A Comparison of DFT, PLP and Cochleagram for Alphabet Recognition

Mark Fanty and Ronald Cole

Oregon Graduate Institute
19600 N.W. Von Neumann Dr.
Beaverton, OR 97006

Malcolm Slaney

Apple Computer Inc.
20525 Mariani Ave
Cupertino, CA 95014

Abstract

The English alphabet is a small but difficult vocabulary for speech recognition, with many fine phonetic distinctions, such as M/N and B/V. We use speaker-independent classification of isolated English letters to evaluate the relative performance of the DFT, Perceptual Linear Predictive analysis, and the cochleagram auditory model. Feedforward neural network classifiers were trained using all three representations on 60 speakers and tested on 60 new speakers. Training and testing data were independently modified by adding two levels of Gaussian noise and babble (20 random letter utterances, attenuated and given random offsets). PLP gave the best results, especially when trained or tested on Gaussian noise.

1 Introduction

The English alphabet is a small but difficult vocabulary for speech recognition, with many fine phonetic distinctions, such as M/N, B/V, B/D and T/G. Because the task requires fine phonetic distinctions, it is ideal for comparing signal representations for computer speech recognition. The authors have previously achieved speaker-independent classification rates of 96% on the English alphabet [1] using DFT and a variety of other features. Classification begins with a broad-category (fricative, closure, stop, sonorant) segmentation of the signal. Spectral and other features are extracted from the various segments of the letter and used by a neural network classifier.

Previous studies have shown the effectiveness of auditory models for speech recognition with degraded and undegraded speech [6, 4, 7]. This paper presents comparative results for three representations when used for whole-word classification with neural networks. The Discrete Fourier Transform (DFT) is included as a baseline. Two auditory models, Perceptual Linear Predictive (PLP) analysis and the cochleagram, are used for the same task.

2 The Representations

2.1 DFT

A 128 point FFT is computed on a 10 msec window (extra points folded back in). The signal was Hanning windowed. The result of the FFT is preemphasized to increase the magnitude of the higher frequencies, then converted to decibels. This yields 64 numbers per frame. Based on previous work, we used all 32 coefficients up to 4 kHz, and compressed the coefficients in the range from 4 kHz to 8 kHz down to 8, yielding 40 numbers per frame.

2.2 Perceptual Linear Predictive Analysis

The PLP speech analysis technique [3] estimates an all-pole autoregressive model of the auditory-like short-term speech spectrum. PLP has been shown to be efficient in suppressing speaker-dependent components in the speech signal, and uses fewer coefficients than the DFT. Compared to the conventional linear predictive (LP) analysis of speech (which estimates an all-pole model of the short-term power spectrum), the number of coefficients needed to obtain comparable recognition performance is typically lower for the PLP.

The auditory-like spectrum is obtained by integrating the short-term power spectrum of speech over simulated critical-band auditory masking curves, resampling the integrated spectrum in approximately 1 Bark intervals, modifying the spectral amplitude by

1058-6393/91 $01.00 © 1991 IEEE

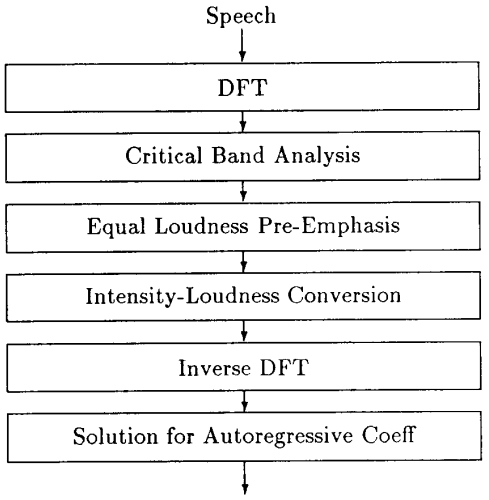

Figure 1: Stages in the PLP analysis.

a simulated fixed equal-loudness curve and compressing it through the cubic root nonlinearity to simulate the intensity-loudness power law of hearing (see Figure 1).

This autoregressive modeling efficiently approximates the spectral peaks in the auditory-like spectrum. The cepstral coefficients of the PLP all pole model are recursively computed, and weighted with an exponential window so all coefficients have a similar range for input to a neural network. Eight cepstral coefficients, including log power, of a seventh order PLP model are produced for each frame.

2.3 Cochleagram

The cochlear model designed by Lyon [5] and described by Slaney [8] converts a sound waveform into a multidimensional vector that represents the information sent from the ear to the brain. This system is illustrated in Figure 1. It is important to remember that the cochlear model used here does not try to accurately model the internal structure of the ear but only to approximate the information contained in the auditory nerve.

The cochlear model in this study uses a relatively simple filter to simulate the response of the outer and middle ears. A cascade of second order filters is used to model the propagation of sound along the basilar membrane. At each point along the cochlea, the basilar membrane responds best to a broad range of frequencies and it is this movement that is sensed by the inner hair cells. The "best" frequency of the cochlea

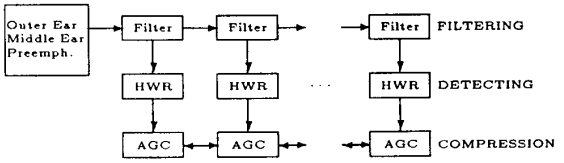

Figure 2: Lyon's Cochlear Model.

varies smoothly from high frequencies at the base to low frequencies at the apex. Inner hair cells only respond to movement of the basilar membrane in one direction. This is simulated in Lyon's model with an array of Half Wave Rectifiers (HWRs) that detect the output of each second order filter. The HWR nonlinearity serves to convert the motion of the basilar membrane at each point along the cochlea into a signal that represents the energy of the acoustic input and retains the fine time structure.

Finally, four stages of Automatic Gain Control (AGC) allow the cochlear model to compress the dynamic range of the input to a level that can be carried on the auditory nerve. The AGC used here also serves to simulate the ear's adaptation to loud sounds.

For these experiments, the 85 outputs per frame are compressed to 42 for efficiency. (Experiments showed little difference in classification accuracy after compression.) The temporal delay of wave propagation in the cochlea is part of the model. The result is that low frequency output for a speech event is delayed with respect to high frequency output. Feature extraction based on boundary locations took the delays into account.

3 Experiments

3.1 Data

A subset of the ISOLET database [2] was used for these experiments. The training set consisted of 60 speakers saying each letter of the English alphabet, in isolation, twice. The test set consisted of 60 different speakers saying each letter twice. Speech was digitized at 16 kHz using 16 bits per sample.

3.2 The Input Features

A rule-based segmenter was used to provide the broad-category segmentation used in all the experiments. This segmentation was fixed at the beginning and used in all experiments. All letters except W have a single sonorant. The feature extraction routine first finds the (longest) sonorant. This is divided into seven equal parts and the spectral representation is averaged over each of those parts. If there is a stop or fricative preceding the sonorant, the spectrum is averaged from three equal parts; if no preceding consonant is present, the spectrum is averaged over the previous 198 msec.

The 240 msec following the sonorant is also divided into three equal parts and the average spectrum is extracted from each interval. A single non-spectral feature is used: the duration of the consonant before the sonorant.

3.3 Network Configurations

Three layer networks, totally connected between layers, were trained and tested using back propagation with conjugate gradient descent. The networks were run through successive sets of 40 or 80 epochs, then tested. Training was halted when performance on the test set leveled off. The PLP networks had 105 inputs; the DFT networks had 521 inputs; the cochleagram networks had 547 inputs. All networks had 48 hidden units and 26 outputs.

3.4 Noise

Three different kinds of noise were added to the signals: Gaussian noise level 1, Gaussian noise level 2, and babble. In each condition, the noise was generated using a different random seed for each utterance (the same seeds were used for each representation) and added to the original signal. Sample values for Gaussian noise were chosen randomly according to a Gaussian probability distribution with mean 0 and standard deviation 500. For noise level 1, the gain was 1.0; for noise level 2, the gain was 2.0. The perceptual effect was to add a fairly strong background hiss.

Babble was the addition of 20 different utterances picked randomly from the training set and added together with random offsets and a gain of 0.1. The perceptual effect was several people talking in the background and one person nearby saying a single letter (the target). Table 1 shows the signal to noise ratio, with the signal being the average power in the undegraded sonorant of the letter and the noise being the average power in the added noise signal.

Table 1: Signal to noise ratio for the three conditions.

noise 1	noise 2	babble
12.5 dB	6.5 dB	14.9 dB

Table 2: Performance of the three representations.

	DFT	PLP	COCH
clean	93.6	93.8	90.9
noise 1	83.0	84.7	79.7
noise 2	74.0	76.4	70.9
babble	79.7	80.0	70.9

3.5 Results

Table 2 compares each representation when trained and tested under identical conditions (e.g. trained with babble added and tested with babble added).

Tables 3 through 5 show the performance, for each representation, under cross-testing. Nets trained under each noise condition were tested on data generated from the other noise conditions. Three patterns emerge: the best performance is usually obtained by training and testing in the same condition—the exception is that that networks trained in the noise 2 condition tested better with noise 1 data; training on noisy speech and testing on undegraded speech produces better results than training on undegraded speech and testing on noisy speech; for the best overall performance across conditions, training on babble is superior (see Table 6).

4 Summary and Conclusions

It is not possible for us to determine which representation is "best," but we conclude, for the task investigated here, PLP is the representation of choice. It has the highest overall accuracy and produces the

Table 3: Using DFT.

train/test	clean	noise 1	noise 2	babble
clean	93.6	33.0	20.0	57.3
noise 1	58.8	83.0	74.4	38.9
noise 2	46.0	78.6	74.0	32.7
babble	87.2	59.7	41.9	79.7

Table 4: Using PLP.

train/test	clean	noise 1	noise 2	babble
clean	93.8	47.4	27.8	55.4
noise 1	68.7	84.7	75.5	46.2
noise 2	53.0	81.6	76.4	36.8
babble	85.6	71.8	58.8	80.0

Table 5: Using Cochleagram.

train/test	clean	noise 1	noise 2	babble
clean	90.9	30.6	19.1	47.2
noise 1	56.9	79.7	69.7	41.2
noise 2	40.5	75.3	70.9	36.5
babble	76.6	61.0	47.3	70.9

least number of features. Its performance is about the same as the DFT when trained and tested on the same conditions, and better when trained on Gaussian noise and tested on clean speech or vice versa.

It is not easy to compare such different representations. For example, the representations have different numbers of coefficients which may require different network configurations to produce optimal results; it is not possible to try all combinations. We ran some additional experiments using the cochleagram representation without improving performance. We ran one of the cochleagram networks with twice as many hidden units. There was little change in the performance. Compressing the cochleagram channels from 85 to 42 was motivated by a desire for smaller networks and by a history of successfully compressing DFT coefficients. A single network was run without compression with little change in performance.

We cannot conclude from this study that the cochleagram is an inappropriate representation for speech recognition. It is designed to mimic the behavior of the first stage of human audition. It may prove very effective in other systems—in particular, those which attempt to more directly mimic higher-level human audition.

Acknowledgements

Research supported by APPLE Computer and ONR grant no. N-00014-91-J1482. We would like to thank Hynek Hermansky and US WEST for making the PLP algorithm available.

Table 6: Average test results across train conditions.

train	DFT	PLP	COCH
clean	51.0	56.1	47.0
noise 1	63.8	68.8	61.9
noise 2	57.8	61.9	55.8
babble	67.1	74.5	64.0

References

[1] R. Cole, Mark Fanty, Yeshwant Muthusamy, and Murali Gopalakrishnan. Speaker-independent recognition of spoken english letters. In *International Joint Conference on Neural Networks*, 1990.

[2] Ronald Cole, Yeshwant Muthusamy, and Mark Fanty. The ISOLET spoken letter database. Technical Report CS/E 90-004, Oregon Graduate Institution, 1990.

[3] H. Hermansky. Perceptual Linear Predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990.

[4] M. J. Hunt and C. Lefebvre. Speaker dependent and independent speech recognition experiments with an auditory model. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989.

[5] R. F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1982.

[6] H. M. Meng and V. W. Zue. A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons. In *International Conference on Spoken Language Processing*, 1990.

[7] Y. K. Muthusamy, R. A. Cole, and M. Slaney. Speaker-independent vowel recognition: Spectrograms versus cochleagrams. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1990.

[8] Malcolm Slaney. Lyon's cochlear model. Technical Report Apple Technical Report #3, Apple Computer Inc., 1988.