1 The Search for Invariant Acoustic Correlates of Phonetic Features

Kenneth N. Stevens
Massachusetts Institute of Technology

Sheila E. Blumstein
Brown University

Editors' Comments

One of the major activities in the study of speech has been the search for the acoustic correlates of perceived phonetic distinctions. The history of this research can be viewed as a three-part story. In the initial period, investigators tended to assume that the waveform of speech contained acoustic properties that mapped onto phonetic structures in a relatively simple one-to-one manner. However, the empirical evidence proved to be just the opposite; invariant properties could not be found in their stead investigators inevitably discovered complex mappings between the acoustic signal and the phonetic percept. Of course, the failure to find acoustic invariance should not be taken to mean that progress in describing the acoustic structure of speech was limited. Indeed, in many respects the search was quite successful as evidenced, for example, by our ability to produce rather intelligible synthetic speech.

In the middle period of the search for acoustic correlates, which extends to the present; researchers, virtually abandoning the assumption of acoustic invariance, continued to obtain evidence for a complex, context-conditioned relation between the acoustic signal and the perceived phonetic structure. However, there are also, at present, investigators who have argued that earlier failures to find invariance were essentially the result of incorrectly characterizing the acoustic information. They have resumed the search for invariance and, in so doing, have initiated the most recent phase of this work. The research of Stevens and Blumstein is one of the most extensive efforts along these lines. In their chapter, they describe their work to discover the invariant acoustic information for perceived distinctions in place of articulation and speculate on possible invariant cues for other phonetic distinctions. In addition, they discuss the significance of their findings for issues related to the development of speech perception and to the nature of processing models of speech.
INTRODUCTION

The nature of the speech perception process in man has been the topic of considerable research and discussion for the past 20 years. The research paradigms devised to study this process, as well as the theoretical approaches taken, have been rich and imaginative. To date, there are three competing theories of the speech perception process. The first and perhaps the most widely accepted argues that the perception of speech depends ultimately on the analysis of the continuous acoustic signal to yield discrete phonetic segments (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). These segments themselves have an intrinsic organization and structure based on underlying features (Chomsky & Halle, 1968; Jakobson, Fant, & Halle, 1963). The relation between the acoustic signal and the phonetic percept is not a direct one. Rather, the perception of phonetic segments requires the extraction of context-dependent cues, which are interpreted phonetically in different ways depending on the nature of the context in which the phonetic segment appears. Such a theory requires an active perceptual mechanism in which identification of phonetic segments depends on some kind of special computation or look-up procedure (cf. Halle & Stevens, 1972; Liberman et al., 1967).

The second theory also proposes that perception depends on the analysis of the speech signal into discrete phonetic features. However, unlike the context-dependent theory, it is hypothesized that the properties of speech can be uniquely and invariantly specified from the acoustic signal itself (Liberman & Stevens, 1979; Cole & Scott, 1974; Fant, 1960; Stevens & Blumstein, 1978) and that these properties are closely related to the distinctive features. The hypothesis is that the speech perception system responds in a distinctive way when a particular sound has these properties so that the process of decoding the sound into a representation in terms of distinctive features can be a fairly direct one. The role of context-dependent cues in the acoustic invariance theory is not denied but is rather considered to provide only alternative or secondary cues to the phonetic dimensions of speech, whereas the invariant properties provide the basic or primary cues. (The notion of primary and secondary cues is elaborated further in the following section.)

The third theory denies neither acoustic invariance nor contextual dependence and their relation to phonetic segments, but proposes that in ongoing speech it may be necessary to have recourse to the identification of acoustic patterns of larger units rather than to features and segments. As a result, word recognition may depend on the extraction of the entire acoustic pattern for a word or syllable as a gestalt without further analysis in terms of component features (Klatt, 1979).

In the past few years, we have focused our research on an examination of the second theory described above, the theory of acoustic invariance. It is the object of this chapter to elucidate this theory, discuss its theoretical and experimental bases, and consider the implications of these findings for a model of speech perception.

The theory of acoustic invariance is based on several major assumptions. The first, of course, is that acoustic invariance corresponding to a particular phonetic category or distinctive feature resides in the acoustic signal. Second, this invariance is not derivable from an analysis of individual components of the acoustic signal, as might be observed in particular regions of an intensity-frequency-time display or spectrogram of speech, but rather is provided by integrated acoustic properties that may encompass several of these components. These properties are sampled at particular points in time, often where there is a rapid change in the amplitude or in the spectrum. Thus, for example, although individual components of the acoustic signal, such as the burst, onset frequencies of particular formants, or directions of formant transitions, do not provide invariant cues to place of articulation in stop consonants (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Delattre, Liberman, & Cooper, 1955; Schatz, 1954), the shape of the spectrum sampled over a particular time interval at the release of the consonant does seem to provide an invariant pattern (Blumstein, Stevens, 1979; Fant, 1960). The spectrum shape includes all the acoustic information within the first 20 msec or so at the release of the consonant and, in this sense, reflects an integrated acoustic property.

We make three further assumptions concerning the theory of acoustic invariance. The nature of these invariant properties and their relation to phonetic segments reflects: (1) the way in which the articulatory mechanism constrains the possible range of speech sounds (i.e., there is evidence that a limited set of articulatory configurations produces stable acoustic patterns; Stevens, 1972); (2) the way in which the perceptual mechanism constrains the possible range of speech sounds (i.e., categorial perception studies have shown that the physical scale along which pairs of sounds differ is not the same as the scale used by the auditory system to judge differences between speech sounds [Liberman, Harris, Hoffman, & Griffith, 1957]); and (3) the way in which the set of classes of speech sounds based on underlying distinctive features are defined. These distinctive features provide the framework for the phonological grammar of the language system. That is, correspondences among speech sounds form natural classes, which in turn help to structure the nature of the linguistic grammar (Halle, 1972; Jakobson, Fant & Halle, 1963).

The theory of acoustic invariance has been elaborated most completely for place of articulation in stop consonants. Considerations from acoustic theory, analysis of acoustic characteristics of natural speech production, and the results of experiments investigating the perception of synthetic speech have contributed to the elaboration of the theory, and we review these findings in the following
section. Nevertheless, a complete theory of speech requires that it can characterize all of the features found in natural language. In this chapter we do not attempt to discuss all of the properties that might be relevant to the formulation of a system of features. We do, however, attempt to go beyond the properties that characterize place of articulation for stop consonants, and we review some of the basic properties relevant to the consonant system in English.

THE NATURE OF SPEECH SOUNDS

Before discussing in detail the various acoustic properties of speech sounds and their perceptual correlates, we should first place boundaries on the characteristics of the sounds that we are considering. All speech sounds appear to have a particular kind of structure in both the temporal and spectral dimensions that distinguish them from nonspeech and musical sounds, and the studies we describe here are concerned primarily with the perception of sounds having this structure.

One common attribute of speech sounds is that the spectra are usually characterized by a series of rather narrow peaks. This property can be observed in the spectra in Fig. 1. Spectra of speech sounds are not flat or monotonic in changing with frequency, but rather they tend to exhibit peaks and valleys, the spectral amplitude in the valleys being 10-30 dB below the amplitude of the spectrum at the peaks. One way of specifying the properties of such a spectrum is to say it contains several narrow peaks, but another way of describing the spectrum is to say it contains valleys or "holes" that are sufficiently deep in relation to the peaks. When this kind of spectral structure is weakened by reducing the amplitudes of the spectral peaks or by filling in the valleys of the spectrum, the speech-like nature of the sound is weakened or disappears (Remez, 1979).

A second universal attribute of speech is that the amplitude of the sound rises and falls. The rises and falls are associated with the syllabic structure, as shown in the example of Fig. 1. Peaks in amplitude occur during vowels when the vocal tract is maximally open, and minima in amplitude occur during consonants when the vocal tract is constricted. The amplitude maxima or minima normally occur at a rate of 3-4 per second.

A third broad property of speech sounds is that there are changes in the short-time spectrum of the sound. These variatons occur as a consequence of movements of the spectral peaks and of changes in the relative amplitudes of the peaks. Often these spectral changes are quite rapid and occur over time intervals of a few milliseconds up to 40-odd milliseconds (as illustrated by the spectra sampled in the vicinity of the [b] release in Fig. 1), but sometimes the spectrum changes more slowly.

FIG. 1.1. Several representations of the acoustic attributes of an utterance in the frequency and time domains. Middle: Spectrogram of the utterance "The big rabbits." Bottom: Variation of amplitude versus time during the utterance. Top: Spectra sampled at various points throughout the utterance as indicated. In the case of [b] at the left, these spectra are shown, and these are obtained by sampling at these points in time 13 msec apart. The spectra are computed for the preemphasized waveform, and are smoothed using a linear prediction procedure.

We suggest that these three attributes provide a set of constraints within which the detailed properties of individual speech sounds are described—an acoustic baseline or "posture," as it were. The implication is that the auditory system is predisposed toward extracting detailed properties from sound signals that have a frequency-time structure of this type. A signal with these three attributes provides the possibility for a wide variety of properties to which the auditory system can respond distinctively. That is, the auditory system may be predisposed to produce a variety of distinctive responses when the properties of the sound are characterized by change or by lack of constancy: changes in spectral amplitude over frequency at a fixed time, changes in amplitude over time, and changes in spectral peaks and valleys over time. As we proceed to discuss the acoustic properties and the perception of different phonetic classes, it will always be understood that we are restricting our consideration to sounds that have these general attributes.
IN Variant PROroperties FOR
PLACE OF ARTICULATION

One of the phonetic dimensions along which consonants in all languages are
classified is the place-of-articulation dimension, i.e., the location of the point of
maximum consonantal constriction in the vocal tract. In most languages, at least
three places of articulation can be distinguished: labial, alveolar, and velar.
Within each of these classes, further subdivisions are made in many languages,
and, in addition, some languages have consonants with constrictions in the
labial and pharyngeal regions.

The results of acoustic analyses have suggested that stop consonants can be
characterized by integrated properties (Fant, 1960; Halle, Hughes, & Radley,
1957; Stevens, 1975). These properties reflect the configuration of acoustic
events that occur at the release of a stop consonant—acoustic events that are a
consequence of a particular place of articulation. It has been implied (Fant,
1960, 1973; Fischer-Jorgensen, 1954, 1972; Stevens, 1975) that the auditory system
responds to these properties in an integrated manner rather than by processing
each of a number of simpler properties and combining these at a later stage.

These issues have been examined more recently in a series of studies focusing on
acoustic theory, acoustic analysis of natural speech, and perception of synthetic
speech for stop and nasal consonants (Blumstein & Stevens, 1979, 1988;
Stevens & Blumstein, 1978). We turn now to a review of this work.

Theoretical Considerations

The shape of the spectrum sampled at the release of a stop consonant for different
burst frequencies and formant starting frequencies can be predicted from the
theory of sound production in the vocal tract (Fant, 1960). This theoretical
analysis can be used as a guide for examining the spectra at the onset for naturally
produced syllables beginning with stop consonants and for interpreting the results
of speech perception experiments. Thus, before discussing data from speech-
production and speech-perception studies, we review briefly this theoretical
background.

When the articulatory structures achieve a particular configuration, the acoustic
cavities formed by these structures have certain natural frequencies or formants. These formants are manifested in the source-harmonic contours at particular
frequencies. When a constriction is made in the vocal tract in the oral cavity, the
frequency of the first formant (F1) is always lower than it is for a vowel. On the
average, the second and higher formants (F2, F3, etc.) occur at regular intervals
in frequency, the average spacing between these higher formants being about 1
Hz for adult male speakers, and somewhat greater for adult female speakers and
children. The spectrum envelope for a sound with the lowered F1 corresponding
to a constricted vocal tract and with F2 and higher formants at their average
frequencies is shown in the upper panel of Fig. 1.2. This envelope assumes that
the acoustical excitation of the vocal tract arises from normal glottal vibration.

Depending on the position and shape of the constriction, however, the fre-
frequencies of the second and higher formants undergo displacements upward and
downward relative to their average values. These shifts in the frequencies of the
spectral peaks are accompanied by changes in the relative amplitudes of these
peaks, such that the gross shape of the spectrum can be influenced by changes in
the formant frequencies.

Shifting of F2, F3, and higher formants downward in frequency (while keeping
F1 at the same frequency) causes a decrease in the amplitudes of the higher
formant peaks in relation to the lower formants, as shown in the upper panel of
Fig. 1.2. Such a downward shift in the formant frequencies occurs when a
constriction is made at the lips, and hence this is the short-time spectrum that is
to be expected at the onset of voicing for a labial consonant immediately after
release of the constriction. The reduction of the amplitudes of the higher formant
peaks arises from the fact that, as the frequency of a given formant Fn shifts
downward, the spectral peaks arising from formants of higher frequencies ride
vertically up and down, so to speak, on the “skirts” of the formant Fn. Thus a
decrease in the frequency Fn causes a reduction of the amplitudes of the higher
formant peaks (Fant, 1956; Stevens & House, 1961). On the other hand, shifting
of F2, F3, and higher formants upward in frequency (while keeping F1 at the
same frequency) causes an increase in the amplitudes of the higher formants
(e.g., F4 and F5) in relation to the lower formants (e.g., F2), as shown in the
same figure. Such a configuration of formants is expected for an alveolar con-
sonant.

Similar shifts in the relative amplitudes of the higher formants, and hence
changes in the gross shape of the spectrum for different places of articulation,
will also occur in the spectrum sampled in the aspirated portion of the sound
following the release of an aspirated voiceless consonant. The theoretical consid-
erations are the same, except that the source of excitation for the vocal tract now
consists of turbulence noise at the glottis, rather than glottal vibration, and, as a
consequence, there is less spectral energy in the region of F1 and possibly F2.
Furthermore, the spectrum in the low-frequency region (in the vicinity of F1)
may be affected by acoustic coupling through the glottis to the trachea (Fant,
Ishizaka, Lindquist, & Sundberg, 1972).

At the release of a syllable-initial alveolar voiced or voiceless consonant in
English, a burst of turbulence noise (sometimes called frication noise) is gener-
ated at the constriction. This burst usually occurs just prior to the onset of voicing
for initial voiced stops, and just prior to the onset of aspiration noise for voiceless
stops. This noise source excites the higher vocal-tract resonances (usually F4 and
F5 and higher) but produces only weak acoustic excitation of the lower formants
(F2 and F3), resulting in a burst spectrum like that shown by the dashed line in
the middle panel of Fig. 1.2. This burst spectrum contributes to the overall
property of the gross shape of the spectrum sampled at the consonant release by
accentuating the amplitudes of the high-frequency peaks in relation to the lower
peaks. The solid line in the figure shows the spectrum for a voiced alveolar stop
consonant sampled at the onset to encompass both the burst and the initial 10 or
so msec at voicing onset. Note that the burst enhances the greater energy in the
higher frequencies (F4 and F5).

For the labial consonant, the spectrum of the burst is relatively flat, because
the frication noise source at the lips tends to excite all of the formants about
equally. The burst is, however, relatively weak, and its influence on the overall
spectrum shape at the onset of the consonant is probably relatively small. As a
consequence of the burst, there could be a tendency toward a somewhat flatter
onset spectrum than the theoretical labial spectrum without a burst, shown in the
upper panel of Fig. 1.2.

Because the resonances of the cavities anterior and posterior to the point of
constriction tend to be roughly equal in the production of velar consonants, the
second and third formants are often relatively close together at onset. The acousti-
cal consequence of the proximity of the two formants is to enhance the amplitudes
of both spectral peaks in relation to the amplitudes of higher formants. The lower
panel of Fig. 1.2 shows the theoretical spectrum envelope at the onset of a voiced
velar stop. As the noise burst at the release of a velar stop consonant excites the
resonance of the vocal-tract cavity anterior to the constriction, it is usually
continuous with either the second or third formant of the following vowel (de-
dpending on whether it is a back or front vowel). Thus the spectrum of the burst
has a relatively narrow peak in the vicinity of the proximate second and third
formants at the onset of voicing. The presence of this burst then enhances the
spectral energy concentration in the mid-frequency region. The lower portion of
Fig. 1.2 shows the theoretical spectra at onset for a voiced velar stop with and
without burst. As in the case of the labial and alveolar consonants, it is expected
that the onset spectrum for a voiceless aspirated velar consonant in natural speech
will be similar to the spectrum shown in this figure, except that the spectral peak

FIG. 1.2 (Opposite page) Top: The dashed line represents the theoretical spec-
trum obtained with a voice source and with a low first-formant frequency and the
second and higher formants located at the neutral positions of 1500, 2500, 3500,
and 4500 Hz. Also shown are theoretical spectra for formant locations correspond-
ing to an alveolar consonant (upward shifts of F2, F3, and F4) and a labial conso-
nant (downward shifts of F2 and higher formants). Middle: The spectrum for voice-
source excitation of formants corresponding to an alveolar consonant is shown (from
the upper panel), together with a theoretical spectrum for a noise burst at the release
of an alveolar consonant (dashed line), and a composite spectrum obtained with both
sources. Bottom: The theoretical spectrum for voice-source excitation of formants
corresponding to a velar consonant is shown, together with the modified spectrum
when a noise burst is present. (Adapted from Stevens & Blumstein, 1978.)
Evidence From Acoustical Measurements

Given the foregoing theoretical considerations, it would be expected that acoustical measurements of short-term spectra in natural speech would reveal distinctive shapes for the various places of articulation. The results of several studies investigating the spectrum of the burst in isolation (Halle, Hughes, & Radley, 1957; Zue, 1976) and of the initial few tens of milliseconds following consonantal release (Fant, 1960; Jakobson et al., 1963; Searle, Jacobson, & Kimble, 1979) have suggested that distinctive patterns for place of articulation can be derived from a short-time spectral analysis. These patterns can be seen in the examples of spectra for several naturally produced voiced and voiceless stop consonants shown in Fig. 1.3—in particular, the gross spectrum shape is diffuse-rising for alveolar consonants, diffuse-falling or flat for labial consonants, and compact for velar consonants. These spectra were obtained by using a window length of 26 msec beginning at the consonantal release and were derived by means of a 14-pole linear prediction algorithm, which emphasized the higher frequencies (Blumstein & Stevens, 1979; Stevens & Blumstein, 1978). The 26-msec window encompasses different portions of the voiced and voiceless stop consonants. For voiced stops, the time window includes the burst as well as some portion of voicing onset (e.g., /b/) or only the burst (e.g., /g/). For voiceless stops, the window includes the initial friction burst and a portion of the aspiration.

Although the obtained spectral shapes for these few samples of natural speech utterances are qualitatively similar to the theoretically derived spectra, it is necessary to determine the extent to which such correspondences exist in a wide variety of utterances produced by different speakers and in different vowel contexts. That is, do invariant acoustic properties reside in natural speech utterances, and if so, are these properties context-independent? In order to address this question, it was necessary to develop a more quantitative measure of the gross spectral shapes corresponding to each place of articulation. To this end, a series of templates was developed in an attempt to reflect each of the spectral properties—diffuse-rising, diffuse-falling or flat, and compact. The configurations of these templates were determined in part from theoretical considerations and in part from an examination of a limited set of consonant-vowel utterances consisting of the initial consonants /bdg/ in the environments of the vowels /i e a o u/ produced by two male speakers.

The three templates are shown on the three panels on the left side of Fig. 1.4. The panels on the right side of the figure illustrate the application of the templates to spectra that fit and to those that fail to fit the templates. In general, these templates specify ranges of acceptable relative amplitudes of peaks in the spectra at consonantal onsets and offsets. Specific details concerning the procedures for fitting the spectra to the templates are given elsewhere (Blumstein & Stevens, 1979).
The diffuse-rising template is represented by two reference lines about 10 dB apart. A spectrum is matched against the template by first adjusting its amplitude such that one peak is tangent to the upper reference line above 2200 Hz, and all other peaks are below that line. The overall requirement of a spectrum, if it is to fit this template, is that at least two spectral peaks minimally 500 Hz apart must lie within the reference lines (the diffuseness requirement), and that there is a general upward tilt of the spectrum with increasing frequency (the rising requirement). At least one peak of energy must fall above 2200 Hz and be higher in amplitude than a lower frequency peak. Thus, the template characterizes as belonging to a single class those spectra having a diffuse-rising spread of spectral energy with no one peak dominating the spectrum. The specific onset frequencies of the individual formants are, within limits, of no consequence. What is critical, however, is the gross shape of the onset spectrum. An example of a spectrum meeting the required characteristics is shown at the right of the diffuse-rising template in Fig. 1.4, superimposed on the template. Examples of spectra that do not have diffuse-rising characteristics are shown in the second panel from the top at the right of the figure. The spectrum of the [g] shows just one prominent peak, and thus does not satisfy the diffuseness requirement. Although the spectrum of [b] is diffuse, in that it contains energy spread over a range of frequencies, the spectral energy distribution is falling with increasing frequency, and thus does not fit within the template.

The property characteristic of labial consonants is diffuse-falling or diffuse-flat, and the template designed to fit these spectral shapes is shown in the middle panel of the left side of Fig. 1.4. A spectrum is matched against the template by adjusting its amplitude such that one peak is tangent to the upper reference line between 1200 and 3600 Hz, and all other peaks in the range are below that line. To fit the requirements of this template, a spectrum must have at least two spectral peaks a minimum of 500 Hz apart falling within the reference lines, one peak falling below 2400 Hz and the other peak falling within the range of 2400-3600 Hz. There is no condition on the amplitude of spectral peaks below 1200 Hz, but peaks above 3600 Hz must be below the upper reference line. An example of a spectrum of a labial consonant with the required characteristics is superimposed on the diffuse-falling template at the right of the figure. Examples of spectra with shapes that do not fit these characteristics are also shown. Although the [d] spectrum is diffuse, i.e., there are several peaks spread out in the frequency domain, the distribution of the spectrum is rising rather than either falling or flat; the [g] spectrum shows one prominent mid-frequency peak and thus is not diffuse.

The property that describes a velar consonant is the presence of a prominent spectral peak in the mid-frequency range. Two spectral peaks that are separated by 500 Hz or less are treated as comprising a single gross spectral peak. A peak is "prominent" if there are no other peaks nearby and if it is larger than adjacent peaks, so that the peak stands out, as it were, from the remainder of the spec-}

trum. This is the sense in which the spectrum is compact. The template shown in the lower left panel of Fig. 1.4 attempts to capture this property. This template consists of an overlapping set of spectral peaks in the mid-frequency range (from 1200-3500 Hz). To meet the requirement for spectral compactness, a single peak in the spectrum of the sound must fit within a matching peak of the template. Further, there can be no other peak in the spectrum projecting through the reference line, nor can there be another peak of the same or greater magnitude falling below 1200 Hz or above 3500 Hz. An example of a spectrum that meets the requirements of the compact template is shown at the right of the compact template in Fig. 1.4. Note that the 1400 Hz peak of the spectrum is matched to a low-frequency peak on the template. Also shown at the right is an example of a spectrum that does not fit the compact template. A second peak (at 2500 Hz) just through the major spectral peak of the template; further, there is an energy peak above 3500 Hz, which is higher in amplitude than the major mid-frequency peak.

Once the individual templates were developed and the particular fitting procedures were determined, a large number of natural speech utterances were collected in order to serve as a data base for determining the extent to which the spectra of initial and final voiced and voiceless stop consonants fit the hypothesized shapes. A total of six subjects, four male and two female, read a list of CV utterances containing five repetitions of each of the stop consonants [p t k b d g] in the context of the five vowels [i e a o u], and a listing of CV utterances containing the same six stop consonants but preceded by the vowels [i e a o u]. These 1800 utterances were tape-recorded in a sound-treated room and subsequently analyzed using the procedures for spectral analysis described earlier. Syllable-initial consonants were analyzed by the procedures shown in Fig. 1.3. For the syllable-final stop consonants, spectra were sampled at two points in the signal: at the point of consonantal closure at the end of the vowel, and at the onset of the burst that occurs at the consonantal release (if the final consonant is, in fact, released). In the measurement of the closure portion of the syllable, the peak of the spectral window was placed at the point of closure. For the release burst, the spectral measurements were analogous to those used for CV onsets with measurements made from the moment of burst release.

The spectra of the 1800 natural CV and VC utterances were individually tested against each template. A conservative strategy was adopted for assessing whether the spectral shapes were accepted or rejected by the particular template. In order to fit the template, the spectrum had to meet all the conditions specified for the template. If it did not fit for any reason, e.g., the shape was clearly wrong or the shape was correct but a peak failed to fit within the reference lines, then the spectrum was rejected. The design of the templates was such that it was possible for some spectra to fit more than one template. (See Blumstein & Stevens, 1979, for further discussion.)

Preliminary data were also obtained from about 110 alveolar and labial nasal consonants produced in consonant-vowel syllables by the same six speakers.
FIG. 1.3. Examples of waveforms and spectra sampled at the release of three voiced and three voiceless stop consonants as indicated. Superimposed on two of the waveforms is the time window (of width 26 msec) that is used for sampling the spectrum. Short-time spectra are determined for the first difference of the sampled waveform (sampled at 10 kHz) and are smoothed using a linear prediction algorithm; i.e., they represent all-pole spectra that provide a best fit to the calculated short-time spectra with preemphasis. (From Blumstein and Stevens, Journal of the Acoustical Society of America, 1979, 66, 1001-1018. Copyright 1979 by the Acoustical Society of America. Reprinted by permission.)

FIG. 1.4. At the left are shown the templates that are used to test whether a spectrum is diffuse-rising (top), diffuse-falling or flat (middle), or compact (bottom). For diffuse-rising template, the arrow indicates that a spectra peak above 2200 Hz is to be fitted to the upper reference line of the template. The vertical markers on the diffuse-falling template indicate regions where spectral peaks should occur (at least one in the range 2000-2400 Hz, and one in the range 2400-3600 Hz) within the reference lines if a spectrum is to match this template. For a spectrum to fit the compact template, a prominent mid-frequency spectral peak must fit entirely within one of the contours of the template. Two panels are shown to the right of each template. In each case, the upper panel shows a spectrum that fits the template, and the lower panel gives examples of spectra that do not fit the template.
Spectra were sampled at the instant of consonant release, using a somewhat shorter time window than that shown in Fig. 1.3. Only the diffuse-rising and diffuse-falling templates were used in this study of the nasals.

The results of applying the three templates to all of these utterances are summarized in Table 1.1. Overall, about 83% of the initial stop consonants and about 77% of the initial nasals were correctly accepted by their respective templates. About 76% of the final-consonant bursts were also correctly accepted, but the spectra sampled at closure for alveolars and velars were rather poorly identified (31 and 52%, respectively). Initial stops and final bursts were generally correctly rejected by the templates (e.g., an alveolar stop was not accepted by the diffuse-falling or the compact templates). Final alveolars and velars, sampled at the closure, tended to have a diffuse-falling spectral shape. Initial alveolar nasals also tended (incorrectly) to fit the diffuse-falling template.

The principal conclusion from Table 1.1 (and from the details of the data from which Table 1.1 was derived) is that the templates effectively described the three types of spectra sampled at the release of stop consonants (including the release of final stop consonants) corresponding to the three different places of articulation. These spectrum shapes as defined by the templates are relatively independent of vowel context and of individual speaker characteristics. Further refinement of the analysis procedures and of the templates would undoubtedly improve the scores given in Table 1.1 for these aspects of the stop consonants.

There may be several reasons for the lack of success of the template-matching procedures for the closure of stop consonants and for initial alveolar nasals. In the case of the closure at offset, the spectral analyses obtained may reflect the fact that final consonants are often deviated prior to closure. Consequently, the point of abrupt amplitude reduction at which the spectrum is sampled may occur several tens of msecs before the closure (corresponding to the articulatory motion from the vowel to the target consonant) rather than at the actual closure itself. There are some perceptual data that are consistent with the observation that unreleased consonants often fail to show invariant acoustic properties. In particular, data from the perception of place of articulation in unreleased stop consonants indicate that identification is a good deal poorer than that obtained for final released stops, although the reported absolute level of identification varies across studies (Halle, Hughes, & Radley, 1957; Malecot, 1958; Wang, 1959).

In the case of both final consonants at closure and initial nasals, the point at which the spectrum is sampled is preceded by low-frequency energy—for the final stop consonant, it is the preceding vowel and for the initial nasal consonant, the nasal murmur. It may be that the spectral representation in the auditory system for a signal with an abrupt onset preceded by silence is different from the representation when the onset (or offset) is preceded by low-frequency spectral energy. The presence of this low-frequency energy may effectively reduce or mask the response of some neural units in the auditory system to the spectrum at the discontinuity, particularly the response to the low-frequency components of the spectrum. As a consequence, this spectral representation would show an attenuation of the low frequencies relative to the spectral representation of the onset when preceded by silence. The auditory representation of the spectra preceded by low-frequency energy would, then, tend to have a more sharply rising characteristic than the measured spectra and would presumably reflect better the diffuse-rising property of the alveolar template. Some support for this line of reasoning is provided by data on single-unit responses of the auditory nerve to stimuli with the acoustic characteristics of nasal-vowel syllables (Delgutte, 1980).

### Evidence From Speech-Perception Studies

The acoustic data we have described in the foregoing section indicate that it is possible to find invariant acoustic properties that can be used to classify stop consonants according to place of articulation. It remains to be determined, however, whether these acoustic properties are utilized by a listener during the perception of these consonants. At least two approaches can be followed to gain an understanding of the acoustic information used by a listener when he is required to identify utterances containing stop consonants. One approach is to

| Initial alveolar stops | 86 | 12 | 15 |
| Initial alveolar-velar closure | 31 | 59 | 29 |
| Initial alveolar-velar burst | 77 | 13 | 17 |
| Initial labial stops | 17 | 81 | 11 |
| Initial labial-velar closure | 11 | 77 | 27 |
| Initial labial-velar burst | 16 | 77 | 12 |
| Initial velar stops | 13 | 8 | 85 |
| Initial velar-velar closure | 17 | 59 | 52 |
| Initial velar-velar burst | 18 | 17 | 75 |
| Initial alveolar nasals | 72 | 67 | |
| Initial labial nasals | 10 | 81 | |
manipulate systematically the properties of the onset spectrum in a consonant-vowel syllable such that, in a series of stimuli, the spectrum changes gradually from a diffuse-falling shape through a diffuse-rising shape through a compact shape. Responses of listeners to this sequence of stimuli would then indicate the ranges of onset-spectrum shapes that give rise to labial, alveolar, and velar responses. A second approach is to strip away from the consonant-vowel stimuli all the information except the attributes that are postulated to provide the appropriate cues for place of articulation, and to determine whether the listeners can identify place of articulation from the sound that remains. This manipulation would involve removing the steady state vowel and portions of the formant transitions in the stimulus, leaving only the initial portion of the sound that contributes to the shape of the onset spectrum. Both of these approaches have been followed in a series of experiments aimed at uncovering the acoustic properties used by a listener in classifying place of articulation for stop consonants.

In the first study (Stevens & Blumstein, 1978), several series of stimuli consisting of stop consonants followed by vowels were synthesized. For each series, the vowel was the same and the starting frequency of the first formant was fixed, but the starting frequencies of the higher formants and the spectrum of an initial burst were manipulated to produce stimuli with a graded continuum of onset spectra. These stimuli were presented to listeners in random order, with a number of replications, and the listeners were instructed to identify the initial consonants as b, d, or g. For each stimulus sequence, three well-defined response regions were usually obtained, with rather sharp changes in the response functions in the vicinity of particular boundary stimuli.

Examples of onset spectra for stimuli that were identified unequivocally as [b], [d], and [g] in the series with the vowel [a] are shown in Fig. 1.5 as spectra 1, 8, and 14, respectively. These spectra clearly have the gross properties of the type respectively described previously as diffuse-falling, diffuse-rising, and compact. Also shown in the figure are two spectra (5 and 10) for which the responses were equivocal. These spectrum shapes do not exhibit strongly any one of the properties defined by the templates in Fig. 1.4. Similar results were obtained with stimulus series containing other vowels.

These data provide support, then, for the notion that the process of identifying place of articulation in these syllables involves detection of the gross shape of the spectrum sampled at onset, and classifying the shape into one of the three broad categories: diffuse-falling, diffuse-rising, and compact.

In the second study (Blumstein & Stevens, 1980), listener responses were obtained to brief stimuli that reproduced acoustic events only in the initial portion of a consonant-vowel syllable. The aim was to determine whether acoustic information in this portion of the syllable was sufficient to give proper identification of consonant place of articulation.

The durations of the synthesized stimuli were in the range 10-46 msec, and the formants during this brief interval followed trajectories appropriate to
consonant-vowel syllables beginning with [b], [d], or [g]. These trajectories moved toward target values corresponding to the three vowels [i, a, u]. Various stimulus conditions were used, including stimuli both with and without bursts. An additional set of stimuli was produced by eliminating the movements of the second and higher formants, so that these formants remained at their frequency values at onset. Results averaged over the three vowel environments and over the straight- and moving-transition conditions are summarized in Fig. 1.6. The performance of listeners in the identification of place of articulation for these stimuli was always well above chance, even when the sounds were as short as 10 msec and when they contained no transitions of the second and higher formants. Thus the responses of the listeners tended to follow the pattern that would be expected if they were basing their identification on the gross shape of the spectrum sampled over the brief duration of the stimuli. However, lack of formant movements and the absence of a burst within this time interval did contribute to a reduction in the overall performance level. Further, there were a few stimuli (particularly velars before the vowel [i]) for which the identification responses were inconsistent. (See Blumstein & Stevens 1980, for further discussion.)

A further finding with these brief stimuli was that listeners were usually able to identify which of the three vowels [i, a, u] the formant trajectories were moving toward, even for stimuli that were so short that they contained only one or two glottal pulses. Apparently, then, at least two kinds of information are packaged within these short stimuli and are accessible to a listener. The gross shape of the spectrum sampled over the initial portion of the stimulus (or perhaps over the entire stimulus duration for the very short stimuli) indicates the place of articulation of the consonant, based on whether it is diffuse-rising, diffuse-flat or falling, or compact. Other attributes of the spectrum, such as the frequencies of the spectral peaks sampled at the termination of the stimulus, are sufficient to enable the listener to determine the identity of the vowel.

In a final experiment with brief synthesized consonant-vowel sounds, several continua of stimuli were produced, spanning the range of acoustic characteristics (at onset) from [b] to [d] to [g] (Blumstein & Stevens, 1980). Three such continua were generated, with formant trajectories moving toward frequencies corresponding to the vowels [i], [a], and [u]. Two stimulus durations (20 and 46 msec) were used. An additional set of stimulus continua was generated in which the second and higher formants remained at their frequency values at onset. Listener responses to these stimuli showed that the continua were for the most part divided into three categories with reasonably sharp boundaries, although there were a few exceptions. A general finding was that, when the spectrum sampled at stimulus onset was not a clear exemplar of one of the three gross shapes postulated for the three places of articulation, then the listeners utilized available information concerning the time course of the formant trajectories. Thus, for example, if the spectrum was diffuse but did not show either a rising or falling slope, then a listener would tend to identify the consonant as [b] if the second formant was rising and as [d] if it was falling.

The results of these experiments with brief stimuli excerpted from the onsets of synthesized consonant-vowel syllables support the hypothesis that information with regard to place of articulation for a voiced stop consonant resides in the initial 10-20 msec of a consonant-vowel syllable. The motions of the formants immediately following the consonantal release do not appear to contribute essential information regarding place of articulation, because elimination of the movements of the second and higher formants does not greatly modify the identification of consonantal place of articulation, given appropriate (i.e., unambiguous) onset spectra. The fact that consonant identification deteriorates only slightly 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 stop consonants.
These observations are consistent, then, with the view that the gross properties of the spectrum sampled over the initial 10–20 msec of a stop consonant provide invariant or primary cues to place of articulation. 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 (Cole & Scott, 1974; Stevens & Blumstein, 1978). These transitions ensure that no further abrupt discontinuities in the spectrum occur following the initial transient at the release. Cues such as the directions of formant motions or frequencies of particular formants at consonantal release also provide information with regard to place of articulation. These cues are secondary in the sense that, for a particular consonantal place of articulation, they depend upon the vowel context. 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 distinction between a diffuse-rising or diffuse-falling shape.

INvariant Properties for Other Phonetic Categories

In the preceding section, we have discussed at some length the evidence indicating that invariant acoustic properties are associated with different places of articulation for stop and nasal consonants. The place-of-articulation dimension was considered in some detail in part because a considerable amount of data on the production and perception of place of articulation for consonants is available and in part because earlier studies concluded that the acoustic cues for different places of articulation show substantial dependence on context.

We now consider the acoustic properties and perceptual correlates of several other consonantal contrasts. The point of view in this discussion is that there are invariant acoustic properties associated with each of these phonetic categories as well as the place-of-articulation categories and that the auditory system is predisposed to detect or to respond selectively to these properties. Some of the material in this section is rather speculative, because data with regard to these phonetic categories are not as extensive as data on place of articulation. However, they provide preliminary evidence supporting the theory of acoustic invariance in speech, and they suggest avenues for future research on these issues.

The Consonant-Vowel Contrast

One of the basic contrasts in language is the contrast between consonants and vowels. From the point of view of articulation, a vowel is produced with a vocal tract that is relatively open, with no narrow constrictions along the length of the tract from just above the glottis to the lips. A consonant is generated with a narrow constriction at some point in the vocal tract. Acoustically, consonants and vowels are distinguished by the nature of the changes that occur in the acoustic spectrum. For consonants, the primary acoustic consequence of a narrow constriction in the vocal tract is the existence of a rapid change in the spectrum as the articulatory structures move toward or away from the constricted configuration (Stevens, 1971). The rapid spectrum changes are preceded and followed by regions in which the spectrum changes relatively slowly. One of these regions may be silence or a region of low acoustic energy (as in the case of stop consonants), but there may be appreciable energy in both regions (e.g., for fricative consonants or nasals). One possible source of the rapid spectrum change is the rapid rise of the first formant that always accompanies an increase in the size of the constriction. Spectral changes can also occur at higher frequencies as a consequence of changes in the source of noise in the vicinity of the constriction or changes in the frequencies of the higher formants. For vowels, the acoustic consequences of a relatively open vocal tract is a well-defined steady-state formant structure with relatively slow changes over time.

Segments that are produced with a narrow constriction along the midline of the vocal tract and exhibit rapid spectrum change are identified by the feature consonantal. Vowels and glides, which do not give rise to a rapid spectrum change, are nonconsonantal. The points in time where rapid spectrum changes occur in consonantal segments can be regarded as special events or landmarks in the signal, and it is suggested that the speech perception strategy used by listeners directs attention to attributes of the signal in the vicinity of these events.

The spectrogram in Fig. 1.7, which shows the sentence “Joe took father’s shoe bench out,” is marked to identify the locations of the events where rapid spectrum changes occur. A time window of width about 20 msec in the vicinity of these events would indicate regions where it is postulated that attributes of the speech signal are sampled. It is these regions and their respective time windows that define the domain for analysis of the acoustic events giving rise to phonetic features such as place of articulation, voicing, etc. In a sense, we are arguing that there are “regions of high information” to which the speech-processing system is directed during ongoing speech perception. These regions are joined to other regions where the spectrum changes are relatively slow. The regions between the slow and rapid spectrum changes constitute the so-called transitions, and these may vary in duration from a few tens of milliseconds to over 100 msec. Of course it is assumed that, in addition to being directed towards the events where rapid spectrum changes occur, the speech-processing system continuously monitors and decodes the signal in the regions between these events.

In Fig. 1.8 the contrast between a consonantal and a nonconsonantal segment is illustrated in terms of the rapidity with which the spectrum changes. In the case of the syllables [ba] and [sa] there are rather abrupt changes in the spectrum in the vicinity of the consonantal release, whereas for [wa] and [Pa] the spectrum
changes are relatively slow. There is an abrupt rise in amplitude at the release of the glottal stop, but the spectrum immediately following this abrupt rise changes very little.

The acoustic attributes that correlate with the identification of the feature consonantal have been examined in a series of speech perception experiments (Liberman, Delattre, Gerstman, & Cooper, 1956; Miller & Liberman, 1978). In these experiments, the rate of movement of the formants at the onset of a synthetic consonant-vowel syllable was manipulated to produce a series of stimuli ranging from those with rapidly moving formants to those with slowly moving formants. Listener responses to the consonant portion of the syllable were obtained. One of the stimulus continua was characterized by rising transitions of the formants, and encompassed a range corresponding to the consonants [b] and [w]. The results showed that fast transitions gave the response [b], which is consonantal, whereas slow transitions yielded the nonconsonantal response [w]. The boundary between these two classes occurred when the duration of the transitions was about 40 msec. The identification of stimuli with transition durations in the vicinity of 40 msec (i.e., stimuli for which the spectrum change is neither rapid nor slow) can be influenced by contextual factors (e.g., vowel duration, as reported by Miller & Liberman, 1978).

As is noted elsewhere in this chapter, the properties of the sound in the vicinity of these consonantal events are used to identify phonetic features of the consonant relating to place of articulation and voicing. The occurrence of a consonantal segment is indicated by the presence of some generalized kind of rapid spectrum change, without specification of the frequency regions or directions or other attributes of the change; certain other phonetic features of the consonant are signaled by the detailed properties of the signal in the vicinity of this rapid spectrum change. On the basis of neuropsychological and psychophysical experiments, evidence is emerging to indicate that components of the auditory system produce a specialized response when a transition is characterized by an abrupt onset or offset of amplitude in a particular frequency range, or by a rapidly changing spectrum (Delgutte, 1980; Kiang, 1975; Kiang, Watanabe, Thomas, & Clark, 1965). Thus, what we know about the response of the peripheral auditory system at the level of the auditory nerve and the cochlear nucleus is not inconsistent with the view that there is a specialized response to stimuli for which there are rapid spectral changes.

Continuant-Abrupt

There is a class of phonetic segments having the common acoustic property that the production of the segment produces an abrupt rise or fall in the amplitude of the sound, whether or not there is a rapid change in spectral shape. This class of abrupt (or stop) segments, includes the set /p t k b d g ć j n n g/. All of these consonants except [?] are a subset of the class of consonantal segments, i.e., they
show an abrupt change in amplitude as well as a rapid spectrum change. (The glottal stop [?]—a nonconsonantal segment—shows only an abrupt amplitude change without an accompanying rapid spectrum change, as shown in Fig. 1.8.) Abrupt consonants are produced by completely blocking the airstream at some point along the midline of the laryngeal and vocal-tract pathways. Consonantal segments, on the other hand, are produced by creating a complete or partial blockage of the airstream at a point in the vocal tract above the larynx. The closure for abrupt segments can occur at the glottis as well as above the glottis, i.e., a glottal stop is included in the class of abrupt consonants, whereas the constriction must be above the glottis for a consonantal segment. Immediately following the release of a glottal stop there is, in fact, not a rapid spectrum change.

The primary acoustic attribute that distinguishes abrupt from continuous segments is that abrupt segments show an abrupt rise in amplitude at the release (or an abrupt fall in amplitude at the implosion). The rise in amplitude occurs at all frequencies; that is, there is no frequency band for which the amplitude is greater before the release than after the release. This property seems to exist for the nasal consonants as well as for the stop consonants, because the spectrum for the nasal murmur is lower in amplitude than the spectrum immediately following the release at essentially all frequencies.

For a fricative consonant, that is, a continuous segment followed by a vowel, there may be a rather abrupt increase in amplitude over some part of the frequency range at the consonantal release, but there always appear to be other frequency regions in which the amplitude decreases, or at least does not increase, at this time. This aspect of fricative consonants can be seen in the example of Fig. 1.8, in which there is an amplitude increase at low frequencies but a decrease in some regions of the spectrum at high frequencies near the point where voicing begins. Furthermore, the amplitude increase at low frequencies is not as abrupt as that for a stop consonant.

Several experiments have examined the perceptual correlates of the feature \textit{continuant}. These experiments have investigated listener responses to stimuli in which the abruptness of an onset or the silent interval preceding the onset were manipulated. These experiments provide an indication of the rate of increase of amplitude at the onset, and of the duration of the interval of low intensity prior to the onset, that are necessary to elicit a special response corresponding to an abrupt or stop consonant.

Experiments in which natural speech stimuli are manipulated in various ways have provided evidence that helps to define the acoustic property associated with the feature \textit{continuant}. In one experiment, it was shown that the syllable [s], for which the onset of the amplitude of the [s] is rather gradual, can be changed to [c?] by removing the initial part of the friction and by leaving an abrupt amplitude rise in the noise (Cutting & Rosner, 1974). The rise time at the perceptual boundary between responses of [s] and [c?] was about 40 msec. In a second experiment, the word \textit{silt} was generated, and the acoustic representation of the word was then manipulated by creating a gap of successively increasing duration between the end of the friction noise in [s] and the onset of voicing in [l] (Bastian, Delattre, & Liberman, 1959). When this duration of silence exceeded about 70 msec, the word was heard as \textit{split}. Apparently, the onset of voicing in the [l] was sufficiently abrupt that a stop consonant was heard if the preceding silent interval was long enough. The suggestion is that about 70 msec of silence or of low amplitude is needed before an abrupt onset if the onset is to be heard as an abrupt consonant. In another class of experiments (Bailey & Summerfield, 1978; Repp, Liberman, Eccardt, & Pesetsky, 1978), the duration of a gap between a [s] or [z] and a following vowel or sonorant was manipulated, to determine the gap duration necessary to elicit the perception of a stop consonant between the fricative and the vowel. It was found that the duration required depended to some extent on the physical characteristics of the onset following the silent interval. If the onset exhibits more of the characteristics of a consonantal onset in terms of the degree of rapid spectrum change, it appears that it is a ‘stronger’ onset, and thus requires less of a preceding silent interval.

Another experiment relevant to the perceptual correlates of the feature \textit{abrupt} utilized nonspeech stimuli with onsets for which the rise time varied from a few msec to several tens of msec (Cutting & Rosner, 1974). These stimuli were categorized systematically by listeners: Those with rise times less than 40 msec sounded like “plucked” musical sounds, whereas those with longer rise times were identified as “bowed” sounds. Within each class, different stimuli along the continuum were essentially nondiscriminable by listeners; two stimuli with rise times of 30 and 50 msec (i.e., stimuli that were classified differently by the listeners) were easily discriminated. The implication of these results is that the auditory system is predisposed to categorize sounds with different rise times into two different classes—those with fast rise times and those with slow rise times. Such a predisposition could form the basis for the phonetic distinction between \textit{abrupt} and \textit{continuant} consonants.

\textbf{Nasal--Nonnasal}

Within the class of abrupt consonants, we can distinguish two categories—nasal and nonnasal. The nonnasal consonants are the stops [p t k ð ɡ چ]. The principal acoustic characteristic that distinguishes nasals from nonnasals is the presence or absence of an appreciable amount of energy within the closure interval, as shown in the spectrograms in Fig. 1.9. In nasal consonants, although the amplitude within this interval is lower than that in the adjacent vowel, it may only be a few dB lower at low frequencies (below about 300 Hz). There is also an appreciable amplitude in the spectrum of the nasal murmur in higher frequency regions. Some low-frequency energy may also be present in the spectrum sam-
Voiced-Voiceless

There is a large number of languages that distinguish voiced from voiceless segments. Thus, for example, in English the consonants [p t k f s] are voiceless, and are in opposition to the voiced consonants [b d g v z]. Voicing is indicated by the presence of low-frequency spectral energy or periodicity in the speech signal due to vibration of the vocal folds, whereas for voiceless sounds there is no such periodicity. In the case of non-consonantal sounds (such as a vowel, or [h], or [w]), detection of voicing is relatively straightforward, and the presence or absence of low-frequency periodicity can be observed in the relatively steady-state or slowly varying time interval in the vicinity of the target configuration for the sound. For consonantal segments, acoustic data from a number of different languages suggest that the voiced feature can be identified by testing for the presence of low-frequency spectral energy or periodicity over a time interval of 20-30 msec in the vicinity of the acoustic discontinuity that precedes or follows the consonantal constriction interval (Lisker & Abramson, 1964). The time interval over which this test for periodicity is made usually extends to the left of the consonantal discontinuity for a consonant preceding a vowel (i.e., a test for prevoicing), or to the right of the consonantal implosion in the case of a syllable-final consonant (where left and right indicate times preceding and following the discontinuity).

This procedure of detecting voicelessness or voicedness is illustrated in Fig. 1.10 for contrasting stop consonants in Spanish and for fricatives in English. We observe that in the 20-30 msec prior to the rapid spectral change near the release of the voiceless stop or the fricative, there is no low-frequency energy or periodicity in the sound, whereas low-frequency energy is present in the case of the voiced consonant.

Stop consonants in English are usually described as having a voiced-voiceless distinction, but the acoustic manifestations of this distinction differ from that previously described, particularly when the consonant occurs in pre-stressed position. For example, when a voiced stop consonant occurs in pre-stressed position, there is frequently no prevoicing, and low-frequency periodicity does not begin until 0-20 msec following consonantal release.

In the case of a voiceless stop consonant in pre-stressed position in English, the consonant is aspirated, i.e., the vocal folds are in a spread configuration at the time the consonant is released, and the onset of low-frequency periodicity due to glottal vibration occurs 30 or more msec after the consonant release (Lisker & Abramson, 1964). In view of these characteristics, it may be appropriate to abandon the phonetic classification of voiced-voiceless for stop consonants in English, and rather to refer to the categories as unaspirated and aspirated, respectively, or to use terms such as tense and lax (cf. Jakobson et al., 1963).

The property that distinguishes an aspirated from an unaspirated stop consonant, then, is the absence of low-frequency periodicity in the 20-odd msec to the
to vocal-fold vibration, and these stimuli were presented to listeners for identification as either voiced or voiceless consonants. In general, the results for Spanish listeners indicate that stimuli are heard as voiceless if there is no low-frequency periodicity immediately preceding the consonantal release and as voiced if there is low-frequency periodicity in this time interval (Williams, 1977). English listeners, on the other hand, hear an unaspirated stop consonant if low-frequency periodicity occurs within about 20 msec after consonantal release, and they hear an aspirated stop if there is no low-frequency periodicity within this time interval (Abramson & Lisker, 1970).

If we examine in somewhat greater detail the location of the boundary between the perception of voiced and voiceless consonants or between aspirated and unaspirated consonants while certain acoustic parameters in the vicinity of the stop–consonantal release are manipulated, it becomes apparent that we need to define more precisely what we mean by the detection of low-frequency spectral energy and periodicity and the time interval within which this detection is to occur. One approach to the study of the acoustic correlates of the voiced–voiceless distinction for stop consonants in prestressed position is to define voice-onset time (VOT) as the time interval from the onset of the burst occurring at the consonantal release to the onset of the first glottal pulse, usually determined from visual observation of a spectrogram. In a number of experiments in which listener judgments were obtained to synthetic stimuli with various voice-onset times, it has been shown that the VOT corresponding to the phonetic boundary varies by a few msec depending on such factors as: (1) the frequency of the first formant at the onset of voicing (a low F1 starting frequency tends to shift the judgments in the direction of the voiced category [Stevens & Klatt, 1974]); (2) the fundamental frequency changes that occur immediately following voicing onset (a sharp drop in F0 biases listener judgments toward the voiceless category [Haggard, Ambler, & Callow, 1970]); (3) the transitions of the second and higher formants and the spectral characteristics of the onset burst (stimuli with formant transitions and bursts appropriate for velar consonants tend to require longer VOTs to be heard as voiceless than do those for alveolars and labials [Abramson & Lisker, 1970]); and (4) the intensity of the aspiration noise relative to that of the vowel (stimuli with larger amplitude aspiration tend to be heard as voiceless [Repp, 1979]).

It has been suggested that each of these effects constitutes an additional independent cue for the perception of the voiced–voiceless distinction (Lisker, Liberman, Ericsson, Dechovitz, & Mandler, 1977). An alternative view, however, is to regard each of these factors as contributing to an integrated acoustic property, As noted earlier, we postulate that the signal has the acoustic property corresponding to voicedness if there is low-frequency spectral energy or periodicity in the signal in a specified time window in the vicinity of the consonantal release. Detailed specification of this property requires that we define two events

---

**FIG. 1.10.** Spectrograms of words containing initial voiceless and voiced stop consonants in Spanish (taco, upper left, and doña, upper right) and initial voiceless and voiced fricative consonants in English (zip, lower left, and zip, lower right). The arrow below each spectrogram indicates the point in time at the consonantal release where the spectrum change is estimated to be most rapid. Low-frequency periodicity occurs immediately to the left of this boundary for voiced consonants but not for voiceless consonants.
in time in the auditory representation of the signal, the time at which the consonantal release occurs, and the time at which the onset of low-frequency spectral energy or periodicity occurs.

Given the limited amount of data that are currently available on the response of the auditory system to transient onsets of various kinds, we can only speculate at present on the nature of the auditory representation of these onsets. For example, if the onset of the first formant (F1) is at a low frequency, the auditory system presumably detects the presence of relatively strong spectral energy at low frequencies, and a judgment of the onset of low-frequency periodicity can be made by examining the spectrum of the first glottal pulse. On the other hand, for an F1 onset at higher frequencies, the low-frequency energy (in the vicinity of the fundamental frequency) may not be detected in the first glottal pulse, and it may be necessary to wait until several glottal pulses have been generated before the presence of low-frequency periodicity can be detected on the basis of spectral energy above the frequency of the fundamental. Thus, a low starting frequency of F1 would be perceived as having an earlier onset of low-frequency periodicity, and consequently would be a positive cue for voicing.

Likewise, a rapidly falling F0 at voicing onset would create a signal in which the first few glottal pulses are aperiodic, and the detection of low-frequency periodicity would then be delayed until the rate of F0 change decreased after these initial glottal pulses. Consequently, a rapid initial fall in F0 would create, in effect, a longer time from consonantal release to onset of low-frequency periodicity, and would tend to shift listener responses toward the voiceless category.

A more intense aspiration noise at the onset of a synthetic voiceless stop consonant could have the effect of producing a slightly earlier onset in the auditory representation than would a weaker aspiration noise, thus creating a longer time from consonantal onset to onset of low-frequency periodicity (Repp, 1979).

An explanation for the increased VOT associated with velar consonants relative to alveolars and labials is somewhat more speculative. A possible explanation is that for a velar there is a delay in the rising of F1 following the consonantal release (or at least F1 rises more slowly), and consequently, with F1 starting effectively at a lower frequency than it does with alveolars and labials, a longer VOT is needed to give a voiceless response, following the argument outlined previously.

The point of this discussion is that it may be possible to postulate an integrated property that classifies consonant segments as voiced or voiceless, independent of whether the segment is a fricative or a stop, and independent of the phonetic environment in which the segment occurs. If the integrated property is suitably defined in terms of detection of the presence or absence of low-frequency periodicity in specified regions of the signal, then several seemingly independent acoustic attributes can be regarded as contributing to the integrated property. It is recognized, however, that situations can be created, particularly with synthetic speech, in which a listener must utilize acoustic attributes that are not encompassed within this integrated property—attributes that might be labeled as secondary. Examples are the vowel shortening that precedes a word-final voiceless consonant (Denes, 1955; Klatt, 1973), the shorter duration of consonant closure for an intervocalic voiced consonant relative to a voiceless consonant, or the influence of the duration of the following vowel on the perception of voicing of an initial consonant (Summerfield, 1975). Close examination of natural speech indicates, however, that the primary property is usually present at the same time as these secondary cues.

**DISCUSSION**

We have attempted to show that for a number of phonetic categories it is possible to specify invariant acoustic correlates or properties that are usually independent of the phonetic context in which the segment appears. In particular, we have argued that invariant properties for the phonetic categories of language reside at various sampling points or regions in the acoustic waveform, and we have elaborated the theory of acoustic invariance most completely for place of articulation in stop consonants and nasals. However, given the analysis procedures and theoretical view discussed throughout this paper, we postulate acoustic invariance for other phonetic dimensions as well, including the features consonantal, continuant, nasal, and voicing. An hypothesis concerning the existence of these invariant properties implies that the auditory system is endowed with mechanisms that respond distinctively when a particular property is present in the acoustic stimulus. Property-detecting mechanisms of the type postulated here integrate a set of different acoustic attributes to yield an invariant or distinctive response in spite of the seeming variability of the individual attributes. We have reviewed some evidence in support of this characteristic of the auditory system.

Do these property-detecting mechanisms develop as a consequence of exposure to the sounds of speech, or are they an innate part of the infant's sound-reception system? We postulate that the mechanisms are innate and, further, that it is these kinds of mechanisms that are needed to get the speech-reception system started. When the infant is presented with speech in which particular phonetic segments occur in a variety of environments and which show considerable acoustic variability, these property-detecting mechanisms help the infant to organize the sounds into a relatively small set of classes in spite of this apparent variability. The property-detecting mechanisms of the type described here provide a framework upon which the speech/language system can be organized, and, we would argue, are critical to the acquisition of language. There is, in fact, evidence that infants are equipped with these property-detecting mechanisms and make use of them in perceiving speech and speechlike sounds as early as one
consonants. Adjustments and optimization of these procedures would undoubtedly increase the accuracy of identification, but would probably still leave some errors. Error-free identification based on primary properties is not, however, a rigid requirement, because, as we have just noted, secondary cues always accompany the primary properties. Thus, in situations where the primary properties yield equivocal identification, the secondary cues can be used to make the final identification. If an infant were equipped with primary property detectors, there might be a small proportion of utterances for which the primary invariant properties corresponding to the different phonetic categories would not be detected. If this percentage of utterances is sufficiently small, however, the infant would be able to identify the majority of speech sounds to which he or she is exposed, and consequently, the ability to acquire language would not be greatly impeded. The utterances for which the primary properties are in evidence would provide the child with stimuli from which he or she could learn appropriate secondary cues that could ultimately be utilized in situations where the primary properties are not present. We cannot at this time, however, specify a minimum level of error in the detection of properties in speech that affect the acquisition of language skills.

The role of property-detecting mechanisms in setting up a phonological system must be distinguished, then, from the role of these mechanisms in ongoing adult speech perception. Property-detecting mechanisms play a crucial role in developing an internalized phonological system that is an essential part of the knowledge of a language user. Ongoing speech perception may, however, require recourse to secondary cues or stored templates for lexical items (Klatt, 1979) as well as to the property-detecting mechanisms. Nevertheless, acquisition of these speech-perception strategies depends on the prior ability to make phonetic classifications based on the detection of acoustically invariant and context-independent properties or features.

REFERENCES


1. INVARIENT ACOUSTIC CORRELATES


Summerfield, A. Q. How a full account of segmental perception depends on prosody and vice versa.
Effects of Speaking Rate on Segmental Distinctions

Joanne L. Miller  
Northeastern University

Editors' Comments

As Stevens and Blumstein (Chapter 1, this volume) have ably demonstrated, the search for acoustic invariance underlying phonetic perception continues and, moreover, does so with realistic expectations of success. At least for one phonetic feature (place of articulation), it was possible to specify properties of the signal that remained constant across changes both in phonetic context and in speaker. There are, however, additional sources of potential variation in the speech waveform, including, for example, the syntactic environment, the stress pattern, and the rate of speech. Only by taking into account all such factors will it ultimately be possible to determine whether invariant acoustic properties corresponding to perceived phonetic distinctions exist, or whether there is some context-conditioned variation. The theoretical consequence of the former is that the invariance for perception resides in the signal and presumably could be detected by relatively simple mechanisms. In the latter case, however, perceptual constancy must derive from the processing system, which would require that the mechanisms of perception be quite complex, perhaps even mediational in nature.

The effects of rate of speech present an interesting case in point. Although it was recognized quite early that variation in speaking rate may well be a source of considerable complexity, it was not until recently that systematic studies were undertaken to discover the consequences of changes in rate on the spectral and temporal properties of speech and the manner in which perceptual constancy is achieved. Miller has reviewed the available literature related to these issues and has presented some of her own findings. In addition, she has attempted to specify at least some of the complications that contextual variables, such as rate of speech, create for those who are concerned with modeling the perceptual system for speech.