Signal Representation, Attribute Extraction and, the Use of Distinctive Features for Phonetic Classification[1]

Helen M. Meng, Victor W. Zue, and Hong C. Leung

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

The study reported in this paper addresses three issues related to phonetic classification: 1) whether it is important to choose an appropriate signal representation, 2) whether there are any advantages in extracting acoustic attributes over directly using the spectral information, and 3) whether it is advantageous to introduce an intermediate set of linguistic units, i.e. distinctive features. To restrict the scope of our study, we focused on 16 vowels in American English, and investigated classification performance using an artificial neural network with nearly 22,000 vowels tokens from 550 speakers excised from the TIMIT corpus. Our results indicate that 1) the combined outputs of Seneff's auditory model outperforms five other representations with both undegraded and noisy speech, 2) acoustic attributes give similar performance to raw spectral information, but at potentially considerable computational savings, and 3) the distinctive feature representation gives similar performance to direct vowel classification, but potentially offers a more flexible mechanism for describing context dependency.

INTRODUCTION

The overall goal of our study is to explore the use of distinctive features for automatic speech recognition. Distinctive features are a set of properties that linguists use to classify phonemes [1,13]. More precisely, a feature is a minimal unit which distinguishes two maximally-close phonemes; for example /b/ and /p/ are distinguished by the feature [VOICE]. Sounds are more often confused in relation to the number of features they share, and it is believed that around 15 to 20 distinctive features are sufficient to account for phonemes in all languages of the world. Moreover, the values of these features, such as [+HIGH] or [-ROUND], correspond directly to contextual variability and coarticulatory phenomena, and often manifest themselves as well-defined acoustic correlates in the speech signal [3]. The compactness and descriptive power of distinctive features may enable us to describe contextual influence more parsimoniously, and thus to make more effective use of available training data.

[1]This research was supported by DARPA under Contract N00014-82-K-0727, monitored through the Office of Naval Research.

In order to fully assess the utility of this linguistically well-motivated set of units, several important issues must be addressed. First, is there a particular spectral representation that is preferred over others? Second, should we use the spectral representation directly for phoneme/feature classification, or should we instead extract and use acoustic correlates? Finally, does the introduction of an intermediate feature-based representation between the signal and the lexicon offer performance advantages? We have chosen to answer these questions by performing a set of phoneme classification experiments in which conditional variables are systematically varied. The usefulness of one condition over another is inferred from the performance of the classifier.

In this paper, we will report our study on the three questions that we posed earlier. First, we will report our comparative study on signal representations. Based on these results, we will then describe our experiments and results on acoustic attribute extraction, and the use of distinctive features. Finally, we will discuss the implications and make some tentative conclusions.

TASK AND CORPUS

The task chosen for our experiments is the classification[2] of vowels in American English. The corpus consists of 13 monothongs /i, ɪ, e, ɛ, æ, a, o, ʌ, ɔ, u, ʊ, ü and ɝ/ and 3 diphthongs /aʸ, ɔʸ, aʷ/. The vowels are excised from the acoustic-phonetically compact portion of the TIMIT corpus [6], with no restrictions imposed on the phonetic contexts of the vowels. For the signal representation study, experiments are based on the task of classifying all 16 vowels. However, the dynamic nature of the diphthongs may render distinctive feature specification ambiguous. As a result, we excluded the diphthongs in our investigation involving distinctive features, and the size of the training and test sets were reduced correspondingly. The size and contents of the two corpora are summarized in Table 1.

[2]It is a classification task in that the left and right boundaries of the vowel token are known through a hand-labelling procedure, and the classifier is only asked to determine the most likely label.

	Training Speakers (M/F)	Testing Speakers (M/F)	Training Tokens	Testing Tokens
I	500 (357/143)	50 (33/17)	20,000	2,000
II	500 (357/143)	50 (33/17)	19,000	1,700

Table 1: Corpus I consists of 16 monothong and diphthong vowels. It is used for investigation of signal representation. Corpus II is a subset of Corpus I. It consists of the monothongs only, and is used for investigation of distinctive features.

For the experiments dealing with distinctive features, we characterized the 13 vowels in terms of 6 distinctive features, following the conventions set forth by others [13]. The feature values for these vowels are summarized in Table 2.

The classifier for our experiments was selected with the following considerations. First, to facilitate comparisons of different results, we restrict ourselves to use the same classifier for all experiments. Second, the classifier must be flexible in that it does not make assumptions about specific statistical distributions or distance metrics, since different signal representations may have different characteristics. Based on these two constraints, we have chosen to use the multi-layer perceptron (MLP) [7]. In our signal representation experiments, the network contains 16 output units representing each of the 16 vowels. The input layer contains 120 units, 40 units each representing the initial, middle, and final third of the vowel segment. For the experiments involving acoustic attributes and distinctive features, the input layer may be the spectral vectors, a set of acoustic attributes, or the distinctive features, and the output layer may be the vowel labels or the distinctive features, as will be described later.

All networks have a single hidden layer with 32 hidden units. This and other parameters had previously been adapted for better learning capabilities. In addition, input normalization and center initialization have been used [8].

SIGNAL REPRESENTATION

Review of Past Work

Several experiments on comparing signal representations have been reported in the past. Mermelstein and Davis [10] compared the mel-frequency cepstral coefficients (MFCC) with four other more conventional representations. They found that a set of 10 MFCC resulted in the best performance, suggesting that the mel-frequency cepstra possess significant advantages over the other representations. Hunt and Lefebvre [4] compared the performance of their psychoacoustically-motivated auditory model with that of a 20-channel mel-cepstrum. They found that the auditory model gave the highest performance under all conditions, and is least affected by changes in loudness, interfering noise and spectral shaping distortions. Later, they [5] conducted another comparison with the auditory model output, the mel-scale cepstrum with various weighing schemes, cepstrum coefficients augmented by the δ-cepstrum coefficients, and the IMELDA representation which combined between-class covariance information with within-class covariance information of the mel-scale filter bank outputs to generate a set of linear discriminant functions. The IMELDA outperformed all other representations.

	i	ɪ	e	ɛ	æ	a	ɔ	o	ʌ	u	ɝ	ʊ	ü
HIGH	+	+	-	-	-	-	-	-	-	+	-	+	+
TENSE	+	-	+	-	-	-	-	+	-	+	-	-	+
LOW	-	-	-	-	+	+	+	-	-	-	-	-	-
BACK	-	-	-	-	-	+	+	+	+	+	+	+	-
ROUND	-	-	-	-	-	-	+	+	-	+	+	+	+
RETROFLEX	-	-	-	-	-	-	-	-	-	-	+	-	-

Table 2: The set of distinctive features used to characterize 13 vowels

These studies generally show that the choice of parametric representations is very important to recognition performance, and auditory-based representations generally yield better performance than more conventional representations. In the comparison of the psychoacoustically-motivated auditory model with MFCC, however, different methods of analysis led to different results. Therefore, it will be interesting to compare outputs of an auditory model with the computationally simpler mel-based representation when the experimental conditions are more carefully controlled.

Experimental Procedure

Our study compares six acoustic representations [9], using the MLP classifier. Three of the representations are obtained from the auditory model proposed by Seneff [12]. Two representations are based on mel-frequency, which has gained popularity in the speech recognition community. The remaining one is based on conventional Fourier transform. Attention is focused upon the relative classification performance of the representations, the effects of varying the amount of training data, and the tolerance of the different representations to additive white noise.

For each representation, the speech signal is sampled at 16 kHz and a 40-dimensional spectral vector is computed once every 5 ms, covering a frequency range of slightly over 6 kHz. To capture the dynamic characteristics of vowel articulation, three feature vectors, representing the average spectra for the initial, middle, and final third of every vowel token, are determined for each representation. A 120-dimensional feature vector for the MLP is then obtained by appending the three average vectors.

Seneff's auditory model (SAM) produces two outputs: the mean-rate response (MR) which corresponds to the mean probability of firing on the auditory nerve, and the synchrony response (SR) which measures the extent of dominance at

the critical band filters' characteristic frequencies. Each of these responses is a 40-dimensional spectral vector. Since the mean-rate and synchrony responses were intended to encode complementary acoustic information in the signal, a representation combining the two is also included by appending the first 20 principal components of the MR and SR to form another 40-dimensional vector (SAM-PC).

To obtain the mel-frequency spectral and cepstral coefficients (MFSC and MFCC, respectively), the signal is pre-emphasized via first differencing and windowed by a 25.6 ms Hamming window. A 256-point discrete Fourier transform (DFT) is then computed from the windowed waveform. Following Mermelstein et al [10], these Fourier transform coefficients are later squared, and the resultant magnitude squared spectrum is passed through the mel-frequency triangular filter-banks described below. The log energy output (in decibels) of each filter, $X_k, k = 1, 2, .., 40$, collectively form the 40-dimensional MFSC vector. Carrying out a cosine transform [10] on the MFSC according to the following equation yields the MFCC's, $Y_i, i = 1, 2, .., 40$.

$$Y_i = \sum_{k=1}^{40} X_k \cos[i(k - \frac{1}{2})\frac{\pi}{40}]$$

The lowest cepstrum coefficient, Y_0, is excluded to reduce sensitivity to overall loudness.

The mel-frequency triangular filter banks are designed to resemble the critical band filter bank of SAM. The filter bank consists of 40 overlapping triangular filters spanning the frequency region from 130 to 6400 Hz. Thirteen triangles are evenly spread on a linear frequency scale from 130 Hz to 1 kHz, and the remaining 27 triangles are evenly distributed on a logarithmic frequency scale from 1 kHz to 6.4 kHz, where each subsequent filter is centered at 1.07 times the previous filter's center frequency. The area of each triangle is normalized to unit magnitude.

The Fourier transform representation is obtained by computing a 256-point DFT from a smoothed cepstrum, and then downsampling to 40 points.

One of the experiments investigates the relative immunity of each representation to additive white noise. The noisy test tokens are constructed by adding white noise to the signal to achieve a peak signal-to-noise ratio (SNR) of 20dB, which corresponds to a SNR (computed with average energies) of slightly below 10dB.

Results

For each acoustic representation, four separate experiments were conducted using 2,000, 4,000, 8,000, and finally 20,000 training tokens. In general, performance improves as more training tokens are utilized. This is illustrated in Figure 1, in which accuracies on training and testing data as a function of the amount of training tokens for SAM-PC and MFCC. As the size of the training set increases, so does the classification accuracy on testing data. This is accompanied by a corresponding decrease in performance on training data. At 20,000 training tokens, the difference between training and testing set performance is about 5% for both representations.

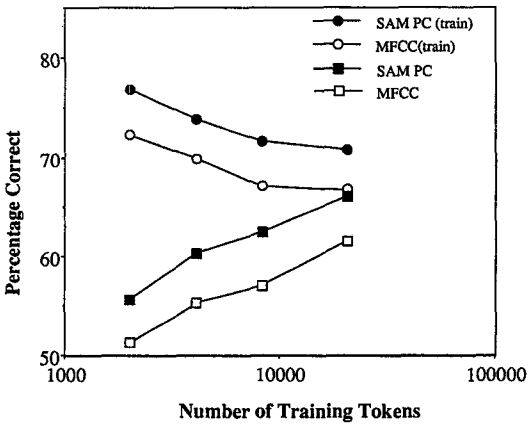

Figure 1: Effect of increasing training data on testing accuracies

To investigate the relative immunity of the various acoustic representations to noise degradation, we determine the classification accuracy of the noise-corrupted test set on the networks after they have been fully trained on clean tokens. The results with noisy test speech are shown in Figure 2, together with the corresponding results on the clean test set. The decrease in accuracy ranges from about 12% (for the combined auditory model) to almost 25% (for the DFT).

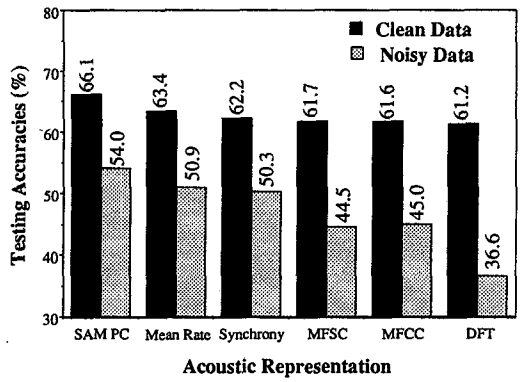

Figure 2: Performance on noisy and clean speech

ACOUSTIC ATTRIBUTES AND DISTINCTIVE FEATURES

Our experiments were again conducted using an MLP classifier for speaker independent vowel classification. Three experimental parameters were systematically varied, resulting in six different conditions, as depicted in Figure 3. These

three parameters specify whether the acoustic attributes are extracted, whether an intermediate distinctive feature representation is used, and how the feature values are combined for vowel classification. In some conditions (cf. conditions A, E, and F), the spectral vectors from the mean-rate response were used directly, whereas in others (cf. conditions B, C, and D), each vowel token was represented by a set of automatically-extracted acoustic attributes. In still other conditions (cf. conditions C, D, E, and F), an intermediate representation based on distinctive features was introduced. The feature values were either used directly for vowel identification through one bit quantization (i.e. transforming them into a binary representation) and table look-up (cf. conditions C and E), or were fed to another MLP for further classification (cf. conditions D and F). Taken as a whole, these experiments will enable us to answer the questions that we posed earlier. Thus, for example, we can assess the usefulness of extracting acoustic attributes by comparing the classification performance of conditions A versus B and D versus F.

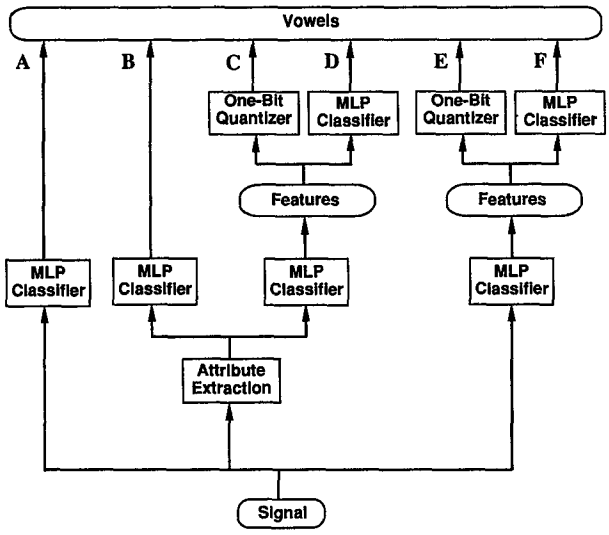

Figure 3: Experimental paradigm comparing direct phonetic classification with attribute extraction, and the use of linguistic features. The mean rate response is chosen to be the signal.

Acoustic Representation

Each vowel token is characterized either directly by a set of spectral coefficients, or indirectly by a set of automatically derived acoustic attributes. In either case, three average vectors are used to characterize the left, middle, and right thirds of the token, in order to implicitly capture the context dependency of vowel articulation.

Spectral Representation Comparative experiments described in the previous section indicate that representations from Seneff's auditory model result in performance superior to others. While the combined mean rate and synchrony representation (SAM-PC) gave the best performance, it may not be an appropriate choice for our present work, since the heterogeneous nature of the representation poses difficulties in acoustic attribute extraction. As a result, we have selected the next best representation - the mean rate response (MR).

Acoustic Attributes The attributes that we extract are intended to correspond to the acoustic correlates of distinctive features. However, we do not as yet possess a full understanding of how these correlates can be extracted robustly. Besides, we must somehow capture the variabilities of these features across speakers and phonetic environments. For these reasons, we have adopted a more statistical and data-driven approach. In this approach, a general property detector is proposed, and the specific numerical values of the free parameters are determined from training data using an optimization criterion [14]. In our case, the general property detectors chosen are the spectral center of gravity and its amplitude. This class of detectors may carry formant information, and can be easily computed from a given spectral representation. Specifically, we used the mean rate response, under the assumption that the optimal signal representation for phonetic classification should also be the most suitable for defining and quantifying acoustic attributes, from which distinctive features can eventually be extracted.

The process of attribute extraction is as follows. First, the spectrum is shifted down linearly on the bark scale by the median pitch for speaker normalization. For each distinctive feature, the training tokens are divided into two classes – [+feature] and [-feature]. The lower and upper frequency edges (or "free parameters") of the spectral center of gravity are chosen so that the resultant measurement can maximize the Fisher's Discriminant Criterion (FDC) between the classes [+feature] and [-feature] [2].

For the features [BACK], [TENSE], [ROUND], and [RETROFLEX] only one attribute per feature is used. For [HIGH] and [LOW], we found it necessary to include two attributes per feature, using the two sets of optimized free parameters giving the highest and the second highest FDC. These 8 frequency values, together with their corresponding amplitudes, make up 16 attributes for each third of a vowel token. Therefore, the overall effect of performing acoustic attribute extraction is to reduce the input dimensions from 120 to 48.

Results

The results of our experiments are summarized in Figure 4, plotted as classification accuracy for each of the conditions shown in Figure 3. The values in this figure represent the average of six iterations; performance variation among iterations of the same experiment amounts to about 1%.

By comparing the results for conditions A and B, we see that there is no statistically significant difference in performance as one replaces the spectral representation by the

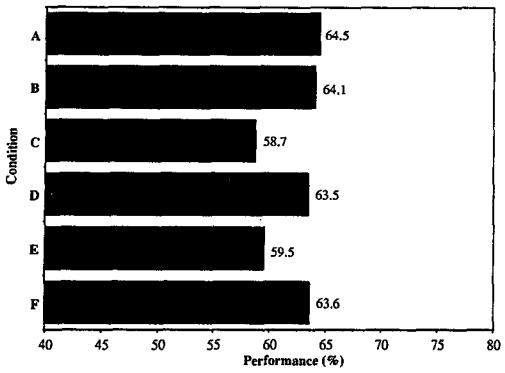

Figure 4: Performance of the six classification pathways in our experimental paradigm

acoustic attributes. This result is further corroborated by the comparison between conditions C and E, and D and F.

Figure 4 shows a significant deterioration in performance when one simply maps the feature values to a binary representation for table look-up (i.e., comparing conditions A to E and B to C). We can also examine the accuracies of binary feature assignment for each feature, and the results are shown in Figure 5. The accuracy for individual features ranges from 87% to 98%, and there is again little difference between the results using the mean rate response and using acoustic attributes. It is perhaps not surprising that table look-up using binary feature values result in lower performance, since it would require that *all* of the features be identified correctly.

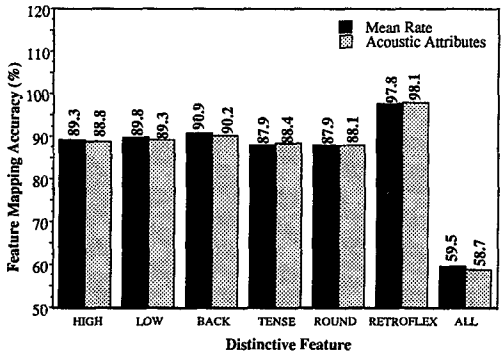

Figure 5: Distinctive features mapping accuracies for the mean rate response and acoustic attributes

However, when we use a second MLP to classify the features into vowels, a considerable improvement (> 4%) is obtained to the extent that the resulting accuracy is again comparable to other conditions (cf. conditions A and F, and conditions B and D).

DISCUSSION

Our results indicate that, on a fully trained network, representations based on auditory modelling consistently outperform other representations. The best among the three auditory-based representations, SAM PC, achieved a top-choice accuracy of 66%.

The MFSC and MFCC representations performed worse than the auditory-based representations and slightly better that the DFT. At first glance, it may appear that the discrepancies are small, since the error rate is only increased slightly (from 33% to 38%). However, previous research on human and machine identification of vowels, *independent* of context, have shown that the best performance attained is around 65% [11]. Looking in this light, the difference in performance becomes much more significant. One legitimate concern may be that principal component analysis has been applied to SAM PC, but not to MFCC. However, the cosine transform used in obtaining the MFCC performs a similar function to principal component analysis. Experiments have been conducted using the first 40 principal components of MFCC, and the classification accuracy (61.3%) shows that principal component analysis has no statistically significant effects on performance. It may also be argued that too many MFCC coefficients have been used, and this may degrade its performance. But further experiments have shown that classification accuracy increases with the number of MFCC used, and using 40 MFCC yielded the highest performance. Therefore, we may tentatively conclude that auditory-based signal representations are preferred, at least within the bounds of our experimental conditions.

Performance on noisy speech for the various representations follows the trend of that on clean speech, with the exception that the range of accuracy increased substantially. The degradation of the SAM representations was least severe - about 12%, whereas the mel-representations showed a drop of 17%. The DFT is most affected by noise, and its performance degraded by over 24%. We believe that training with clean speech and testing with noisy speech is a fair experimental paradigm since the noise level of test speech is often unknown in practice, but the environment for recording training speech can always be controlled.

Our investigation on the use of acoustic attributes is partly motivated by the belief that these attributes can enhance phonetic contrasts by focusing upon relevant information in the signal, thereby leading to improved phonetic classification performance when only a finite amount of training data is available. The acoustic attributes that we have chosen are intuitively reasonable and easy to measure. But they are by no means optimum, since we did not set out to design the best set of attributes for enhancing vowel contrasts. Nevertheless, their use has led to performance comparable to the direct use of spectral information. With an improved understanding of the relationship between distinctive features and

Pathway	A	B	C	D	E	F
Connections	4288	1984	1760	2400	4064	4704

Table 3: Sizes of the networks in our experimental paradigm.

their acoustic correlates, and a little more care in the design and extraction of these attributes, it is conceivable that better classification accuracy can be obtained.

Another advantage of using acoustic attributes is a saving on run-time computations through reduction of input dimensions. Table 3 compares the total number of connections in the one or more MLP within each condition in our experimental paradigm. With a small amount of preprocessing, the use of acoustic attributes can save about half of the computations required by the direct use of spectral representation.

One potential source of discrepancy in our experiments has to do with pitch normalization, which was not performed on the mean-rate response. However, a pitch-normalized spectral center of gravity measure was used to extract acoustic attributes, since it can eliminate singularities that complicate the search for a maximum FDC value in the optimization process. However, this advantage is obtained sometimes at the expense of getting a lower FDC value, thus leading to poorer performance. While we do not feel that pitch normalization has any significant effect on the outcome of our experiments, further experiments are clearly necessary.

To introduce a set of linguistically motivated distinctive features as an intermediate representation for phonetic classification, we first transform the acoustic representations into a set of features, and then map the features into vowel labels. While one may argue that such a two-step process is inherently sub-optimal, we nevertheless were able to obtain comparable performance, corroborating the findings of Leung [7]. Such an intermediate representation can offer us a great deal of flexibility in describing contextual variations. For example, all vowels sharing the feature [+ROUND] will affect the acoustic properties of neighboring consonants in predictable ways, which can be described more parsimoniously. By describing context dependencies this way, we can also make use of training data more effectively by collapsing all available data along a given feature dimension.

Figure 5 shows that performance on some features is worse than others, presumably due to inadequacies in the attributes that we use. For example, performance on the feature [TENSE] should be improved by incorporating segment duration as an additional attribute. When a second classifier is used to map the feature values into vowel labels, a 4-5% accuracy increase is realized such that the performance is again comparable to cases without this intermediate feature representation. This result suggests that the acoustic-phonetic information is preserved in the *aggregate* of the features, and that the subsequent performance recovery may be a consequence of the redundant nature of distinctive features, as well as the ability by the second classifier to capture various contextual effects.

Based on the results of our experiments, we may tentatively conclude that the auditory-based representations are preferred. Furthermore, the use of acoustic attributes can significantly reduce run-time computation for vowel classification with no cost to accuracy. Finally, the introduction of an intermediate representation based on distinctive features can potentially provide us with a flexible framework to describe contextual variations and make more effective use of training data, again at no cost to classification performance.

REFERENCES

[1] Chomsky N. and M. Halle, *Sound Pattern of English*, Harper & Row, 1968.

[2] Duda, R.O. and P.E. Hart, "Pattern Classification and Scene Analysis", a Wiley-Interscience publication, 1973.

[3] Fant, G., "Manual of Phonetics, ch. 8, edited by Bertil Malmberg", North-Holland Publishing Company, 1970.

[4] Hunt, M. and C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model", *Proc. ICASSP-88*, New York, 1988.

[5] Hunt, M. and C. Lefebvre, "A Comparison of Several Acoustic Representation for Speech Recognition with Degraded and Undegraded Speech", *Proc. ICASSP-89*, 1989.

[6] Lamel, L.F., R.H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus", *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, February 1986.

[7] Leung, H.C. *"The Use of Artificial Neural Networks for Phonetic Recognition,"* Ph.D. Thesis, MIT Depart. of Elect. Engin. and Comp. Sci., Cambridge, MA, 1989.

[8] Leung, H.C. and V.W. Zue, "Phonetic Classification Using Multi-Layer Perceptrons", *Proc. ICASSP-90*, 1990.

[9] Meng, H.M. and V.W. Zue, " A Comparative Study of Acoustic Representations of Speech for Vowel Classification using Multi-Layer Perceptrons", *Proc. ICSLP-90*, Kobe, 1990.

[10] Mermelstein, P. and S. Davis, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech and Signal Processing*, August 1980.

[11] Phillips, M.S., "Speaker Independent Classification of Vowels and Diphthongs in Continuous Speech, *Proc. of the 11th International Congress of Phonetic Sciences,* Estonia, USSR, 1987.

[12] Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing", *J. of Phonetics*, January 1988.

[13] Stevens, K.N., Unpublished course notes for *Speech Communications*, Department of Electrical Engineering and Computer Science, MIT, Spring term, 1989.

[14] Zue, V.W., J.R. Glass, M.S. Phillips, and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the SUMMIT System", *Proc. ICASSP-89*, Scotland, 1989.