INVARINANCE AND VARIABILITY IN SPEECH: INTERPRETING ACOUSTIC EVIDENCE

Kenneth N. Stevens
Dept. of Electrical Engineering and Computer Science, Division of Health Sciences and Technology, and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
stevens@speech.mit.edu

ABSTRACT
The acoustic signal is the medium that is probably the most used in research on human speech production and perception and for characterizing the universal structures and units in terms of which the phonological forms of language are stored. Acoustic analysis of the speech signal also provides information which helps to characterize how particular distinctive features in a given language are implemented. Another use of acoustic analysis is to describe and to interpret differences in the way particular individuals produce an utterance. In this paper, examples are given to illustrate how interpretation of the acoustic signal or acoustic-articulatory models might provide insight into (1) possible defining acoustic/perceptual and articulatory correlates for universal distinctive features, (2) enhancements of these defining attributes in various contexts in particular languages, and (3) characteristics of this implementation by particular talkers. At the level of features, the possible defining acoustic and articulatory bases for the features [back] and [continuant] as revealed by acoustic/articulatory models are discussed. At the level of language differences, acoustic consequences of enhancement of the perceptual saliency of the voicing feature and of alveolar stops in English are described. Examples of talker differences that are discussed are: (1) individual characteristics of vocal-fold vibration, and (2) implementation of phrase boundaries.

1. INTRODUCTION
Over the last 50 or 60 years, beginning with the work of Chiba and Kajiyama (1941) and Gunnar Fant (1960), we have learned a great deal about the relation between flow-generated sound sources and articulatory shapes on the one hand and the properties of the acoustic signal that is generated with these configurations on the other hand. Our models of the mechanisms of sound source generation due to airflow through constrictions with yielding walls is less well understood, although in recent years there is some progress in this area. We are still far from having good models of how speakers plan an utterance in terms of a discrete linguistic representation and how this plan is converted into a continuous speech signal. And, models of the reverse process that a listener must use to extract these discrete units from the speech signal are still in a primitive state. A principal reason for this lack of progress is the variability in the attributes of the acoustic signal corresponding to the linguistic units. For a particular type of linguistic unit, whether it is a distinctive feature, a phoneme, a syllable, or a word, this variability depends on the context in which the unit occurs, the speaker, and the state of the speaker.

This paper examines some sources of both invariance and variability, primarily through acoustic data and through models that relate articulatory and acoustic levels. The aim is to begin to
develop some principles that govern the variability. The discussion here will be limited to variability in the acoustic representation of distinctive features, and the influence of the phonetic and prosodic context in which these features occur. Three levels of variability (or invariance) will be discussed: (1) What are the properties of the speech production and perception systems that define the set of phonological contrasts or distinctive features that are used in language? (2) In a given language, how is the acoustic manifestation of these features embellished or enhanced in different contexts? (3) How do individual speakers contribute further sources of variability in these acoustic attributes?

2. ACOUSTIC/PERCEPTUAL/ARTICULATORY BASES FOR DISTINCTIVE FEATURES

There are a number of examples for which manipulation of an articulator through a range of values results in an acoustic parameter that varies in a nonmonotonic way, such as that schematized in Figure 1 (Stevens, 1989). In this articulatory/acoustic relation there are two regions (I and III) in which the acoustic parameter is relatively insensitive to changes in the articulatory parameter, and in which the difference in the acoustic parameter between the two regions is assumed to be perceptually salient. Within region II, the acoustic parameter is sensitive to small changes in articulation, and thus this region could be considered to have a somewhat unstable acoustic output.

![Figure 1. Schematic representation of the relation between an acoustic and an articulatory parameter as the articulatory parameter is manipulated through a range of values.](image)

It is hypothesized that each distinctive feature or possible phonemic contrast in spoken language is grounded in an articulatory/acoustic or articulatory/perceptual relation of this kind. For each distinctive feature there is a defining acoustic or perceptual correlate and a defining articulatory correlate. That is, the value (+ or -) of each distinctive feature is based on the presence or absence of a particular articulatory attribute together with the presence or absence of an acoustic attribute that has a distinctive perceptual consequence. In a sense, then, these defining articulatory and acoustic attributes are invariant. As observed in Section 3 below, however, there are contexts in which some of these defining attributes may be obscured.
We consider here two examples of such “quantal” articulatory/acoustic relations, one related to the distinctive feature [continuant] (a feature that applies to consonants that are [-sonorant]), and the other to the distinctive feature [back].

For the acoustic and articulatory correlates of the feature [continuant], we examine the mechanical, aerodynamic, and acoustic aspects of the production of a narrow constriction in the vocal tract (Stevens, 1998). The articulator forming the constriction has a compliant surface, and this articulator is displaced toward the palatal surface or, in the case of a labial consonant, toward the upper lip. The glottal opening is assumed to remain fixed as the cross-sectional area of the constriction is decreased, and the subglottal pressure is constant. When the cross-sectional area of the constriction is large compared with the glottal opening, the pressure drop across the constriction is small, and the turbulence noise at the constriction is also small. As the constriction area becomes equal to or less than the glottal area, the intraoral pressure increases and creates a force on the articulator that forms the constriction, and this force causes a displacement of the articulator surface. The turbulence noise at the constriction increases to a maximum when the constriction area is about one-half of the glottal area. For this constriction size, the amplitude of the turbulence noise is relatively insensitive to changes in the constriction area. For further reductions in the constriction size, the airflow decreases, and the turbulence noise decreases to zero when a complete closure is formed. Thus there are two stable regions where the noise amplitude is relatively insensitive to the constriction area: one corresponding to a fricative consonant with the feature [+continuant] and one corresponding to the [-continuant] stop consonant. In addition, at the time of release when there is complete closure, it can be shown that there is an abrupt increase in airflow and in turbulence noise. The intraoral air pressure behind the constriction creates a force on the articulator which contributes to the time course of the acoustic onset. Thus the constriction formed by a stop consonant must be made with sufficient force that the intraoral pressure does not cause a release prior to the active downward movement of the articulator forming the closure. When the release occurs from the constriction for the [+continuant] configuration, the onset of turbulence noise is somewhat less abrupt. Thus the defining acoustic correlate of [-continuant] is an abrupt increase in noise amplitude, preceded by silence (at least at high frequencies), whereas for [+continuant] there is continuous noise followed by a rapid decrease in the noise amplitude.

Stop and fricative consonants are members of a larger class of segments with the feature [+consonant]. The defining acoustic attribute for this feature is an abrupt discontinuity in the acoustic signal, usually across a range of frequencies. The defining articulatory attribute is the formation of a constriction in the oral cavity that is sufficiently narrow to create such an acoustic discontinuity. This description applies to both [-sonorant] and [+sonorant] consonants.

As a second example we consider the articulatory/acoustic relation that potentially helps to define the feature [back] for vowels (Hanson and Stevens, 1995; Chi and Sonderegger, 2004). We examine the spectrum prominence corresponding to the second formant as the tongue body is manipulated from a backed to a fronted position, much like the sequence of tongue body positions that normally occur in an utterance of the diphthong [Ai]. During this forward displacement of the tongue body, the second natural frequency F2 of the vocal tract passes through the second natural frequency of the airway below the glottis, which we will call F2T, for the second tracheal resonance. For adult speakers, F2T has been observed to be in the range 1400 to 1600 Hz, and it is relatively constant for a given speaker (Cranen and Boves, 1987;
There is some acoustic coupling between these two natural frequencies $F_2$ and $F_{2T}$, and evidence for this coupling can usually be observed in the acoustic signal when $F_2$ is close to $F_{2T}$. As $F_2$ passes through $F_{2T}$, the spectrum prominence corresponding to $F_2$ often does not move smoothly, but exhibits a discontinuity or abrupt jump in frequency. Thus there tends to be a range of values of $F_2$ within 100 Hz or so where the frequency of the spectrum prominence is unstable. It appears that languages avoid vowels with $F_2$ in or close to this region (Chi and Sonderegger, 2004), and put the $F_2$ their vowels on one side or the other of this region; corresponding to [+back] vowels for lower $F_2$ and [-back] vowels for higher $F_2$. Thus there appears to be a dividing line between two regions with a low $F_2$ for a backed tongue body position and a high $F_2$ for a fronted tongue body position.

Quantal relations of the type schematized in Figure 1 have also been shown for a number of other distinctive features, including features of place for stops and fricatives, the feature [strident], the feature [nasal] and features relating to voicing for obstruents (Stevens, 1989, 2003).

3. LANGUAGE-SPECIFIC IMPLEMENTATION OF FEATURES

In Section 2 above, we gave examples of “quantal” relations between articulatory movements and acoustic consequences of these movements. It is hypothesized that these kinds of relations between articulations and their acoustic outcomes define a universal inventory of distinctive features. A given language selects a subset of these features to specify the contrasts that are used in the language. For speakers of the language, the lexicon is assumed to be represented in terms of sequences of bundles of these distinctive features. In planning an utterance, a speaker selects a sequence of words described in terms of these sequences of feature bundles; the features that define these words, together with some additional prosodic and other information, provide the basis for calculating a sequence of articulatory movements. Although the defining articulations for the features play some role in organizing the articulatory gestures for the utterance, many additional actions, not specified directly by the defining articulations, are introduced. These additional gestures supplement or enhance the perceptual saliency of the defining acoustic correlates. The enhancing gestures are language dependent, and it is hypothesized that they are often introduced to enhance the saliency of a feature when it occurs in a particular context (Keyser and Stevens, 2001; see also Kingston and Diehl, 1995). We consider here a few examples of enhancing gestures, taken mostly from English.

We have proposed elsewhere (Stevens, 2003) that the defining acoustic attribute for place of articulation for stop consonants is the spectrum of the release burst. This spectrum shape is determined essentially by the natural frequency of the portion of the vocal tract anterior to the constriction. This natural frequency may correspond to the second, third, or fourth or higher natural frequency of the entire vocal tract. As is well known, the transitions of the formants in the vowel adjacent to a stop consonant, particularly the $F_2$ transitions, can also provide cues for place of articulation for the consonant. In the case of labial consonants, $F_2$ in the vowel immediately adjacent to the consonant is always equal to or less than $F_2$ in mid vowel, and is always less than $F_2$ for an adjacent alveolar or velar consonant. The starting $F_2$ frequency in the vowel immediately adjacent to a velar consonant is higher than the $F_2$ for an adjacent alveolar or labial consonant when the following vowel is a front vowel, and tends to be intermediate between the $F_2$ for an alveolar and the $F_2$ for a labial consonant when the following vowel is a back vowel (cf. Sussman et al., 1991). In fact, the starting $F_2$ frequency for
Alveolar stop consonants tend to be less variable than it is for labials and velars, and is close to the second subglottal resonance or just above this resonance for the speaker. This F2 frequency for an alveolar consonant is determined by the positioning of the tongue body posterior to the alveolar constriction. It is hypothesized that this positioning of the tongue body for an alveolar stop consonant is selected so that the F2 transition for the alveolar is distinct from that for a labial and a velar. That is, the tongue body positioning for an alveolar consonant in English is adjusted so as to maximize the perceptual contrast of the alveolar with respect to the labial and velar stops. This tongue body positioning, then, is considered to be an enhancing gesture for the alveolar stop consonant.

Another example of an enhancing gesture is the glottal spreading, extending into a following vowel, that is implemented to strengthen the perceptual saliency of the feature [-voice] when a stop consonant occurs in pretonic position. It is assumed that the defining gesture for the feature [-voice] is a stiffening of the vocal folds, and possibly of the vocal-tract walls. (The feature [-voice] has sometimes been called [+stiff vocal folds] or “tense”.) This gesture inhibits glottal vibration, and extends that inhibition into the first few tens of milliseconds following the consonant release. This delay in onset of glottal vibration has been called Voice Onset Time (VOT). Another enhancing gesture for the feature [-voice] that is often implemented for alveolar stop consonants in syllable-final position is glottalization, i.e., adducting the vocal folds so that glottal vibration is inhibited. This kind of glottalization sometimes occurs with other syllable-final voiceless stop consonants, but less frequently than for alveolars. Since it occurs more frequently for alveolar stops, it can also be regarded as an enhancing gesture for alveolars.

A number of other examples of enhancing gestures in English and in other languages have been described in Keyser and Stevens (2001); an example in Mandarin Chinese is given in Stevens et al. (2004).

Most of the gestures that are introduced in order to enhance the perceptual saliency of the defining attribute of a particular distinctive feature also create additional acoustic attributes in the speech signal. An example is the tongue body adjustment that is hypothesized to enhance the saliency of the defining acoustic attribute for alveolar stop consonants. Thus for each distinctive feature this combination of enhancing and defining gestures can create a number of acoustic cues that are available to a listener for identifying the feature.

Of some interest, perhaps, is the observation that, while in some contexts in casual speech the defining gesture for a feature may be obliterated due to overlap with gestures from an adjacent segment, the enhancing gestures remain robust, and are resistant to obliteration as a consequence of gestural overlap. Thus, for example, in a casually produced sequence like “I can’t go”, it may happen that no closure is made with the tongue tip against the hard palate. But much of the F2 transition in the latter part of the vowel /æ/ has the character of an alveolar transition, indicating a fronted tongue body gesture. This fronting of the tongue body is an enhancing gesture, and it remains in the sound to provide a cue for the alveolar place, even though the defining attribute (the spectral characteristic of the noise burst) is not present. A number of examples of this kind can be cited.

This distinction between the defining articulatory gestures for a feature and the enhancing gestures may have some relevance for models for lexical access as they might apply to speech...
recognition systems. Normally a set of acoustic cues is used to identify a particular feature, and knowledge of which cues are more robust to the effects of gestural overlap would be helpful in the selection of the inventory of cues and their relative importance in various contexts.

4. SPEAKER-SPECIFIC IMPLEMENTATION OF FEATURES

As has just been noted, one type of variability in the acoustic manifestation of distinctive features arises because articulatory gestures in addition to the defining gestures for features may be introduced to enhance their perceptual saliency. As is well known, variability is also speaker-dependent. Some of the speaker-dependent influences are a consequence of anatomical differences between speakers, such as vocal-tract and vocal-fold dimensions and mechanical properties of tissue surfaces. Perhaps a more pervasive source of variability comes from learned speech habits; some are due to local dialects and some are unique to the individual. Quantification of these speaker-specific attributes is especially important if one wishes to synthesize the speech of different individuals using a formant synthesizer or an articulatory-based synthesizer. Synthesizing a complete sentence produced by a talker, using copy synthesis with a formant synthesizer, is a painstaking process if every detail is to be correct so that the talker on which the synthesis is based is indistinguishable from the original. This is particularly true of many female talkers, and the synthesis exercise gives one some insight into the kind of detail that is needed to characterize a particular talker (Klatt and Klatt, 1990). Furthermore, it has been shown experimentally that a listener may need to take into account the attributes of the talker when processing the speech of that talker. (See, for example, Mullennix et al., 1989; Nygaard and Pisoni, 1998.)

The examples to be discussed here will focus mostly on the acoustic consequences of individual differences in controlling the vocal folds and other laryngeal structures, particularly in the vicinity of prosodic boundaries, including the effect of changes in lung pressure and transglottal pressure on the sound sources generated by these structures.

One type of speaker variability that has been reasonably well documented is the waveform and spectrum of the glottal source for phonation of stressed vowels (Hanson, 1997; Hanson and Chuang, 1999). These studies have shown that there is a large range of spectrum tilt in the glottal source across both male and female speakers. The difference (in dB) between the amplitude H1 of the first harmonic and the amplitude A3 of the spectrum prominence corresponding to the third formant is in the range 15 to 35 dB for a group of 20-odd adult female speakers and 5 to 23 dB for a group of adult male speakers. In general, greater high-frequency noise is apparent in the spectrum of vowels at high frequencies for speakers with greater spectrum tilt. A similar range of spectrum shapes was observed by Gobl (1989) who measured the characteristics of phonations identified by terms such as modal, breathy, whispery and creaky.

The glottal source, as described by these measures, is also influenced by the prosodic context of the vowel, and the magnitude of some of these influences also appears to be speaker dependent (cf. Klatt and Klatt, 1990; Fant et al., 1991; Stevens, 1994). One of the strongest of these effects occurs in vowels in utterance-final position, where there is often a sharp increase in spectrum tilt. This reduction in high frequency amplitude appears to be caused by a combination of a reduced lung pressure and a spreading of the glottis that occurs near the end of an expiration (Stevens, 1994). Additional modifications of utterance-final vowels include
irregular glottal vibration that occurs occasionally for some speakers and frequently for others (Slifka, 2000).

Some of the cues that are used by listeners in identifying the distinctive features for consonants and vowels involve examining the peaks in the spectrum of the frication noise for a consonant in relation to the spectrum amplitude in the same frequency range for an adjacent vowel, or the spectrum in one frequency region of a vowel in relation to that in another frequency region (for example, in assessing nasalization in a vowel) (Chen, 1997). Cues such as these are sensitive to the spectrum tilt of the vowel, which may be quite different from one talker to another. A listener may need to adjust to these speaker-dependent influences in extracting and weighting the cues for particular features. This adjustment may be in place when listening to familiar speakers, but an adaptation period may be necessary for unfamiliar speakers.

Further investigation of these and other individual differences is clearly relevant to the development of models for human lexical access. And in applications such as speech recognition, speaker recognition, and speech synthesis, there is much to be gained by examining the details of how speakers are different.

5. SUMMARY
We have attempted to classify some of the sources of variability in the acoustic manifestation of different features when they occur in various phonetic, syllabic, lexical, and prosodic contexts, and with different talkers. The point of view we are taking as a starting point is that there are invariant or defining acoustic and articulatory correlates for each distinctive feature. A distinctive feature, in a sense, owes its existence to a particular convergence of an articulatory action giving rise to an acoustic and perceptual result that is relatively insensitive to perturbations in articulation. Configurations leading to an acoustic result that is sensitive to small perturbations in articulation are avoided.

Additional articulatory gestures are usually added to the defining gestures, and the acoustic consequences of these additional gestures are to enhance or to make more salient the identification of the feature. The enhancing gestures that are added may depend on the context in which the feature occurs. In some contexts and in some speech styles the defining gestures may be absent from the speech stream due to overlap with gestures for an adjacent segment. However, enhancement gestures are observed to be resistant to the influence of such processes. Enhancing gestures can lead to substantial variability in the acoustic attributes that provide evidence for a distinctive feature.

When an utterance is produced by two different talkers, there will be differences in the acoustic parameters that describe the two utterances. Although some normalization, such as scaling of formant frequencies, is possible, there may be differences in vocal tract and laryngeal dimensions and properties of surface tissues that influence some of the acoustic attributes that are used as cues to distinctive features. Thus the strength of some cues that contribute to the identification of a feature may vary from one talker to another.

Not discussed in this paper is the contribution of additive noise or reverberation to variability in the acoustic cues to the distinctive features. If we can classify the various acoustic cues in
terms of their spectral and temporal properties, it should be possible in the future to determine how to weight these cues based on knowledge of the intruding noise or reverberation.

ACKNOWLEDGEMENTS
Conversations and collaboration with a number of colleagues and students is gratefully acknowledged. This research was supported in part by grants DC00075 and DC02978 from the National Institutes of Health.

REFERENCES