# 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 nonsyllabic 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 nonsyllabic 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

| | voiced | sonorant | nonsyllabic | nasal |
|---|---|---|---|---|
| voiced fricatives,stops,affricates | + | − | + | − |
| unvoiced fricatives,stops,affricates | − | − | + | − |
| semivowels | + | + | + | − |
| nasals | + | + | + | + |
| vowels | + | + | − | − |

**Table 1:** Features which characterize various classes of consonants

*Supported by a Xerox Fellowship

| | stop | high | back | front | labial | retroflex |
|---|---|---|---|---|---|---|
| /w/ | − | + | + | − | + | − |
| /y/ | − | + | − | + | − | − |
| /r/ | − | − | | − | − | + |
| light /l/ | + | − | | − | − | − |
| dark /l/ | − | − | + | − | − | − |

**Table 2:** Features for discriminating between the semivowels

| Feature | Acoustic Correlate | Parameter | Property |
|---|---|---|---|
| Voiced | Low Frequency Periodicity | Energy 200-700 Hz | High* |
| Sonorant | Comparable Low & High Frequency Energy | Energy Ratio $\frac{(0-300)}{(3700-7000)}$ | High |
| Nonsyllabic | Dip in Energy | Energy 640-2800 Hz | Low* |
| | | Energy 2000-3000 Hz | Low* |
| Stop | Abrupt Spectral Change | 1st Difference of Bandlimited Energies (positive & negative) | High |
| High | Low F1 Frequency | F1 − F0 | Low |
| Back | Low F2 Frequency | F2 − F1 | Low |
| Front | High F2 Frequency | F2 − F1 | High |
| Labial | Downward Transitions for F2 and F3 | F3 − F0 | Low* |
| | | F2 − F0 | Low* |
| Retroflex | Low F3 Frequency & Close F2 and F3 | F3 − F0 | Low |
| | | F3 − F2 | Low |

**Table 3:** Parameters and Properties

*Relative to a maximum value

[0,1]. A value of 1 means we are confident that the property is present, while a value of 0 means we are confident that it is absent. Values between these extremes represent a fuzzy area indicating our level of certainty that the property is present/absent.

### Control Strategy

Phonotactic constraints are used heavily in the recognition system. These constraints state that semivowels almost always occur adjacent to a vowel. Therefore, they are usually prevocalic, intervocalic or postvocalic. For recognition, these contexts map into three types of places within a voiced sonorant region. First the semivowels can be at the beginning of a voiced sonorant region, in which case they are prevocalic. Second, the semivowels can be at the end of a voiced sonorant region, in which case they are postvocalic. Finally, the semivowels may be further inside a voiced sonorant region. We refer to these semivowels as intersonorant, and one or more may be present within such a region. Semivowels of this type can be either intervocalic or in a cluster with another sonorant consonant such as the /y/ in "banyan." Although there is one overall recognition strategy, there are modifications for these contexts.

The recognition strategy for the semivowels is divided into two steps: detection and classification. The detection process marks certain acoustic events in the vicinity of times where there is a potential influence of a semivowel. In particular, we look for minima in the mid-frequency energies and we look for minima and maxima in the tracks of F2 and F3. Such events should correspond to some of the features listed in Tables 1 and 2. For example, an F2 minimum indicates a sound which is more "back" than an adjacent segment(s). Thus, this acoustic event will occur within most /w/'s and within some /l/'s and /r/'s.

Once all acoustic events have been marked, the classification process integrates them, extracts the needed acoustic properties, and through explicit semivowel rules decides whether the detected sound is a semivowel and, if so, which semivowel it is. An example of this process is illustrated with the word "flour-

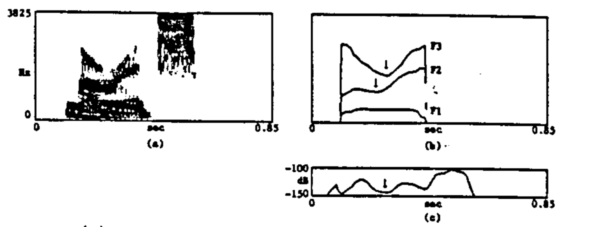

**Figure 1:** (a) Spectrogram of the word "flourish," (b) formant tracks and (c) Energy 640 Hz to 2800 Hz.

ish" shown in Figure 1. As can be seen, several acoustic events signal the presence of the intervocalic /r/. These events include an energy dip, a small F2 dip and a strong F3 dip. Given the energy dip marked in part c, the recognition system will extract the surrounding energy maxima corresponding to syllabic nuclei. These latter points are used to define a region for further analysis of the detected sound. Among the various events, the F3 dip is the most prominent one which gives some clue to the identity of the detected sound. Thus, it is in a small region surrounding the time of this event that the formant based properties are extracted. In addition, it is between the time of the F3 dip and the surrounding energy peaks that we characterize the rate of spectral change to determine its degree of abruptness.

Once the properties listed in Table 3 are extracted for the detected sound, the control strategy, on the basis of the types of events marked, decides which semivowel rules to apply. Again, since there is a strong F3 dip, the /r/ rule is applied first. The only other semivowel which is expected to sometimes have a sizeable F3 dip is the labial sound /w/. Thus, the /w/ rule is applied if the /r/ rule receives a low score (< 0.5).

Rules for integrating the properties were written for each of the semivowels. In addition, because they are acoustically similar, a rule was written for identifying a class that could be either /w/ or /l/. Across contexts, the rules are similar. However, well known acoustic differences between allophones such as the closer spacing between F2 and F1 for sonorant-final /l/'s as opposed to sonorant-initial /l/'s are accounted for. Additionally, within the rules, primary versus secondary cues are distinguished. For example, the /r/ rule states that if the detected sound is retroflexed, classify it as an /r/. However, if the sound is "maybe" retroflexed, look at other cues before making a decision.

Since the value of each property lies between 0 and 1, the score of any rule within the fuzzy logic framework is also in this range. Thus, we consider a sound to be classed as a semivowel if the result of a rule is greater than or equal to 0.5.

### Recognition Results

The overall recognition results are given in Table 4 for each of the data bases. The term "nc" in the table means that one or more semivowel rules was applied, but the score(s) was less than 0.5. The term "others" refers to flaps, voiced /h/'s and sonorant-like voiced consonants.

As can be seen, there is quite a bit of confusion between /w/ and /l/. However, the degree to which they are confused varies considerably with context. For example, when they are prevocalic and are not preceded by a consonant, the system correctly classifies 80% of the /w/'s in Database-1 and 67% of the /w/'s in Database-2. Likewise, it correctly classifies 63% of the /l/'s





| | w | l | r | y | nasals | others | vowels | |
|---|---|---|---|---|---|---|---|---|
| # tokens | 369 | 540 | 558 | 222 | 464 | 508 | 2385 | |
| undetected(%) | 1.4 | 3.3 | 2.6 | 2.9 | 24 | 81.5 | | |
| w(%) | 52 | 7.5 | 3.4 | 0 | 1 | 1 | 1 | |
| l(%) | 9.1 | 55.7 | 0 | 0 | 11 | 3.3 | 5.5 | |
| w-l(%) | 31.4 | 30.4 | 0 | 0 | 3 | .8 | 2 | Database-1 |
| r(%) | 4 | .2 | 90 | 0 | 2 | .6 | 6 | |
| y(%) | 0 | 0 | 0 | 93.7 | 6 | 1.4 | 8.6 | |
| nc(%) | 2 | 3 | 4.7 | 4.9 | 53 | 11.4 | 39 | |
| # tokens | 181 | 274 | 279 | 105 | 232 | 135 | 1184 | |
| undetected(%) | 1.7 | 1.5 | 4.3 | 2.8 | 24 | 69 | | |
| w(%) | 48 | 3.6 | 1.9 | 0 | 5 | 0 | 1 | |
| l(%) | 12.7 | 57.7 | 0 | 0 | 7 | 6 | 5 | |
| w-l(%) | 29 | 33.8 | 0 | 0 | 3 | 1 | 4 | Database-2 |
| r(%) | 3.5 | .4 | 91.3 | 0 | 3 | 2 | 4 | |
| y(%) | 0 | 0 | 0 | 84.9 | 3 | 3 | 10 | |
| nc(%) | 6.7 | 2.9 | 4.3 | 13.3 | 55 | 19 | 42 | |
| # tokens | 28 | 40 | 49 | 23 | 44 | 121 | 350 | |
| undetected(%) | 3.6 | 7.5 | 0 | 4 | 50 | 73 | | |
| w(%) | 46 | 10 | 0 | 0 | 15 | 0 | 2 | |
| l(%) | 21.6 | 52.6 | 0 | 0 | 13 | 2.5 | 9 | |
| w-l(%) | 21.6 | 24.7 | 0 | 0 | 0 | 0 | 4 | Database-3 |
| r(%) | 7.1 | 0 | 89.8 | 0 | 5 | 2.5 | 15 | |
| y(%) | 0 | 0 | 0 | 78.5 | 0 | 5 | 9 | |
| n-c(%) | 0 | 5.1 | 10.2 | 17.2 | 17 | 17 | 62 | |

**Table 4:** Overall Recognition Results

in Database-1 and 76% of the /l/'s in Database-2. This context is not covered in Database-3. However, 71% of the prevocalic /w/'s adjacent to unvoiced consonants in Database-3 were classified correctly. Considering the many differences between Database-3 and the other corpora which include coverage of contexts, coverage of dialects, recording methods and transcription biases, the results across data bases are 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). It is this issue which is addressed in the remainder of the paper.

### Discussion

This research has highlighted several 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

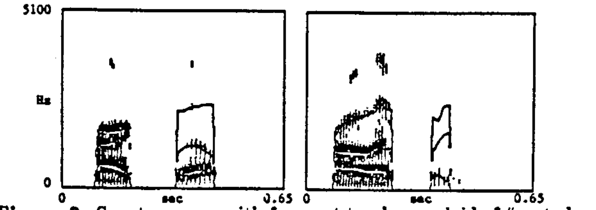

**Figure 2:** Spectrograms with formant tracks overlaid of "cartwheel" (left) and "harlequin" (right).

with the words "cartwheel" and "harlequin." In each instance, it appears as if the underlying /r/ and adjacent vowel combine such that their acoustic realization is an r-colored vowel. The occurrence of such feature assimilation is predicted by the syllable structure theory as explained by Selkirk [8]. This syllable structure is shown in Figure 3, where the onset consists of any syllable-initial consonants, the peak consists of either a vowel or vowel and sonorant, and the coda consists of any syllable-final consonants. Selkirk states that when /l/ or /r/ is followed by a consonant which must occupy the coda position, it becomes part of the peak. Thus, the structure for the first syllable in "cartwheel" is as shown in Figure 4. Since the /ɑ/ and /r/ both occupy the syllable peak, we might expect some type of feature assimilation to occur. If it is true that a vowel and /r/ in this context will always overlap to form an r-colored vowel, then no exception is needed in the phonotactic constraints of semivowels for words like "snarl" where the /l/ is "supposedly" separated from the vowel by the /r/. Instead, the constraints can simply state that semivowels must always be adjacent to a vowel.

When a postvocalic /l/ or /r/ is not followed by a syllable-final consonant, Selkirk states that it will tend to be in the coda although it has the option of being part of the peak. This option was clearly exercised across the speakers in Database-1 and Database-2. As an example, consider the two repetitions of the word "carwash" shown in Figure 5. As in the word "harlequin," the /ɑ/ and /r/ in the word "carwash" on the left appears to be one segment in the sense that retroflexion extends over the entire vowel duration. However, in the repetition on the right, the /ɑ/ does not appear to be retroflexed. Instead, there is a clear downward movement in F3 which separates the /ɑ/ and /r/ and thus the /r/ appears to be syllable-final.

We dealt with this feature spreading phenomenon in the recognition system by considering it a correct classification if the vowels in words like "cartwheel," "harlequin" and "carwash" were labeled /r/. This seemingly "disorder" was allowed since the vowel's and following /r/'s appear completely assimilated.

Allowing this "disorder" at the acoustic level means that the ambiguity must be resolved at or before lexical access. There is at least one example in the data bases where a seemingly prevocalic /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

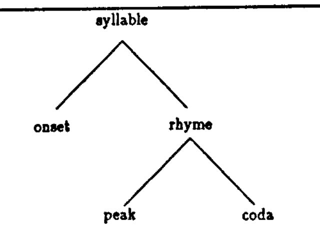


**Figure 3:** Tree structure of syllable.

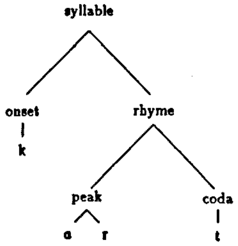


**Figure 4:** Tree structure of syllable "cart."

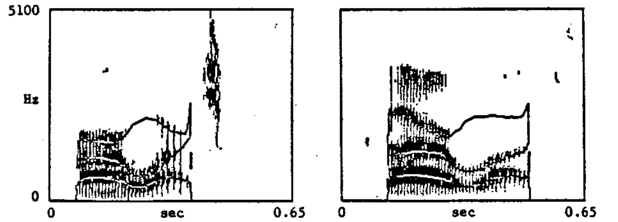


**Figure 5:** Spectrograms with formant tracks overlaid of two repetitions of "carwash."

determining whether an r-colored vowel is underlyingly a vowel followed by /ɪ/ or a vowel preceded by /ɪ/.

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 misclassifications 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 /ɪ/, the beginning of the first /ɔ/ 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 misclassifi-

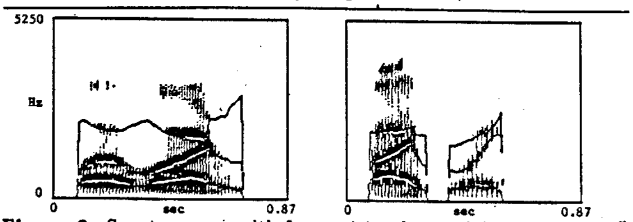


**Figure 6:** Spectrograms with formant tracks overlaid of "forewarn" (left) and "fibroid" (right).

| | lexical representation | | realization #1 | | realization #2 |
|---|---|---|---|---|---|
| | ɑ | ɪ | ɑ | ɪ | ɑʳ |
| high | − | − | 0 | 0 | 0 |
| low | + | − | 1 | 0 | 1 |
| back | + | ± | 1 | 1 | 1 |
| retroflex | − | + | 0 | 1 | 1 |

**Table 5:** Lexical Representation vs. Acoustic Realizations of /ɑɪ/.

cations 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 /ɑ/ and postvocalic /ɪ/ 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 /ɑ/ and /ɪ/ 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

[1] Lamel, L., Kassel, R., and Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. Speech Recog. Workshop*, CA., 1986.

[2] Cyphers, D., Kassel, R., Kaufman, D. Leung, H., Randolph, M. Seneff, S., Unverferth, J., Wilson, T. and Zue, V. "The Development of Speech Research Tools on MIT's Lisp Machine-Based Workstations," *Proc. Speech Recog. Workshop*, CA, 1986.

[3] Jakobson, R., Fant, G. and Halle, M., "Preliminaries to Speech Analysis," *MIT Acoustics Lab. Tech. Rep. No. 13*, 1952.

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

[5] Espy-Wilson, Carol Y., "An Acoustic-Phonetic Approach to Speech Recognition: Application to the Semivowels," Doctoral Dissertation, MIT, to be completed in June 1987.

[6] Lehiste, I., "Acoustic Characteristics of Selected English Consonants," *Report No. 9*, U. of Mich., Comm. Sci. Lab., 1962.

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

[8] Selkirk, E.O., "The Syllable," *The Structure of Phonological Representations (part II)*, ed. van der Hulst, H. and Smith N., Dordrecht: Foris Publications, 1982.