MODELS OF PHONETIC RECOGNITION II: AN APPROACH TO FEATURE-BASED RECOGNITION

K.N. Stevens

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

Abstract An approach to speech recognition is proposed in which phonetic features are identified as acoustic properties in the speech signal, and lexical items are accessed directly without explicitly labeling phonetic segments. A possible advantage of such an approach is that a feature representation shows minimal modification as a consequence of the deletions and assimilation phenomena that occur in natural speech. Problems of determining acoustic correlates of features and of representing lexical items in terms of features are discussed.

In this paper I would like to argue that there are advantages to be gained by using phonetic features as primary units for identifying words. I hope to show that variability that occurs from speaker to speaker and from context to context can be taken into account in a natural way if features are used for representing utterances and if they form the building blocks for larger units by means of which utterances are identified.

Before discussing some of the advantages of features, and the structure of a speech recognition procedure based on features, let me first review some of the basic ideas underlying the concept of features.

Features and their Acoustic Correlates

A feature is a minimum unit in terms of which lexical items are represented (Jakobson, Fant, and Halle, 1963; Chomsky and Halle, 1968). Words that have different meaning (except for homonyms) have a different representation in terms of binary features. Thus, for example, the words *mill* and *bill* are differentiated on the basis of one of the features that characterize the initial segment -- in this case the feature *sonorant*. (Other features, such as *nasal*, may also play a role in this distinction. This concept of redundancy in the feature representation is discussed below.) It appears that about 20 features are needed to perform this function in language. Each lexical item is assumed to be represented in the mind of a speaker/listener in terms of patterns of features (with some further structure to this pattern).

Associated with each feature there is an acoustic correlate. This acoustic correlate, or property, is assumed to give rise to a pattern of response in the auditory system that is qualitatively different or distinct from the response pattern associated with other features. The property associated with each feature can be present in the sound with different degrees of strength. Features have articulatory correlates as well as acoustic or perceptual correlates, but in this paper our principal concern is with the acoustic correlates.

The acoustic properties that qualify as correlates of phonetic features tend to be *relational* and not absolute. Thus, for example, acoustic parameters such as the overall intensity of a component of the signal or the frequency of a particular spectral prominence, divided arbitrarily into two classes by a fixed intensity or frequency, would not qualify as the bases for the acoustic correlates of phonetic features. Parameters such as these show large interspeaker differences for the same utterance. Furthermore, there is no evidence to indicate a natural perceptual boundary or qualitative shift in the pattern of auditory response at an absolute intensity or an absolute frequency. On the other hand a property such as the frequency of one formant in relation to another could lead to qualitatively different auditory response pattern depending on whether the spacing between the two formants was greater or less than a critical value. (See, for example, Chistovich, Sheikin, and Lublinskaja, 1979.) Through proper selection of properties that describe spectral relationships, these properties can be speaker independent, since they do not depend on the speaker's vocal tract length or average fundamental frequency. Properties defining features can also be relational in the time domain. Thus, for example, a qualitatively different auditory pattern could result from an abrupt rise in spectrum amplitude in a broad frequency region as opposed to an abrupt fall in amplitude. In this case the relevant property is relational in the same sense that the amplitudes of spectral components at one time are interpreted in relation to the amplitudes of these components at an adjacent time.

There is a tendency for groups of features to be implemented more or less simultaneously, and consequently these features are naturally organized into segments. For example, within 10-20 msec of the release of a stop consonant, the sound contains properties identifying the features *continuant* and *sonorant* as well as the features related to place of articulation. In general, however, each feature is not specified for every segment. (For a discussion, see Halle, 1985, and references cited therein.) Sometimes just one feature might show a change at a point in time at which no other feature shows evidence for a change (e.g., the feature *continuant* in the initial consonant in /ča/, or the feature *high* in the vowel in /se/). On the other hand, some features may be defined for some segments, with no specification of these features for intervening segments. Thus, for example, in the word *banana*, the features indicating backness and high pitch are specified only on the second vowel and not in the other vowels, which are unstressed and reduced.

An important characteristic of the representation of an utterance in terms of features is that the representation usually has more features than the minimum number that are needed to distinguish the utterance from possible competitors. That is, there is *redundancy* in the feature representation. A consequence of this redundancy is that there is room for variability in the acoustic representation of an utterance. Not all features need to be marked in the signal, and the acoustic properties associated with these features can be present with different degrees of strength (Stevens, Keyser, and Kawasaki, 1986).

Situations often arise in which one or more features of one segment spread to a nearby segment, resulting in a change of some features of the segment, a specification of features that were previously unspecified, or even a deletion of the segment. Examples are: in *miss you* /s/ becomes [š], taking the palatal feature of the adjacent [j]; in *at the*, the sequence /t#ð/ can become [t̪], i.e., a dental *t*; in *sit close* in rapid speech, /t/ can lose its place features but retain the stop feature; in *tree*, the initial /t/ takes on the retroflex feature of the next segment. In many cases the spreading of features is allowed because there is redundancy in the feature description of a segment, and changing one or more features does not lead to misidentification of a lexical item. These assimilation phenomena often occur when there are two or more adjacent consonants, and they can occur within words or at the boundaries between morphemes or words. They appear to follow certain general principles, and linguists are working on models of feature organization that capture these principles in a natural way. (See, for example, Clements, 1985 and Halle, 1985.) The point is, however, that if the feature is used as a basic unit of representation these sources of variability in the speech signal can be accounted for in a rather natural manner.

Features, Variability, and Invariance

From the above discussion we can identify two principal sources of variability when an utterance such as a word is produced by different speakers with different speaking styles and in various contexts. One kind of variability arises mainly because different speakers have different vocal-tract sizes and shapes, and because talkers may use various speaking rates. This source of variability can be accounted for by proper specification of the acoustic correlates of the features. In particular, the acoustic properties should be relational so that they are insensitive to vocal tract size and speaking rate. Considerable progress has been made in specifying these acoustic properties, but much work remains to be done in this area. This research can be guided by an understanding of the psychophysics and physiology of hearing, and of theories of speech production.

The second source of variability arises because a speaker may modify the feature description that underlies an utterance or may make adjustments in the strength with which a feature is implemented. In some situations this modification is dictated by rules specific to the language, and in other cases the changes are optional and are influenced by speaking style. These modifications in the feature description appear to be capable of specification in terms of spreading of features across segments, such that features in one segment are changed as a consequence of particular feature values in an adjacent segment. The spreading can lead to changes in or elimination of one feature or groups of features.

Another source of interspeaker variability, which we shall not consider here, arises when different dialects are involved. Usually, however, it is possible to describe the phonetic differences between dialects in terms of a small set of rules operating on features.

Toward a Model for Feature-Based Recognition

How might a listener make use of features in decoding an utterance given the acoustic signal? Or, given the theme of this conference, how might we implement these ideas in a speech recognition system? The point of view we take here is that there are two stages to this process. The first stage is to identify the properties in the signal from which estimates of the features are made, and the second stage is to identify the lexical items from these properties. We imagine that testing for each property is carried out continuously through the speech signal. Most of the properties achieve maximum values or degrees of strength at particular points in time in the speech signal. These peak values of the properties define events in time within the signal. Some properties, however, maintain approximately constant strength over longer time intervals, and thus are identified with regions of time rather than with events in time. An example is the feature *voiced*, for which the acoustic correlate is the presence of low-frequency periodicity. (Other features are often active, and hence other properties are often present in the signal, when the feature *voiced* is implemented in English.) Also, there are some interrelationships between properties so that some properties cannot be extracted unless other properties are present. Thus the continuous speech signal is characterized by a series of signal streams, one corresponding to each property that is the acoustic correlate of a feature. For the most part, these signal streams consist of marks indicating brief time intervals or events, and these marks are labeled with the strength of the property. There is a tendency for these events corresponding to some groups of features to be approximately aligned, for example in the vicinity of a stop-consonant release.

We shall not discuss in detail the next stage of processing in which lexical items are accessed on the basis of these signal streams. Probably the most difficult and important problem to be solved is to determine a proper structure for the lexicon so that it can be accessed from these signal streams (or modified versions of these signals), given that these signals reflect the effects of redundancies and spreading phenomena of the type discussed above. There are several requirements for this structure: (1) in the feature representation, the notion that some features are redundant should be indicated in some manner; (2) while some features are aligned within the same segment, the representation should be structured to allow some flexibility in this alignment, possibly along lines of the tiered structure proposed by phonologists; (3) features or feature groups that are susceptible to spreading should be indicated so that assimilation phenomena may be accounted for in a natural manner.

[Supported in part by grants from the National Institute of Neurological and Communicative Disorders and Stroke and from the National Science Foundation.]

References

Chistovich, L.A., Sheikin, R.L., and Lublinskaja, V.V. (1970) "Centres of gravity and spectral peaks as the determinants of vowel quality," in B. Lindblom and S. Ohman (eds.), *Frontiers of Speech Communication Research*, Academic Press, London, pp. 143-157.

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

Clements, G.N. (1985) "The geometry of phonological features," *Phonology Yearbook*, *Vol. 2*, Cambridge University Press, Cambridge.

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

Halle, M. (1985) "Speculations about the representation of words in memory," in V.A. Fromkin (ed.), *Phonetic Linguistics*, Academic Press, New York, pp. 101-114.

Stevens, K.N., Keyser, S.J., and Kawasaki, H. (1986) "Toward a phonetic and phonological theory of redundant features," in J. Perkell and D.H. Klatt (eds.), *Variability and Invariance in Speech Processes*, Erlbaum, Hillsdale, NJ.