2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

IMPLEMENTATION OF A MODEL FOR LEXICAL ACCESS BASED ON FEATURES

Kenneth N. Stevens, Sharon Y. Manuel, Stefanie Shattuck-Hufnagel, and Sharlene Liu*

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge MA 02139
*Also at Dept. of Communication Disorders & Sciences, Wayne State University, Detroit, MI 48202

ABSTRACT

A feature-based model for accessing words in a lexicon from measurements on the speech signal is proposed, and the implementation of some components of the model is described. Words are represented in a lexicon in terms of hierarchies of distinctive features arranged in matrix form. Acoustic measurements are made on an utterance in order to estimate the pattern of features implemented by the speaker, and access to lexical items is based on these hypothesized features. The acoustic measurements are of two kinds: (1) they identify and classify critical locations in the signal, and (2) they indicate the activity of articulatory structures in the vicinity of these critical locations. Proposed procedures for representing the lexicon and for accessing words from the lexicon have the potential of being relatively insensitive to speaking style and to context.

I. INTRODUCTION

This paper is a progress report on the development and implementation of a model for lexical access based on a representation of lexical items in terms of hierarchies of distinctive features arranged in matrix form [1]. The features for each segment in the lexical representation are organized into classes, and particular types of acoustic properties must be extracted from the signal in order to identify the features within each class.

The model specifies a set of acoustic measurements that are made on the acoustic signal for an utterance, and from these measurements, estimates are made of the features implemented by the speaker. From these features, the sequence of lexical items is derived.

Three issues need to be considered in developing this model: (1) how are lexical items represented, (2) what transformation of the signal is required to derive the features implemented by the speaker, and (3) by what process are lexical items accessed given these estimates of the features.

Before discussing the details of these three issues, we motivate our lexical access model by reviewing the possible advantages of a feature-based approach to lexical access. By features, we mean something akin to the hierarchies of distinctive features proposed by linguists. In this terminology, a feature is an abstract entity that represents and distinguishes among lexical items. The acoustic manifestation of a feature is indirect, and, as we shall observe, there may be a set of acoustic properties that contribute to the identification of a given feature. However the acoustic measurements that lead to the identification of each feature are clearly defined and well-motivated in acoustic theory.

There are two principal reasons for using hierarchies of features to represent lexical items. One motivation is that, when a word is embedded in a sentence, the modifications that are likely to occur can usually be described in terms of relatively simple set of processes that modify individual features or complexes of features that are describable in a natural way using the hierarchy. A second potential advantage of features lies in the nature of the process of extracting information about the features from the acoustic signal. The analysis procedures can be organized in a hierarchical fashion such that at each step in the process only a limited set of acoustic measurements needs to be made. The analysis can be modularized so that at any stage only relevant acoustic information is being sought.

II. THE LEXICAL REPRESENTATION

The items in the lexicon are described in terms of hierarchies of binary distinctive features. The feature hierarchies are selected on the basis of two criteria. One criterion is that there must be a well-defined set of acoustic measurements from which each feature can be identified if that feature is implemented in an utterance. A second criterion is that any context-induced change that occurs in the production of an utterance should be capable of being described in terms of simple modifications of individual features or groups of features. The features we have selected are slight modifications of the features that have been used in recent years by phonologists and others [2,3,]. The lexical representation of a word specifies the features that would be implemented if the word were uttered in a citation form. When the word is produced in context or in a casual style, there is the potential for certain of the features in the lexical representation to be modified. Our present strategy for dealing with this type of variability is to mark these features as being modifiable.

An example of the lexical representation of a word is given in Table 1. This example contains a partial list of the features that are used for representing all items. (We currently have a lexicon of 250 words in this form.) The representation of a word is basically a matrix of feature values, with the columns representing segments. The features for each item are organized into three categories. The first category consists of *articulator-free* or stricture features. The feature [+consonantal] designates a consonant that is produced with a narrow constriction in the vocal tract, leading to an abrupt discontinuity in the spectrum of the sound when this constriction is formed or is released. Several additional articulator-free features specify whether or not the constriction forms a complete closure (±continuant) or whether or not there is pressure buildup behind the constriction (±sonorant). Each of these features has well-defined acous-

Table 1 Partial list of distinctive features used to classify lexical items (left column). Values of features representing the word **ten** are given in the remaining columns. Feature values marked with M are subject to potential modification when the word is pronounced in certain contexts.

	t	ɛ	n
consonantal	+	-	+
sonorant	-		+
continuant	-		-
lateral			-
strident			
lips			
round			
tongue blade	+		+M
anterior	+		+M
distributed	-		-M
tongue dorsum		+	
high		-	
low		-	
back		-	
spread glottis	+		
constricted glottis	-		
stiff vocal folds	+		
slack vocal folds	-		
nasal			+

tic correlates. Other features (lateral, strident) can be used to further specify the class of consonant. For vowels, which are [-consonantal], no other articulator-free features need to be specified [4].

A second category of features in Table 1 consists of *articulator-bound features that indicate the major articulator* involved in forming the segment, and how that articulator is to be placed or configured. Thus, for example, the initial consonant of the word in Table 1 is produced with the tongue blade, and the features [+anterior, -distributed] specify the placement and shaping of the tongue blade. For vowels, the tongue body is always specified as the major articulator.

A third category of features for each segment lists *articulator-bound features that specify the adjustment of articulators other than the major articulator* noted above. In the example in Table 1, the initial consonant is voiceless aspirated, and hence the laryngeal adjustment is specified by [+spread glottis] and [+stiff vocal folds]. In the case of the final consonant, the feature [+nasal] indicates that the soft palate is lowered.

The organization of the features into classes in the manner shown in Table 1 leaves many entries in the table blank. For example, for segments that are [-consonantal] no other articulator-free features are relevant since there are no contrasts involving these features. Or, for most consonants produced by the tongue blade in English, no specification of tongue-body features is required.

Another point with regard to the lexical representation is that the features for a given consonantal segment (such as a segment represented by the symbol /t/) may be characterized by a somewhat different set of values when it is in postvocalic position in the syllable coda than when it is in prevocalic position. Thus the initial and final segments in the word **tot** have somewhat different specifications for certain laryngeal features.

As noted above, some of the features in the lexical representation of a word may be modified when the word is produced in certain contexts, or when it is produced in a casual style. In the case of the word **ten** in Table 1, the features for the prestressed stop consonant /t/ and the stressed vowel /ɛ/ are rarely if ever modified. However, certain features of the syllable-final /n/ are susceptible to modification in some contexts. For example, in the casually spoken sequence **ten cats**, the nasal consonant may not be produced with a tongue-blade closure, but rather the velar closure for the following consonant may be assimilated to the nasal consonant. This is an example of a general rule involving syllable-final consonants that are [-continuant] and are produced with the tongue-blade. In this case, then, the feature [+blade] (together with [+anterior] and [-distributed]) is represented with M, indicating susceptibility to modification. Needless to say, a major task in the implementation of the lexicon is to indicate the features that are prone to modification. Features of segments in the vicinity of unstressed vowels are particularly susceptible to change and sometimes to deletion [5].

III. ACOUSTIC ANALYSIS LEADING TO FEATURE HYPOTHESES

We consider now the process of extracting acoustic properties from the acoustic waveform of an utterance in order to make estimates of which features were implemented by the speaker. Some initial components of the acoustic analysis procedures that we will describe have been implemented and have undergone preliminary evaluation. For others, the proposed acoustic measurements have not been formalized into a set of algorithms. Rather, they are based on theoretical models and on data derived from analysis of a number of utterances.

In general, the analysis process involves several steps. First, the signal is scanned to determine the locations of acoustic events or landmarks, and to classify these acoustic events. This initial step has two functions: (1) it indicates that a segment was implemented and identifies the articulator-free features for the segment, and (2) it identifies a region in the signal around which additional acoustic observations are to be made in order to estimate the articulator-bound features for the segment. This process of estimating the articulator-bound features is the second step in the acoustic analysis. The particular measurements that are required to estimate the articulator-bound features depend on the articulator-free features.

The principles governing the selection of acoustic properties that lead to hypotheses about the features stem from the observation that for consonants, the context or the speech style can introduce significant variability into the acoustic manifestation of a certain features for rather small modifications in the articulation. Thus, for example, the extent of vocal-fold vibration in the constricted interval for a voiced obstruent consonant in syllable-final position can be quite variable depending on small adjustments in the glottal configuration. Or, the spectrum and time course of a burst at the release of a particular stop consonant is dependent on the rate of release and on the following vowel. The processing of the signal, therefore, should be carried out in such a way that this surface acoustic variability is suppressed and the aspects of the sound relating to the articulatory invariance for the feature in question are highlighted. The transformation on the signal should utilize any acoustic information that is relevant for determining the essential articulatory attributes relating to the particular consonantal feature. Finding this transformation requires that basic information about articulatory-acoustic relations should be incorporated in the analysis, particularly for articulatory adjustments involving time-varying consonantal constrictions.

We use the spectrogram of the utterance in Fig. 1 to illustrate the analysis procedures that are being employed to estimate the various articulator-free and articulator-bound features. The first step is to identify the landmarks. In the case of consonants, these are the points in time either when an articulator in the oral cavity makes a narrow constriction or when such an articulator moves away from a constricted state. These locations are indicated by vertical lines on the spectrogram. An analysis procedure that provides a first approximation to these events involves dividing the frequency range into several bands, determining the short-time average of the energy in each band, and examining the rate of change of this energy. When this rate of change exceeds a given threshold in one or more frequency bands, an event is tentatively marked. Care must be taken in the selection of frequency bands, window length for spectrum calculation, averaging time for energies, and time over which the first difference is calculated. Selection of these parameters is based on known properties of the articulator movements at these boundaries, taking into account some aspects of auditory processing of the sound. Certain additional requirements are placed on these hypothesized acoustic discontinuities, such as a constraint that there are sufficient changes in the formant frequencies at the event to indicate that a constriction has been formed or released. This requirement eliminates the apparent discontinuity at voicing onset at time 630 ms in Fig. 1, since the high first formant at voicing onset and the lack of formant transitions indicate that an articulator is not being released from a constricted position in the oral cavity at this time. We have implemented an algorithm that locates landmarks of the type shown in Fig. 1, and it appears to identify essentially all the

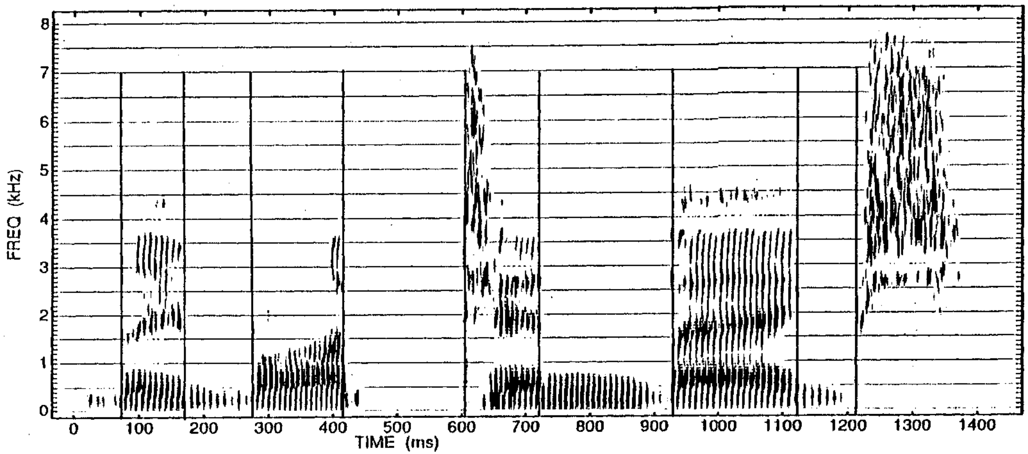

Fig. 1 Spectrogram of the utterance "They bought ten bags." The vertical lines indicate times of implosion or release of articulators when consonants are produced.

landmarks when tested with a small labeled database of utterances. Some refinement of the procedure is still needed to weed out unwanted acoustic discontinuities.

Each of these landmarks signifies that a [+consonantal] segment has been implemented, and can be labeled according to whether it is a release or an implosion. In some cases, a sequence of an implosion and a release is associated with the same articulator, and thus is a consequence of a single segment (or a geminate).

In the initial analysis stage, a second set of landmarks is identified in regions that lie between a release and an implosion — time intervals in which the vocal tract is relatively unconstricted. These landmarks are in regions where the vocal-tract is maximally open, and they occur at locations where the first-formant frequency is highest, or, equivalently, near where the low-frequency amplitude is highest. Landmarks are also located at points where there is a maximum constriction during a glide, as well as at certain points in consonant clusters. In this paper our principal concern is with the consonantal landmarks, and hence we will not expand further on the acoustic analysis that is performed in the vicinity of the landmarks that designate vocalic regions and glides.

Having designated, then, the consonantal landmarks, we now examine the acoustic data to further classify the segment as [±sonorant] and [±continuant]. In the case of [+sonorant] segments, we further determine whether [+lateral] is being implemented, whereas for [+continuant, -sonorant] segments, we estimate the value of the feature [strident]. The acoustics of these various classes of consonants is reasonably well understood, and hence the battery of acoustic measurements needed to identify these articulator-free features is relatively straightforward.

Up to this point, we have identified and classified nine consonantal landmarks in the utterance in Fig. 1. The next step is to identify the features indicating the major articulator that is involved in forming the constriction and the features designating the adjustment of the secondary articulator or articulators. Information about these features resides in the vicinity of the consonantal landmarks (e.g., reference [6]). We shall describe the analysis process for just two of the landmarks — those located at 600 ms and at 730 ms in Fig. 1.

The landmark at 600 ms has already been identified as having the features [-sonorant,-continuant] and as a point of consonantal release. For such a landmark, the active secondary articulator must be the larynx, and the following questions are asked about this articulator: What is the configuration of the glottis? and What is the state of the vocal-fold stiffness or slackness? In more conventional terms, the answer to these questions indicate whether the consonant is voiced or voiceless.

With regard to the glottal configuration, several acoustic properties point to the feature [+spread glottis]. These include the presence of aspiration noise, the delay in onset of vocal-fold vibration, and the breathy voicing during the initial few glottal periods. Measurement of the fundamental frequency ($F0$) at the onset of vocal-fold vibration at about 640 ms indicates a fall in $F0$ of about 30 Hz during the first 2-3 glottal periods. This pattern is evidence for the feature [+stiff vocal folds][4].

Several acoustic measurements lead to the hypothesis that the major articulator associated with the landmark at 600 ms is the tongue blade. These properties are evident in the spectra sampled during the few tens of ms following the consonantal release, shown in Fig. 2. Immediately following the release, the burst spectrum shows a rising characteristic with a peak at about 4 kHz, indicating a length of about 2.2 cm for the cavity anterior to the constriction, i.e., consistent with an alveolar constriction. The duration of frication noise (before aspiration begins) is about 5 ms, indicating a rapid release of the constriction (as opposed to a slower release that would be observed for a velar stop consonant). The $F2$ and $F3$ peaks in the aspiration (at 1.75 and 2.65 kHz), followed by $F2$ and $F3$ values of 1.8 and 2.5 kHz near the vowel onset also provide evidence for a constriction formed by the tongue blade [7].

Turning next to the landmark at 730 ms, we recall that this landmark designates a consonantal implosion with the features [+sonorant, -continuant, -lateral]. This designation automatically identifies the segment as [+nasal]. Verification of this designation comes from the shape of the spectrum immediately following the landmark, as well as from evidence for nasalization in the time interval immediately preceding the implosion. The frequencies of $F2$ and $F3$ immediately preceding the implosion, together with the spectrum shape immediately following the landmark, point to the tongue blade as the major articulator.

Similar procedures are followed to estimate the articulator-bound features associated with major and secondary articulators at each of the consonantel landmarks.

In the relatively simple procedures given above for making acoustic measurements for estimating the features, unequivocal decisions are made concerning each feature. Situations can arise, however, in which the acoustic data are ambiguous. One such case is where noise is present, and this noise may mask the

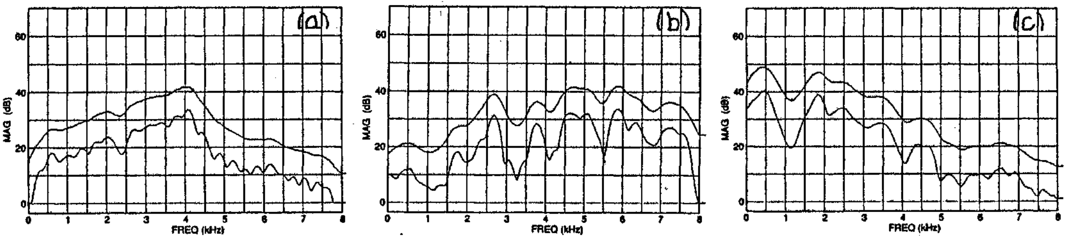

Fig. 2 Spectra sampled at three points in the vicinity of the release of the stop consonant at a time of 600 ms in Fig. 1. Each spectrum is obtained by first calculating a series of discrete Fourier transforms with a time window of 6 ms, spaced 1 ms apart. Each displayed spectrum is an average of several such adjacent spectra:(a) spectrum sampled over 4 ms at consonant release; (b) spectrum sampled over 4 ms during aspiration, 9 ms after release; (c) spectrum sampled over 9 ms shortly after onset of glottal vibration. The upper curve in each panel is a smoothed version of the actual averaged spectrum.

acoustic attributes that signal the implementation of a particular feature [8]. Under these circumstances, it may be necessary to leave this feature unspecified.

IV. ACCESSING THE LEXICON

We have described a process whereby the values of the features are estimated based on an ordered series of acoustic measurements. This feature representation must now be matched against lexical representations of the type shown in Table 1. In many cases there may be an exact match of the features determined from the acoustic data to the features in the representation of a lexical item. As noted above, a match can also be obtained as long as one or more of the estimated features fail to match the corresponding features in the lexicon as long as those features are designated as modifiable. The acceptance of such a match must be regarded as tentative, depending on the features that are found in the preceding or following context.

It may frequently happen, particularly in noisy or casually produced speech, that the acoustic data are not sufficient to identify some features. The matching procedures are defined in such a way that a lexical item can be returned when there is only a partial specification of the feature pattern, with some features remaining unspecified. In these situations, of course, more than one word might provide a match to the hypothesized pattern of features.

The question of obtaining valid matches to words when the hypothesized feature matrix spans a word boundary has not yet been addressed. As the size of the lexicon becomes large, problems may arise as a consequence of multiple matchings with different word boundaries. Some use of higher-level linguistic information may be necessary to resolve these problems.

These and other problems involved in accessing the lexicon from a hypothesized pattern of features cannot be examined in detail until more experience has been obtained in deriving features from analysis of the sound and in comparing these features with those specified in the lexicon.

V. DISCUSSION

We are attempting to make a clear separation between two basic problems in lexical access. One problem is how to extract from the acoustic signal the essential attributes that describe what features were actually implemented by a speaker in producing an utterance. Achievement of this goal requires that, at least for consonantal segments, the acoustic measurements provide a clear description of the relevant articulatory configurations and movements in the vicinity of certain acoustically-determined landmarks in the signal. Several diverse acoustic properties may contribute to identification of the features for consonantal segments. For nonconsonantal segments when there is not a narrow constriction in the vocal tract, the acoustic data can provide relatively direct estimates of the features.

The second, and probably more elusive, problem is how to access items in the lexicon when a speaker produces a word with some features modified relative to their representation in the lexicon. Our current method of labeling some features as modifiable is one step toward solution of this problem, but in the long run more sophisticated approaches may need to be followed.

ACKNOWLEDGEMENT

The research described in this paper was supported by grant IRI-8910561 from the National Science Foundation.

REFERENCES

1. K. N. Stevens. "Models of phonetic recognition II: An approach to feature-based recognition." In P. Mermelstein (ed.) *Symposium on Units and Their Representation in Speech Recognition*, 12th International Congress on Acoustics. Montreal, 1986.

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

3. M. Halle. "Features." In W. Bright (ed.) *Oxford International Encyclopedia of Linguistics.* New York: Oxford University Press, 1991.

4. M. Halle and K. N. Stevens. "Knowledge of language and the sounds of speech." In J. Sundberg, L. Nord, and R Carlson (eds.) *Music. Language, Speech and Brain.* London: MacMillan, 1991, pp.1-19.

5. S.Y. Manuel, S. Shattuck-Hufnagel, M. Huffman, R. Carlson, and S. Hunnicutt. "Studies of vowel and consonant reduction." *This Volume.*

6. K. N. Stevens. "Evidence for the role of acoustic boundaries in the perception of speech sounds." In V. Fromkin (ed.) *Phonetic Linguistics.* New York: Academic Press, 1985, pp. 243-255.

7. D. Kewley-Port. "Measurement of formant transitions in naturally produced stop consonant-vowel syllables." *J. Acoust. Soc. Am. 72*, pp.379-389, 1982.

8. G.A. Miller and P. E. Nicely. "Analysis of perceptual confusions among some English consonants." *J. Acoust. Soc. Am. 27*, pp.338-352, 1955.