A feature-based semivowel recognition system

Carol Y. Espy-Wilson

Electrical, Computer and Systems Engineering Department, Boston University, 44 Cummington Street, Boston, Massachusetts 02215 and Research Laboratory of Electronics, Room 36-545, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 8 June 1992; accepted for publication 25 March 1994)

A recognition system based on linguistic features was developed for the semivowels /w j r l/ in American English. The features of interest are *sonorant, syllabic, consonantal, high, back, front*, and *retroflex*. Acoustic correlates and events related to these features were used to detect and classify the semivowels. The recognizer was tested across semivowels occurring in a wide range of phonetic environments. The corpora included polysyllabic words and sentences spoken by males and females of several dialects. The results show that a feature-based approach to recognition is a viable methodology. Fairly consistent overall recognition results were obtained. Across the test data, acoustic events were detected within 97% of the semivowels and classification rates were 62% for /w/, 74% for /l/ (/w/ and /l/ were often confused), 90% for /r/ and 84% for /j/.

PACS numbers: 43.70.Fq, 43.72.Ne, 43.70.Jt

INTRODUCTION

A system for recognizing the class of sounds known as the semivowels /w j r l/ in American English was developed to demonstrate the viability of a feature-based approach to recognition. Recognizing the semivowels is a particularly challenging problem since the semivowels, which are acoustically similar to the vowels, almost always occur adjacent to a vowel. Furthermore, the spectral changes between the semivowels and adjacent vowels are often quite gradual so that acoustic boundaries are usually not apparent. In this respect, recognition of the semivowels is more difficult than the recognition of other consonants.

In a feature-based approach to recognition, speech-specific information consisting of the acoustic correlates of the linguistic features which comprise a phonological description of the speech sounds is used. Research into recognition systems of this type which attempt to explicitly extract the linguistic information from the speech signal and discard the components which are extra-linguistic is presently suffering in comparison to probabilistic approaches such as hidden Markov models (HMM) (Levinson *et al.*, 1983; Rabiner *et al.*, 1983; Jelinek, 1985; Lee, 1988) and pseudo-neural networks (e.g., Elman and McClelland, 1986; Waibel, 1989) which attempt to make the linguistic–extralinguistic differentiation implicitly by training models on large databases. The appeal of statistically based systems is that they can be automatically trained and the success they have achieved in limited speech recognition/understanding tasks suggests that they are able to extract statistical regularities from simple signal representations such as cepstral coefficients and their time derivatives. However, as Zue (1985) and Makhoul and Schwartz (1985) suggest, further improvements in these systems will depend on the successful incorporation of speech knowledge (which gets at the phonetically relevant information) into these frameworks.

A variety of efforts have been made over the past several years to combine speech knowledge and probabilistic frameworks. For example, in their segment-based recognition system, Phillips and Zue (1992) use acoustic–phonetic knowledge to design generalized algorithms which, given sufficient training data, automatically measure acoustic attributes felt to be important for phonetic contrasts. In contrast, Deng and Erler (1992) use a general representation of the speech signal in terms of cepstral coefficients, but incorporate speech knowledge into their HMM-based recognizer by modeling subphonemic units called microsegments. Although such automatic training methods have many advantages, they are limited by the amount of training data available and the signal representation provided.

The need of a better signal representation coupled with several advances that have been made in recent years suggest that another look into feature-based recognition is warranted. The advances include an improved understanding of the distinctive features and the relations between them (Stevens and Keyser, 1989), a better idea of the acoustic correlates of the features (Stevens, 1980; Espy-Wilson, 1992) and the development of theories of hierarchical structures for the representation of lexical items in terms of features (Clements, 1985; Sagey, 1986).

In addition to these recent gains, there are several other reasons why this approach to recognition is desirable. First, the acoustic properties for features can be defined in relational terms so that much of the cross speaker and contextual variability disappears. Second, a feature-based approach provides a framework for understanding other sorts of variability that can occur. For example, the semivowel recognition system discussed in this paper recognized the /b/ in one token of the word "disreputable" as a /w/. While the acoustic manifestation of this /b/ is grossly different from it's canonical form, an analysis in terms of features showed that this particular /b/ differs from a canonical [b] in terms of only two features, a shift from $-$*sonorant* to $+$*sonorant* and, consequently, a shift from $-$*continuant* to $+$*continuant*. Thus using features as the basic unit for recognition provides a simple mechanism for capturing and handling the gross acoustic changes that occur.

 0001-4966/94/96(1)/65/8/$6.00

TABLE I. Mapping of features into acoustic properties.

Feature	Acoustic correlate	Parameter	Property
Sonorant	Comparable low- and high-frequency energy	Energy ratio $\frac{(0-300)}{(3700-7000)}$	High[a]
Nonsyllabic	Dip in midfrequency energy	Energy 640–2800 Hz	Low[a]
		Energy 2000–3000 Hz	Low[a]
Consonantal	Abrupt amplitude change	First difference of adjacent spectra	High
High	Low F_1 frequency	F_1-F_0	Low
Back	Low F_2 frequency	F_2-F_1	Low
Front	High F_2 frequency	F_2-F_1	High
Retroflex	Low F_3 frequency and close F_2 and F_3	F_3-F_0	Low
		F_3-F_2	Low

[a]Relative to a maximum value within the utterance.

Finally, in a feature-based approach to recognition, an acoustic-event-oriented as opposed to a segment-oriented scheme can be used for recognition. In traditional approaches to recognition, either the speech signal is segmented into phoneme-sized pieces to which labels are assigned, or labels are assigned on a frame-by-frame basis. Sounds like the semivowels pose a problem for such approaches since there are often no obvious acoustic boundaries between them and adjacent sounds. In contrast, the system discussed in this paper identifies specific acoustic events around which acoustic properties for features are extracted. This event-oriented approach led to the detection of select acoustic landmarks which signaled the presence of 97% of the semivowels. Another advantage offered by an event-oriented approach is that there is no underlying assumption that sounds are nonoverlapping. Thus it allows for the possibility of recognizing sounds that are completely or partially coarticulated.

I. METHOD

A. Stimuli

To develop the recognition system, a database of 233 polysyllabic words containing semivowels in a variety of phonetic environments was selected from the 20 000-word Merriam–Webster Pocket dictionary. The semivowels occur adjacent to voiced and unvoiced consonants, as well as in word-initial, word-final, and intervocalic positions. The semivowels occur adjacent to vowels which are stressed and unstressed, high and low, and front and back. A more detailed discussion of this database is given in Espy-Wilson (1992).

To test the recognition system, the same database and a small subset of the TIMIT database (Lamel *et al.*, 1986) was used. In particular, the sentences "She had your dark suit in greasy wash water all year" (sentence-1) and "Don't ask me to carry an oily rag like that" (sentence-2) were chosen since they contain several semivowels in a number of contexts. However, note that many of the contexts represented in the polysyllabic words are not included in the sentences.

B. Speakers and recordings

The polysyllabic words were embedded in the carrier phrase "_ pa." The final "pa" was added in order to avoid glottalization and other types of utterance-final variability. The speakers were recorded in a quiet room with a pressure-gradient close-talking noise-cancelling microphone (part of a Sennheiser HMD 224X microphone/headphone combination). They were instructed to say the utterances at a natural pace.

For development of the recognition algorithms, each word was spoken once by two females and two males. The females are from the northeast and the males are from the midwest. For testing, each word was spoken once by two additional speakers, one female and one male from the same geographical areas cited above. The speakers were students and employees at the Massachusetts Institute of Technology. All are native speakers of English and reported no hearing loss.

In addition to the latter database, we also tested the recognition system on 14 repetitions of sentence-1 (6 females and 8 males) and 15 repetitions of sentence-2 (7 females and 8 males). The speakers cover 7 U.S. geographical areas and an "other" category used to classify talkers who moved around often during their childhood. Like the words in the other databases, these sentences were recorded using a close-talking microphone.

C. Segmentation and labeling

The polysyllabic words were excised from their carrier phrase after they were digitized with a 6.4-kHz low-pass filter and a 16-kHz sampling rate, and pre-emphasized to compensate for the relatively weak spectral energy at high frequencies (a particular issue for sonorants). To facilitate analysis and recognition, the words were segmented and labeled by the author with the help of playback and displays of several attributes including LPC and wideband spectra, the speech signal and various bandlimited energy waveforms. The Merriam–Webster Pocket dictionary provided a baseline phonemic transcription of the words. However, modifications of some of the labels were made based on the speakers' pronunciations (for more details, see Espy-Wilson, 1992).

D. Feature analysis

The features chosen to separate the semivowels as a class from other sounds are *sonorant, nonsyllabic*, and *nasal*. The features selected to distinguish among the semivowels

TABLE II. Acoustic events which may signal the presence of semivowels.

Semivowel	Acoustic events				
	Energy dip	F_2 dip	F_2 peak	F_3 dip	F_3 peak
w	×	×		×	×
y	×		×		×
r	×	×		×	
l	×	×			×

are *consonantal, high, back, front,* and *retroflex*. A more detailed discussion of these features are given in Espy-Wilson (1992).

An acoustic study (Espy-Wilson, 1992) was carried out in order to supplement data in the literature (e.g., Lehiste, 1962) 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 I. (The acoustic properties in Table I have been refined since the development of the semivowel recognition system. For a discussion, see Espy-Wilson, 1992.) To minimize their sensitivity to speaker, speaking rate and speaking level, all of the properties in Table I are based on relative measures as opposed to absolute ones such as the frequencies and amplitudes of spectral prominences. The relative 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 midfrequency 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 F_2 and F_1^1.

To quantify the properties, we used a framework motivated by fuzzy set theory (De Mori, 1983), which assigns a value within the range [0,1]. A value of 1 means that the property is definitely present, while a value of 0 means that it is definitely absent. Values between these extremes represent a fuzzy area indicating the level of certainty that the property is present/absent.

E. Recognition strategy

The recognition strategy for the semivowels is divided into two steps: detection and classification. The detection process marks certain events signaled by changes in the parameters listed in Table I. The process begins by finding all sonorant regions within an utterance. Next, certain acoustic events are marked within the sonorant regions on the basis of substantial energy change and/or substantial formant movement. Results of an acoustic study (Espy-Wilson, 1992) showed that the events listed in Table II usually occur within the designated semivowel(s). Dip detection[2] is performed within the time functions representing the midfrequency energies to locate all nonsyllabic sounds. In addition, dip detection and peak detection[3] are performed on the tracks of F_2 and F_3^4. An F_2 dip should be found in sounds which are produced with a more "back" articulation than adjacent sounds. An F_2 peak should be found in sounds which are produced with a more "front" articulation than adjacent sounds. Retroflexed and some labial sounds should contain an F_3 dip. Finally, an F_3 peak should occur in the semivowels /l/ and /j/. In addition, some /w/'s which are in a retroflexed environment may also contain an F_3 peak since F_3 is generally higher in a /w/ than it is in an adjacent retroflexed sound. The methodology used to detect these acoustic events differs, depending upon whether the semivowels are prevocalic, postvocalic, or intersonorant. For more details, see Espy-Wilson (1987).

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. At this time, by combining all the relevant acoustic cues, the recognizer can correctly classify the semivowels while the remaining detected sounds should be left unclassified.

Specific rules were applied to integrate the extracted acoustic properties for identification of the semivowels in prevocalic, intersonorant, and postvocalic contexts.[5] Given the acoustic similarity between many /w/'s and /l/'s, a /w–l/ rule was included for sounds that were judged to be equally likely a /w/ or /l/. The rules are dependent upon context so that they capture well known acoustic differences due to allophonic variation. For example, the prevocalic /l/ rule states that the /l/ can have either a gradual or abrupt offset, allowing for the possibility of an abrupt rate of spectral change between the /l/ and following vowel. Several researchers (Joos, 1948; Fant, 1960; Dalston, 1975) have observed a sharp spectral discontinuity between /l/ and following stressed vowels and they attribute this to the rapid release of the tongue tip from the alveolar ridge in the production of a prevocalic /l/. On the other hand, the postvocalic /l/ rule requires that the rate of spectral change between the /l/ and preceding vowel be gradual since alveolar contact is often not realized or is realized only gradually in the production of a postvocalic /l/.

The properties in the rules are combined using fuzzy logic. In the fuzzy logic framework, addition is analogous to a logical "or" and the result of this operation is the maximum value of the properties being considered. Multiplication of two or more properties is analogous to a logical "and." In this case, the result is the minimum value of the properties being operated on. Since the value of any property is between 0 and 1, the result of any rule is also between 0 and 1. We chose 0.5 to be the dividing point for classification. That is, if the sound to which a semivowel rule is applied receives a score greater than or equal to 0.5, it will be classified as that semivowel.

II. RESULTS

The semivowel recognizer is evaluated by comparing its output with the hand transcription. Given that the placement of segment boundaries is subjective, some flexibility is used in tabulating the detection and classification results. A semivowel is considered detected if an energy dip and/or one or more formant extrema is placed somewhere between the beginning (minus 10 ms) and end (plus 10 ms) of its hand-

TABLE III. Overall percent recognition results for the semivowels. "nc" are those sounds which were detected but "not classified" as a semivowel by any of the rules.

Detection					Classification				
				Original					
	w	l	r	j		w	l	r	j
No. tokens	369	540	558	222	No. tokens	369	540	558	222
detected	99	97	97	96	undetected	1	3	3	4
Energy dip	47	51	36	35	w	52	8	3	0
F_2 dip	97	83	46	0	l	9	56	0	0
F_2 peak	0	0	1	92	w-l	31	30	0	0
F_3 dip	41	10	95	2	r	4	0	90	0
F_3 peak	21	54	1	78	j	0	0	0	94
					nc	2	3	5	5
				New speakers					
	w	l	r	j		w	l	r	j
No. tokens	181	274	279	105	No. tokens	181	274	279	105
detected	98	99	96	98	undetected	2	1	4	2
Energy dip	49	59	44	41	w	48	4	2	0
F_2 dip	93	85	49	0	l	13	58	0	0
F_2 peak	0	0	1	95	w-l	29	34	0	0
F_3 dip	37	7	90	0	r	4	0	91	0
F_3 peak	30	69	2	87	j	0	0	0	85
					nc	7	3	4	13
				TIMIT					
	w	l	r	j		w	l	r	j
No. tokens	28	40	49	23	No. tokens	28	40	49	23
detected	96	93	100	96	undetected	4	7	0	4
Energy dip	61	89	61	57	w	46	10	0	0
F_2 dip	93	83	65	0	l	22	53	0	0
F_2 peak	0	0	0	91	w-l	22	25	0	0
F_3 dip	47	36	94	0	r	7	0	90	0
F_3 peak	61	50	4	70	j	0	0	0	79
					nc	0	5	10	17

transcribed region by the appropriate detection algorithms. The arbitrarily chosen 10-ms margin was not always large enough to include all of the detected acoustic events occurring during the semivowels. Thus, for about 1% of the semivowels, further corrections were made when tabulating the detection results.

A. Semivowel recognition results

The overall recognition results for the databases are compared in Table III [more detailed recognition results according to various contexts are given in (Espy-Wilson, 1987)]. The database used to develop the recognition system is referred to as "original." "New speakers" refer to the words contained in original which were spoken by new speakers. Finally, TIMIT refers to the sentences taken from the TIMIT corpus. On the left side of the table are the detection results which are given separately for each database. The top row lists the semivowels. The following rows show the actual number of semivowels that were transcribed (No. tokens), the percentage of semivowels that contain one or more acoustic events during their hand-transcribed region (detected), and the percentage of semivowels that contain each type of acoustic event marked by the detection algorithms. For example, the detection table for original states that 97% of the transcribed /w/'s contained an F_2 dip within their segmented region.

The classification results for each database are given on the right side of the table. As before, the top row lists the semivowels. The following rows show the number of semivowel tokens transcribed, the number which were undetected (the complement of the corresponding number in the detection results) and the percentage of those semivowel tokens transcribed which were classified by the semivowel rules. For example, the results for original show that 90% of the 558 tokens of /r/ which were transcribed were correctly classified. The term "nc" (in the bottom row) for "not classified" means that one or more semivowel rules was applied to the detected sound, but the classification score(s) was less than 0.5. A semivowel which is considered undetected may show up in the classification results as being recognized. Thus the numbers in a column within the classification results may not always add up to 100%.

When comparing the recognition results of the three databases, the differences between TIMIT and the other corpora should be kept in mind. Recall that the speakers of the original and new speakers databases were from two dialect regions whereas the speakers of the TIMIT database were from several geographical areas. In addition, the sparseness of the semivowels in TIMIT affects the recognition results. Several of the contexts represented in original and new speakers where the semivowels receive high recognition scores don't occur in TIMIT.

TABLE IV. Percent recognition of other sounds as semivowels. The results for original, new speakers, and TIMIT are lumped together.

	Nasals	Others	Vowels
No. tokens	740	764	3919
undetected	26	78	
w	3	1	1
l	10	4	6
w-l	3	1	3
r	2	1	6
j	5	1	9
nc	51	14	42

1. Detection results

In spite of the differences between the databases, the detection results are fairly consistent. The results from all three databases show the importance of using formant information in addition to energy measures. Across contexts, F_2 minima are most important in locating /w/'s and /l/'s, F_3 minima are most important in locating /r/'s, and F_2 maxima are most important in locating /j/'s. When in an intervocalic context, however, the detection results using only energy minima compare favorably with those using the cited formant minimum/maximum.

Due to formant transitions between semivowels and adjacent consonants and the placement of segment boundaries in the hand transcription, there are a few events listed in the detection results which, at first glance, appear strange. For example, several /r/'s that are adjacent to a coronal consonant such as the first /r/ in the word "foreswear" contain both an F_3 dip and an F_3 peak. The F_3 peak is due to the rise in F_3 between the /r/ and the /s/.

2. Classification results

The classification results are also fairly consistent across the databases. The results for /r/ and /j/ are much better than those for /w/ and /l/. A considerable number of /w/'s and /l/'s are classified as /w–l/ in all three databases. No one measure used in the recognition system provides a good separation between these sounds; however, in several contexts, the system is able to correctly classify these sounds at a rate better than chance. For example, lumping the results of the databases together, 76% of word-initial /w/'s and 67% of word-initial /l/'s are classified correctly. If we assign half of the /w–l/ score to /w/ and /l/, the correct classification rates change to 85% for /w/ and 73% for /l/.

B. Consonants called semivowels

Table IV shows that many nasals are called semivowels. One main reason for this confusion is the lack of a parameter which captures the feature *nasal*. The main cue used for the nasal-semivowel distinction is the abruptness of the spectral change between the sonorant consonant and adjacent vowel(s). This property accounts for the generally higher misclassification of nasals as /l/ as opposed to one of the other semivowels.

In addition to the nasals, several voiced obstruents are lenited (the process whereby a consonant is produced with a weakened constriction, cf. Catford, 1977) so that they are recognized as semivowels. The latter sounds are grouped into a class called "others" and their recognition results are shown in Table IV. An example of this type of confusion is shown in Fig. 1 where the intervocalic /b/ in "disreputable" was classified as /w–l/. As can be seen from the spectrogram in Fig. 1, the /b/ is realized as a sonorant consonant. In addition, the formant frequencies are acceptable for a /w/ and an /l/. Finally, the rate of spectral change between the /b/ and the surrounding vowels is gradual.

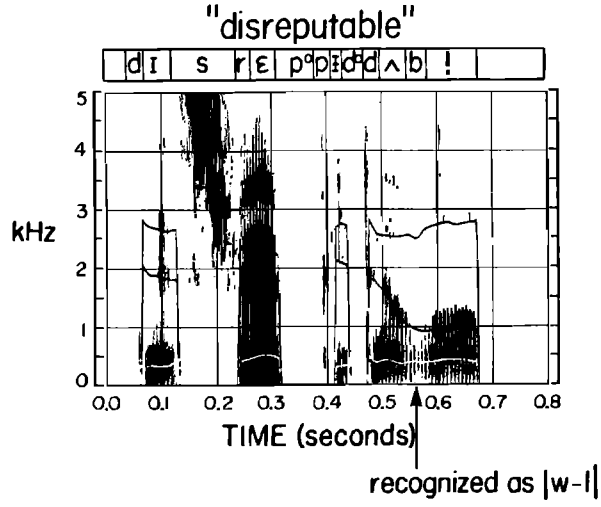

FIG. 1. Wideband spectrogram of the word "disreputable" which contains a lenited /b/ that was recognized as /w-l/.

C. Vowels called semivowels

The classification results for the vowels are also given in Table IV. No detection results are given for the vowels since different portions of the same vowel may be detected and labeled a semivowel. For example, across several speakers, the beginning of the /ɔɪ/ in "flamboyant" was classified as either /w/, /l/, or /w–l/ and the offglide was classified as a /j/. When this phenomenon occurs, the vowel shows up in the results as being misclassified twice. Thus the vowel statistics for the databases may not add up to 100%.

Most of the misclassifications of vowels are of the type described above. They occur because of contextual influence as in the case of the beginning of the /ɔɪ/ in "flamboyant" which resembles a /w/ due to the influence of the preceding /b/. They also occur when diphthongs like the /ɔɪ/ in "flamboyant" are followed by another vowel. In this case, the offglide of the diphthong is often recognized as a semivowel. Other errors occurred because semivowels which are in the underlying transcription were not hand labeled, but were detected and classified by the recognition system. Such an example is the /l/ in the word "stalwart." Finally, some errors occur because of coarticulation as in the word "harlequin" shown in Fig. 2. The lowest point of F_3 which signals the /r/ is at the beginning of what is labeled /ɑ/, suggesting that the articulation of these two sounds overlap. However, given the discrepancy in time between where the /r/ is recognized and where it appears in the transcription, the recognition of the /r/ is classified as an error.

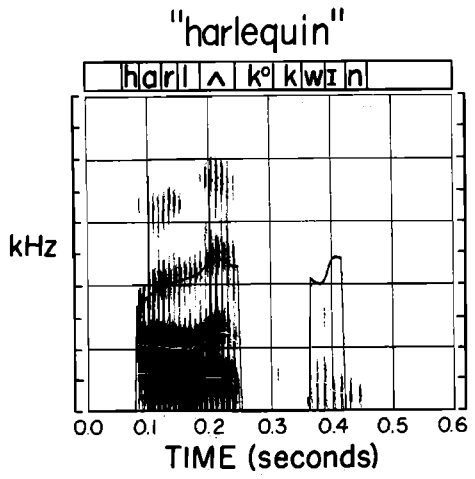

FIG. 2. Wideband spectrogram of the word "harlequin." The vowel /a/ and /r/ are co-articulated so that the beginning of what is transcribed as the /a/ is recognized as an /r/.

III. SUMMARY AND DISCUSSION

The framework developed for the feature-based approach used in the recognition system for semivowels is based on three key assumptions. First, it assumes that the abstract features have acoustic correlates which can be reliably extracted from the physical signal. The acoustic properties are derived from relative measures as opposed to absolute measurements so that they are less sensitive to speaker differences, speaking rate, and context. Second, the framework is based upon the general idea that the acoustic manifestation of a change in the value of a feature is marked by a specific event in the appropriate acoustic parameter(s). An acoustic event can be a minimum, maximum or an abrupt spectral change in a parameter. Finally, the framework is based on the notion that these events serve as landmarks for when the acoustic correlates for features should be extracted to classify the sounds in the signal.

The results obtained show that the feature-based framework is a viable methodology for speaker-independent continuous speech recognition. Fairly consistent recognition results were obtained for the three corpora which include polysyllabic words and sentences which were spoken by males and females of several dialects. Thus one major conclusion that can be drawn from these data is that much of the across-speaker variability disappears if relative measures are used to extract the acoustic properties for features.

As a comparison, the baseline semivowel recognition results obtained from a three-state HMM with Gaussian mixture observation densities (Lamel and Gauvain, 1993) are shown in Table V. These data were obtained using 61 context-independent phone models which are mapped to 39 phones for scoring (Lee and Hon, 1989). The data are based on the core test set of the TIMIT database which contains 8 sentences from each of 24 speakers. A similar table of results is shown in Table VI for the feature-based semivowel recognition system where we have pooled the data across the databases new speakers and TIMIT (the results for original were not included since this database was used to develop the recognition system). Half of the /w–l/ score was assigned to the scores for /w/ and /l/. (This reassignment makes comparison easier and is reasonable since the sounds assigned to this category were felt to be equally likely a /w/ or /l/.)

TABLE V. Percent recognition of semivowels by HMM system. "Other" are those semivowels which were detected, but classified as some sound other than a semivowel. (Note: The overall phone accuracy of the HMM-based system of 60% increases substantially to 71% with context-dependent models and phone bigram probabilities as phonotactic constraints. Specifically, the recognition rates increase to 80% for /w/, 81% for /l/, 82% for /r/ and 58% for /j/.)

	w	l	r	j
No. tokens	144	291	270	50
detected	87	82	79	86
w	61	4	0	0
l	5	58	0	0
r	1	1	56	0
j	0	0	0	56
other	20	19	23	30

In comparing the recognition results of Tables V and VI, two differences should be kept in mind. First, the results of the HMM system are based on TIMIT sentences whereas the semivowel recognition results are based on a smaller and different set of TIMIT sentences, and polysyllabic words. Second, the HMM system was designed for a closed class problem, that is to distinguish between 39 possible phone labels. The feature-based system, on the other hand, was designed to recognize semivowels in unrestricted speech, which is an open class problem.

The HMM system detects 60% of the semivowels, and 71% of those detected are correctly classified. In addition, 3% (156 tokens out of a total of 4999) of nonsemivowel sounds are recognized as semivowels. The feature-based system detects 92% of the semivowels (this number does not include those semivowels listed in Table VI as "nc") and correctly classifies 85% of those detected. However, it classifies 22% of the nonsemivowel sounds as semivowels. (Recall from the discussions in Secs. II B and II C that many of these "misclassifications" are in fact reasonable.) Since the HMM and feature-based systems are operating at different correct recognition/false recognition trade-off points, it is difficult to compare accuracy. However, these data suggest that sounds can be adequately detected on the basis of acoustic events and that a signal representation in terms of the

TABLE VI. Percent recognition of semivowels by the feature-based system. Data across the databases new speakers and TIMIT are pooled and the /w-l/ scores are incorporated in /w/ and /l/ scores. "nc" are those semivowel which were detected but "not classified" as a semivowel by any of the rules.

	w	l	r	j
No. tokens	209	314	328	128
detected	98	98	97	98
w	62	21	2	0
l	28	74	0	0
r	4	0	91	0
j	0	0	0	84
nc	6	3	4	14

acoustic correlates of features is extracting the relevant information from the speech signal.

With these differences in mind, a comparison of the recognition results show that the feature-based recognition system does substantially better in terms of detection and classification of the semivowels. These data suggest that sounds can be adequately detected on the basis of acoustic events and that a signal representation in terms of the acoustic correlates of features is extracting the relevant information from the speech signal.

While the feature-based recognition results are encouraging, an analysis of the errors has brought forth several issues, which need to be addressed to improve and extend the feature-based approach and to appropriately evaluate its performance. First, phenomena such as coarticulation and lenition make the present hand-transcription procedure inadequate for evaluating phoneme recognition performance. As in the case of "harlequin" discussed in Sec. III C, speech sounds often overlap, at least to some extent, so that some of the strongest acoustic evidence for a feature that is distinctive for a particular sound may occur outside of the region transcribed for that sound. Present hand-transcription techniques don't allow for such overlap so that matching in such cases is problematic. In the case of lenition, the often large acoustic change that accompanies weakened consonants is not reflected in the hand transcription. Thus the evaluation of misclassifications resulting from lenition is not straightforward.

Second, an analysis of the insertions and other misclassifications show that, in many cases, errors occur because decisions about the underlying phonemes are being made too early in the recognition process. The portions of the waveforms recognized as semivowels do look like semivowels. Thus to improve recognition results would require that contextual influences and feature changes due to coarticulation and lenition be taken into account before labeling is done. One possibility is to not integrate the extracted acoustic properties to make a decision about what the sounds are within a word before lexical access. Instead, lexical access can be performed directly from the extracted acoustic properties. In this way, the underlying sounds within a word are not known until the word has been recognized.

Finally, there appears to be a hierarchy of features which may govern not only the appropriate acoustic property for features, but also what features are applicable during different portions of the waveform. In particular, the acoustic correlates of some or all of the features may differ depending upon whether the sound is *syllabic* so that the vocal tract is relatively open, or *nonsyllabic* so that the vocal tract is more constricted. For example, the acoustic correlate used in this study for the feature *high* is one often associated with vowels. However, this property which should separate the liquids /l r/ from the glides /w j/ grouped all of the semivowels together. Along this same line, the features *back* and *front* were used to help distinguish among the semivowels. However, given that the semivowels are usually *nonsyllabic*, the features *labial* and *coronal* should probably be used in addition to or instead of the more vowel-like features. The latter features may also be more desirable because they are based on spectral shape as opposed to formant frequencies. The tracking of formant frequencies is often problematic during consonants since, due to the constricted vocal tract, the formants may merge or be obscured by nearby antiresonances.

IV. CONCLUSIONS

In conclusion, the feature-based approach to recognition shows much promise. However, a great deal of work still needs to be done to understand and reliably extract all of the acoustic correlates of the linguistic features; to specify all of the feature changes that can occur and in what domains; to define an appropriate lexical representation for the words in the lexicon; and to develop strategies that can match a feature-based representation of the lexical items with the extracted properties for features.

ACKNOWLEDGMENTS

This work was supported in part by a Xerox Fellowship and NSF Grant No. BNS-8920470. It is based in part on a Ph.D. thesis by the author submitted to the Department of Electrical Engineering and Computer Science at MIT. The author gratefully acknowledges the encouragement and advice of Professor Kenneth N. Stevens. I also want to acknowledge the helpful comments of Corine Bickley, Caroline Huang, and Melanie Matthies, which greatly improved the quality and clarity of this paper. Special thanks to Lori Lamel and Jean-Luc Gauvain for providing confusion matrices obtained from their HMM-based recognition system.

[1]The formants were tracked automatically (Espy-Wilson, 1987) during the detected sonorant regions of the waveforms. The algorithm was based on peak-picking of the second difference of the log-magnitude linear-prediction spectra.

[2]Dip detection in a particular signal is performed by finding minima relative to adjacent maxima. A dip in a signal can occur at (1) the beginning of the signal in which case the signal rises substantially from some low point, (2) the end of the signal in which case the signal falls substantially from some high point, and (3) inside the signal in which case the signal consists of a fall followed by a rise. For further details see Espy-Wilson (1987).

[3]Peak detection in a particular waveform is performed by inverting the waveform and applying the dip detection algorithm.

[4]Before peak detection and dip detection are performed, missing frames in the formant tracks are filled in automatically by an interpolation algorithm and the resulting formant tracks are smoothed.

[5]The prevocalic semivowels rules are: /w/=(very back)+(back)(high +maybe high)(gradual onset) (maybe close F_2F_3+not close F_2F_3); /l/ =(back+mid)(gradual offset+abrupt offset)(maybe high+nonhigh+low) (maybe retroflex+not retroflex) (maybe close F_2F_3+not close F_2F_3); /w-l/=(back) (maybe high) (gradual offset) (maybe close F_2F_3+not close F_2F_3); /r/=(retroflex) (close F_2F_3+maybe close F_2F_3)+(maybe retroflex) (close F_2F_3) (gradual offset) (back+mid) (maybe high+nonhigh+low) and /j/=(front)(high+maybe high) (gradual offset+abrupt offset). The intersonorannt semivowels rules are /w/=(very back)+(back)(high+maybe high)(gradual onset)(gradual offset)(maybe close F_2F_3+not close F_2F_3); /l/=(back+mid)(maybe high+nonhigh+low)(gradual onset+abrupt onset)(gradual offset+abrupt offset)(maybe retroflex+not retroflex) (maybe close F_2F_3+not close F_2F_3); /w-l/=(back) (maybe high) (gradual onset) (gradual offset) (maybe close F_2F_3+not close F_2F_3); /r/=(retroflex) (close F_2F_3+maybe close F_2F_3)+(maybe retroflex) (close F_2F_3) (gradual onset) (gradual offset)(back+mid)(maybe high+nonhigh+low); /j/ =(front)(high+maybe high) (gradual onset) (gradual offset). The postvocalic semivowel rules are: /l/=(very back+back) (gradual onset) (not retroflex)(not close F_2F_3)(maybe high+nonhigh+low) and /r/=(retroflex) (close F_2F_3)+(maybe retroflex) (close F_2F_3)(maybe high+nonhigh +low)(back+mid) (gradual onset).

Catford, J. C. (**1977**). *Fundamental Problems in Phonetics* (Indiana U.P., Bloomington, IN).
Clements, G. N. (**1985**). "The geometry of phonological features," Phonol. Year. **2**, 225–252.
Dalston, R. M. (**1975**). "Acoustic characteristics of english /w,r,l/ spoken correctly by young children and adults," J. Acoust. Soc. Am. **57**, 462–469.
De Mori, Renato (**1983**). *Computer Models of Speech Using Fuzzy Algorithms* (Plenum, New York).
Deng, L., and Erler, K. (**1992**). "Structural design of hidden Markov model speech recognizer using multivalued phonetic features: Comparison with segmental speech units," J. Acoust. Soc. Am. **92**, 3058–3067.
Elman, J. L., and McClelland, J. L. (**1986**). "Exploiting lawful variability in the speech wave," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 360–380.
Espy-Wilson, C. Y. (**1987**). "An acoustic–phonetic approach to speech recognition: Application to the semivowels," Technical Rep. No. 531, Research Laboratory of Electronics, MIT.
Espy-Wilson, C. Y. (**1992**). "Acoustic measures for linguistic features distinguishing the semivowels in American English," J. Acoust. Soc. Am. **92**, 736–757.
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Netherlands).
Jakobson, R., Fant, G., and Halle, M. (**1952**). "Preliminaries to speech analysis," MIT Acoustics Lab. Tech. Rep. No. 13.
Jelinek, F. (**1985**). "The Development of an experimental discrete dictation recognizer," Proc. IEEE **73**, 1616–1625.
Joos, M. (**1948**). "Acoustic phonetics," Lang. Suppl. **24**, 1–136.
Lamel, L., Kassel, R., and Seneff, S. (**1986**). "Speech database development: Design and analysis of the acoustic–phonetic corpus," Proc. Speech Recog. Workshop, CA.
Lamel, L., and Gauvain, J. (**1993**). "Cross-lingual experiments with phone recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process.
Lee, K. (**1988**). *Automatic Speech Recognition: The Development of the SPHINX System* (Kluwer Academic, Norwell, MA).
Lee, K., and Hon, H. (**1989**). "Speaker-independent phone recognition using hidden Markov models," IEEE Trans. Acoust. Speech Signal Process **37**, 1641–1648.
Lehiste, I. (**1962**). "Acoustical characteristics of selected english consonants," Report No. 9, University of Michigan, Communication Sciences Laboratory, Ann Arbor, Michigan.
Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (**1983**). "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," Bell Syst. Technol. J. **62**, 1035–1074.
Makhoul, J., and Schwartz, R. (**1985**). "Ignorance modeling," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 344–345.
Phillips, M., and Zue, V. (**1992**). "Automatic discovery of acoustic measurements for phonetic classification," Proc. Second Intern. Conf. Spoken Lang. Proc. **1**, 795–798.
Rabiner, L., Levinson, S., and Sondhi, M. (**1983**). "On the application of vector quantization and hidden Markov models to speaker independent isolated word recognition," Bell Syst. Technol. J. **62**, 1075–1105.
Sagey, E. C. (**1986**). "The representation of features and relations in nonlinear phonology," doctoral dissertation, Massachusetts Institute of Technology, Department of Linguistics.
Stevens, K. N. (**1980**). "Acoustic correlates of some phonetic categories," J. Acoust. Soc. Am. **68**, 836–842.
Stevens, K. N., and Keyser, J. K. (**1989**). "Primary features and their enhancement in consonants," Language **65**, 81–106.
Waibel, A. (**1989**). "Phoneme recognition using time-delay neural networks," IEEE Trans. Acoust. Speech, Signal Process. **37**, 328–339.
Zue, V. W. (**1985**). "The use of speech knowledge in automatic speech recognition," Proc. IEEE **73**, 1602–1615.