A PHONETICALLY BASED 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

A phonetically based approach to speech recognition uses speech specific knowledge obtained from phonotactics, phonology and acoustic phonetics to capture relevant phonetic information. Thus, a recognition system based on this approach can make broad classifications as well as detailed phonetic distinctions. This paper discusses a framework for developing a phonetically based recognition system. The recognition task is the class of sounds known as the semivowels. The recognition results reported, though incomplete, are encouraging.

INTRODUCTION

The spectrogram reading experiments conducted by Zue and Cole in 1979 [1] showed that the wealth of phonetic information contained in the speech signal can be extracted by applying explicit rules. As a result of this research, a phonetically based approach to speech recognition is a viable alternative to traditional pattern matching techniques.

The development of a phonetically based recognition system requires three major steps. First, features needed for classifying the sounds of interest must be identified and translated into acoustic properties which can be quantified. Second, algorithms must be developed to reliably and automatically extract these acoustic properties from the speech signal. Finally, a control strategy for integrating these properties which are present with various degrees of strength must be developed.

In this paper, we discuss the design of a phonetically based system for recognizing the class of sounds known as the semivowels /w,l,r,y/. By recognition we mean both detection and classification. We constrain the problem to the recognition of nonsyllabic voiced semivowels within polysyllabic words.

The semivowels have some properties which make recognition of them particularly challenging. First, of the consonants, the semivowels are acoustically most similar to the vowels. They are produced orally without complete closure of the vocal tract and without any frication. The constriction required to articulate them does not inhibit spontaneous voicing and, therefore, they are sonorant consonants. Furthermore, as is true for the vowels, the semivowels often have a steady state. Second, due to phonotactic constraints, a semivowel must occur adjacent to a vowel except for the /rl/ cluster in words like "snarl". As a consequence of these two properties, acoustic boundaries between the semivowels and vowels are usually not apparent from a spectrogram. In this respect, recognition of the semivowels is more difficult than recognition of other consonants. This point is illustrated in Figure 1 where a spectrogram of the word "demoralize" is given. The acoustic properties change rather abruptly between the stop consonant /d/ and the vowel /iʸ/ and between the nasal consonant /m/ and the surrounding vowels /iʸ/ and /ɔ/. On the other hand, the changes are more gradual between the semivowel /r/ and the surrounding vowels /ɔ/ and /ə/, or between the semivowel /l/ and the surrounding vowels /ə/ and /ɑʸ/, although there is a degree of abruptness at the release of the /l/.

*Supported by a Xerox Fellowship

DATA BASE

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 a variety of contexts. The semivowels occur in clusters (with one another, nasals, stops, fricatives and /s/-stops), in word initial, word final and intervocalic positions. They also occur adjacent to vowels which are stressed and unstressed, high and low and front and back.

Two repetitions of each word were recorded by two males and two females. For acoustic analysis, one token of each word from each speaker was hand transcribed using Spire [2].

FEATURE EXTRACTION

To recognize the semivowels, features are needed for separating the semivowels as a class from other sounds and for dis-

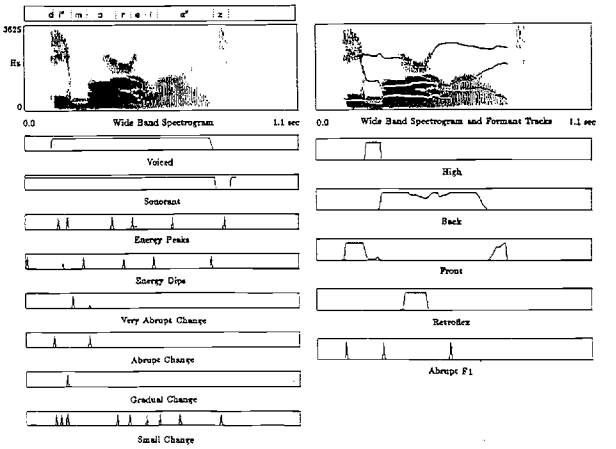

Figure 1: Acoustic properties of the word "demoralize"

	voiced	sonorant	nonsyllabic
voiced fricatives, stops, affricates	+	−	+
unvoiced fricatives,stops,affricates	−	−	+
devoiced semivowels and nasals	−	+	+
voiced nasals and semivowels	+	+	+
vowels	+	+	−

Table 1: Features which characterize various classes of consonants

	stop	nasal	high	back	front	retroflex
nasals	+	+	+			
/w/	−	−	+	+	−	−
/y/	−	−	+	−	+	−
/r/	−	−				+
light /l/	+	−				−
dark /l/	−	−	−	+	−	−

Table 2: Features for separating semivowels from nasals and for discriminating between the semivowels

tinguishing between the semivowels. Shown in Tables 1 and 2 are the features needed to make these classifications. 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.

Based on the features in Table 1, we note that the class of segments of interest in this study is identified by a "+" for each of the three features. Furthermore, the nasals, which are also sonorant consonants, are the only other sounds so defined. Thus, based on these features, the voiced sonorant consonants should be discriminated from other speech sounds. The features in Table 2 are needed to separate the semivowels from the nasals and to identify the individual semivowels. The feature high is defined slightly differently from the conventional way, but its acoustic correlate is a low first formant frequency. Two allophones of /l/ are specified, light /l/ and dark /l/. Dark /l/ occurs in syllable final position whereas a light /l/ occurs in other syllable positions.

Considerable research (e.g. [3]) has been done studying the perceptual and acoustic properties of the semivowels. This research has indicated the pattern of formant frequencies that distinguish between the semivowels. This can be seen in Figure 2 where the frequency difference between F2 and F1 and between F3 and F1 are shown for the intervocalic semivowels spoken by one of the female talkers. The formant tracks were automatically extracted and the formant frequencies were measured in the middle of the transcribed semivowels. The frequency difference between F3 and F1, which is one of the acoustic properties used to detect retroflexion, separates /r/ from the other semivowels. In addition, the frequency difference between F2 and F1, which is used to extract the features front and back, clearly separates /y/ from the other semivowels and partially separates /w/ from /l/.

Table 3 contains the present mapping of the features listed in Tables 1 and 2 into measurable acoustic properties. Based on our present knowledge of acoustic phonetics, some features can be extracted more reliably than others. For example, the feature sonorant can be reliably extracted using energy based

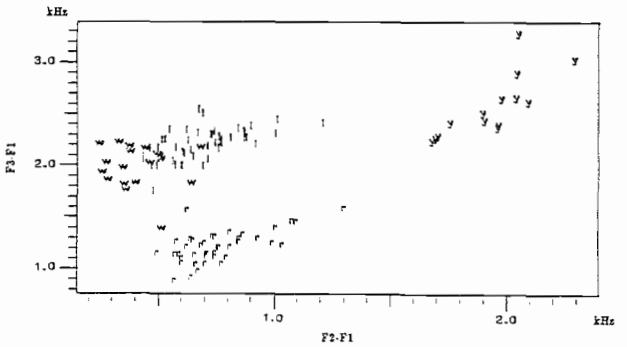

Figure 2: Difference in Formant Frequency Values for Intervocalic Semivowels

parameters while the feature high is not as easy to extract since reliable computation of formant tracks can be difficult.

As indicated in Table 3, no absolute thresholds are used to quantify the acoustic properties. Instead, we used relative measures which tend to make the properties independent of speaker, recording level, and recording environment. These measures are based on anchor points within the specific utterance being analyzed. Each measure examines an attribute in one speech frame in relation to another frame, or, within a given frame, examines one part of the spectrum in relation to another. Thus, for example, in the extraction of the feature voiced, which is based on the bandlimited energy 200 Hz to 700 Hz, the classification of each speech frame within the utterance was based on the energy in that frame with respect to the maximum energy measured within the entire utterance. On the other hand, in the extraction of the feature sonorant, the energies in the frequency bands 0 Hz to 300 Hz and 3700 Hz to 7000 Hz were compared on a frame by frame basis.

The framework provided by fuzzy set theory [6] is used to quantify each property into the range [0,1]. A value of 1 means we are confident that the property is present. Conversely, a value of 0 means we are confident that the acoustic property is absent. Values in between these two extremes represent a fuzzy area with the value indicating our level of certainty that the property is present/absent.

As an example of how this framework is applied, consider the quantification of the acoustic property used to extract the feature nonsyllabic. The acoustic correlate of this feature is significantly less energy in the consonant regions than in the vowel

Feature	Parameter	Property
Voiced	Energy 200-700 Hz	High (Relative)
Sonorant	Energy Ratio (0-300)/(3700-7000)	High
Nonsyllabic	Energy 640-2800 Hz	Low (Relative)
	Energy 2000-3000 Hz	Low (Relative)
Stop	1st Difference of Total Energy	High
	1st Difference of F1	High
Nasal	Nasal Formant	Present
High	F1 - F0	Low
Back	F2 - F1	Low
Front	F2 - F1	High
Retroflex	F3 - F1	Low

Table 3: Parameters and Properties for Feature Extraction

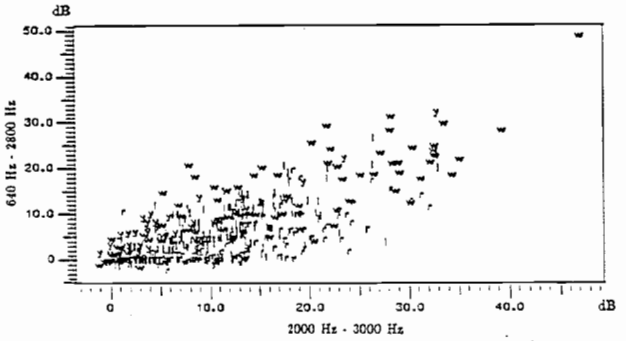

Figure 3: Energy Dip in Intervocalic Semivowels

regions. In an attempt to define this property of "less energy" more precisely, we selected the bandlimited energies 640 Hz to 2800 Hz and 2000 Hz to 3000 Hz and examined their effectiveness in identifying the presence of intervocalic semivowels. For each of these parameters, the differences between the minimum energy within the semivowels and the maximum energy within the adjacent vowels was measured. The smaller of these two differences indicates the significance of the energy dip. A scatter plot of the range of values of this energy dip for the two bandlimited energies is shown in Figure 3. We also looked at the range of values of energy dips within the vowel regions. Less than 1% of the vowels contained an energy dip. Furthermore, these energy dips tended to be less than 2 dB.

Based on these data, this property was quantified into the regions shown in Figure 4. An energy dip of 2 dB or more definitely indicates a nonsyllabic segment. If an energy dip between 1 dB and 2 dB is measured, we are uncertain as to whether a nonsyllabic segment is present or not. Finally, energy dips of less than 1 dB are not indicative of a nonsyllabic segment.

Examples of the quantified properties for the word "demoralize" are shown in Figure 1. The energy based properties are shown on the left and the formant based properties are shown on the right. The automatically extracted formant tracks are overlaid on the second spectrogram. Note that the feature stop was quantified into the properties "very abrupt change", "abrupt change" "gradual change" and "small change".

Some properties, such as the voiced property, define a region(s) within the speech signal whereas others, such as the energy dip property, mark instants of time. The amplitude of each property indicates our level of certainty that the property is present. Some of the properties are extracted independently of one another. However, the knowledge gained from some of the properties is used in the extraction of other properties. For example, the formants are only tracked within the region specified by the voiced property.

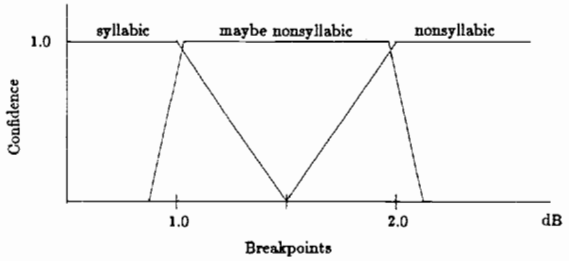

Figure 4: Quantification of Energy Dip Property

CONTROL STRATEGY

To classify the semivowels, rules were written integrating the extracted acoustic properties. At the writing of this paper, the formant tracker developed for this recognition system was being refined. Therefore, some of the more detailed phonetic distinctions could not be made. However, using the energy based properties, a rule was written for the detection of sonorant consonants. Having detected these segments, another rule, which is also incomplete without formant information, was written to separate the semivowels from other sonorant consonants or imposters. In general, the imposters include the nasals only, although other voiced consonants are sometimes sonorant when in intervocalic position. To compensate somewhat for missing formant information, the properties "murmur", "maybe murmur" and "no murmur" were used to help separate nasals from semivowels. These properties were based on the ratio of the bandlimited energy 0 Hz to 400 Hz to the bandlimited energy 400 Hz to 1000 Hz.

The rules written to recognize the semivowels are:

SONORANT CONSONANT = (NONSYLLABIC)(VOICED)
(SONORANT)

SEMIVOWEL = (SONORANT CONSONANT)
{[(NO MURMUR) + (MAYBE MURMUR)]
(NOT STOP LIKE) +
(NO MURMUR)(STOP LIKE)}

IMPOSTERS = (SONORANT CONSONANT)
{(MURMUR) + (VERY STOP LIKE) +
(MAYBE MURMUR)(STOP LIKE)}

where "very stop like", "stop like" and "not stop like" are also rules which are based on the spectral change properties displayed in Figure 1.

From the sonorant consonant rule, the energy dips detected must occur within a voiced and sonorant region. Having detected such an energy dip, the spectral change properties are extracted from the region surrounding the energy dip. In the case of intervocalic semivowels, this region is defined by the energy peaks that are adjacent to the energy dip. These energy peaks are required to be in a voiced and sonorant region as well. That is, they should occur within the adjacent vowels. For prevocalic and postvocalic semivowels, this region is defined by the adjacent energy peak and the beginning or end of the voiced sonorant region, respectively.

In the recognition rules, the addition or "+" operation of properties is analogous to a logical "or". The result of this operation is the maximum value of the properties being operated on. The multiplication of properties is analogous to a logical "and". The result of this operation is the minimum value of the properties being operated on. Thus, since the value of each property must be in the range [0,1], the scores obtained by applying these rules must also be in the range [0,1].

Note that a rule was also written to classify imposters. This rule and the semivowel rule are applied to each detected sonorant consonant. Thus, an imposter which classifies as a semivowel will hopefully classify as an imposter with a higher score. To

	semivowels	nasals	other imposters
total	445	84	77
detected	411	83	60

Table 4: The number of possible intervocalic sonorant consonants containing an energy dip.

classify as a semivowel or an imposter, the scores obtained from these rules must be greater than 0.5.

RECOGNITION

This section contains preliminary results for the detection and classification of intervocalic sonorant consonants. The convex hull algorithm developed by Mermelstein[4] was used to detect these intervocalic energy dips within the energy bands mentioned in Table 3.

Detection of Intervocalic Sonorant Consonant

To qualify as occurring within an intervocalic sonorant consonant, the detected energy dip must be surrounded by energy peaks which also occur in voiced and sonorant regions. As mentioned earlier, such energy peaks should occur within the adjacent vowels. Table 4 summarizes the results obtained with this strategy. The row labeled "total" is the number of such intervocalic consonants transcribed in the data base. As mentioned earlier, the other imposters includes other voiced consonants which are sometimes sonorant when in intervocalic position. In this study, we defined the "other imposters" to be /h/, /ɾ/, /v/ and /ð/ when, according to the transcription of the words in the data base, they occurred in intervocalic position.

All the nasals and other imposters were classified as nonsyllabic. That is, an energy dip was always contained within their transcribed region. The 17 undetected imposters, unlike the semivowels and nasals, were also classified as either unvoiced, nonsonorant or both. One misclassification of the nasal /m/ (in the word "demoralize") occurred because it was preceded by a short /ə/ (15 msec) such that the preceding energy peak occurred at the release of the /d/ burst, an unvoiced region.

Of the 34 undetected semivowels, 94% were not classified as nonsyllabic. That is, these semivowels were not found through energy dip detection. Of those not classified as nonsyllabic, 91% occurred after a stressed vowel and before an unstressed vowel, such as the /ɹ/ in "chlorination". The remaining semivowels not found through energy dip detection occurred before and after unstressed vowels, such as the /l/ in "musculature". Thus, these semivowels do not have the nonsyllabic property. In the final design of the recognition system, redundancy will be built in such that other features, such as retroflex, can be used to identify semivowels independently of the nonsyllabic property.

In addition to the consonants listed in Table 4, a few other voiced consonants which are sometimes sonorant when in intervocalic position and a few vowels also contained energy dips. However, we will not discuss these segments here.

Classification of Intervocalic Segments

Table 5 summarizes the results obtained by applying the classification rules to the detected intervocalic segments. Given that no formant information was used, the results are encouraging. Only 7% of the semivowels were classified as imposters. Of these semivowels, 54% were/l/'s which are often stop-like due to the release of the tongue from the roof of the mouth in their production before a vowel. With formant information, particularly the frequency of F1, we expect to do considerably better. Shown in Figure 5 is a multi-histogram of the frequency of F1 of the intervocalic semivowels and nasals. As was true in Figure 2, these data were automatically extracted from the words spoken by one of the female talkers. There is little overlap in the distribution of the F1 frequency values of the nasals and the distribution of the F1 frequency values of the /l/'s.

classification	identification	
	semivowels	imposters
semivowel	379	34
imposter	27	169
not class.	5	0

Table 5: Classification of Intervocalic Segments

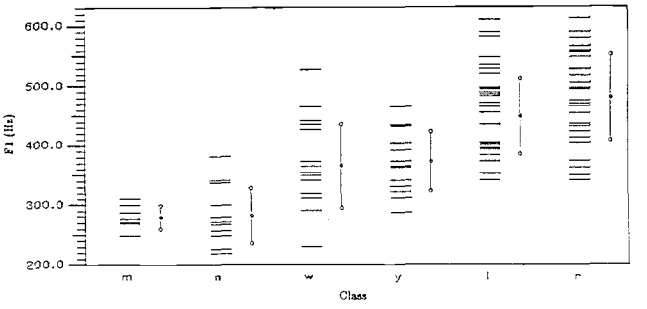

Figure 5: Frequency of F1

Of the imposters classified as semivowels, 53% were nasals. The energy change between these nasals and adjacent vowels was too gradual for the nasals to be considered stop-like. Furthermore, the murmur property did not distinguish these nasals even though many of them can be distinguished from the semivowels on the basis of the frequency of F1.

SUMMARY

In this paper, we develop a framework for a phonetically based approach to speaker independent speech recognition. First, we identify features for classifying the speech sounds of interest. Second, we develop algorithms to automatically extract the acoustic correlates of these features. Finally, the extracted properties are integrated in a set of recognition rules. Given that all of the necessary acoustic properties are not used, the recognition results obtained are encouraging.

REFERENCES

[1] Zue,V.W. and Cole, R.A., "Experiments on Spectrogram Reading", *ICASSP*, pp. 116-119, April 1979.

[2] Cyphers, D. S., "Spire: A Speech Research Tool", S. M. Thesis, MIT, 1985.

[3] Lehiste, I., "Acoustic Characteristics of Selected English Consonants," *Report No. 9*, Univ. of Mich., Comm. Sciences Lab., 1962.

[4] Mermelstein, P., "Automatic Segmentation of Speech into Syllabic Units," *J. Acoust. Soc.Am.*, vol. 58, pp. 880-883, 1975.

[5] Fant, Gunnar, *Acoustic Theory of Speech Production*. The Hague: Mouton & Co., 1960.

[6] DeMori, Renato, *Computer Models of Speech Using Fuzzy Algorithms*. New York and London: Plenum Press, 1983.