upon individual components of the speech signal, and has assumed, in particular, that bursts and transitions provide, at some level, independent cues to place of articulation. Implicit in this approach is the notion that the perceptual mechanism operates on the speech signal in much the same way as the eye examines attributes of the sound spectrogram. The conventional spectrogram, for example, has a different appearance for bursts, which are produced by noise or by transient excitation of the vocal tract, and for transitions toward the vowel, which are often produced by voiced excitation of the vocal tract, and extend over a time interval of a few tens of milliseconds.

An alternative approach concerning cues for consonantal place of articulation views the speech signal in terms of its gross short-time spectral characteristics, independent of whether the spectra are produced by a noise burst, vowel formant transitions, or both. The work of Fant (1960), Jakobson et al. (1963), and Halle et al. (1957) in particular, has suggested that spectral analysis of the speech signal results in distinctively different patterns corresponding to individual phonetic features. For example, the results of the Halle et al. study revealed that short-time spectra of bursts in isolation showed three classes of patterns associated with the three places of articulation—labial, alveolar, and velar. This finding was further documented in a more extensive study by Zue (1976).

The notion that spectral analyses reveal distinct patterns for individual phonetic features suggests that there may indeed be an invariance of acoustic patterns which can characterize place of articulation independent of the following vowel. This concept of invariance has been discussed by Cole and Scott (1974), who proposed a model of speech perception "in which invariant and transitional cues are integrated in the direct perception of whole syllable units..." (p. 348).

The concept of invariance has been investigated recently in a series of studies using both synthetic and real speech. In the first study (Stevens and Blumstein, 1978) a series of synthetic continua representing the phonetic categories [b d g] in the context of the vowels [a i u] was analyzed in terms of the spectral character-
istics of the stimuli representing each phonetic category. Results indicated that the spectrum obtained by sampling the first 20-odd ms at consonantal release seems to exhibit invariant properties for each place of articulation for stop consonants independent of vowel context. This invariance can be characterized in terms of a diffuse-rising spectrum for [d] (i.e., the spectral peaks are distributed throughout the frequency range, and the amplitudes of the higher-frequency peaks are greater than those of the lower-frequency peaks), a diffuse-falling or diffuse-flat spectrum for [b], and a compact mid-frequency spectral energy peak for [g].

In a second study, (Blumstein and Stevens, 1979), we attempted to provide a more quantitative measure of the degree to which invariant patterns were derivable for spoken voiced and voiceless stop consonants across vowel contexts. To this end, we developed a set of templates characterizing a diffuse-falling or diffuse-flat spectral pattern, a diffuse-rising pattern, and a compact pattern, and determined the extent to which the onset spectra of natural CV utterances consisting of [b d g] and [p t k] in the environment of five different vowels could be correctly classified by those templates. Over 80% of the utterances were correctly classified according to this measure. Thus, the spectrum at onset does seem to produce an invariant gross shape for place of articulation independent of the vowel context. Both the burst and the initial part of the transition contribute to this invariance and together form a single integrated acoustic property (Stevens and Blumstein, 1978; Stevens, 1975). In this view, the onset spectra provide the primary context-independent cues to place of articulation. The formant transitions seem to provide secondary context-dependent cues, and form the acoustic material that links the abrupt onset to the steady-state vowel (Cole and Scott, 1974; Stevens and Blumstein, 1978).

Nevertheless, although such acoustic invariance may be derivable from the acoustic signal, it is not clear whether these onset spectra alone, independent of the transitions and steady-state vowel information, can provide sufficient information for the perception of place of articulation. Recent work (Winitz et al., 1972; LaRiviere et al., 1975) suggests that this may indeed be the case. In these investigations of initial stop consonants, it was shown that presentation of the burst alone or the aperiodic portion of the consonant in natural speech utterances, results in good identification of the appropriate place of articulation as well as of the phonetic quality of the following vowel (Ohde and Sharf, 1977).

In this study, we have attempted to determine whether just the onset of a synthetic consonant–vowel syllable can provide cues to the perception of place of articulation for voiced stop consonants. To this end, we synthesized a series of stimuli which systematically varied along specific parameters, and we presented these stimuli to listeners for judgments of consonantal place of articulation and vowel quality. We were particularly interested in the following questions: (1) Do the brief onset portions of consonant–vowel syllables provide sufficient perceptual cues for accurate place-of-articulation identification, and if so, what is the minimal size of the window required? To this end, we synthesized stimuli containing a 5- or 10-ms noise burst followed by a brief voiced interval containing formant transitions with onset characteristics appropriate to the consonants [b d g] followed by the vowels [ai u] (experiment 1). By varying the voicing interval from 45 ms down to 10 ms, we could determine the perceptual effects of varying degrees of onset information on consonant identification. (2) To what extent are listener responses influenced by the presence of formant motions as well as by stimulus characteristics at onset? In order to answer this question, we generated a second set of stimuli containing a noise burst and formant onset frequencies equivalent to the consonants [b d g] in the environment of [ai u] but without moving transitions (experiment 1). (3) To what extent are integrated properties consisting of the noise burst in combination with the onsets of the formant transitions relevant to the perception of the place-of-articulation dimension? To examine this question, we generated additional sets of stimuli equivalent to the first two series but containing no noise bursts (experiment 2). (4) Will a continuum of stimuli consisting only of brief onsets ranging from those appropriate for [b] to [d] to [g] be perceived in terms of discrete phonetic categories as is found with synthetic CV syllables? To this end, four sets of [b d g] continua were generated in the context of each of the vowels [ai u] containing brief voicing intervals of 46 and 20 ms duration with moving and straight transitions (experiment 3). (5) Do brief stimuli consisting of 10-46 ms of voicing carry phonetic information other than consonantal place of articulation? In particular, are cues to the following vowel present in these short stimuli? This question was addressed by obtaining vowel as well as consonant responses to the brief synthetic stimuli (experiment 4).

I. EXPERIMENT I

A. Stimuli

Stimuli in all the experiments were generated using a computer simulation of a terminal analog speech synthesizer in which the tuned circuits for the vowel-generating portion were connected in cascade (Klatt, 1972). The sampling rate for the synthesizer output was 10 kHz and the output was low-pass filtered with a cutoff frequency of about 4800 Hz. Two sets of stimuli were generated for experiment 1: a set containing a noise burst followed by a voiced portion of various lengths with formant transitions, and another set in which the formant motions were removed from the basic set.

The basic stimulus set consisted of a 5- or 10-ms noise burst followed by initial piecewise-linear transitions in which the onset frequencies of the first four formants were appropriate to the consonants [b d g] and transitions were moving toward frequencies appropriate to the vowels [ai u]. The durations of the transitions of the second and higher formants of the basic set of stimuli were 40 ms. The burst location and the trajectories of the synthesizer parameters for the formants for a typical stimulus are shown in Fig. 1. The
pulses. The voicing duration is defined to be 20 ms, although of the initial noise burst. The vertical lines indicate the effective termination times for the four voicing durations used in the experiment (shown at the top). The dashed lines at the right show how the formant trajectories would continue for a full consonant-vowel syllable. This particular stimulus is produced with two glottal pulses. The voicing duration is defined to be 20 ms, although the final portion of the decay following the last glottal pulse continues beyond 20 ms, as the time scale indicates.

Formant parameter values used in this study were chosen from a set of stimuli generated in an earlier experiment in which subjects were required to label a series of acoustic continua varying along the place-of-articulation dimension. The particular stimuli used in the present study are equivalent to the best exemplars of each of the phonetic categories [b d g] in the environment of the vowels [a i u] (Stevens and Blumstein, 1978). The formant frequencies appropriate for the vowels [a i u] are given in Table I. The starting frequencies of the transitions of formants one to four at stimulus onset for each of the consonants [b d g] in the context of the vowels [a i u] are also given in the table. In the case of the consonants in the environments of the vowels [a] and [u], the F4 transitions probably play no significant role in perception and thus were not varied. The starting frequency, duration, and slope of the F1 transitions differed slightly across stimuli, with the shortest F1 durations occurring for the labial consonants and the longest for the velar consonants.

The fundamental-frequency contour for all stimuli started at 103 Hz and rose in a piecewise-linear manner to 125 Hz after 30 ms of voicing (or to a lower interpolated frequency if the voicing duration was less than 30 ms). In all stimuli in the basic set, a brief noise burst was inserted prior to the onset of voicing. The noise burst was produced by passing a gated 5- or 10-ms burst of white noise through a simple-tuned circuit. The [b] burst for all stimuli was inserted 5 ms prior to the onset of voicing and was produced by exciting a resonator centered at the frequency of F2 at voicing onset. The [d] burst was inserted 10 ms prior to the onset of voicing and excited F4 in the environment of [a], and F4 and F5 in the environment of [i] and [u]. The [g] burst was inserted 15 ms prior to voicing onset and excited F2 in the environment of [a] and [u] and F3 in the environment of [i]. In all cases, the values of the formant frequencies for the resonators excited by the noise burst were equal to the frequencies of the formants at the onset of voicing, as given in Table I. The duration of the burst was 5 ms for the [a] and [u] stimuli and 10 ms for the [i] stimuli. The amplitude of the burst was adjusted to be 0–5 dB less than the amplitude of the adjacent formant in the following vowel (Bush, 1977), and the same noise burst was used as the excitation for each of the stimuli generated.

These nine stimuli consisting of [b d g] in the environment of [a i u] formed the basic stimulus set from which the remaining stimulus series was derived. For each of the stimuli, the duration of the glottal excitation of the synthesizer was systematically varied to generate one, two, three, and five glottal pulses, producing a total of 36 stimuli—four different durations for each of the consonants [b d g] in the environment of [a i u]. A typical waveform for this set of stimuli is shown at the bottom of Fig. 1. If we define the effective duration of the output from the last glottal pulse to be 10 ms (although, as Fig. 1 shows, the actual duration in which there is a significant amplitude is somewhat
light lines. Stimulus durations and burst characteristics are
lines. The trajectories for the moving-transition stimulus
the same as in Fig. 1. All of these stimuli contained moving transitions.
In order to assess the perceptual effects of onset information independent of formant motions, we generated a second set of stimuli containing the same burst frequency and onset characteristics of the previous set but no moving transitions. This type of stimulus is shown schematically in Fig. 2. In all of these stimuli, the frequencies of the second and higher formants were constant throughout and were equal to the formant frequencies at voicing onset for each of the consonants [b d g] in the environment of the vowels [a i u] (see Table I). The movement of F1 was retained in this series of stimuli in order to maintain the stop-like quality of the stimuli. The length of the voiced interval was varied as in the previous series from one to five glottal pulses, producing another 36 stimuli.

B. Subjects
Thirty-four students at Brown University served as subjects in this experiment. All were native English speakers and had no known hearing impairments.

C. Procedure
Subjects were divided into two groups of 17 each, one group receiving the stimuli containing moving transitions and the other receiving the stimuli containing straight transitions. As all 36 stimuli could not be
stored in the computer for presentation on one test tape, the stimuli were divided in half and two test tapes were constructed per test group. The stimuli on each tape were selected so that there would be, as nearly as possible, an equal occurrence of consonant and vowel classes across the different durations. Each test tape contained two random orders of five presentations per stimulus, each part preceded by two practice trials. In the actual test, then, there were ten presentations of each stimulus. The interstimulus interval was 3 s, and after each block of ten stimuli there was a 5-s silent interval. The order of presentation of the test tapes was counter-balanced across subjects. Each group received all of the test tapes in one session, which lasted approximately 1 h. Short breaks were given at the end of each part. Subjects were required to identify each stimulus as [b], [d], or [g] by writing the appropriate initial consonant on the answer sheet provided.

D. Analysis of results
The upper, middle, and lower panels of Fig. 3 show percent identification for the [b], [d], and [g] stimuli, respectively. Inspection of the graphs indicates that, with the exception of several [gl] stimuli, overall identification performance is well above chance (the 29-ms moving [g] stimulus was incorrectly generated by the computer and hence the data point for that stimulus is missing), and indeed these brief stimuli seem to provide minimal but sufficient cues for the identification of place of articulation across different vowel contexts. Nevertheless, there are clearly performance differences across these various stimuli.

In order to assess the various contributions of moving versus straight transitions and transition duration across the three vowel environments, three separate three-way analyses of variance were conducted, one for each place of articulation. The Newman-Keuls test was used for post-hoc comparisons.

Analyses for the [b] stimuli indicated significant main effects for transitions (F(1, 39) = 3.84, p < 0.06) due to better performance with moving transitions than straight transitions, and a significant main effect for vowels (F(2, 62) = 9.11, p < 0.001) due to better performance for [b] identification in the context of the vowels [a] and [i] in comparison to [u]. All other comparisons were nonsignificant. Thus, although stimuli with moving transitions were better identified than those with straight transitions, the duration of the transitions did not have a significant effect on identification performance.

Results for the [d] stimuli (middle panel of Fig. 3) indicated significant main effects for transitions (F(1, 23) = 8.54, p < 0.01) and duration (F(3, 66) = 3.22, p < 0.03), and significant interactions for transition X vowel (F(2, 44) = 4.52, p < 0.02), transition X duration (F(3, 66) = 2.85, p < 0.04), vowel X duration (F(3, 183) = 2.80, p < 0.01), and transition X vowel X duration (F(6, 183) = 3.08, p < 0.01). The significant effect for transitions was due to better performance on the moving than the straight transitions. However these results were conditioned by both duration and vowel. For [a] and [i] stimuli, the difference in performance between moving and straight transitions was
found with the longer duration stimuli only. Thus, for \([a]\) there was a significant difference between moving- and straight-transition stimuli at the 46- and 29-ms durations, and for \([i]\) at the 46-ms duration. For \([u]\), moving transitions were better perceived than straight at all durations. Effects for duration were attributable overall to better identification performance at 20 ms than at 46 ms. This was significant, however, only for the straight transitions: for the vowel \([a]\), durations of 10, 20, and 29 ms were better than 46; for \([i]\), 10 and 20 were better than 46; and for \([u]\), 20 was better than 10.

For the \([g]\) stimuli (lower panel of Fig. 3), analysis of variance revealed a significant main effect for vowels \((F_{(2, 64)} = 33.32, p < 0.001)\), duration \((F_{(3, 96)} = 23.26, p < 0.001)\), and interactions for transition \(X\) vowel \((F_{(2, 64)} = 10.28, p < 0.001)\), transition \(X\) duration \((F_{(3, 96)} = 13.96, p < 0.001)\), and transition \(X\) vowel \(X\) duration \((F_{(6, 192)} = 22.00, p < 0.001)\). The vowel effect was due to better performance with the \([u]\) and \([a]\) vowel stimuli than with the \([i]\) stimuli. Differences between moving and straight transitions were significant only for the \([i]\) vowel and at all transition lengths. Unlike the \([d]\) stimuli, the 46-ms transition durations were better perceived than the shorter durations. This difference occurred only for the vowel \([i]\), but was due to both the moving and the straight transitions, indicating that the formant movements were not responsible for the improved performance.

E. Discussion

The results of this experiment indicate that the brief onset portions of consonant-vowel stimuli provide sufficient information for perception of place of articulation across vowel contexts. Moving transitions enhance performance for all places of articulation, although transition movement interacts differently with durations across the three places of articulation.

For the labial consonants, the effects of duration on perception were small. The somewhat lower performance for the \([bu]\) stimuli can, perhaps, be attributed to the fact that the burst for the \([b]\) stimuli was not entirely natural. That is, the spectral energy in the burst was concentrated in a narrow frequency range (the vicinity of \(F2\)) rather than being distributed over a wider frequency range. In the case of \([bu]\), the second-formant starting frequency was rather low, giving rise to a burst in the range normally found for the consonant \([g]\) in the syllable \([gu]\).

In the case of alveolar consonants, performance for the straight-transition stimuli tended to decrease with increasing duration. This finding suggests that the subject may indeed be tracking movement of transitions when these transitions are sufficiently long, and that the longer durations of the straight-transition stimuli are providing conflicting or inappropriate cues to place of articulation. For these longer straight-transition stimuli, identification errors were usually substitutions of \([b]\) for \([d]\), i.e., the consonant was still perceived as diffuse.

Unlike the alveolar consonants, the velars show enhanced performance for the longer duration stimuli, except for those stimuli whose responses are already close to 100%. The effect, most marked for the \([gl]\) stimuli, was found for both moving and straight transitions, indicating that it is not formant movements which are responsible for improved performance at the longer durations. Rather, it appears that a longer time is necessary to generate an auditory representation that is interpreted unambiguously as a velar. It is as though the auditory representation of a spectrally compact, mid-frequency prominence does not exhibit the compactness property immediately at stimulus onset, and this property takes a longer time to build up or to develop. Thus the short-duration \([g]\) stimuli are sometimes interpreted as being diffuse rather than compact. These anomalous responses include a fairly equal distribution of \([b]\)'s and \([d]\)'s.
Nevertheless, the major finding of this experiment is that brief stimuli with a voicing interval as short as 10–29 ms (together with the initial burst) provide cues to place of articulation across vowel contexts. Further, and perhaps more importantly, although transition movement does facilitate performance, it is not a necessary condition for reliable place-of-articulation perception.

II. EXPERIMENT 2

The aim of experiment 2 was to determine if onset information provided by the transitions within the initial few glottal periods without the burst could also cue place of articulation. Previous acoustic analysis (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979) has shown that the presence of the burst does not change certain gross attributes of the overall spectral shape of synthetic stimuli containing only transitions, but rather serves to enhance the property intrinsic to the particular place of articulation. The perceptual consequence, then, of listening to stimuli containing various durations of voicing onset in transition-only stimuli might be that subjects are able to identify the appropriate place of articulation, but less accurately than in experiment 1, when burst information was also provided.

A. Stimuli

The stimuli used in experiment 2 corresponded almost exactly to those used in experiment 1, except that there was no burst of noise preceding the onset of voicing. Thus, there were two sets of stimuli, the first varying in duration from 10 to 46 ms with initial piecewise-linear transitions, for which the onset frequencies of the first four formants were appropriate to the consonants [b d g] and the offset frequencies of these transitions were appropriate to the vowels [a i u] (see Table I). Unlike experiment 1, stimuli of 90 ms voicing duration were also synthesized to serve as a basis for comparison of performance for the short stimuli with that for a consonant–vowel stimulus containing a brief steady vowel. For the second set of stimuli, the onset frequencies at the four durations were also appropriate to the consonants [b d g] in the environment of the vowels [a i u], but the frequencies of the second and higher formants were constant throughout.

B. Subjects

Eleven subjects at Brown University, all native speakers of English and with no known hearing impairments, served as subjects.

C. Procedure

The test tapes were similar in organization to those used in experiment 1. The moving- and straight-transition stimuli were presented as two separate subtests, and within each subtest the stimuli were divided in half and two test tapes were constructed. There were 45 different stimuli for the moving-transition subtest: the three consonants [b d g] in the context of the vowels [a i u] at five durations of voicing (10, 20, 29, 46, and 90 ms). For the straight-transition subtest, there were 36 different stimuli: the three consonants in the context of the three vowels, but at only four durations ranging from 10 to 46 ms. Each subtest contained two random orders of five presentations per stimulus, each part preceded by two practice trials. There was a 3-s interstimulus interval, and after each block of ten stimuli there was a 5-s silent interval. Presentation of the tests was counterbalanced across subjects. All subjects listened to both tapes, and were required to identify each stimulus as [b], [d], or [g].

D. Analysis of results

Results for consonant identification are shown in Fig. 4. In order to determine whether performance level for the no-burst shortened stimuli was equivalent to that for full-syllable CV transition-only stimuli, a correlated t-test was conducted comparing overall

---

FIG. 4. Percent correct identification for the consonants in stimuli without bursts with four different durations (experiment 2). See legend for Fig. 3.

S. E. Blumstein and K. N. Stevens: Perceptual invariance 653
percent identification for the 90 and 46 ms moving-transition stimuli. Results showed a nonsignificant difference ($r = 0.27$), indicating that the performance level for the 46-ms stimuli with negligible steady-state vowel information was equivalent to that for CV stimuli containing steady-state information.

Inspection of overall performance level on the shortened stimuli indicates that, in general, subjects are able to identify place of articulation for the brief stimuli even without an initial burst. There are several exceptions, however. Performance for straight-transition [bi] stimuli was below chance, as was performance for the 10- and 20-ms moving-transition [gi] stimuli. The [bi] stimuli tended to be identified as [d], whereas both [d] and [b] responses were made to the [gi] stimuli. A comparison of overall performance level on the stimuli without bursts (Fig. 4) and those in experiment 1 with bursts (Fig. 3) indicates that, with the exception of the [bu] stimuli, performance level is relatively higher for all stimuli containing bursts than for comparable stimuli without bursts. Nevertheless, onset information provided by transitions only is usually sufficient for correct place-of-articulation identification.

The effects of moving versus straight transitions and of stimulus duration was assessed by conducting three separate three-way repeated measures (transition $X$ duration $X$ vowel) analyses of variance, one for each place of articulation. The Newman–Keuls test was used for all post-hoc comparisons.

Results for the [bi] stimuli indicated significant main effects for vowels ($F_{(4,10)} = 16.15, p < 0.001$), transitions ($F_{(4,10)} = 85.31, p < 0.001$), and duration ($F_{(4,30)} = 4.43, p < 0.01$), and a significant interaction for transition $X$ duration ($F_{(4,30)} = 7.03, p < 0.002$). The vowel effect was due to significantly better performance for identification of the [a] and [u] stimuli in comparison to the [i] stimuli. Moving-transition stimuli were significantly better perceived than were straight-transition stimuli. The duration effect was due to overall better performance with 20-ms stimuli than with 10-ms stimuli, although the effects of duration were restricted to the moving stimuli only, with significantly better identification for 20 ms than 10, 29, and 46 ms moving-transition stimuli, and with 46- and 29-ms moving stimuli better than 10-ms moving-transition stimuli.

Analysis for the [d] stimuli indicated a significant main effect for duration ($F_{(4,30)} = 3.36, p < 0.03$), due to better performance on 20-ms vs 446-ms stimuli, and a vowel $X$ duration interaction ($F_{(4,60)} = 2.60, p < 0.03$) due to effects of duration occurring for the [a] vowel stimuli only. Although there was not a significant main effect for transitions as there was for the alveolar stimuli in experiment 1, there was a tendency for moving-transition stimuli to be better perceived than straight transitions ($F_{(4,10)} = 3.62, p < 0.09$).

For the [g] stimuli, there was a significant main effect for duration due to better performance on the 46- than the 20- and 10-ms stimuli ($F_{(4,30)} = 6.91, p < 0.02$), and significant interactions for vowel $X$ transition ($F_{(4,20)} = 7.46, p < 0.004$), and vowel $X$ transition $X$ duration ($F_{(4,10)} = 3.00, p < 0.01$). The failure to find a main effect for transitions was due to a reversal of performance for [gi] stimuli in which straight transitions were perceived better than were moving transitions. Interaction effects indicated better performance for moving versus straight transitions for 10-ms [a] vowel stimuli, and 46- and 29-ms [u] vowel stimuli. In contrast, for the [i] vowel stimuli, straight transitions were better than moving transitions at durations of 46, 20, and 10 ms. Comparison of effects of duration showed the following significant effects. For [i] vowel stimuli, 46- and 29-ms moving transitions were better perceived than 20- and 10-ms moving transitions, 46-ms straight-transition stimuli were better than 29- and 10-ms stimuli, and 20-ms straight-transition stimuli were better than 10-ms stimuli. For [u] vowel stimuli, 46- and 29-ms moving-transition stimuli were better than 20- and 10-ms moving-transition stimuli. All other comparisons across vowels, durations, and transitions were nonsignificant.

E. Discussion

The results of experiment 2 indicate that brief portions of the onsets of CV stimuli without bursts still provide sufficient information for identification of place of articulation across vowel contexts for most of the stimuli. As in experiment 1, although stimuli with moving transitions tend to be better perceived than those with straight transitions, subjects still can identify the straight-transition stimuli better than chance (with the exception of the [bi] and [gi] stimuli). This result suggests that subjects need not rely upon secondary cues, in particular, directions of formant motions, for the identification of place of articulation, although when these cues are available, they can be used in making a phonetic decision. The poor performance for the straight-transition [bi] stimuli may be due to the fact that the burst is necessary for the perception of labial consonants in the environment of this vowel, in order to provide the requisite downward-sloping onset spectrum. In the absence of a burst, the listener utilizes the secondary cue of the rising formant transitions to indicate a labial response. If the transitions are missing as well, this cue is lost, and [b] responses drop to a low level.

The effects of stimulus duration on identification responses for the no-burst stimuli are similar to those for the stimuli with bursts. For the [bi] and [d] stimuli, the responses tend to be slightly better for the shorter than for the longer stimuli, whereas the opposite is the case for the [g] stimuli.

A summary of the results of experiments 1 and 2, showing the effects of the stimulus duration and of the burst on identification performance, is presented in Fig. 5. Mean data for each consonant are obtained by averaging performance over vowels and over the moving- and straight-transition conditions. In the case of the [d] and [g] stimuli, the presence of the burst seems to enhance the properties intrinsic to each place of articulation (Stevens and Blumstein, 1978), resulting in overall higher identification performance. However, removal of the burst does not change the gross attri-
III. EXPERIMENT 3

On the basis of experiments 1 and 2 and other data, we have postulated that the primary cue for place of articulation is the gross shape of the short-time spectrum sampled at the onset. It has been further postulated that, under circumstances where the primary spectral cues are weakened or modified, a listener has the capability of resorting to secondary cues, some of which may be context dependent, in making a phonetic identification (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979). In experiment 3 we attempted to determine the effects on perception of place of articulation when the strength of the spectral cues was varied in a systematic way.

The strength of the spectral cues was manipulated by generating several series of stimuli in which the acoustic characteristics varied along a continuum that spanned the range from [b] to [d] to [g]. In this way, stimuli with acoustic characteristics intermediate between those corresponding to the different phonetic categories were synthesized. The primary acoustic property associated with a given phonetic category (such as a property specifying the gross shape of the onset spectrum) would presumably be weaker for these intermediate stimuli. An aim of this experiment was to determine how the listener makes use of secondary cues such as formant trajectories or stimulus durations in making a phonetic identification when the stimuli contain vowel primary cues.

### A. Stimuli

Twelve place-of-articulation synthetic continua were generated: four [b d g] continua for each of the three vowel series [a], [i], and [u]. As in experiment 1 of this study, the basic stimulus set consisted of a 5-ms noise burst followed by brief, glottal excitation of a set of formant resonators. There were initial 40-ms piecewise-linear-formant transitions in which the onset frequencies of the first four formants were in the ranges appropriate to the consonants [b d g] and the offset frequencies of these transitions were appropriate to the vowels [a i u]. The parameter values for the continua were interpolated on a linear scale from the set of stimuli used in experiment 1. Thus, the starting frequencies of F2, F3, and, in the case of [i] stimuli, F4 as well, were systematically manipulated through a series of values to range across the best exemplars of the phonetic categories [b d g] in the environment of the vowels [a], [i], and [u]. The formant starting frequencies for the continua are shown in Table II.

The bursts for these continua were generated by appending a 5-ms noise segment to the starting frequencies of the formants and systematically adjusting the amplitude of the spectral peak associated with one formant in relation to that associated with another formant. For the [a] series, the burst for stimulus 14 (at the [g] end of the continuum), the burst was produced by exciting a resonator centered at 3600 Hz with a 5-ms segment of white noise, beginning 10 ms prior to the onset of voicing. Thus, the burst frequency was continuous with F4 of the adjacent vowel. For stimulus number 14 (at the [g] end of the continuum), the burst was produced by exciting a resonator centered at 1810 Hz (the starting frequency of the F2 transition) with a 5-ms segment of white noise beginning 15 ms prior to voicing onset. Between stimuli 8 and 14, both F4 (which remained at 3600 Hz in the series) and F2 (which varied from one stimulus to the next) were experimentally manipulated.

### TABLE II. Starting frequencies for F2, F3, and F4 (in Hz) for the synthetic continua of consonant–vowel syllables.

<table>
<thead>
<tr>
<th>[ba da ga]</th>
<th>[bi di gi]</th>
<th>[bu du gu]</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>F3</td>
<td>F2</td>
</tr>
<tr>
<td>1</td>
<td>900 2000 1</td>
<td>1500 2600 3200</td>
</tr>
<tr>
<td>2</td>
<td>1010 2110 2</td>
<td>1833 2633 3317</td>
</tr>
<tr>
<td>3</td>
<td>1120 2220 3</td>
<td>1867 2667 3433</td>
</tr>
<tr>
<td>4</td>
<td>1240 2340 4</td>
<td>1900 2700 3550</td>
</tr>
<tr>
<td>5</td>
<td>1350 2450 5</td>
<td>1933 2733 3667</td>
</tr>
<tr>
<td>6</td>
<td>1470 2570 6</td>
<td>1967 2767 3783</td>
</tr>
<tr>
<td>7</td>
<td>1580 2680 7</td>
<td>2000 2800 3900</td>
</tr>
<tr>
<td>8</td>
<td>1700 2800 8</td>
<td>2067 2833 3817</td>
</tr>
<tr>
<td>9</td>
<td>1800 2900 9</td>
<td>2133 2867 3733</td>
</tr>
<tr>
<td>10</td>
<td>1870 2970 10</td>
<td>2196 2925 3650</td>
</tr>
<tr>
<td>11</td>
<td>1960 3040 11</td>
<td>2267 2993 3567</td>
</tr>
<tr>
<td>12</td>
<td>1980 3067 12</td>
<td>2293 3025 3483</td>
</tr>
<tr>
<td>13</td>
<td>1980 3067 13</td>
<td>2400 3000 3400</td>
</tr>
<tr>
<td>14</td>
<td>1980 3067 14</td>
<td>2467 3067 3325</td>
</tr>
</tbody>
</table>

[FIG. 5. Data from Figs. 3 and 4 averaged across vowels and across moving-and straight-transition conditions. The solid lines show responses to stimuli with bursts (from experiment 1, Fig. 3), and the dashed lines show responses to stimuli without bursts (from experiment 2, Fig. 4).]
cited by a 5-ms segment of noise, but the relative amplitudes of the bursts were systematically reduced from the endpoint amplitudes of stimuli 8 and 14 by 4–5 dB for each stimulus step. Similar interpolation procedures were used for generating bursts for stimulus 1–8. For stimulus 1, the noise burst excited F2 and began 5 ms prior to voicing. This procedure for generating a stimulus continuum with varying noise bursts and formant transitions has been described in detail elsewhere (Stevens and Blumstein, 1978). The bursts for the [budu gu] series were generated in the same manner as the [badu go] series. For the [bidi gi] series, the burst for the endpoint [g] stimulus (number 13) was generated by exciting the third rather than the second formant.

This basic set of stimuli consisting of 40-ms moving transitions was produced by excitation of the moving-formant resonators by five glottal pulses (as in the 46-ms stimuli of experiments 1 and 2). Three other continua were derived from this basic set. In the second set, the duration of the stimulus was shortened, and was obtained by excitation of the resonators with two glottal pulses, giving an effective duration of 20 ms. The last two sets of continua consisted of the same durations, burst frequencies, and onset characteristics as the first two continua, but with no moving transitions (except for the F1 transition). As in experiment 1, the frequencies of the second and higher formants of these stimuli were constant throughout and were equal to the formant frequencies at voicing onset as given in Table II.

B. Subjects

Ten subjects from Brown University participated in this experiment. All were native speakers of English and had no known hearing impairments.

C. Procedure

Six stimulus tapes were constructed—two each for the vowel series [a], [i], and [u]. Within each series, each tape consisted of stimuli of equal duration containing both moving and straight transitions. Thus, the 46-ms moving- and straight-transition stimuli comprised one tape and the 20-ms moving- and straight-transition stimuli comprised the other. In the [a] series, there were 14 stimuli in each continuum. Thus, there was a total of 28 different stimuli for each of the two tapes of the [a] series, and 26 different stimuli for each of the two tapes of the [i] and [u] series. Each test tape consisted of two random orders of five presentations per stimulus, each part being preceded by two practice trials. The interstimulus interval was 3 s, and after each block of 10 stimuli there was a 5-ms silent interval. Subjects were tested in two sessions, the order of the tapes being counterbalanced across vowels and durations. Short breaks were given at the end of each tape and between each part within the test tape. Subjects were required to identify each stimulus as [b], [d], or [g] by writing the appropriate initial consonant on the answer sheet provided. Prior to each test tape, the subject heard two repetitions of the particular series to be tested.

All subjects participated in all of the test series. However, one subject was unable to label the 46-ms moving-transition [a] stimuli consistently. Consequently, his scores were eliminated from the entire [a] series. Thus, nine subjects completed the [a] series in its entirety, and ten subjects the [i] and [u] series.

D. Labeling functions

Figure 6 shows mean labeling functions from the twelve test continua. Inspection of the graphs for the [a] stimuli indicates that the labeling functions are reasonably discrete for all conditions, and most especially for the 20-ms moving- and straight-transition stimuli. There are some [g] intrusions in the [b] region (stimuli 3–5), particularly for the 46-ms stimuli. These [g] responses are presumably due to the presence of a spectral component of the burst in the frequency region 1120–1350 Hz (the starting point of the F2 transition for stimuli 3–5), and this component tends to give the impression of a narrow spectral prominence in a mid-frequency region appropriate for the “compact” consonant [g]. This effect is greater for the longer stimuli, in accordance with the finding, noted earlier, that a longer time seems to be necessary to give a stronger impression of compactness.

For the [i] stimuli, the place-of-articulation categories are fairly well defined, with the exception of the [b] category for the stimuli with 46-ms straight transitions. In this case, the subjects tend to perceive the entire [b]–[d] ranges as diffuse-acute (i.e., as [d]), although there are some [g] intrusions for the first two or three stimuli.1 The identification functions for the [i] stimuli did not have very sharp boundaries, suggesting that, for a range of stimuli lying between phonetic categories, there were conflicting cues to place of articulation, as discussed in more detail later.

In the case of the [u] stimuli, for which the responses are shown in the lower panel of Fig. 6, there are well-defined [d] and [g] categories, although the [g] region for the 46-ms moving-transition stimuli is rather narrow, and the boundary for these stimuli is not sharp. Neither the 46-ms nor the 20-ms series shows a well-defined [b] region (in contrast to the results of experiment 1 in the upper right panel of Fig. 4), and the [g] intrusions for stimuli 1–3 are greater for the 46-ms stimuli than for the 20-ms stimuli. To the extent that there is a tendency toward [g] responses in the [bu] region, it is not unexpected that this tendency should be greater for the longer stimuli, as indicated above in connection with Fig. 4. Again, the [g] intrusions are presumably caused by the presence of the mid-frequency burst with a spectral prominence in the vicinity of the F2 starting frequency—a burst spectrum that is probably not a good representation of the acoustic characteristics of a natural labial stop consonant. The number of [g] intrusions in the [b] region for the vowel [u] is greater than those in the [b] region in the continuum with the vowel [a], even though the starting frequencies for F2 and F3 are roughly the same in this region for the two continua. The difference in responses must be attributed to the differences in the F1 transitions. The frequency extent of this transition is large for the [ba]
stimuli, and the transition is rapid, thus providing a secondary cue to labiality, whereas in the case of the [bu] stimuli the extent of the F1 transition is much less, and the rate of F1 transition does not help to distinguish labials from velars.

In summary, then, the results indicate that relatively discrete labeling functions can be obtained for stimuli containing information from the burst and the initial 2–5 glottal pulses of a consonant–vowel syllable. The briefer stimuli seem to lead to the most well-defined response categories, whereas the consistency of the responses, particularly in the [b] category, tends to be reduced for the longer stimuli.

E. Phonetic boundaries

We now examine in more detail the responses to stimuli that are located near phonetic boundaries. As noted above, we are especially interested in the pattern of responses to these stimuli since the primary spectral properties corresponding to any one of the phonetic categories are only weakly defined, whereas there are still secondary properties which are potentially available to a listener.

We consider first the responses to stimuli in the region between [bu] and [da] (upper panel of Fig. 6). For this boundary region, there is little difference between the phonetic boundaries for the various conditions of long and short stimuli and straight and moving transitions. For stimulus 5, which is closest to the phonetic boundary, the F2 and F3 transitions are almost flat, and, consequently, no significant difference in responses to straight— and moving-transition stimuli would be expected. There is a tendency for stimulus 5 to receive more intrusive [g] responses for the longer duration, for reasons already indicated.

At the [da]–[ga] boundary, there appears to be a difference in identification responses to the 20 and the 46-ms stimuli. For stimulus 10, for example, responses to the 20-ms stimuli are about equally divided between [d] and [g], whereas they are predominantly [g] when the stimulus duration is 46 ms. The starting frequencies of F2 and F3 for stimulus 10 are sufficiently close (790 Hz) that this proximity contributes to the mid-frequency spectral prominence characteristic of a compact consonant. The longer duration is sufficient to enhance the compactness property (as suggested above), and thus to shift the responses in the direction of the velar consonant.

Turning next to the [bi]–[di] boundary (middle panel of Fig. 6), we see some differences in responses to stimulus 4 and 5 depending on the stimulus duration and on the presence or absence of formant movements. For stimulus 5, for example, responses to the 46-ms moving-transition condition are equally divided between [b] and [d]. This stimulus has a rising F2 and F3 and a
s. When all three of these formants rise (stimuli 1-4), however, the predominant response is [b]. For the stimulus continuum with straight transitions, responses to stimulus 5 (46 ms) become predominantly [d]. If it is postulated that rising transitions tend to bias the responses toward [b] in the absence of strong primary cues from the gross shape of the onset spectrum, then this result is not unexpected. By the same token, [d] responses to the 20-ms versions of stimulus 5 would tend to increase, since the transitions are too brief to provide cues based on transition directions. Figure 6 indeed shows that there are more [d] responses to the 20-ms stimulus 5 than to the 46-ms version with moving transitions.

The [di]-[gi] boundary seems to be located close to stimulus 10 for all four versions of the stimulus. Because the starting frequencies for F2, F3, and F4 are close to those of the vowel, the formant transitions for the stimuli with moving transitions are minimal, and hence there is little difference in responses for the straight- and moving-transition stimuli. A tendency toward increased [g] responses for the longer version of this stimulus might have been expected, but this did not occur. (As noted earlier, however, the longer stimuli within the [g] region were more consistently identified as [g] than were the shorter stimuli.)

At the [bu]-[du] boundary, we note some differences in responses across the different stimulus conditions, and these differences can be seen most clearly by examining the responses to stimuli 4 and 5. The spectrum at onset for stimulus 5, for example, has the diffuse-rising property, but only weakly, as shown in Fig. 7, and, thus, based on this spectral property alone there are somewhat more [d] than [b] responses. The F2 and F3 transitions are slightly falling for the versions of stimulus 5 with moving transitions. Thus for the longer stimulus (46 ms) with moving transitions the [d] responses are enhanced. Straight transitions, on the other hand, are neutral with respect to the [b]-[d] distinction, and hence there are fewer [d] responses. Likewise, the shorter versions of stimulus 5 also show fewer [d] responses than the 46-ms moving-transition stimulus, presumably because these stimuli are too short to provide an indication to the listener of the directions of formant motions.

Differences in responses to the various stimulus continua are also observed in the vicinity of the [du]-[gu] boundary. For item 9 on the continuum, for example, the responses to the longer, straight-transition stimulus are about equally divided between [d] and [g]. For this stimulus, the spectrum at onset is neither strongly diffuse nor strongly compact, since F2 and F3 are at intermediate, neutral positions close to 1500 and 2500 Hz, respectively, and thus the primary cues provided by the onset spectrum are equivocal. If formant transitions are inserted, however, there is a falling F2 and a slightly falling F3, and the secondary cues provided by these transitions shift the responses toward [d]. Formant transitions probably do not provide significant cues for the brief 20-ms stimulus 9, and on this basis one might expect the responses to be similar to the longer straight-transition stimuli. However, according to the hypothesis advanced earlier, brevity of the stimuli would prevent the buildup of a compact spectral representation, and would tend, therefore, to reduce the number of [g] responses for this boundary stimulus—a prediction that is supported by the data for stimulus 9 in the lower panels of Fig. 6.

The interpretation of the responses for stimuli near the phonetic boundaries of the various continua can be summarized, then, as follows. When the spectrum at onset does not provide clear-cut information about place of articulation, listeners resort to other, secondary, cues based on formant motions. Rising trajectories of formants 2 and 3 tend to provide [b] responses, falling transitions [d] responses, and spreading formants [g] responses. The data also support the notion that the auditory representation of an onset with a mid-frequency spectral prominence is weakened if the duration of the stimulus is decreased, and consequently there tend to be more [g] responses for a 46-ms stimulus than for a 20-ms stimulus.

IV. EXPERIMENT 4

We have postulated that, for the brief stimuli used in experiments 1 and 2, the gross attributes of the spectrum at the stimulus onset remain invariant for a particular place of consonant articulation independent of context, and we have implied that the auditory system is predisposed to interpret the stimuli in terms of these gross attributes. The detailed characteristics of these brief stimuli are, however, dependent on the following vowel as well as on the consonant, as the descriptions in Table I indicate. We were interested in determining, therefore, whether listeners could derive vowel as well as consonant information from stimuli of the type used in experiment 1.

![Figure 7](image_url)

FIG. 7. The curve shown by the solid line is the short-time spectrum sampled at the onset of stimulus 5 on the [bu du gu] continuum, for which responses were about equally divided between [b] and [d]. The dashed line represents the onset spectrum for stimulus 7 on the same continuum, which was unanimously identified as [d]. Spectra are calculated in the manner described by Stevens and Blumstein (1978); they are pre-emphasized at the high frequencies, and are smoothed by a linear prediction algorithm.
FIG. 8. Vowel responses to the stimuli used in experiment 1. There was a closed set of responses [a], [i], or [u]. Each panel is labeled with the intended syllable, as in Fig. 1. Solid lines represent data for moving-transition stimuli, and dashed lines for straight-transitions stimuli.

A. Stimuli

The stimuli and test tapes were identical to those used in experiment 1. That is, the stimuli contained bursts at the onset, there were four different durations for both the moving-transition and the straight-transition conditions, and there were tokens of [b d g] in the environment of the vowels [a i u].

B. Subjects and procedure

Fifteen subjects who were trained in phonetics listened to the stimuli. They were asked to identify the vowel in each stimulus as [a], [i], or [u]. Five of the subjects listened to both the moving- and the straight-transition tapes; the remaining ten listened to either the moving-transition stimuli or the straight-transition stimuli.

C. Results and discussion

The results of the vowel identification experiment are shown in Fig. 8. For the [a] vowel stimuli, labeling is quite poor for the 10-ms duration for both moving and straight transitions. These short stimuli were usually identified as [u] rather than [a]. The labeling improves dramatically, however, when the stimuli are longer. For the remaining vowels, identification is very good across all voicing durations for both moving and straight transitions, with somewhat better performance for stimuli with moving transitions. This ability of listeners to identify vowels from brief initial segments drawn from syllables beginning with stop consonants has also been observed by Winitz, et al. (1972) and by Ohde and Sharf (1977).

The trajectories of $F_1$ and $F_2$ for the various moving- and straight-transition stimuli are shown in Fig. 9. Each line on the figure represents the variation in $F_1$ and $F_2$ throughout one of the 46-ms stimuli, i.e., up to the time that the vowel target is reached in the case of stimuli with moving transitions. Also shown on the figure are the average $F_1$ and $F_2$ values for the vowels [a], [i], and [u] spoken by an adult male, as reported by Peterson and Barney (1952). The $F_1-F_2$ trajectory for the [ba] stimuli, for example, begins at the point where $F_1 = 200$ Hz and $F_2 = 900$ Hz. For the stimuli with moving transitions, both $F_1$ and $F_2$ rise in frequency until the point $F_1 = 720$ Hz, $F_2 = 1070$ Hz is reached, and this occurs 20 ms after the onset of voicing. At this time the transition of $F_1$ is completed, and $F_1$ remains at 720 Hz while $F_2$ continues to increase for another 20 ms until the vowel target at $F_2 = 1240$ Hz is reached.

For a voicing duration of 10 ms (the approximate duration of the briefest stimulus, which is generated with one glottal pulse), the trajectory for moving-transition [ba] reaches the point $F_1 = 450$ Hz, $F_2 = 990$ Hz. (The trajectory actually continues beyond this...
The trajectories for [da] and [ga] stimuli of duration 20 ms or more all approach the [ə] target sufficiently closely that they are identified almost unanimously as belonging to the vowel [ʊ] rather than to [i] or [u]. The 10-ms [da] and [ga] trajectories terminate at positions that are rather remote from [ə], and hence there are some intruding [u] responses. The fewest [ə] responses are obtained for the 10-ms straight-transition [ga] stimulus, for which the final F1 and F2 values are approximately 330 and 1640 Hz, respectively, leading to predominantly [u] responses (together with a few [i] responses).

In the case of the stimuli with vowels [i] and [u], the terminal values of the F1 – F2 trajectories approach the appropriate vowels in most cases, and, hence, correct vowel identification is obtained. Possible exceptions are the straight-transition [bi] and [du], for which the trajectories terminate at F1 and F2 values that are intermediate between [i] and [u]. In Fig. 8 the few intrusions obtained for the [bi] stimulus (F2 = 1800 Hz) are [u], and the intrusions for the [du] stimulus (F2 = 1600 Hz) are [i]. The value of F2 at the boundary between [i] and [u] responses appears to be at about 1700 Hz. This conclusion is consistent also with the result, noted earlier, that the 10-ms straight-transition [ga] stimulus, for which F2 = 1640 Hz, yielded predominantly [u] responses.

The main conclusion to be drawn from the vowel identification experiments is that listeners are able to extract both consonant and vowel information from these brief stimuli that can be as short as one glottal pulse (together with an initial burst). One way of interpreting this finding is that the brief stimulus signals the identity of the syllable, which is processed by the listener as a unitary percept or a single event. Having identified the syllable, the listener is then able to identify the identity of the syllable, i.e., the features of both the consonant and the vowel.

It is important to observe, however, that the identification of the initial segment of the syllable as a consonant is determined by the fact that the first formant is rising from a relatively low frequency. (In the case of a voiceless consonant, the consonantal feature is signalled by the absence of energy in the first-formant region rather than by the rapidly rising F1, i.e., the rapid change in F1 and the absence of F2 apparently gives rise to identification of the same phonetic feature.) If the first formant had been fixed in frequency (as were the second and higher formants in the straight-transition stimuli), then the impression of the listeners would have been that of a brief isolated vowel with no initial consonant.

V. GENERAL DISCUSSION

The results of experiments 1 and 2 show that information with regard to place of articulation for a voiced stop consonant resides in the initial 10–20 ms of a consonant–vowel syllable. The motions of the formants immediately following the consonantal release, although contributing to place-of-articulation identification, are not essential, since eliminating movements of the second and higher formants still results in good identification performance of consonantal place of articulation. The fact that consonant identification remains good (although somewhat reduced in level) when the initial noise burst is removed, supports the hypothesis that both the burst and the attributes of the sound at voicing onset contribute to a more global acoustic property that is a cue for place of articulation for voiced stop consonants.

These observations are consistent with the view that the gross properties of the spectrum sampled over the initial 10–20 ms of a stop consonant provide the principal cues to place of articulation for stop consonants. In this view, the transitions of the formants from the release of the consonant to the vowel provide the acoustic material that links the transient events at the onset to the slowly varying spectral characteristics of the vowel. These transitions ensure that no further abrupt discontinuities in the spectrum occur following the initial transient at the release.

The particular spectral properties that are associated with different places of articulation have been discussed elsewhere. For velar consonants, the spectrum has a compact prominent spectral peak in the mid-frequency range, whereas for labials and alveolars there is a diffuse spread of spectral energy. In the case of alveolars, this diffuse spectrum has a rising slope, whereas for labials it is flat or falling. A result that emerges from the present study is that a velar tends...
to be identified with fewer errors if the duration of the stimulus is longer than 10–20 ms, suggesting that a longer time is necessary to build up a representation of a "compact" onset spectrum in the auditory system. Labials and alveolars, on the other hand, are identified well even when the stimulus is very short.

Secondary cues such as directions of formant motions or frequencies of particular formants at consonantal release also provide information with regard to place of articulation. This use of secondary cues is most clearly seen when the primary attributes of the onset spectrum are equivocal, so that the spectrum does not demonstrate strong unambiguous properties such as compactness or diffuseness, or is neutral! with respect to the grave–acute distinction. In experiment 3, for example, there were some stimuli along the continuum for which the attributes of the onset spectrum did not provide a clear and unambiguous indication of place of articulation. For these stimuli, the directions of formant motions, or the stimulus duration, were apparently used by the listeners to resolve the ambiguity with regard to place of articulation.

Experiment 4 showed that the brief stimuli used in this study provide cues not only to consonantal place of articulation but also to the vowel. Consonantal place of articulation seems to be cued by the gross shape of the short-time spectrum sampled at the onset of the stimulus, whereas vowel information is carried by the formant frequencies at the termination of the stimulus. Other features are signalled by these stimuli as well. For example, the fact that the consonant is identified as a stop rather than as a continuant is presumably a consequence of the relatively abrupt onset of the stimulus. The feature of voicing is carried by the rapidly rising first formant at the onset of the stimulus. Thus a number of phonetic features of the segments in the syllable are packaged within a brief interval of 20–50 ms, and each of these features is signalled by different attributes of the stimulus. One way of looking at the auditory processing of the stimulus is to imagine that a number of auditory detectors respond selectively to different properties, such as compactness of onset spectrum, abruptness of onset, etc. This constellation of properties leads to the identification of the syllable.

On the basis of this study and of related research, it is possible to make the following hypotheses concerning the perception of speech: (1) In the stream of speech, abrupt onsets and offsets provide markers to indicate points in time where acoustic information relevant to consonantal place of articulation is sampled. (2) This acoustic information resides for the most part in the 10–20 ms time interval immediately adjacent to the onset or offset. (3) The gross shape of the spectrum in this region provides the essential perceptual cues for place of articulation across vowel contexts. If the auditory system is predisposed toward performing these kinds of operations leading to invariant acoustic correlates of phonetic features, then a relatively simple and direct model of speech perception at the segmental level can be hypothesized, both for the adult listener and for the developing infant.

**ACKNOWLEDGMENTS**

This work was supported in part by the National Institute of Health, grant NS-04332. Doctor Blumstein’s participation was supported in part by the John Simon Guggenheim Foundation and the Radcliffe Institute. The assistance of Laura Hall and Pamela Hamilton in running the experimental subjects is acknowledged with thanks.


