Estonia, USSR, 1987, Vol. 4.

RELATIONAL PROPERTIES AS PERCEPTUAL CORRELATES OF PHONETIC FEATURES

KENNETH N. STEVENS

Research Laboratory of Electronics and
Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge MA 02139 USA

Abstract

The view taken in this paper is that the apparent variability of the acoustic correlates of phonetic features is reduced if these correlates are described in relational terms. That is, the acoustic properties are specified in relation to the spectral and temporal context in which they occur. We present a number of examples of vocalic and consonantal features for which there appear to be advantages in specifying the acoustic correlates in relational terms.

1. Introduction

Some acoustic attributes of speech sounds produced by different speakers and in different contexts show a great deal of variability. This variability is especially evident if the properties are specified in terms of absolute measurements such as the frequencies of spectral prominences, times between acoustic events, or amplitudes of particular regions in the frequency-time representation of speech. The view taken in this paper is that much of this variability tends to disappear if the properties that constitute acoustic correlates of the phonetic features are defined in a relational sense. The term *relational* is taken to mean that an attribute at a particular frequency and time in the speech stream is specified in relation to the context in frequency and in time in which this attribute occurs.

2. Three frequency regions

In what follows, we shall consider examples of acoustic properties of classes of speech sounds that appear to be described most naturally in relational terms. The descriptions are usually in terms of relations between spectral prominences or periodicities at a particular point in time, or relations between spectrum amplitudes in particular frequency ranges in adjacent time regions. These relational properties will refer to spectral characteristics that are observed within broad regions of the audible frequency spectrum. In identifying the properties in different broad frequency regions, we are implying that the sound can be processed in the auditory system in different ways in these frequency regions. That is, we are assuming that the capabilities of the auditory system for processing sound in these frequency regions may be different, although we recognize that some aspects of the processing are common to all frequency regions. The edges of these frequency regions are not well defined and there may be some overlap between the regions. Before discussing in detail examples of the various relational properties, we will review briefly what the different basic frequency regions are, and what are some of the bases for selecting these particular regions.

The lowest frequency region extends up to about 800 Hz, and usually encompasses the frequency range of the first formant for adults. Within this frequency range, the frequency resolution as defined by the critical bandwidths, or by the bandwidths of the psychophysical or physiological tuning curves, tends to be independent of frequency, and is less than 100 Hz. The shape of the tuning curves is more or less symmetrical in this range and there are no low-frequency tails on the tuning curves [1]. The time resolution is poorer than it is at higher frequencies, as expected on the basis of the narrower frequency resolution.

Above the low-frequency region, the critical bandwidths increase with frequency, and the tuning curves are characterized by low-frequency tails. That is, the auditory filters in this frequency range can respond to low-frequency energy in the sound if it is sufficiently large. The time resolution at these higher frequencies is good, and the auditory filters respond within a millisecond or two to an abrupt increase in amplitude.

This high-frequency region is divided into two parts. The lower part, which we call the midfrequency region, encompasses the normal range of variation of the second and third formants for vowels. The bandwidths of these formants for vowels tend to be narrower than the bandwidths of the auditory tuning curves in this midfequency region. Within both the low- and midfrequency ranges, there tends to be synchrony of firing of individual auditory-nerve fibers

to pure tones and to the frequencies of the first three formants for fibers whose characteristic frequencies are in the vicinity of one of these frequencies [2, 3, 4].

At much higher frequencies, probably above 3000 or 3500 Hz, the frequency resolution is much poorer than in the mid- and low-frequency range. The synchrony of firings of auditory-nerve fibers, either at their own characteristic frequencies or at the frequencies of nearby spectral prominences, is much less evident in this frequency range [2]. The frequencies of the formants for vowels do not contribute significantly to vowel identification at these frequencies. Spectral energy in this frequency range contributes primarily to the identification of particular consonant features, and the important aspect of the auditory-nerve response is probably its strength rather than the temporal characteristics of the nerve firings [5]. There are only about 5 critical bands relevant to speech processing in this high-frequency range, out of a total of about 22 such bands when they are spaced in such a way that one critical band separates adjacent filters.

3. Some relational properties for vowels

3.1 $F0$ contours

In describing the fundamental frequency ($F0$) variations in tone languages, it is common to use labels such as high, mid, and low tones. These labels are generally considered, however, to apply in a relational sense. A vowel with a high tone, for example, is regarded as having, at some point within the vowel, a maximum in $F0$ in relation to the $F0$ in nearby regions, either within the same vowel or within adjacent sonorant regions. Thus in a monosyllabic word with a high tone, the $F0$ is higher near the midpoint of the vocalic nucleus than it is near the beginning and end of the vowel. The value of $F0$ depends, however, on the individual talker, as well as the position of the vowel in the sentence unless the word is spoken in isolation. A listener presumably interprets this concave downward contour as being qualitatively different from a contour that is falling or is concave upward. When interpreting $F0$ contours, a listener seems to be examining $F0$ at one point in time in relation to $F0$ at other points—that is, the listener is making use of relational properties. Experiments have shown that the range of frequency variation of individual tones in a tone language (Mandarin) can be reduced drastically without modifying their perceptual integrity as long as the contour shape is preserved, i.e., as long as the relational aspects of the contour are maintained [6]. Thus $F0$ contours provide examples of acoustic properties that need to be defined in relational terms.

3.2 Relations between formant frequencies

Experiments of Chistovich and her colleagues [7] have suggested that some aspects of the auditory processing of a vowellike sound with two spectral prominences are qualitatively different depending on the frequency spacing between the prominences. This conclusion is based on experiments in which subjects are asked to match a vowellike sound with an adjustable single spectral prominence to be similar in quality to a two-prominence test stimulus. When the spacing between the two prominences is less than about 3-4 Bark, listeners tend to adjust the frequency of the matching stimulus to be between the frequencies of the two prominences of the test stimulus. For a greater spacing, a best match is obtained when the frequency of the matching prominence is equal to one or other of the prominences of the test stimulus. Our own (unpublished) experiments, using slightly different stimulus characteristics, have led to results that are consistent with those of Chistovich et al. Syrdal [8] has applied this concept of a critical formant spacing to an examination of the analysis and perception of vowels in English. She has shown that back vowels tend to have an $F2 - F1$ difference that is within or close to this critical range, whereas for front vowels it is the $F3 - F2$ spacing that is less than 3-4 Bark. One can conclude that vowel perception is based on a relational property among the spectral prominences for the vowel: the second formant in relation to the first, or the second formant in relation to higher spectral prominences—usually $F3$.

3.3 Breathy vowels

A different kind of relational property has been shown to distinguish between breathy and nonbreathy vowels in languages that use this feature contrastively. The amplitude of the fundamental component of the vowel in relation to the spectrum amplitude of the first formant is greater for breathy than for nonbreathy vowels [9, 10]. The perceptual relevance of property has been shown by Bickley [10]. While a quantitative specification of this property, valid across all vowel heights, has yet to be developed, it is clear that it is the relation between the amplitude of the fundamental component and that of higher components that contributes to the identification of the breathiness feature.

3.4 Formant contours

When a vowel is produced in the context of consonants, the formant frequencies usually vary with time throughout the vowel, and there may not be a time interval in which the formants remain relatively fixed. Several experiments have compared the identification of vowels characterized by time-varying formant frequencies and those with steady formants [11, 12, 13]. A general outcome of these experiments is that the identification of a vowel with time-

varying formant frequencies cannot be predicted from the formant frequencies sampled at a particular point within the vowel such as the midpoint or the point where one of the formants reaches a maximum or a minimum value. Identification of the vowel is dependent on the entire contour. Preliminary data of Huang [12] and of Di Benedetto [13] suggest that the processing may be different for the first formant (i.e., within the low-frequency region defined above) and for the second formant. When the first formant ($F1$) has a concave-downward shape, the equivalent vowel height is that of a steady vowel with a lower $F1$ than the maximum $F1$ on the contour, i.e., the listener extracts some kind of average frequency from the contour. The $F2$ contour tends to be identified with a steady vowel for which $F2$ is beyond the maximum or minimum $F2$ on the contour, particularly when the contour is concave upward. Again we can conclude that the listener interprets the extreme values of the formants in relation to the formant contour preceding and following these maximum or minimum values.

4. Some relational properties for consonants

4.1 Acoustic correlate of sonorancy

A consonantal segment with the feature [+sonorant] is characterized by continuity of the spectrum amplitude at low frequencies in the region of the first and second harmonics—a continuity of amplitude that extends into an adjacent vowel without substantial change. This property is a consequence of the fact that there is essentially no obstruction to the airflow in the airways above the larynx. The vocal folds can therefore continue to vibrate in a normal manner, so that the low-frequency amplitude in the radiated sound remains unchanged. The perceptual salience of this low-frequency continuity has been examined in a limited way through experiments in which the spectrum amplitude at low frequencies was manipulated in the consonant region of a synthetic consonant-vowel stimulus (unpublished research of S.E. Blumstein and K.N. Stevens). This manipulation resulted in a continuum in which the consonant was heard as a prevoiced stop at one end and a nasal consonant (either [m] or [n] in two different continua) at the other. The [+sonorant] nasal consonant was heard when the amplitude of the lowest spectral prominence in the consonant was equal to or greater than the amplitude in the same frequency region in in the adjacent vowel. When the low-frequency energy in the consonant region was weaker, listeners tended to hear the consonant as a stop rather than a nasal. The correlate of sonorancy appears to involve a relation between the low-frequency spectrum amplitudes in the consonant and vowel regions.

4.2 Acoustic correlate of anteriority

A consonant with the feature [-anterior] is produced with a constriction at some point along the vocal tract posterior to the alveolar ridge. When the constriction is located in this region, the lowest front-cavity resonance of the vocal tract will usually be either $F2$ or $F3$. Such a consonant will have a spectrum that contains one or more spectral prominences in the midfrequency region as defined in section 2 above. The acoustic correlate of [-anterior] is that the spectrum amplitude of at least one of these prominences is approximately the same as the amplitude of the spectral prominence at the same frequency in the adjacent vowel. That is, there is continuity of the spectrum amplitude for one or more spectral prominences at the release of the consonant into the adjacent vowel. The evidence for this property relating midfrequency spectrum amplitudes before and after the consonant release is derived from acoustic analysis data and from perceptual experiments with synthetic consonant-vowel stimuli [14, 15, 16].

In separate perceptual experiments with stop and fricative consonants, the amplitude of a spectral prominence in the consonant region was manipulated and listeners were asked to identify the consonant. Results for the fricative experiment have been described previously [17]. For the fricatives, the stimuli at the ends of the continuum were /ʒ/ ([-anterior]) and /s/ ([+anterior]), whereas for the stops they were /g/ and /d/ or /k/ and /t/ [16]. In all cases, the [-anterior] consonant was heard when the amplitude of the midfrequency spectral prominence in the frication noise was equal to or greater than the corresponding peak at the vocal onset, whereas the [+anterior] cognate was heard when the noise peak was weaker. It is suggested, then, that the acoustic correlate of the feature [-anterior] can be described as a property indicating the relation between spectrum amplitudes in the frication noise and in the vowel.

4.3 Acoustic correlate of coronality

Consonants that are classified as [+coronal] are often described as having a spectrum that rises with increasing frequency, so that there is substantial energy in the spectrum in the higher-frequency range (as defined in section 2 above) [18, 19]. Acoustic analysis of a variety of stop, nasal, and fricative consonants, together with some limited perceptual experiments [15, 17, 20], have led to a revised formulation of this property in relational terms. The property specifies that the high-frequency spectrum amplitude in the vicinity of the consonantal release should exceed the high-frequency amplitude observed after the onset of the following vowel. The spectrum at the release is regarded as a spectrum that would be represented in the peripheral auditory system in the sense that it is governed by the

frequency resolution and adaptation characteristics of the peripheral auditory system.

This specification of the acoustic correlate of [+coronal] as a relational property involving changes in high-frequency energy is supported in part by the results of perceptual experiments in which the high-frequency spectrum amplitude of a burst is manipulated in synthetic consonant-vowel syllables [20]. Except for this high-frequency energy, the spectrum of the burst had no major prominences. The consonant was identified as an alveolar stop (i.e., having the feature [+coronal]) when the high-frequency amplitude of the burst was within about 5 dB of the high-frequency amplitude at the onset of the vowel (assuming neutral formant transitions). If one takes into account the overshoot in the response at the auditory nerve following an abrupt onset, this burst amplitude might be expected to yield a high-frequency response at onset that equals or exceeds the response in the following vowel. When the high-frequency spectrum amplitude of the burst was weaker, the consonant was heard as a labial. We conclude that the feature [+coronal] is based on a relational property, this time involving the high-frequency spectrum amplitude in the consonant region in relation to that in the adjacent vowel.

5. Concluding remarks

The examples that have been presented here could be expanded to include a variety of other acoustic properties. These examples provide support for a view that some advantage might be gained if the acoustic correlates of phonetic features were expressed in relational terms, particularly relations between acoustic events at nearby points in time. For the most part, these relational properties are probably manifested in the listener's auditory pathway at a level somewhat higher than the level of the auditory nerve. The representation of the speech stream at the level of the auditory nerve, however, provides a first step in the series of transformations that ultimately lead to the relational properties from which phonetic features can be derived.

6. Acknowledgements

The preparation of this paper was supported in part by grants from the National Institute of Neurological and Communicative Disorders and Stroke and the National Science Foundation.

7. References

1. N.Y.S. Kiang, T. Watanabe, E.C. Thomas, and L.F. Clark. **Discharge patterns of single fibers in the cat's auditory nerve.** Cambridge MA: MIT Press (1965).

2. D.H. Johnson. *The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones.* **J. Acoust. Soc. Am., 68,** 1115-1122 (1980).

3. E.D. Young and M.B. Sachs. *Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers.* **J. Acoust. Soc. Am, 66,** 1381-1403 (1979).

4. B. Delgutte and N.Y.S. Kiang. *Speech coding in the auditory nerve: I. Vowel-like sounds.* **J. Acoust. Soc. Am., 75,** 866-878 (1984).

5. B. Delgutte and N.Y.S. Kiang. *Speech coding in the auditory nerve: III. Voiceless fricative consonants.* **J. Acoust. Soc. Am., 75,** 887-896 (1984).

6. V.W. Zue. Unpublished research.

7. L.A. Chistovich and V.V. Lublinskaya. *The "center of gravity" effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli.* **Hearing Research, 1,** 185-195 (1979).

8. A.K. Syrdal. *Aspects of a model of the auditory representation of American English vowels.* **Speech Communication, 4,** 121-135 (1985).

9. P. Ladefoged. *The linguistic use of different phonation types.* In **Vocal Fold Physiology: Contemporary Research and Clinical Issues.** D. Bless and J. Abbs, eds. San Diego: College-Hill Press, 351-360 (1983).

10. C. Bickley. *Acoustic analysis and perception of breathy vowels.* **Speech Communication Group Working Papers I.** MIT, Cambridge MA , 71-81 (1982).

11. B. Lindblom and M. Studdert-Kennedy. *On the role of formant transitions in vowel identification.* **J. Acoust. Soc.Am., 42,** 830-843 (1967).

12. C.B. Huang. **Perceptual correlates of the tense/-lax distinction in General American English.** SM thesis, MIT, Cambridge MA (1985).

13. M.G. Di Benedetto. **An acoustical and perceptual study on vowel height.** Ph.D. thesis, University of Rome (1987).