A SEMIVOWEL RECOGNITION SYSTEM

Carol Y. Espy-Wilson

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

Abstract

We discuss a framework for an acoustic-phonetic approach to speech recognition. The recognition task is the class of sounds known as the semivowels (w,l,r,y) and the results obtained across several data bases are fairly consistent. We discuss some issues which were manifested by this work. These issues include feature spreading, the assignment of phonetic labels and lexical representation.

Introduction

We have developed a framework for an acoustic-phonetic approach to speech recognition. Such an approach consists of four basic steps. First, the features needed to recognize the sound(s) of interest must be specified. Second, acoustic correlates of the features must be determined. Third, algorithms to extract the properties must be developed. Finally, the properties must be integrated for recognition.

In this paper, we discuss briefly the application of the above mentioned steps to the development of a recognizer of voiced and non-syllabic semivowels of American English. In addition, we discuss some issues brought forth by this work. These issues include feature spreading and how it can possibly be explained with a theory of syllable structure, how feature spreading affects lexical access, and if and when phonetic labels should be assigned to acoustic events.

Corpora

The initial step in this research was the design of a data base for developing and testing the recognition algorithms. We chose 233 polysyllabic words from the 20,000 word Merriam Webster Pocket dictionary. These words contain the semivowels and other similar sounds in many different contexts. The semivowels occur in clusters with voiced and unvoiced consonants and they occur in word initial, word final and intervocalic positions. The semivowels are also adjacent to vowels which are stressed and unstressed, high and low, and front and back.

For developing the recognition algorithms, the data base was recorded by two males and two females. We refer to this corpus as Database-1. Two corpora were used to test the recognition system. Database-2 consisted of the same polysyllabic words spoken by two new speakers, one male and one female. Database-3 consisted of a small subset of the sentences in the TI data base [1]. In particular, we chose two sentences which contained a number of semivowels. One sentence was said by 6 females and 8 males. The other sentence was said by 7 females and 8 males. The speakers covered 8 dialects.

Several tools described in [2] were used in the transcription and analysis of the data bases. Database-1 and Database-2 were transcribed by the author and Database-3 was segmented and labelled by several experienced transcribers.

Features, Properties and Parameters

To recognize the semivowels, features are needed for separating the semivowels as a class from other sounds and for distinguishing between the semivowels. Shown in Tables 1 and 2 are the features needed to make these classifications. The features listed are modifications of ones proposed by Jakobson, Fant and Halle [3] and by Chomsky and Halle [4]. In the tables, a "+" means that the speech sound(s) indicated has the designated feature and a "-" means the speech sound(s) does not have the designated feature. If there is no entry, then the feature is not specified or is not relevant.

An acoustic study [5] was carried out in order to supplement data in the literature (e.g., [6]) to determine acoustic correlates for the features. The mapping between features and acoustic properties and the parameters used in this process are shown in Table 3. As indicated, no absolute thresholds are used to extract the properties. Instead, we used relative measures which tend to make them independent of speaker, speaking rate and speaking level. The properties are of two types. First, there are properties which examine an attribute in one speech frame relative to another speech frame. For example, the property used to capture the non-syllabic feature looks for a drop in either of two mid-frequency energies with respect to surrounding energy maxima. Second, there are properties which, within a given speech frame, examine one part of the spectrum in relation to another. For example, the property used to capture the features front and back measures the difference between F2 and F1.

To quantify the properties, we used a framework, motivated by fuzzy set theory [7], which assigns a value within the range

Table 1: Features which characterize various classes of consonants

<table>
<thead>
<tr>
<th>voiced fricatives, stops, affricates</th>
<th>unvoiced fricatives, stops, affricates</th>
<th>nasals</th>
<th>vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiced</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>sonorant</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>nonsyllabic</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>nasal</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>voiced</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>sonorant</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>nonsyllable</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>nasal</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Supported by a Xerox Fellowship
sent. Values between these extremes represent a fuzzy area indicated that the property is present/absent. Values between these extremes represent a fuzzy area indicated that the property is present/absent.

Control Strategy
Phonetic constraints are used heavily in the recognition system. These constraints are based on the assumption that a sound almost always occurs adjacent to a vowel. Therefore, if a sound is not adjacent to a vowel, it is not present.

Table 2: Parameters and Properties

<table>
<thead>
<tr>
<th>Feature</th>
<th>Acoustic Correlate</th>
<th>Parameter</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound</td>
<td>Low Frequency</td>
<td>F0</td>
<td>High</td>
</tr>
<tr>
<td>Nasal</td>
<td>Low F0 Frequency</td>
<td>F3 - F2</td>
<td>Low</td>
</tr>
<tr>
<td>Stop</td>
<td>High F0 Frequency</td>
<td>F3 - F2</td>
<td>Low</td>
</tr>
<tr>
<td>Room</td>
<td>Energy Ratio</td>
<td>F3 - F0</td>
<td>High</td>
</tr>
<tr>
<td>Bifront</td>
<td>Low F0 Frequency</td>
<td>F3 - F2</td>
<td>Low</td>
</tr>
<tr>
<td>Front</td>
<td>High F0 Frequency</td>
<td>F3 - F2</td>
<td>Low</td>
</tr>
<tr>
<td>Labial</td>
<td>Low F0 Frequency</td>
<td>F3 - F2</td>
<td>Low</td>
</tr>
<tr>
<td>Closed</td>
<td>Low F0 Frequency</td>
<td>F3 - F2</td>
<td>Low</td>
</tr>
</tbody>
</table>

Table 2: Features for discriminating between the semivowels

Figure 1a: Spectrogram of the word "fountain." (b) Formant tracks (c) Energy 400 Hz to 2800 Hz.

Figure 2: Spectrograms with formant tracks overlaid of "cartwheel" (left) and "harlequin" (right).

The overall recognition results are given in Table 4 for each of the databases. The term "acoustic" in the table means that one or more acoustic properties were used, and the score(s) was less than 0.5. The term "others" refers to rules, voiced /f/ and /s/ and consonant-like voiced contexts. As can be seen, there is a quite a bit of confusion between /f/ and /s/. However, the degree to which they are confused varies considerably with context. For example, when they are prevoiced and are not preceded by a vowel, the system correctly classifies 80% of the /f/ in Database-1 and 67% of the /s/ in Database-2. Likewise, it correctly classifies 63% of the /s/ in Database-1 and 76% of the /f/ in Database-2. This contrast is not covered in Database-3. However, 51% of the prevocalic /f/ and /s/ were misclassified in Database-3. These were classified. Considering the differences between Database-3 and the other corpora which include coverage of contexts, coverage of dialects, recording methods and processing bases, the results across databases are not quite consistent.

From Table 4 we see that there are several "misclassifications" of nasals, vowels and other sounds as semivowels. It is important to note, however, that the system has no method for detecting the feature "nasalization." Therefore, the distinction between nasals and semivowels lies mainly in the abruptness of spectral change surrounding the detected sounds. As in the case of the nasals, some misclassifications of vowels and other sounds as semivowels can be eliminated by including other features in the recognition system and by refining the parameters. However, the avoidance of other confusions is not straightforward (in addition, some of the misclassifications do not appear to be errors of the system, but errors in the transcription).

Discussion
This research has highlighted some interrelated issues which are important to any recognition system based on an acoustic phonetic approach. One such issue relates to the spreading of one or more features of a sound to a nearby segment, thereby resulting in a change of some of the features of the segment and possibly a merging of the two segments. Although examples of this phenomenon occurred with several features, we will discuss it in the context of the feature retroflexion which appears highly susceptible to spreading. Examples are illustrated in Figure 2.

The vowels /r/ and adjacent vowel merged to form an r-colored vowel. If this is so, then there does not appear to be a clear method for either /f/ or /s/ to be considered as an r-colored vowel. Instead, there is a "disorder" at the acoustic level means that the vowels are not analyzed. Instead, there is a "disorder" at the acoustic level means that the vowels are not analyzed. Instead, there is a "disorder" at the acoustic level means that the vowels are not analyzed. Instead, there is a "disorder" at the acoustic level means that the vowels are not analyzed. Instead, there is a "disorder" at the acoustic level means that the vowels are not analyzed.
determining whether an r-colored vowel is underlingly a vowel followed by /r/ or a vowel preceded by /l'/.

This ambiguity as well as the fact that some vowels and other voiced consonants are classified as semivowels raises the issue of whether or not phonetic labels should be assigned before lexical access. In other words, is the representation of items in our lexicon in terms of phonetic labels or features?

If we assume that lexical items consist of a sequence of phonetic labels, then it is clear from an analysis of the misclassification made in the semivowel recognition system that context must be considered before phonetic labels are assigned. That is, some sounds are misclassified because contextual influences caused them to have patterns of features which normally correspond to a semivowel. For example, consider the word "forewarn" shown in Figure 6. Because of the labial F2 transition and the downward F3 transition arising from the adjacent /r/, the beginning of the first /o/ was classified as a /w/. It is clear in cases like this that if phonetic labels are going to be assigned, context should be considered before it is done. The issue then becomes, how much context needs to be considered. For example, consider the word "fibroid" also shown in Figure 6 which has a fairly steady state F3 frequency of about 1900 Hz. We have observed that in words like this where a labial consonant is preceded by a normally non-retroflexed vowel and followed by a retroflexed sound, the first vowel can be totally or partially retroflexed. Such feature spreading is not surprising when we consider that the intervening labial consonant does not require a specific placement of the tongue.

If, instead of phonetic labels, lexical items are represented as matrices of features, it may be possible to avoid misclassification due to contextual influences and feature spreading since we are not trying to identify the individual sounds before lexical access. For example, consider the comparison given in Table 5 of what may be a partial feature matrix in the lexicon for an /a/ and postvocalic /l'/ with property matrices for these segments in the words "carwash" shown in Figure 6. The lexical representation is in terms of binary features whereas the acoustic realizations are in terms of properties whose strengths as determined by fuzzy logic lie between 0 and 1.

Acoustic realization #1 and the lexical representation are a straightforward match. (Assume a simple mapping strategy where property values less than 0.5 correspond to a "−" and property values greater than or equal to 0.5 correspond to a "+.") However, the mapping between acoustic realization #2 and the lexical representation is not as obvious. It may be possible for a metric to compare the two representations directly since the primary cues needed to recognize the /a/ and /l'/ are unchanged. On the other hand, we may need to apply feature spreading rules before using a metric. The rules can either generate all possible acoustic manifestations from the lexical representation or generate the "unspread" lexical representation from the acoustic realization.

Determining the mapping between features and properties which have varying degrees of strength is an important and difficult problem which may give insights into the structure of the lexicon. The solution to this problem will require a better understanding of feature assimilation in terms of what features are prone to spreading, and in terms of the domains over which spreading occurs. Resolution of these matters is clearly important to an acoustic-phonetic approach to speech recognition.

REFERENCES


Table 5: Lexical Representation vs. Acoustic Realizations of /ar/.