SPEAKER-INDEPENDENT VOWEL RECOGNITION: COMPARISON OF BACKPROPAGATION AND TRAINED CLASSIFICATION TREES

R. A. Cole*, Y. K. Muthusamy*, L. Atlas, T. Leen*, and M. Rudnick***

Dept. of Computer Science & Engineering*
Oregon Graduate Center
19600 NW Von Neumann Dr.
Beaverton, OR 97006

Dept. of Electrical Engineering, FT-10**
University of Washington
Seattle, WA 98195

ABSTRACT

A series of experiments compare performance of trained classification trees to multi-layer feedforward networks on speaker-independent vowel recognition using information in a single spectral slice. The vowel stimuli are exemplars of 12 monophthongal vowels of American English taken from all phonetic contexts in spoken utterances. The training set consists of 342 vowel tokens provided by 320 speakers, and the test set consists of 137 tokens provided by a different 100 speakers. The classification trees and neural classifiers are trained and tested on identical data. In addition, experiments are performed to determine the most effective way to present vowel information for classification. Classification performance is compared using (a) spectral coefficients from the DFT; (b) spectral coefficients from the pitch-synchronous DFT (PS-DFT); (c) features describing the six largest peaks in the spectrum; and (d) features derived from principal component analysis of the spectra, using an unsupervised neural network.

The results show that neural nets trained with backpropagation produce better results than classification trees in all comparable experimental conditions. Spectral coefficients from the DFT and PS-DFT produce equivalent performance. Relative to these results, using measurements of spectral peaks produces inferior performance, while using principal component analysis to eliminate redundancy in the spectrum slightly improves performance while greatly reducing the size of the network. Possible reasons for the superior classification performance obtained by the backpropagation networks is discussed, and the results are compared to other speaker-independent vowel classification studies.

1. INTRODUCTION

At a recent meeting of the Acoustical Society of America, the session devoted to automatic speech recognition was entitled "Neural Networks and Other Techniques." It may be an exaggeration to imply that neural networks are the dominant approach to computer speech recognition, but their popularity is undeniable, and researchers have reported recognition results on limited task domains that are comparable, and in some cases superior, to those obtained with more traditional techniques [1, 2, 3].

Recent advances in classification with neural networks have been paralleled by advances in other techniques, and it is important to examine the most promising of these. In this paper, we compare the performance of backpropagation (BP) networks and trained classification trees (CT) on an important real-world problem – speaker-independent classification of vowel sounds in natural continuous speech.

The mathematical details of the CT technique are presented in a book by Breiman, Friedman, Olshen, and Stone [4]. Given training data in the form of a set of multidimensional vectors { **x** }, the top node of the binary classification tree is built by choosing a threshold for one of the variables of { **x** }. This threshold is chosen to maximize the class separability of the data which passes down the two descendent branches of the top node. The input space is then divided for the two descendent nodes and each of these nodes continue to split within a subset of the original input space. These splits (and tree growth) continue recursively until all training data are assigned to a leaf node with a unique and accurate class label.

In order to reduce the expected ill-effect of overfitting the training data, a clever "pruning" criteria is then used, reducing the size of the tree and allowing for the best possible performance outside the training set. The final tree, if necessary, will fit arbitrary non-linear decision regions.

One of the attractive features of the recursive classification tree technique is the ability to accurately utilize very different types of input variables. For example, if the input patterns consist of both ordinal and categorical data, appropriate splits could still be made. Another possible advantage is the flexibility to have detailed fits in part of the input space while maintaining much smoother fits in other regions. This property, which would be indicated by trees which were asymmetrical (after pruning), could be very useful for classification problems which have non-uniform behavior in their input variables. The last, and perhaps most important, advantage of the classification trees is the careful use of advanced statistics in the design of the final classifier. For example, careful considerations have gone into the techniques used for estimation of true probability of error.

0073-1129/90/0000/0132$01.00 © 1990 IEEE

In order to evaluate the performance of BP and CT, we chose a very difficult task: speaker-independent classification of vowels excised from continuous speech. Speaker-independent vowel recognition is difficult because of the many sources of variability that influence the physical realization of a vowel. The two main sources of variation are the length of the speaker's vocal tract [5] and the phonetic context in which the vowel occurs. For example, the second vowel formant of [ih] in "kick" and "Lil" may differ by as much as 1000 Hz in the two contexts. Additional sources of variation include speech rate, syllable stress and word stress.

To make the task even more difficult, in the present experiments, the classifiers were presented only with the information in a single spectral slice. The spectral slice was taken from the center of the vowel, where the effects of coarticulation are least apparent.

There are several interesting reasons to investigate vowel recognition using a single spectral slice. First, experiments using individual spectra provide an excellent way to compare different representations of speech (e.g., DFT, LPC, auditory models). The most important feature of any speech representation is how well it preserves phonetic information. Classification using a single spectral slice provides one measure of this capability. We compare classification using two representations of speech, the DFT and pitch-synchronous DFT (PS-DFT). Second, the present experiments address the question of how best to present information about the spectrum to a classifier. Should the classifier be presented with the complete set of spectral coefficients, or will some processing scheme produce better classification results? We compare classification using the complete set of spectral coefficients verses a processing scheme that captures information about the spectral peaks in the spectrum. A second preprocessing scheme identifies features in the input spectra that are information-rich. This principal component analysis does not rely on specific knowledge of the speech signal. Instead features in the raw spectra that show the largest statistical variation are identified, and the original spectra are encoded in terms of these features. Finally, experiments using a single spectral slice allow us to examine how performance may be improved by adding additional features to the spectrum. In the present studies, these features include estimates of the pitch, duration and relative amplitude of the vowel.

2. EXPERIMENTS

2.1. Stimuli

The stimuli for the experiments consisted of featural descriptions of the 12 monophthongal vowels of English, shown in Table 1. The vowels were excised from all phonetic contexts in utterances of the TIMIT database, a standardized acoustic phonetic corpus of continuous speech, displaying a wide range of American dialectical variation [6,7]. The diphthongs /oy/, /ay/, /ey/, /aw/ were excluded because they are characterized by spectral change, and are therefore inappropriate for experiments using information from a single spectral slice.

Table 1. The 12 Vowel Classes

Phone	Example	Phone	Example
/iy/	b*ea*t	/ah/	b*u*tt
/ih/	b*i*t	/uw/	b*oo*t
/eh/	b*e*t	/uh/	b*oo*k
/ae/	b*a*t	/ao/	b*ough*t
/ix/	rose*s*	/aa/	c*o*t
/ax/	th*e*	/er/	b*ir*d

In one set of experiments, each vowel was represented by a set of features describing the information in a spectral slice near the center of the vowel, as shown in Figure 1. In a second set of experiments, the spectral information was augmented by three additional features providing estimates of the fundamental frequency, duration and relative amplitude of the vowel. The feature set used in these experiments is described in greater detail in sections 2.2 and 2.3.

In all experiments, the training set consisted of 342 exemplars of each vowel provided by 320 speakers, for a total of 4104 vowel spectra. The test set consisted of 137 exemplars of each vowel, provided by a different set of 100 speakers, for a total of 1644 vowel spectra.

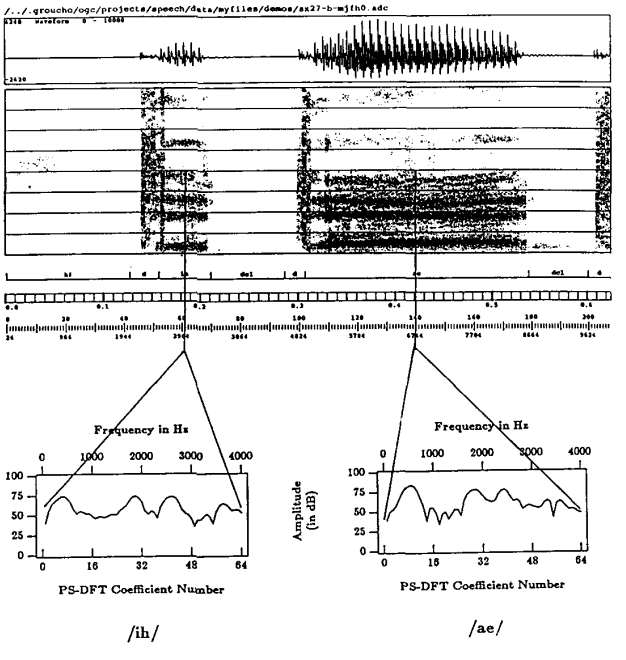

Figure 1. Waveform, pitch-synchronous spectrogram and the individual vowel spectra for a portion of the utterance "Did dad do acacemic bidding?".

The number of tokens of each vowel class was determined by the number of tokens in the least frequent class. It was found that the vowel */uh/* had the least number of tokens, 342 in the training set, and 137 in the test set. (In comparison, the corresponding figures for the most frequent vowel class */ix/* were 5798 and 1809 respectively). For each of the remaining 11 vowel classes, 342 tokens were selected by iterating through all the 320 speakers, picking one token at random from each speaker, until 342 different tokens were obtained. This procedure ensured that there was wide across-speaker variation in the tokens selected. A similar procedure was followed in creating the test set (137 tokens per class).

2.2. Spectral Representations

Three spectral representations were compared: (a) Constant-increment discrete Fourier transform (DFT), (b) Pitch-synchronous discrete Fourier transform (PS-DFT), and (c) Information about the six largest peaks in the pitch-synchronous spectrum.

DFT. A 256-point real DFT was computed on each utterance, with a 10 ms Hamming window and 3 ms increment, yielding 128 spectral coefficients every 3 ms frame of the utterance. Since the important information about vowel identity is found below 4 kHz., only the first 64 spectral coefficients (0–4 kHz) were used. Using the hand-segmented phonetic transcriptions provided in the TIMIT database, the center frame of each vowel token was located, and the first 64 spectral coefficients corresponding to this frame were extracted. The coefficient values were normalized to lie between 0 and 1 in order to train the neural networks. Normalization was done by computing the "relative value" of each coefficient with respect to the maxima and minima in the 64 coefficients, as shown in (1)

$$\textit{normalized value} = \frac{(X - min)}{(max - min)} \qquad (1)$$

where X is the value of any spectral coefficient, *max* is the value of the largest of the 64 spectral coefficients, and *min* is the value of the smallest of the 64 spectral coefficients.

PS-DFT. The main difference between the constant-increment DFT and the PS-DFT is that the latter usually provides better resolution of formants (resonant frequencies of the vocal tract) in voiced portions of speech. This can be seen in Figure 2, which shows the waveform and spectrograms of the same utterance using the DFT and PS-DFT.

The PS-DFT was computed for every pitch period with a window that began with the zero-crossing before the pitch period, and extended to the zero crossing before the following pitch period. Pitch periods were located automatically using a neural network classifier that locates pitch periods about 98% of the time [8]. We located the pitch period closest to the center frame of each vowel token. When no pitch period was found within 20 msec of the center frame (less than 0.1% of the time), the default DFT window (10 ms) and increment (3 ms) were used. As in the case of the DFT, the first 64 spectral coefficients corresponding to the center frame of each vowel were extracted. The average (normalized) pitch-synchronous spectra for the 12 vowel categories are shown in Figure 3.

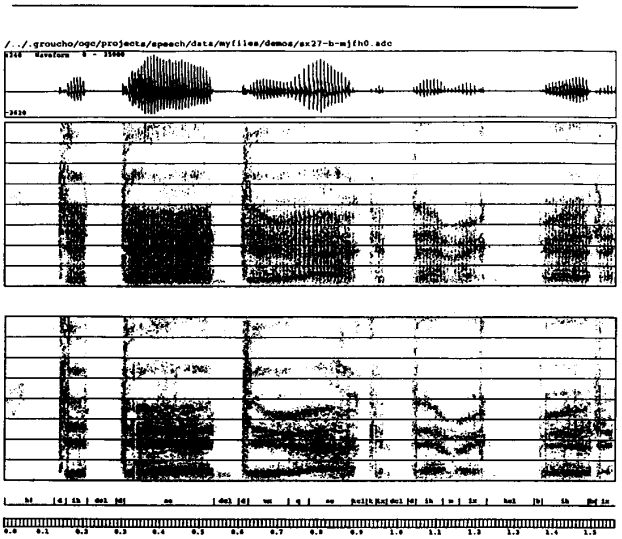

Figure 2. The waveform, DFT (top) and PS-DFT for the utterance "Did dad do academic bidding?". Note that the formants are much clearer in the PS-DFT than in the DFT.

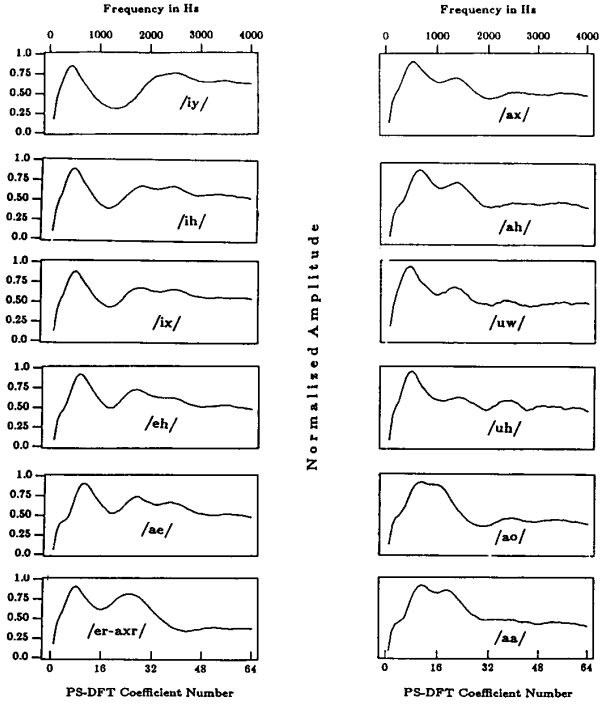

Figure 3. The average normalized pitch- synchronous spectra for the 12 vowel classes. Note the similarity between the spectra of the pairs /ix/ - /ih/, /ax/ - /ah/ and /ao/ - /aa/.

2.3. Dimension Reduction Strategies

Spectral Peaks It is well known that formant frequencies contain the most important information about vowel identity [9]. Formant tracking is very difficult, but formant information is contained within the peaks of the spectrum. We reasoned that superior classification performance might be obtained by training a classifier with information about the largest peaks in the spectrum, since the important information is retained using a smaller set of features.

Information about the six largest peaks in the pitch-synchronous spectrum consisted of (a) the frequency location (coefficient number) of the peak, (b) the magnitude of the peak (in dB), (c) the frequency location of the 3 dB fall before the peak, and (d) the frequency location of the 3 dB fall after the peak. The last two features provide information about the width of the peak, which should be especially useful when two formants merge to form a single (broad) peak. Thus, there were 24 numbers associated with each vowel label. Figure 4 shows the pitch-synchronous spectrum of a single /iy/ token with the six biggest peaks and the 3 dB fall-off points marked on the spectrum.

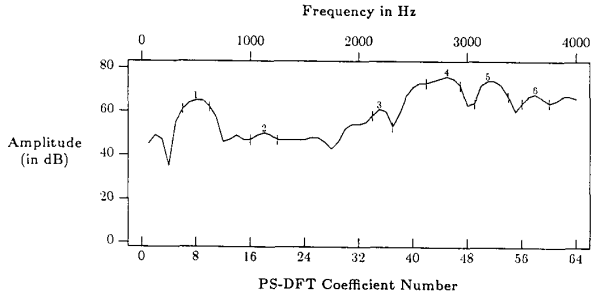

Figure 4. Pitch-synchronous Spectrum of a single /iy/ token showing the locations of the six largest spectral peaks and the 3 dB fall-off points on either side of the peak

Spectral Principal Components The use of spectral peaks as input features relies on expert knowledge of the problem (i.e. identification of the formants as distinguishing spectral characteristics). This procedure, though intuitively promising, provides no quantitative measure of the information captured in the tokens chosen. In this section we describe an alternate approach which makes use of statistical properties of the input signal, and provides estimates of the information captured by the features constructed. Our purpose here is to outline the approach and give the basic results as applied to speech recognition. A more detailed exposition will appear elsewhere.

The alternate strategy is to chose a set of feature tokens based on statistical properties of the input signal. The goal is to reduce the size of the input vector by eliminating redundant degrees of freedom in the signal, without discarding information. Intuitively, the amplitudes of the spectral components for neighboring channels will be highly correlated, and thus redundant. One approach to eliminating this redundancy is to form linear combinations of the original features that are statistically uncorrelated. These linear combinations are adopted as new basis features.

When the original data is expressed in terms of the new basis, an inherent ordering emerges; across the ensemble of input data, the amplitude variance is different from one feature to the next. Those features for which the amplitude variance is large carry the most information. Conversely, features for which the amplitude variance is small carry little information and can be discarded without significant loss. The dimension of the input space can thus be reduced by retaining only those statistically uncorrelated features which carry the most information.

Principal component analysis (PCA) is a linear transformation technique for constructing statistically uncorrelated feature vectors [10]. In PCA, each input vector is expanded in the basis of eigenvectors of the input covariance matrix. Related techniques include the Karhunen- Loeve transformation, and singular value decomposition. [11]

In traditional implementations of PCA, the covariance matrix of the data sample is estimated and then diagonalized using standard algebraic techniques. Singular value decomposition applied to the matrix of data vectors achieves the same goal. In either case, rather extensive matrix calculations are involved. Recent advances in stochastic approximation theory provide means for recursively estimating the eigenvectors of the input covariance, without explicit construction of the covariance matrix itself. [12] These techniques can handle data spaces of large dimension, and can operate in real time on an incoming data stream without requiring large data storage resources.

More intriguing, these stochastic approximation algorithms map quite naturally onto unsupervised neural networks. Oja [13] showed that a single model neuron that develops according to a Hebbian learning rule performs a limited PCA. The activity of the mature neuron is proportional to the component of its input along the eigenvector of largest variance. Sanger [14] extended Oja's work to a two-layer unsupervised network that performs a complete PCA. The activity of an output neuron is proportional to the component of the input along a specific eigenvector. Each neuron corresponds to a different eigenvector. The network orders the outputs according to decreasing variance, and we retain only the first few, information-rich nodes.

We implement Sanger's approach to encode spectral slice signals. We present the PCA network with the DFT from the 4104 vowel exemplars in the training set. During this learning phase, the network discovers the statistics of the input ensemble. For comparison, we have performed a PCA on exemplars from the same data set, using a standard algebraic algorithm. Because of computational limitations, we are only able to perform the algebraic PCA on the first 1000 vectors of the training set. Figure 5 shows the leading 20 eigenvalues from the neural network and algebraic calculations. The network results agree well with those from the algebraic algorithm. Eigenvectors were also well estimated.

The progressive drop in the magnitude of the eigenvalues indicates that information is concentrated in the highest-order principal components. This condition allows for dimension reduction by truncating the series of feature tokens retained. The results of applying this procedure to vowel classification are given in section 3.

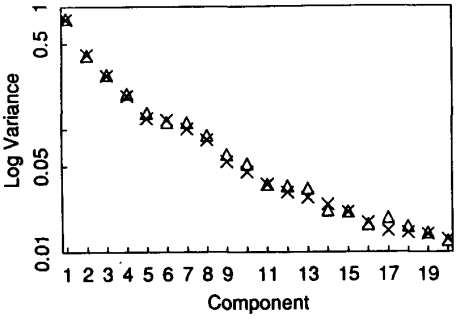

Figure 5. The leading 20 eigenvalues of the vowel-slice covariance matrix. The results from the algebraic calculation are denoted by crosses, while the neural network estimates are denoted by triangles.

2.4. Additional Features

In a second series of experiments, we examined the effect of adding three additional features, providing estimates of pitch (P), duration (D) and amplitude (A), to the information in a single spectral slice. The additional features, described below, were computed for both the training and test utterances and appended to the corresponding feature vectors in the PS-DFT and spectral peaks representations. (The DFT representation was not considered in this series of experiments as the results with DFT were comparable to those with PS-DFT).

Median Pitch. The fundamental frequency (F0) of a sound is positively correlated to the length of the vocal tract, and hence to the formant frequencies. In general, female speakers have shorter vocal tracts (and therefore higher formants) and higher F0 than male speakers. F0 should therefore be useful as a feature for normalizing differences in formant frequencies (realized as differences in spectral shape) across speakers.

Pitch peaks were located automatically using the neural network classifier used in the first series of experiments. The median pitch was calculated based on the 10 pitch peaks closest to the center of the vowel. The median F0 for female speakers for these vowels was 207 Hz and 119 Hz for male speakers. Figure 6 displays the median F0 for each of the 12 vowel classes averaged over all speakers. The dotted lines encompass the range defined by one standard deviation from the mean.

Duration. The duration estimate was taken from the phonetic transcriptions provided in the TIMIT database. As Figure 7 shows, many of the spectrally similar, perceptually confusable vowel pairs can be distinguished by their duration. In particular, the reduced vowels /ix/ and /ax/ can be distinguished from their spectrally similar counterparts /ih/ and /ah/, respectively.

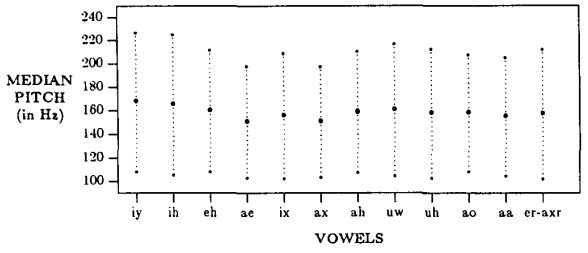

Figure 6. Average Median Pitch - Training Set (342 tokens per class)

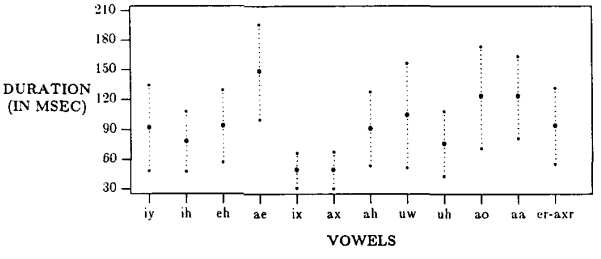

Figure 7. Average Duration - Training Set (342 tokens per class)

Relative Amplitude. The amplitude estimate was based on the peak-to-peak amplitude computed in a 10 msec window in the filtered waveform between 0 and 700 Hz. The relative amplitude of the vowel was the maximum peak-to-peak amplitude in a 30 ms window around the center of the vowel, divided by the maximum peak-to-peak amplitude in a larger window, extending 300 ms behind and 250 ms ahead of the vowel. Figure 8 shows the filtered waveform and 0–700 Hz peak-to-peak amplitude for an utterance from TIMIT-1. Figure 9 shows the average relative amplitude for the 12 vowel tokens. It can be seen that the spectrally similar vowel pairs /ix/–/ih/ and /ax/–/ah/ have different amplitudes.

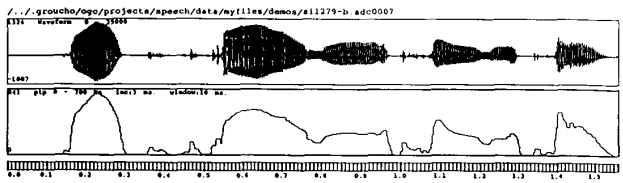

Figure 8. The filtered waveform and peak-to-peak amplitude (0-700 Hz) for the utterance "Bricks are an alternative".

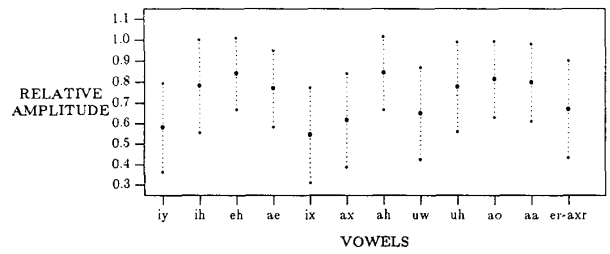

Figure 9. Average Relative Amplitude - Training Set (342 tokens per class)

2.5. Procedure

The neural network classifiers were fully connected feedforward networks (no recurrent links), as shown in Figure 10. The number of input units of the network was determined by the number of features used in the experiment (e.g., 64 for DFT and PS-DFT). All of the networks had 12 output units, corresponding to 12 vowel categories.

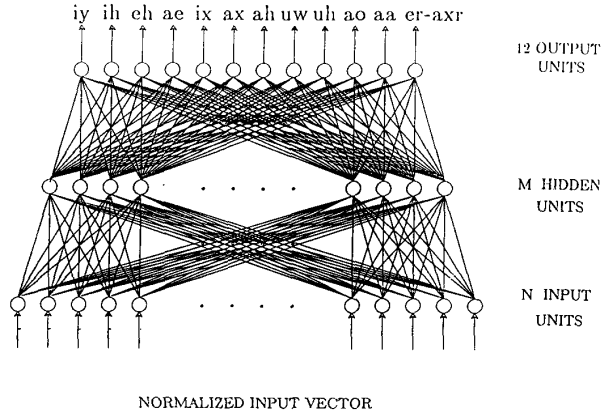

Figure 10. Illustration of a multi-layer feedforward network. In our experiments, M and N took on several values.

The number of hidden layers and the number of units in each hidden layer (one or two) was determined experimentally. We parametrically investigated network configurations with one and two hidden layers and different numbers of hidden units in each layer. The network configuration that produced the best results for each condition were: DFT: 64-40-12; DFT + PDA: 67-16-8-12; PS-DFT: 64-32-12; PS-DFT + PDA: 67-16-12; Spectral Peaks: 24-16-8-12; Spectral Peaks + PDA: 24-16-8-12.

The networks were trained using backpropagation with conjugate gradient optimization [15]. The procedure for training and testing a network proceeded as follows: The network was trained on 100 iterations through the 4104 training vectors. The trained network was then evaluated on the training set and the 1644 test vectors. This process was continued and the performance of the network on the training and test vectors was recorded after every 100 iterations through the training set. The training was stopped when the network had converged; convergence was observed as a consistent decrease or leveling off of the classification percentage on the test data over successive sets of 100 iterations, as shown in Figure 11. Typically, the networks converged after about 1100-1200 iterations and took around 77-90 hours to converge on a Sequent Symmetry.

Classification trees were trained and tested using the CART program. CART is a commercially available classifier [4] that is widely-used in the statistics community. The identical training set and test set was used to compare classifiers.

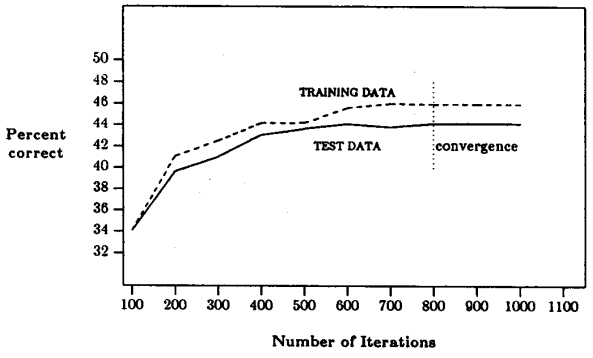

Figure 11. Performance of a 24-16-12 network on the Spectral Peaks input

3. RESULTS

The classification performance of the BP network and classification tree (CT) are shown for the different experimental conditions in Tables 2 and 3. It can be seen that:

- the neural net classifiers produced superior performance in all conditions. The best performance by a neural net (DFT+PDA) was 56%, compared to 42% for the best classification tree (Spectral Peaks +PDA).
- DFT and PS-DFT representations yielded about the same level of accuracy for both NNs and CTs,
- principal components of the DFT (PC-DFT) yielded a slight improvement over DFT with an 83% reduction in the network size, 512 vs 3040 weights,
- for NNs, the poorest performance was observed using spectral peaks; for CTs, performance was comparable for all representations,
- the addition of median pitch, duration and relative amplitude improved classification performance by about 10% for the NNs and about 4% for the CTs for all representations.

Analysis of confusion matrices (not shown) revealed that the main benefit of adding PDA estimates was a reduction in the number of confusions between /ix/ – /ih/, /ao/ – /aa/, and /ax/ – /ah/.

Table 2. Performance Comparisons of the BP-NN and CT on the Three Representations

Representation	Neural Network	Classification Tree
DFT	47.38%	34.0%
PC-DFT	48.48%	NA
PS-DFT	47.08%	37.22%
Spectral Peaks	35.40%	37.41%

Table 3. Performance Comparisons of BP-NN and CT on the Three Augmented Representations

Representation	Neural Network	Classification Tree
DFT with PDA	56.02%	39.29%
PS-DFT with PDA	54.87%	40.0%
Spectral Peaks with PDA	45.01%	41.85%

4. LISTENING EXPERIMENTS

In order to better interpret these results, we have initiated listening experiments using the vowel tokens in the training and test sets. These experiments will allow us to compare human identification performance with that of our classifiers. Although many vowel recognition experiments have been performed [16], none have used a large number of monophthongal vowels excised from continuous speech from a variety of contexts, with sufficient training on this type of classification task. We believe that a lengthy training procedure is necessary, since subjects are not used to hearing segments removed from fluent speech. Training on a large set of tokens, with feedback on each trail, provides a fair estimate of listeners' classification abilities on the test set. The results of these listening experiments will be presented in a subsequent paper.

5. DISCUSSION

5.1. Comparison of BP and CART

We find it very interesting that, in all experiments, a neural net trained with backpropagation outperformed CART by a significant amount. There are several possible reasons for the superior performance of the BP networks, all of which we are currently investigating. One advantage may stem from the ability of BP to easily find correlations between large numbers of variables. CART is best suited for finding significant properties of single features (i.e. those whose range is associated with a given class). Although it is possible for CART to form arbitrary nonlinear decision boundaries, the efficiency of the recursive splitting process may be inferior to BP's nonlinear fit. Another relative disadvantage of CART may be due to the successive nature of node growth. For example, if the first split that is made for a problem turns out, given the successive splits, to be sub-optimal, there is no means of changing the first split to be more suitable.

CART also has a technique available to automatically design hyperplane splits on some of the input variables. We did use this option for many of our comparisons and found that there was no significant performance improvement. It should be noted, however, that we have yet to do a thorough study to find the most appropriate application of the hyperplane splitting with CART.

We feel that the careful statistics used in CART could also be advantageously applied to BP. The superior performance of BP is not yet indicative of best performance and it may turn out that careful application of statistics may allow advancements in the BP technique. It also may be possible that there would be vowel representations that would cause better performance for CT than for BP. This continued interplay between BP and CART is an important part of our research in progress.

5.2. Implications for Speech Recognition

How much information does a single spectral slice convey about the identity of a vowel, and what is the best way to encode this information for presentation to a neural network? Our results show that the information in a single spectral slice enables neural networks to distinguish between 12 spectrally confusable vowel classes with an accuracy of about 47% (as opposed to 8.33% by chance). This performance can be improved to about 56% by the addition of features that capture important information in the vicinity of the spectral slice.

Insights about the amount of information in a single spectral slice can be obtained by comparing the present results to human listening performance, and to recognition experiments using sequences of spectra. Although our own listening experiments are not yet complete, Phillips [16] presented listeners with segments excised from continuous speech from a set that included 19 vowel sounds. The average listener to listener agreement on the labels was about 65%.

Table 4 summarizes recent speaker-independent vowel recognition studies. It is interesting to note that, with just a single spectral slice, we obtained recognition accuracies of 47% for 12 vowels, compared to 60% obtained by Leung and Zue with *three* averaged spectra, for a larger set of 16 monophthongal vowels and diphthongs. With the addition of duration, median pitch and relative amplitude, our results improve to 56%, compared to 77% obtained by Leung and Zue with the addition of duration and phonetic context.

The comparison of representations suggest that, if only a single spectral slice is available for classification, spectral coefficients provide a better representation than spectral peak features alone. However, our results with spectral compression using PCA suggest that reducing redundancy can improve accuracy. Classification accuracy using principal components of the DFT is slightly better than the full complement of spectral coefficients. Experiments are now underway to determine the statistical validity of this result.

More importantly, PCA achieves a significant reduction in the complexity of the classifier network without loss of accuracy; a five-fold reduction in the number of weights relative to the full DFT representation. Since the training time for backpropagation is found to scale as a polynomial in the number of network weights [17] the reduction in complexity is attractive from a computational viewpoint. As larger problems are tackled with neural networks, compact, information-rich data representations will become an absolute necessity. Furthermore, studies of VLSI implementation for neural computers [18] indicate that hardware costs will increase as the cube of the node fan-in. For a feedforward classifier network this implies cubic scaling with the number of input features. Thus, hardware implementations will require compact data representations as well.

Table 4. Speaker-Independent Vowel Recognition Studies

Study	Classification Approach	Stimuli	Training tokens/ speakers	Test tokens/ speakers	Representation	Results
Lee and Hon (1988)	Context-sensitive HMM phone models	All sonorants in 2830 utterances	approx. 30,000/357 (from 2830 utterances)	6061/20	LPC cepstral coefficients	65.71%
Seneff (1987)	Line-formants	16 monophthongal vowels and diphthongs	2135/288	2135/288	Synchrony Spectrogram	48.5%
Phillips (1986)	Pairwise Bayesian classifiers	19 monophthongal vowels and diphthongs	approx. 10,000/100 (from 1000 utterances)	approx. 1600/20 (from 160 utterances)	Feature measurements (formants, duration, pitch, etc.)	48%
Robinson (1989)	Several classifier models	11 monophthongal vowels	88/8	77/7	LPC-derived log area ratios	56% (nearest neighbor classifier)
Leung and Zue (1988)	Multi-layer Perceptrons	16 monophthongal vowels and diphthongs	20,000/500	2,000/50	Average spectra from each third of vowel (the synchrony spectrogram) and Additional features	60% (spectra alone) 77% (with additional features)
Present Results (1989)	Multi-layer Perceptrons	12 monophthongal vowels	4104/320	1644/100	1 spectral slice (DFT and PS-DFT) and Additional features	47.38% (slice alone) 56.02% (additional features)

References

1. A. Waibel, T. Hanazawa, K. Shikano, G. Hinton, and K. Lang, "Speech recognition using time-delay neural networks," TR-1-0006, ATR Interpreting Telephony Research Laboratories (1987).
2. R. Watrous, "Speech Recognition using Connectionist Networks," Doctoral Dissertation, University of Pennsylvania (1988).
3. R.P. Lippmann, "Review of Neural Networks for Speech Recognition," *Neural Computation* 1(1) pp. 1-38 MIT Press, Cambridge, Massachusetts., (Spring 1989).
4. L. Breiman, J.H. Freidman, R.A. Olsen, and C.J. Stone, *Classification and Regression Trees,* Wadsworth and Brooks Inc., Monterey, California (1984).
5. G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *Journal of the Acoustical Society of America* **24** pp. 175-184 (1952).
6. W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specification and status," pp. 93-100 in *Proceedings of the DARPA Speech Recognition Workshop,* (February, 1986).
7. L. Lamel, R. Kassel, and S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," pp. 100-110 in *Proceedings of the DARPA Speech Recognition Workshop,* (February, 1986).
8. R. A. Cole, E. Barnard, M. Vea, and F. Alleva, "Classification of pitch periods using expert knowledge and neural net classifiers," *Journal of the Acoustical Society of America* **84** p. S60 (A) (1988).

9. D. H. Klatt, "Prediction of perceived distance from critical band spectra: A first step," pp. 1278-1281 in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing,* (May, 1982).

10. Watanbe, Satosi, "Karhunen-Loeve Expansion and Factor Analysis, Theoretical Remarks and Applications," pp. 645-660 in *Transactions of the 4th Prague Conference onf Information Theory, Statistical Decision Functions and Random Processes,* (1965).

11. Broomhead, D.S., King, G.P., "Extracting qualitative dynamics from experimental data," *Physica* **D 20** p. 217 (1986).

12. Oja, E., Karhunen, J., "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *J. of Math. Anal. and Appl.* **106** pp. 69-84 (1985).

13. E. Oja, "A simplified neuron model as a principal component analyzer," *J. Math Biology* **15** pp. 267-273 (1982).

14. Sanger, T., "An optimality principle for unsupervised learning," in *Advances in Neural information Processing Systems 1,* ed. Touretzky, D.S.,Morgan Kauffmann (1989).

15. E. Barnard and D. Casasent, "Image processing for image understanding with neural nets," in *International Joint Conference on Neural Nets,* (1989). (Submitted for publication.)

16. M. Phillips, "Speaker-independent classification of vowels and diphthongs in continuous speech," in *Proc. of the 11th International Congress of Phonetic Sciences,* , Estonia, USSR (1987).

17. G. E. Hinton, "Connectionist learning procedures," CMU-CS-87-115, Carnegie-Mellon University (1987).

18. D. Hammerstrom, "A Connectivity Analysis of Recursive, Auto-Associative Connection Networks," Computer Science/E-86-009, Dept. of Computer Science/Engineering, Oregon Graduate Center (1986).