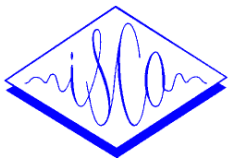

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A Comparative Study of Acoustic Representations of Speech for Vowel Classification Using Multi-Layer Perceptrons[1]

Helen M. Meng and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

ABSTRACT

This paper describes a study comparing several signal representations for context-independent vowel classification. It forms the first step in our investigation for a distinctive-feature-based approach to phonetic recognition. Six different signal representations were investigated. They include the outputs of Seneff's Auditory Model (SAM), the mel-scale representations and the conventional Fourier Transform. To strive towards a fair and meaningful comparison, the mel-frequency filters were carefully designed to resemble the filters of SAM and the dimensionality of the feature vectors were constrained to be equal. The representations were compared on the basis of classifying 16 vowels in American English. Experiments with speech degraded by adding white noise were also conducted. Our results are based on over 22,000 vowel tokens excised from 2,750 sentences spoken by 550 speakers. The combined Synchronous and Mean Rate responses from SAM outperformed all the other representations with both undegraded and noisy speech, yielding top-choice accuracies of 66% and 54% respectively.

INTRODUCTION

One of the most critical and as yet unsolved problems in automatic speech recognition is the transformation of the continuous speech signal to a discrete representation for accessing words in the lexicon. To successfully achieve this process, several important issues must be properly addressed. First, how should the speech signal be represented so that relevant acoustic information is preserved or enhanced? Second, should the resulting signal representation be used directly, or should one attempt to extract acoustic attributes that may better signify phonetic contrasts? Third, what should be the inventory of units for describing the lexical items? Fourth, what acoustic modelling procedures should be employed for the lexical units? Finally, what algorithms should be used to detect and identify these lexical units?

The overall goal of our study, of which this paper describes a part, is to explore the use of distinctive features for automatic speech recognition. Linguists generally believe that phonemes can be represented by a small set of basic linguistic units - distinctive features [1]. 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. 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.

Phonological and phonetic research conducted over the past three decades have resulted in a wealth of information, albeit incomplete, on the acoustic correlates of distinctive features. Before we can begin to contemplate how these acoustic attributes may be captured automatically, we must first select an optimal signal representation. One way to achieve this goal is to compare phoneme or word recognition performance of a recognizer using a variety of input representations. We may infer that the representation which gives the best performance should also be the most suitable for use in defining and quantifying acoustic attributes, from which distinctive features can eventually be extracted.

Several experiments on comparing signal representations have been reported in the past. Mermelstein and Davis [2] compared five representations, namely the mel-frequency cepstral coefficients (MFCC), the linear frequency cepstral coefficients, the linear prediction cepstrum, the linear prediction spectrum, and the reflection coefficients. On the task of recognizing monosyllabic words spoken continuously by two speakers, 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 [3] compared the performance of their psychoacoustically-motivated auditory model with that of a 20-channel mel-cepstrum. The first eight discriminant functions obtained by applying linear discriminant analysis on the two auditory model outputs were compared with 8 unweighted MFCC (C_1 to C_8). Experiments conducted include speaker-dependent and -independent, connected and quasi-isolated word recognition, with undegraded, noisy and spectrally tilted speech. 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, Hunt and Lefebvre [4] 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 tests conducted were similar to those in the previous comparison. The IMELDA outperformed all other representations.

In summary, these studies generally show that the choice of parametric representations is very important to recognition performance, and auditory-based representations generally yield bet-

ter 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.

This paper describes a comparative study of six acoustic representations on the task of vowel classification using an artificial neural net (ANN) classifier. Three of the representations are obtained from the auditory model proposed by Seneff [5]. 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 signal representations, the effect of increasing training data on the robustness of the results, and the tolerance of the different representations to additive white noise.

EXPERIMENTAL PROCEDURES

To strive towards a fair comparison of the various signal representations, we restricted the ANN classifier to have the same architecture throughout the experiments. All input feature vectors were measured at the same points in the speech signal, and the dimensionalities of the input vectors were all identical.

Signal Processing

The speech signal is sampled at 16 kHz and a spectral vector is computed once every 5 ms. Three feature vectors, representing the average spectra for the initial, middle, and final third of every vowel token, are determined for each representation. These vectors attempt to crudely capture the dynamic characteristics of vowel articulation. All the acoustic representations result in a 40-dimensional feature vector covering a frequency range of slightly over 6 kHz.

Seneff's auditoy 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. This is done by appending the first 20 principal components of the MR and SR to form another 40-dimentional 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 [2], 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 [2] on the MFSC according to the following

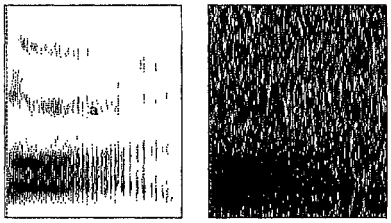

Figure 1: Wideband spectrograms showing clean and noisy speech for the vowel /a/

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, C_0, is excluded to reduce sensitivity to overall loudness.

In order to achieve as fair a comparison as possible, 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.

To obtain the Fourier transform representation, a 256-point DFT is computed in the same manner as described above. The resulting spectrum is then cepstrally smoothed and down-sampled 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 of 20dB, which corresponds to a signal-to-noise ratio (computed with average energies) of slightly below 10dB. Figure 1 shows wideband spectrograms of one of the test tokens before and after noise corruption.

Task and Corpus

Comparisons of the various signal representations are based on the task of classifying 16 American English vowels using tokens excised from the acoustic-phonetically compact portion of the TIMIT database [6]. It is a classification task in that the boundaries of the vowel token are given, and the classifier is only asked to determine the most likely label. The 16 vowels include 13 monothongs /i, ɪ, e, ɛ, æ, a, o, ʌ, ɔ, u, ʊ, ü, ɝ/ and 3 diphthongs /aʸ, ɔʸ, aʷ/. No restrictions were imposed on the phonetic contexts in which they may appear. The training data consist of over 20,000 tokens, excised from 2,500 continuous sentences spoken by 500 speakers. The testing data consist of about 2,000 tokens, excised from 250 continuous sentences spoken by 50 speakers. The size and contents of the corpus are summarized in Table 1.

Training Speakers (M/F)	Testing Speakers (M/F)	Training Tokens	Testing Tokens
500 (357/143)	50 (33/17)	20,000	2,000

Table 1: Corpus used for the experiments

ANN Classifier

The classifier used for our experiment is an artificial neural network based on multi-layer perceptrons (MLP). The details of the classifier have been described elsewhere [7]. The network contains 16 output units representing each of the 16 vowels. It has a single hidden layer consisting of 32 hidden units. The input layer contains 120 units, 40 units each representing the initial, middle, and final third of the vowel segment. Input normalization and center initialization [8] were adopted for better learning capabilites, and a weighted mean squared error criterion was used.

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, classification performance improves as more training tokens are utilized. This is illustrated in Figure 2, in which we display test set accuracies for the six different acoustic representations, using 2,000 and 20,000 training tokens. For a fully trained network, the classification accuracies for different acoustic representations differ by about 5%, with the auditory-based representations consistently yielding better results than others.

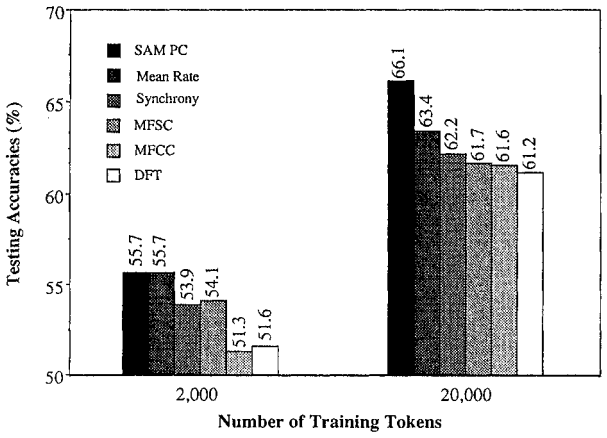

Figure 2: Performance of the six signal representations for 2,000 and 20,000 training tokens

In order to get some ideas about the robustness of the various representations, we also determined for each experiment the classification performance on training data. Figure 3 shows accuracies on training and testing data as a function of the amount of training tokens for the combined auditory representation and the popular mel-frequency cepstral coefficients. As the size of the training set increases, so does the classification accuracy on testing data. This is accompanied by a corresponding decrease

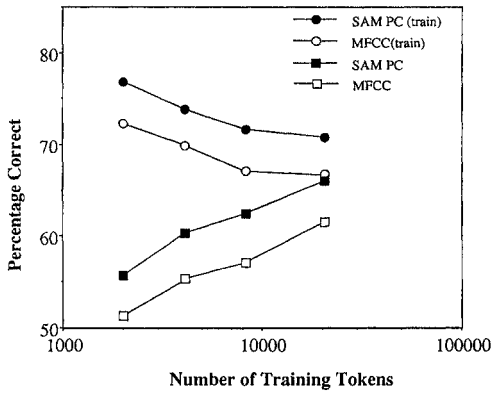

Figure 3: Effect of increasing training data on testing accuracies

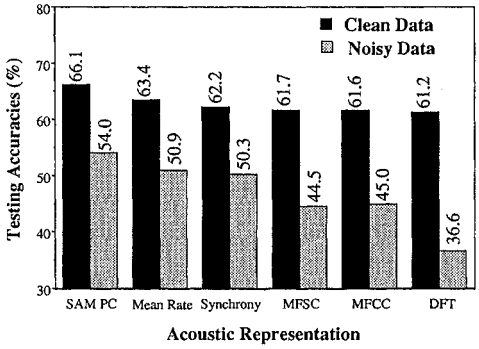

Figure 4: Performance of the different representations on noisy speech

in performance on training data. At 20,000 training tokens, the difference between training and testing set performance is about 5% for both representations.

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 4, together with the corresponding results on the clean test set. The decrease in classification accuracy ranges from about 12% (for the combined auditory model) to almost 25% (for the DFT).

DISCUSSION

Our results indicate that, on a fully trained network, acoustic 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%, which is comparable to those reported in the literature. For example, Leung [7] reported a classification accuracy of 64%, with the same network and the same data set, when synchrony and mean-rate responses were used.

When the two outputs of SAM are used separately, the performance typically drops by 3-4%, with the mean-rate response performing better than the synchrony response. This result is somewhat surprising, since the generalized synchrony detector (GSD) in SAM has the property of enhancing spectral peaks, whose locations are important for correct vowel identification. Apparently the mean-rate response also preserves the necessary acoustic information for vowel identification. It is also possible that the GSD algorithm over-sharpens the peaks in some cases, thus making the network unduely sensitive to amplitude variations at formant locations.

The MFSC and MFCC representations performed similarly on the fully trained network, 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 indentification of vowels, *independent* of context, have shown that the best performance attained is around 65% [9]. 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. It may also be argued that too many MFCC coefficients have been used, and this may degrade its performance. Further experiments may be necessary to resolve this issue. Nevertheless, we may tentatively conclude that auditory-based signal representations are preferred, at least within the bounds of our experimental conditions.

As illustrated in Figures 2 and 3, the relative performances of the six representations remained fairly stable as more training data were used. Overall, classification accuracy improved by an average of 9% as the training data increased ten-fold. The training accuracies, on the other hand, decrease as expected with more training, suggesting that the network began to extract relevant acoustic cues for phonetic distinction, rather than memorizing individual differences among tokens. The accuracies converge to less than 2% for DFT and over 5% for SR. With additional training data, we expect that the test set accuracy can continue to improve. However, as shown in Figure 3, it is not very likely that relative performances will change.

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%. The fact that the SAM representations are more immune to noise can be gleaned from Figure 5, which shows the clean and noisy versions of the same vowel in Figure 1. 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.

In summary, we conducted a set of vowel classification experiments that compare the relative merits of six acoustic representations and found that SAM based representations consistently yield superior performance. We are now pursuing other issues related to the acoustic to lexical transformation. Specifically, we would like to determine whether one should use the signal representation directly, or attempt to extract acoustic attributes that may better signify phonetic contrasts. We will also explore the possibility of introducing distinctive features as an intermediate lexical representation [10].

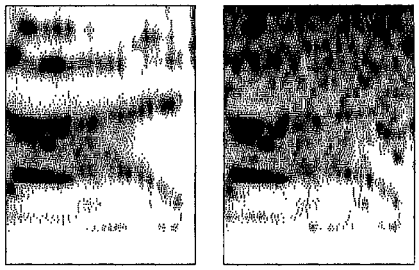

Figure 5: Synchrony spectrograms showing clean and noisy speech for the vowel /a/

ACKNOWLEDGEMENTS

We would like to acknowledge the help provided by Hong C. Leung, who generously offered his expertise in ANN, both in ideas and programs.

REFERENCES

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

[2] 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*, vol. ASSP-28, No.4, August 1980.

[3] Hunt, M. and C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model", *Proc. ICASSP-88*, New York, pp.215-218, 1988.

[4] Hunt, M. and C. Lefebvre, "A Comparison of Several Acoustic Representation for Speech Recognition with Degraded and Undegraded Speech", *Proc. ICASSP-89*, pp.262-265, 1989.

[5] Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing", *Journal of Phonetics*, vol.16, no.1, pp,55-76, 1988.

[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, pp. 100–109, February 1986.

[7] Leung, H.C. *"The Use of Artificial Neural Nets for Phonetic Recognition,"* P.h.D. Thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, May 1989.

[8] Leung, H. C. and V.W. Zue, "Phonetic Classification Using Multi-Layer Perceptrons", *Proc. ICASSP-90*, Albuquerque, pp.525-528, 1990.

[9] Phillips, M. S., "Speaker Independent Classification of Vowels and Diphthongs in Continuous Speech", *Proc. ICPhS-87*, Tallinn, pp.240-243, 1987.

[10] Meng, H. M., *"Implementing a Novel Representation of Speech - the Use of Distinctive Features for Automatic Speech Recognition"*, S.M. Thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, December 1990.