MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

A COMPARISON OF SIGNAL-PROCESSING FRONT ENDS FOR AUTOMATIC SPEECH RECOGNITION

C.R. JANKOWSKI, JR.
H-D.H. VO
R.P. LIPPMANN
Group 24

TECHNICAL REPORT 1002

18 JULY 1994

LEXINGTON MASSACHUSETTS

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Advanced Research Projects Agency under Air Force Contract F19628-90-C-0002.

The ESC Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Gary Tattungian
Administrative Contracting Officer
Contracted Support Management

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document when it is no longer needed.

ABSTRACT

The first stage of any system for automatic speech recognition (ASR) is a signal-processing "front end" that converts a sampled speech waveform into a more suitable representation for later processing. Several front ends are compared, three of which are based on knowledge about the human auditory system. The performance of an ASR system with these front ends was compared to a control mel filter bank (MFB)-based cepstral representation in clean speech and with speech degraded by noise and spectral variability. Using the TI-105 isolated word data base, it was found that auditory front ends performed comparably to MFB cepstra, sometimes slightly better in noise. With MFB cepstral recognition error rates ranging from 0.5% to 26.9%, depending on signal-to-noise ratio (SNR), auditory models could perform as high as four percentage points better. With speech degraded by linear filtering, where MFB cepstra showed error rates ranging from 0.5% to 3.1%, auditory outputs could improve performance by as much as 0.4% for conditions with high baseline error rates. This performance gain comes at a significant computational expense—approximately one-third real time for MFB cepstra as opposed to as much as over 100 times real time for auditory models. These results disagree with previous studies that suggest considerably more improvement with auditory models. However, these earlier studies used a linear predictive coding (LPC)-based control front end, which is shown to perform significantly worse than MFB cepstra under noisy conditions (e.g., 2.7% error rate with mel-cepstra vs. 25.3% with LPC at 18-dB SNR). Data-reduction techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) were also evaluated. PCA provided no gain in noise and slight gain with spectral variability. LDA on MFB energies improved performance for more difficult spectral variability conditions. LDA provided significant performance improvement (as much as 4.7% word error with LDA compared with 94.8% for mel-cepstra) with speech degraded by both noise and spectral variability when the LDA is trained on examples of corrupted speech.

ACKNOWLEDGMENTS

The authors wish to thank Charles Wayne for his support of this work on behalf of the Advanced Research Projects Agency. During this study, many of those involved with the formulation and testing of the evaluated front ends were most helpful. These include Stephanie Seneff of MIT's Spoken Language Systems Group, Oded Ghitza of AT&T Bell Laboratories, and Richard Stern and Yoshiaki Ohshima of Carnegie-Mellon University.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1. INTRODUCTION

This work is concerned with systems for automatic speech recognition (ASR). Such systems convert a continuous speech waveform produced by a microphone or telephone receiver into a linguistic message representing a speaker's meaning. In almost any "real-world" ASR system, degradation of the speech signal and addition of noise make this task significantly more difficult.

Figure 1 is an overall block diagram of such an ASR system. An initial signal-processing stage, or "front end," converts the noisy and degraded speech waveform into a representation more suitable for further processing. A pattern-classification module compares this intermediate representation to models that have been computed for relevant linguistic units such as words or phonemes (speech sounds). Finally, in more complex systems, a linguistic module further constrains the set of possible system outputs using higher level knowledge about the grammar of the task for which the ASR system was designed.

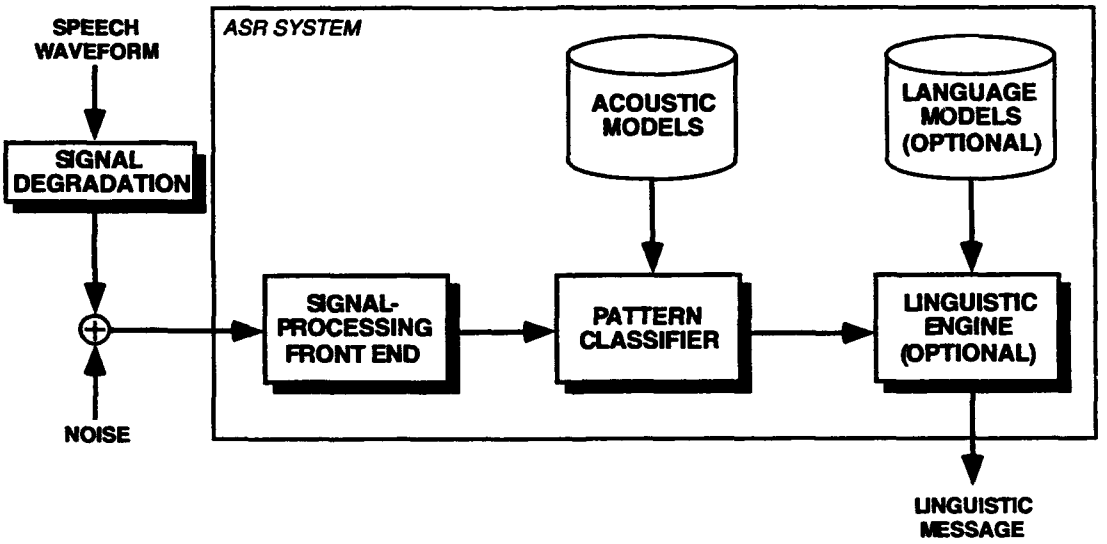

Figure 1. Block diagram of a generic automatic speech recognition (ASR) system.

The concern here is only with how the first stage, the front end, affects overall system performance. This signal-processing module is required for two reasons:

- The data rate of a sampled speech waveform (typically 8000–16,000 samples per second) is prohibitively high for pattern classifiers to process the waveform directly. A common rate for the intermediate representation is 100 samples per second. This does *not* correspond to an 80–160 times reduction in overall data rate, because the intermediate representation typically consists of several parameters per sample. A data reduction of 3–6 times is typical for systems in this study.
- There is much redundancy in the unprocessed speech waveform. For optimum pattern classification performance, earlier system components should reduce this redundancy as much as possible so that the useful information content of the intermediate representation is maximized.

Traditionally, ASR systems have employed front ends based on standard signal-processing techniques such as filter banks, linear predictive coding (LPC), or homomorphic analysis ("cepstra"). ASR systems based on dynamic programming or probabilistic techniques might use these features directly in a pattern classifier, while acoustic-phonetic-based systems use these parameters to derive more linguistically relevant features.

There has also been interest in front ends based on known properties of the human auditory system. Some of these front ends remain linear but with parameters that more closely correspond to the auditory system (e.g., filter bank bandwidths increasing with frequency) [1], although most of the auditory-based front ends are quite nonlinear because many physiological and/or perceptual processes in the auditory system are known to be quite nonlinear [6][18].

In addition, recent work has focused upon the use of traditional data-reduction techniques such as linear discriminant analysis (LDA) to automatically generate new features with maximum classification power for a given feature vector size. Conversely, these techniques can be viewed as providing the minimum feature vector size for a given recognition performance level. Such techniques have been shown to be successful in recognition tasks, especially when speech is degraded by noise or spectral tilt [7].

Front ends based on the auditory system and pattern classification have both been shown to outperform more traditional signal-processing schemes in ASR tasks [6][7][18][21]. In these evaluations, a particular front end is typically compared against one control representation for a given ASR task. The control front end, ASR task, and speech corpus all differ.

In this work several front ends are compared, including a single high-performance "control" representation, and the performance of an ASR system based on them is evaluated, performing the same recognition task. Two of these front ends are based on the human auditory system, while another is derived from pattern-classification techniques.

2. FRONT ENDS

This section describes the front ends that were compared in this study. The indicated references provide more detailed information.

2.1 MFB CEPSTRA

The "control" front end is a mel filter bank (MFB)-based cepstral transformation [1]. In this front end the speech waveform is windowed every 10 milliseconds, and a Fourier transform is computed for each windowed waveform segment. In the frequency domain, each waveform segment is then processed with a filter bank. The center frequencies of the filters are spaced on a linear scale from 100 to 1000 Hz and on a logarithmic scale above 1000 Hz. In the nonlinear region, each center frequency is 1.1 times the previous center frequency. Each filter's frequency response has a triangular shape, with the magnitude response equal to zero at the center frequencies of the adjacent filters as shown in Figure 2.

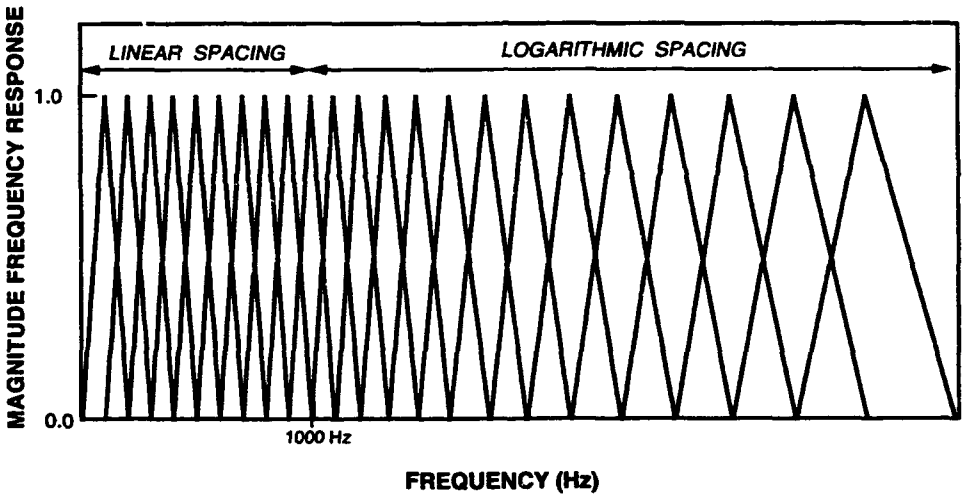

Figure 2. Linear filters for mel filter bank (MFB) front end.

A vector of log energies derived from the filter bank is then processed by an inverse cosine transform [1], creating a vector of MFB cepstral coefficients ("mel" comes from the *mel-scale*, a mapping from acoustic frequency to perceptual frequency). The cepstral coefficients are then passed to the later stages of the speech recognizer. The vectors of log filter bank energies and MFB cepstral coefficients are computed every 10 milliseconds. On a SPARCstation 2 workstation, the MFB cepstral front end operates in roughly one-third real time.

2.2 SENEFF AUDITORY MODEL

The first "test" front end for this work is an auditory-based scheme proposed by Seneff [18]. The first stage is a bank of time-domain infinite-impulse-response (IIR) linear filters. These have been carefully designed to match physiological data on the response of a cat's basilar membrane to acoustic stimuli [2].

The second stage of Seneff's front end models the transduction stage of signal processing in the inner ear, or the translation of basilar membrane motion into auditory nerve firing patterns. An efficient approximation to a half-wave rectifier, followed by an algorithm modeling short-term adaptation, a low-pass filter, and finally an automatic gain control (AGC), perform this transformation.

The third stage of Seneff's model has two branches. The "mean rate" branch simply processes each of the channel outputs from the second stage with a low-pass filter. This filtered output models the average firing rate of the auditory nerve fibers corresponding to a given channel. The second "synchrony" branch uses a "generalized synchrony detector" (GSD) to measure the extent that a second-stage channel's output is periodic with the characteristic period of that channel ($1/f$, where f is the center frequency). The higher auditory processing centers in the brain might make use of both rate and synchrony information during recognition of speech. For all evaluations with Seneff's model, there are two separate sets of results: one for the "mean rate" branch and one for the "synchrony" branch.

On a SPARCstation 2 workstation, Seneff's model operates in approximately 40 times real time. This is primarily due to the time domain nature of all processing stages in the model, which are quite computationally expensive when compared to the frequency-domain techniques used in the MFB cepstral model.

2.3 EIH AUDITORY MODEL

A second auditory front end is the Ensemble Interval Histogram (EIH) model developed by Ghitza [6]. The EIH model begins with a physiologically based time-domain linear filter bank much like the first stage of Seneff's model.

A bank of "level crossing detectors" then processes each of the filter bank channels. Each level crossing detector has an amplitude threshold and records the times when the first stage output crosses the amplitude threshold while exhibiting a positive slope, subject to the constraints that only the last 20 crossings will be recorded and no crossings occurring more than 40 milliseconds before the current time will be kept. Figure 3 schematically illustrates the operation of level crossing detectors for levels L_1 to L_4 with Δt shown for L_1.

Each first-stage channel has 12 level crossing detectors, corresponding to 12 logarithmically spaced amplitude thresholds. The detectors record the frequencies corresponding to the times between the positive crossings, using the relation $f = 1/(\Delta t)$, and accumulate this data in frequency histograms. These histograms are then combined across levels into a single histogram for each channel. An EIH is obtained by combining these histograms across all channels. An EIH can be calculated at any desired sampling rate; 100 samples per second was chosen to have one consistent sampling rate for each front end.

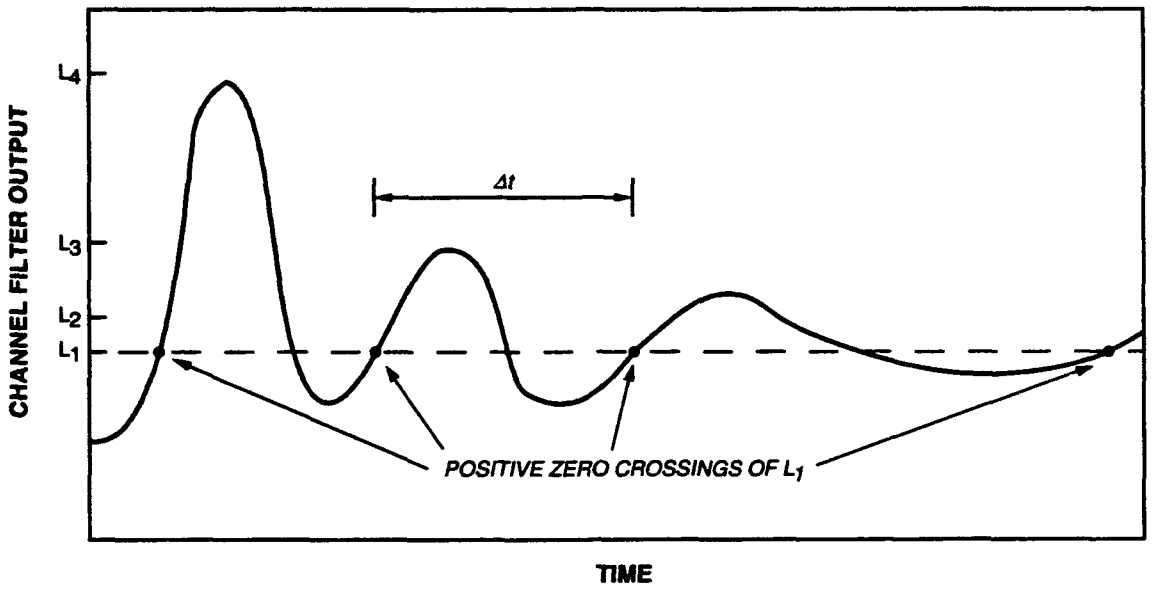

Figure 3. Level crossing detectors from the EIH model.

On a SPARCstation 2, the EIH model runs in approximately 120 times real time. Approximately 40% of the total computation time is devoted to first-stage linear filtering, and another 40% is used to upsample the first-stage filter outputs by a factor of eight before the level crossing detectors. This upsampling is necessary to obtain reasonable frequency accuracy at high frequencies.

2.4 DATA-REDUCTION TECHNIQUES

Some of our experiments use techniques to reduce the amount of data the speech recognizer receives from the front end. Two data-reduction techniques—principal components analysis and linear discriminant analysis—are described here.

2.4.1 Principal Components Analysis

Principal components analysis (PCA) is a linear transformation on an input feature space, producing a modified feature space, according to

$$\vec{x}'_i = \hat{A}\vec{x}_i \ ,$$

where $\vec{x}_i$ is the i^{th} input feature vector, $\vec{x}'_i$ is the corresponding transformed feature vector, and $\hat{A}$ is a *transformation matrix*. $\hat{A}$ is determined such that the individual elements of all $\vec{x}'_i$ are uncorrelated; i.e., the covariance matrix of the set of transformed feature vectors $\vec{x}'_i$ should have nonzero elements only on

the main diagonal. Computing $\hat{A}$ is relatively straightforward; the rows of $\hat{A}$ are the eigenvectors of the covariance matrix for $\hat{x}_i$ [22].

2.4.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a linear transformation on the input feature space, as is PCA. The calculation of the transformation matrix $\hat{A}$ is different, though; while PCA uncorrelates the input features, LDA maximizes some measure of *class separability*. Fukunaga [5] contains more details on this procedure. Because LDA uses a measure of separability between classes of input data, each input feature vector must be associated with a class before $\hat{A}$ can be calculated. LDA is thus a *supervised* procedure.

Two methods are used here for assigning each input feature vector a "class"; they are referred to as the "clustering" method and the "TIMIT" method.

- **Clustering Method:** For this technique, a hidden Markov model (HMM) speech recognizer first processes the speech used to "train" the LDA front end. During recognition, for each word the HMM system will produce a maximum likelihood state sequence, or a most probable mapping from input frame to HMM state. Starting with one cluster for each HMM state, *Leader clustering* [23] is then applied to all speech frames to reduce the number of clusters. Each cluster can then be considered a class and the LDA procedure can be calculated accordingly. The number of clusters can be specified, depending on the application.
- **TIMIT Method:** This method uses the phonetic labels of the TIMIT database [3] as classes for the LDA calculation. When using LDA on TI-105 mel filter bank energies, it is unnecessary to estimate a spectral transformation across data bases, as a constant additive vector does not change an LDA rotation matrix. For the auditory model outputs, though, such a transformation is required due to the nonlinear nature of the auditory front ends.

Hunt and Lefebvre [7] applied a linear-discriminant-based signal-processing scheme called IMELDA (Integrated MEL-scale linear Discriminant Analysis) to a mel-scale filter bank and showed significant performance improvements in noise and with spectral shape. These results were with a digit task, using the word as the class for the LDA procedure.

3. ISOLATED WORD EXPERIMENTS

This section describes a collection of experiments performed with the various front ends using an isolated-word speech recognizer.

3.1 EXPERIMENTAL CONDITIONS

3.1.1 Data Base

The speech data base used was the TI-105 isolated-word data base [17], with a vocabulary of 105 command-type words. Eight speakers (five male and three female) spoke five training tokens of clean speech and two testing tokens of clean speech for each vocabulary word.

3.1.2 Recognition System

An isolated-word HMM recognition system was used for evaluation. Each word was modeled as a sequence of eight states. While in each state, the probability density function for an observation vector was a multivariate normal distribution with a diagonal covariance matrix. This covariance matrix was shared across all HMM states in the system so that only the mean of the probability distribution distinguished one state from another. These diagonal and tied characteristics of the covariance matrix were chosen because of the relatively small amount of training data available. Paul [10] and Rabiner [16] provide much detailed information on HMMs and HMM speech recognition systems.

The recognizer is run in speaker-dependent mode; all results shown are overall word error rates averaged across all eight speakers.

3.1.3 Front End Processing

When using the MFB cepstral front end, the HMM recognizer directly used cepstral vectors, along with measures of change, as input feature vectors. The Seneff and EIH models required additional processing. An inverse cosine transform converted each auditory model "pseudo-spectral" vector to a cepstral-like representation. Instead of converting log filter bank energies to cepstral coefficients, as in the MFB cepstral front end, Seneff and EIH outputs were converted to cepstra. The use of a diagonal covariance matrix in the recognition system suggested this additional processing; the raw Seneff and EIH features are probably not uncorrelated, as a diagonal covariance matrix assumes. Pols [13] has suggested that the MFB cepstral transform performs a crude principal components analysis, which has been shown to uncorrelate the input features.

For the MFB cepstra and mean-rate front ends, the recognizer used 12 cepstral coefficients and 13 cepstral first-difference coefficients, as absolute energy was not used. For synchrony and EIH, the system used 24 cepstral coefficients and 25 first-difference coefficients. More coefficients were used for synchrony

and EIH because these two representations produce sharper spectral peaks. For a rational transfer function of the form

$$X(z) = \frac{\prod_{k=1}^{M_i}\left(1 - a_k z^{-1}\right)}{\prod_{k=1}^{N_i}\left(1 - c_k z^{-1}\right)} ,$$

i.e., zeros at a_k and poles at c_k all inside the unit circle, the complex cepstrum [15] takes the form

$$\hat{x}[n] = \sum_{k=1}^{N_i} \frac{c_k^n}{n} - \sum_{k=1}^{M_i} \frac{a_k^n}{n} \qquad n > 0 .$$

For models such as synchrony and EIH that resolve spectral peaks highly, the "poles" are closer to the unit circle; the cepstrum—simply the real part of the complex cepstrum described above—should therefore decay slower and require more coefficients to preserve similar information.

3.1.4 Noise

The first set of experiments evaluated the performance of front ends and data-reduction techniques under noisy conditions. In real-world applications, speech-recognition systems would most certainly operate under noisier conditions than are typical in the laboratory.

Instead of adding white or pink noise, speech "babble" was added from a NATO data base of recorded noise [19]. The babble was recorded by placing a microphone in a common area where many people congregate and carry on conversations. The speech babble should simulate possible real-word noise conditions more accurately than artificially generated noise.

Noise was added to the isolated word utterances at various signal-to-noise-ratios (SNRs). The SNR was measured by calculating the ratio of speech energy to noise energy, where each energy was measured by averaging the square of the signal level across the entire utterance. This is the same measurement technique Ghitza used in evaluating the EIH model [6].

3.1.5 Spectral Variability

Experiments were also conducted with speech processed by linear filters to approximate the spectral variability that a recognizer might see under real-world conditions. Three sets of conditions were evaluated.

- **Head Shadow:** As a talker rotates his/her head in the horizontal plane, one can reasonably model the effect on the listener as a linear filter [4]. Both 90° and 180° head shadow was modeled.
- **Talking Style:** As a talker varies his/her talking style, a change is seen in the long-term spectral characteristics of the speech. The overall changes in talking style cannot be modeled merely with a linear filter; these experiments only attempt to characterize typical spectral variability. The long-term spectral differences between "normal" speech and speech spoken as both "soft" and "loud" were calculated using the Lincoln-style isolated-word stressed-speech data base [8].
- **Recording Conditions:** Speech recorded at different sites and under different recording conditions exhibit varied long-term spectral characteristics. Here the difference between the TIMIT speech data base [3] and a pilot corpus for the *Wall Street Journal* data base [12] was calculated.

Figure 4 shows smoothed plots of the frequency responses used for these various conditions.

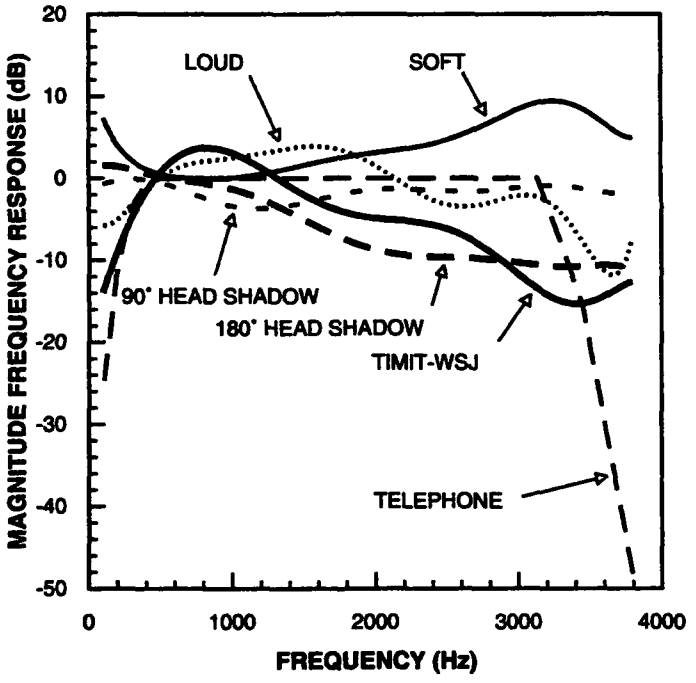

Figure 4. Smoothed frequency responses for various spectral variability conditions.

3.1.6 Multistyle Training

Some experiments used multistyle training, whereby a recognizer is trained with samples of corrupted (e.g., noisy, filtered) speech as well as "clean" speech. This technique has been successful in recognizing stressed speech [8]. For LDA, all four possible training permutations were evaluated under different conditions: training the LDA transformation matrix on clean and multistyle speech, and subsequently training the HMM recognizer on clean and multistyle speech. The results of these evaluations are reported for various experimental conditions and results described below. When using multistyle training, for each training utterance a sample of clean speech as well as a sample at each testing SNR or linear filtering condition was including in the training set.

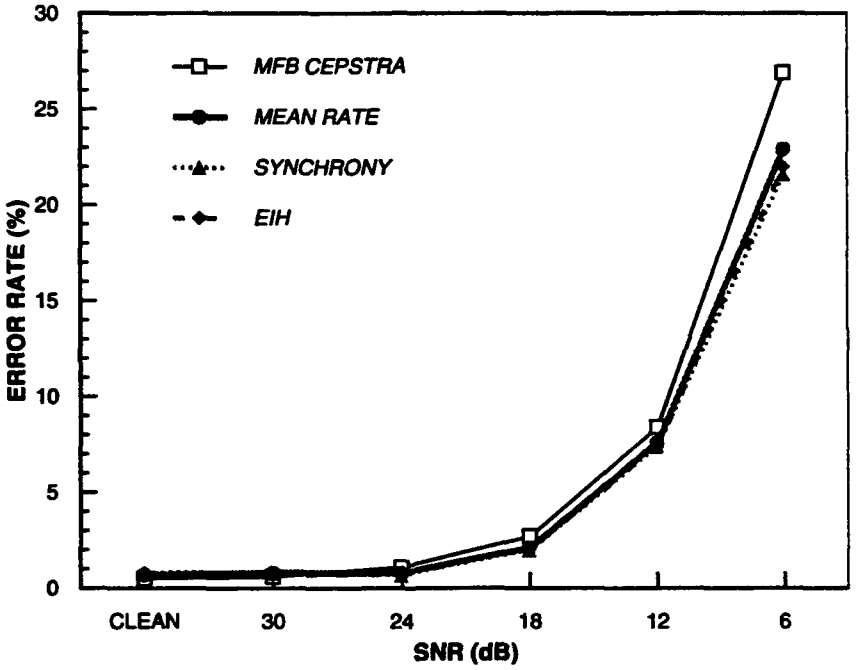

Figure 5. Error rates of MFB cepstra and auditory front ends in noise.

3.2 RESULTS IN NOISE

The HMM recognizer was trained with clean speech and tested with noisy speech for the first set of experiments in noise. Figure 5 shows the error rates of the recognizer with various front ends as a function of SNR. The following can be seen:

- Considering the disparate processing methods, all front ends perform quite similarly overall.

- For clean speech and 30-dB SNR, all auditory models perform similarly to the MFB cepstral representation.
- Below 30-dB SNR, the auditory models perform slightly better than the MFB cepstral representation. For all these cases, the difference between the auditory models and the MFB cepstral front end is from 0.6 to 4 percentage points (depending on the SNR), which exceeds a binomial standard deviation of the MFB cepstral error rate.
- The performance of all front ends degrades considerably at very high noise levels, making the usefulness of the system questionable.

A second set of experiments evaluated the effectiveness of multistyle training on the various front ends with noisy speech. Figure 6 shows the error rate averaged across all front ends, both without and with multistyle training. The individual front ends exhibited trends similar to those seen in Figure 5. From this figure it is clear that:

- For SNR of 24 dB or higher, multistyle training results in a slight performance improvement.
- For SNR of 18 dB or lower, multistyle training provides substantially greater performance.

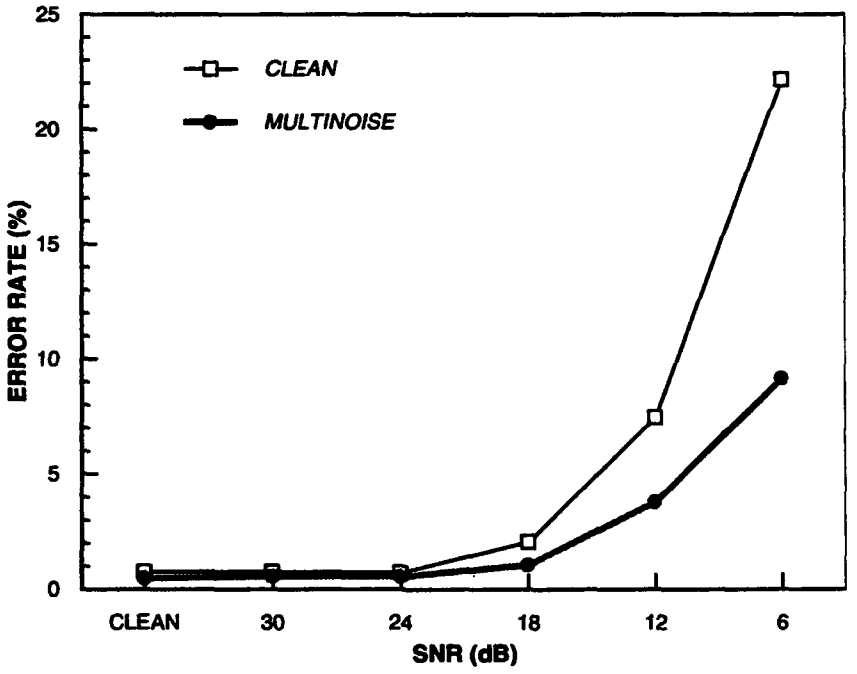

Figure 6. Error rates with clean and multistyle training in noise, averaged across front ends.

These results are consistent with Lippmann et al. [8], which indicated a performance improvement using multistyle training.

Several experiments tested the effectiveness of data-reduction techniques on front ends with noisy speech. The first experiments used principal components analysis (PCA). For PCA, multistyle speech was used to both derive the PCA transformation matrix and to train the HMM speech recognizer. PCA provided minimal improvement in noise compared to standard multistyle training.

In another set of data reduction experiments, linear discriminant analysis (LDA) was applied to the mel filter bank outputs, using the "clustering" technique described previously to separate the training data into classes. Numbers of clusters and the training procedure were both evaluated and it was determined that:

- The recognizer achieved best performance when no clustering was performed on the input states—the raw states from the HMM recognition were used as classes.
- The recognizer performed best when both the LDA transformation matrix was derived and the speech recognizer trained with multistyle speech data.

From Figure 7 it is clear that LDA performed worse than MFB cepstra for all SNRs. Experiments using LDA with the "TIMIT" technique for generating class information showed similar results across all front ends.

3.3 RESULTS WITH SPECTRAL VARIABILITY

Figure 8 shows the recognition results for the various front ends trained in clean speech and tested with the various spectral variability conditions. The following points are worth noting:

- The difference between the performance of the front ends was extremely small (never more than 1% difference between the error rates for best and worst front ends).
- For the conditions with low baseline error rates (e.g., "clean" and "90° head shadow"), the MFB cepstra outperforms the auditory models by a small amount.
- For other conditions, the synchrony and EIH outputs slightly outperform MFB cepstra front end, which has slightly better results than the mean-rate output.

The results of using multistyle speech for training is shown in Figure 9, where results are averaged across all front ends. As with multistyle training with noisy speech, multistyle training resulted in increased performance for every linear filtering condition.

Principal components analysis was then applied to multistyle training data; the recognition results are reported in Figure 10. Unlike the PCA results in noise, the spectral variability results show a small but consistent improvement using PCA.

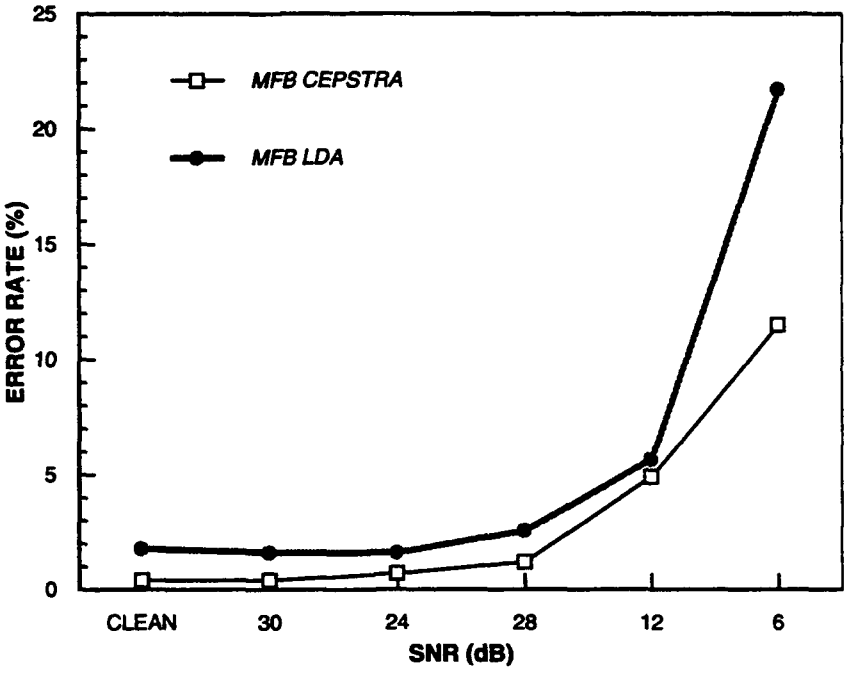

Figure 7. Error rates for MFB cepstra and LDA on mel filter bank energies in noise.

Using the clustering technique for class generation, LDA was then used with the filtered speech. Before full recognition experiments were run, it was discovered that:

- The recognizer performed best when LDA used 326 clusters for class. This is in contrast to the results in noise, which showed best results with no clustering.
- The recognizer also performed best when the LDA transformation matrix was derived from multistyle speech data but the speech recognizer was trained on clean data.

Figure 11 shows a comparison of MFB cepstra and LDA on mel filter bank energies, as described previously. For clean speech and the head shadow conditions MFB cepstra is better; LDA performs better for all other conditions. This is similar to results obtained across front ends with linear filtering; for the more difficult conditions—as indicated by error rate with clean speech—alternative processing led to improved performance.

LDA was also performed on various front ends, using the TIMIT technique for generating class. The results proved disappointing; MFB cepstra outperformed LDA for all linear filtering conditions. This is believed to be due to a failure in the TIMIT technique of generating LDA class. To justify using one database (TIMIT) to generate class data for a transformation of another database (TI-105), it had to be assumed that for mel-scale filter bank log energies, the major difference between the feature vectors of the data

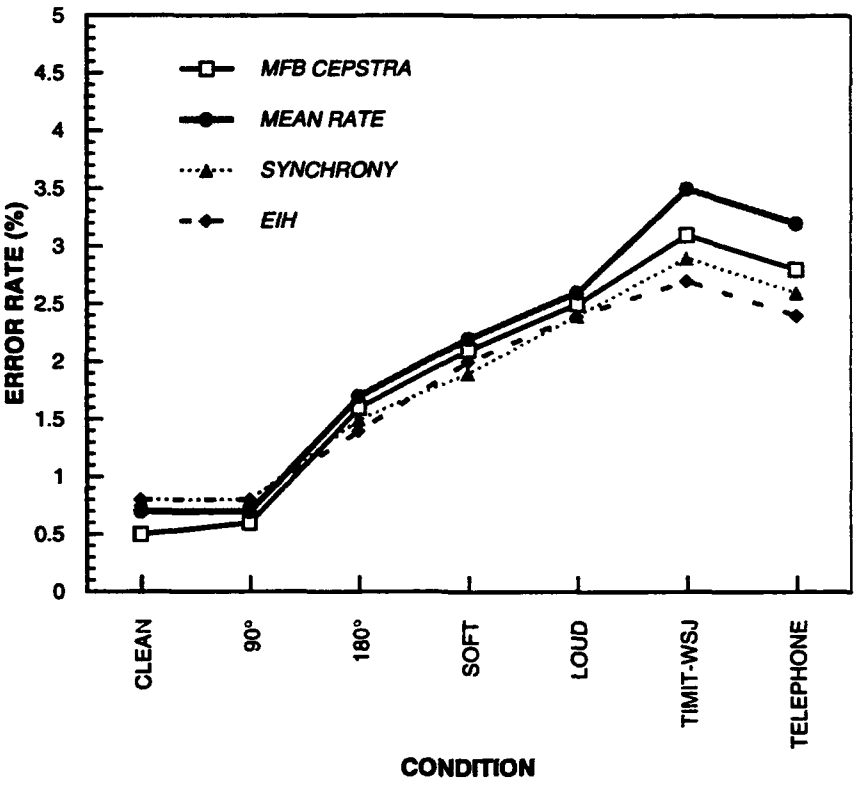

Figure 8. Error rates for MFB cepstra and auditory front ends with spectral variability.

bases can be approximated as a constant difference, or roughly a linear filtering operation between the speech of the two data bases. In this case, the LDA transformation matrix remains the same, so no explicit cross-data base transformation is necessary. On the other hand, the nonlinear nature of the auditory models demands a normalization between data bases. From the poor results using this method of calculating LDA, it appears that linear filtering alone cannot explain the major acoustic difference between the data bases, and that the normalization used with the auditory models was not sufficient.

3.4 RESULTS WITH NOISE AND SPECTRAL VARIABILITY

Experiments were also conducted using LDA on mel filter bank energies in the presence of both noise and spectral variability. The methods for training the LDA and the recognizer that were used in the above experiments were combined; i.e., the LDA parameters were calculated using both multinoise and multicondition speech, while the speech recognizer was trained only on multinoise speech. No clustering was done on the HMM states, so there was a total of 1050 classes. Time dictated testing using only a few noise and spectral variability conditions; Figure 12 shows the results; a dramatic improvement in error rate

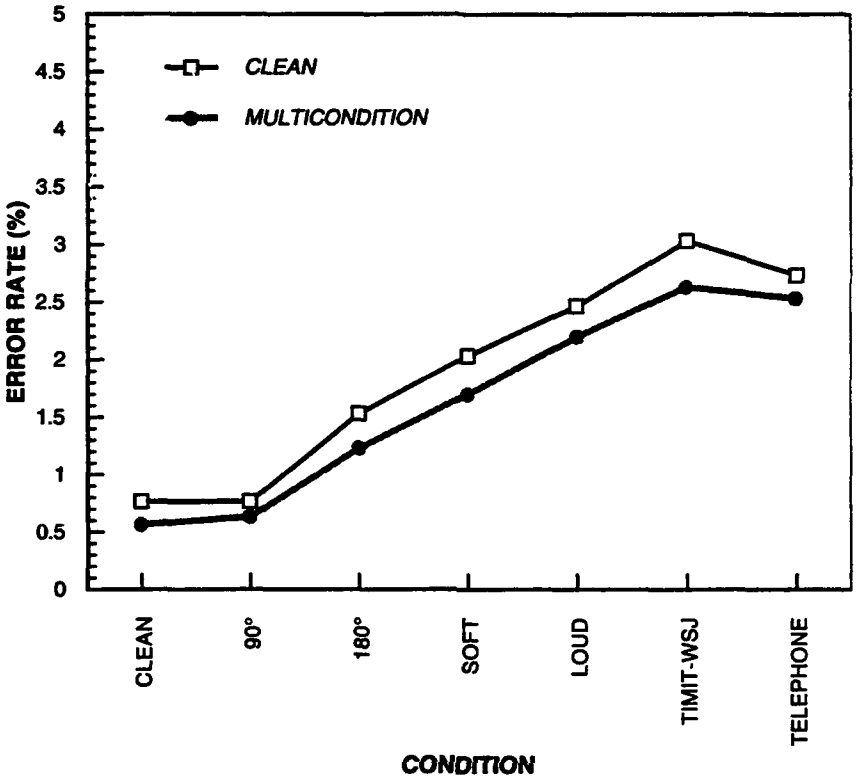

Figure 9. Error rates for clean and multistyle speech with spectral variability, averaged across all front ends.

for conditions with linear filtering. With no linear filtering, MFB cepstra outperforms LDA, but with all other spectral variability conditions, LDA results in significantly better recognition performance for all noise levels. These results are consistent with Hunt and Lefebvre's work [7] suggesting that LDA offers significant performance improvement with corrupted speech when LDA parameters can be trained using corrupted speech.

3.5 VALIDATING THE MFB CEPSTRAL FRONT END

Before comparing the Seneff and EIH auditory models to the MFB cepstral representation, several errors in the existing implementation of the EIH algorithm were corrected. During this process, parameters of the EIH algorithm were set to provide "reasonable" performance on samples from the test corpus. It was also decided to verify that the existing parameters of the MFB cepstral front end were reasonable choices for the test corpus and recognition system.

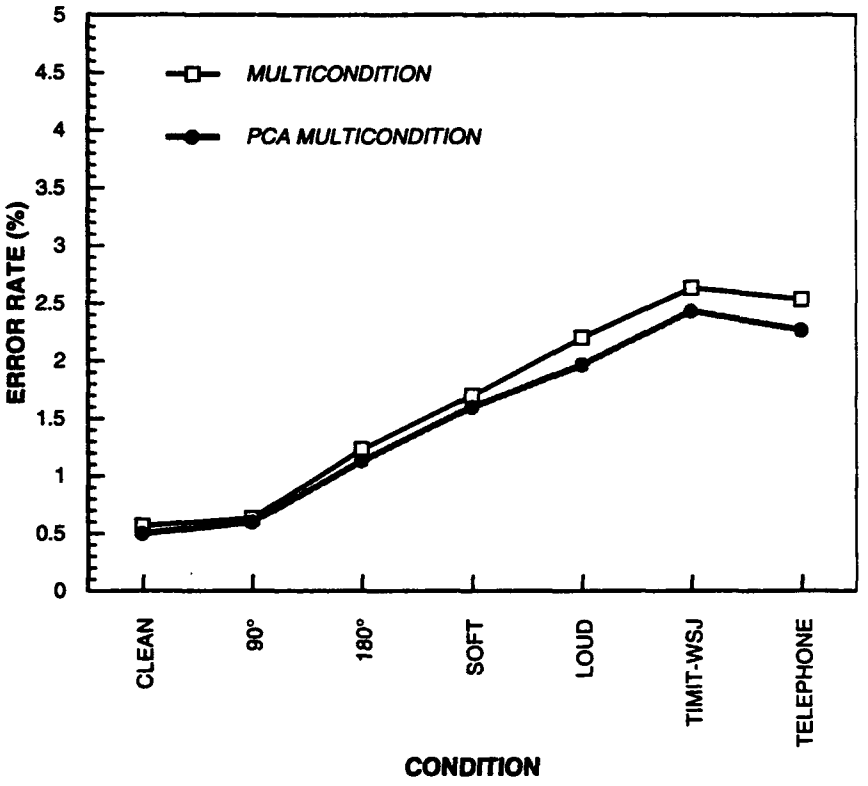

Figure 10. Error rates without and with PCA with spectral variability, averaged across all front ends.

The following parameters of the MFB cepstra front end were varied:

- **Number of filters:** The number of filters in the linear filter bank were modified; 13, 16, 24 (default), 31, and 47 filters were used. This was varied by changing the filter spacing in both the linear and nonlinear frequency regions. The edges of the filter responses remained at the center frequencies of the adjacent filters so that filter bandwidth decreased as number of filters increased.
- **Number of cepstra:** The number of cepstral coefficients was varied; 10, 13 (default), and 16 cepstra were used before calculating first differences and dropping the first cepstral coefficient corresponding to raw energy. This resulted in overall feature vector sizes of 19, 25, and 31.

The HMM recognizer, using an MFB cepstral front end with these various parameters, was tested across all noise conditions and spectral variability conditions. Not surprisingly, the MFB cepstral front end responded quite differently to different noise levels and spectral variability conditions, but overall, the

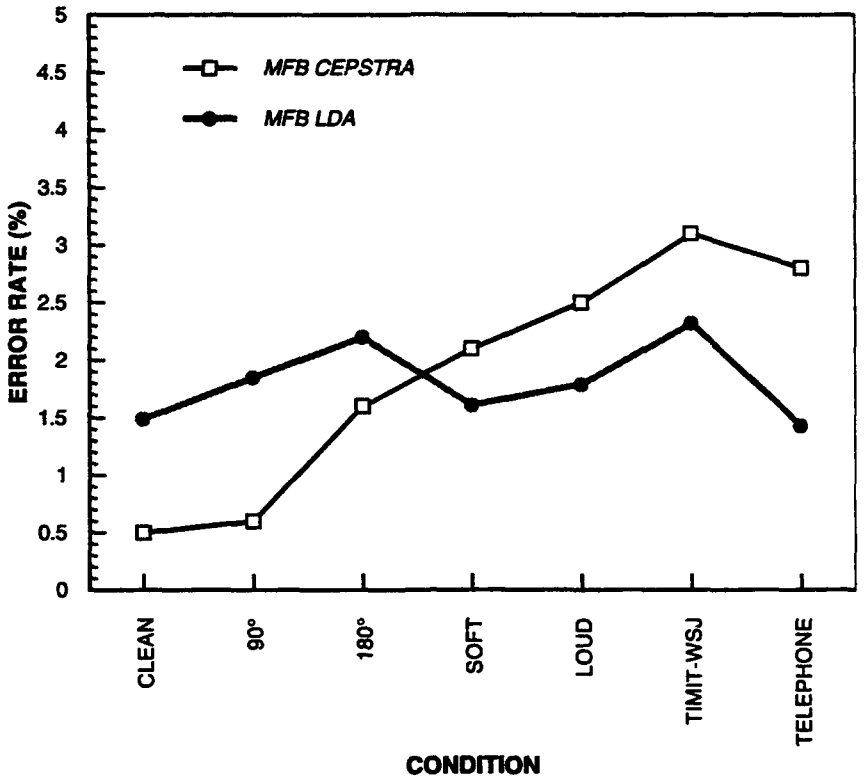

Figure 11. Error rates for MFB cepstra and LDA on mel filter bank energies with spectral variability.

default choice of parameters (24 filters and 13 cepstral coefficients) provided good, and usually the best, performance across conditions.

3.6 COMPARISON OF RESULTS WITH OTHER SITES

Other sites have obtained different results than those shown here. In particular:

- Ghitza at AT&T Bell Laboratories [6] showed significant performance improvement with an isolated word recognition task using the EIH model.
- Carnegie-Mellon University (CMU) [21] reported similar significant improvement using both the mean-rate and synchrony branches of Seneff's auditory model with a continuous speech data base and recognizer.

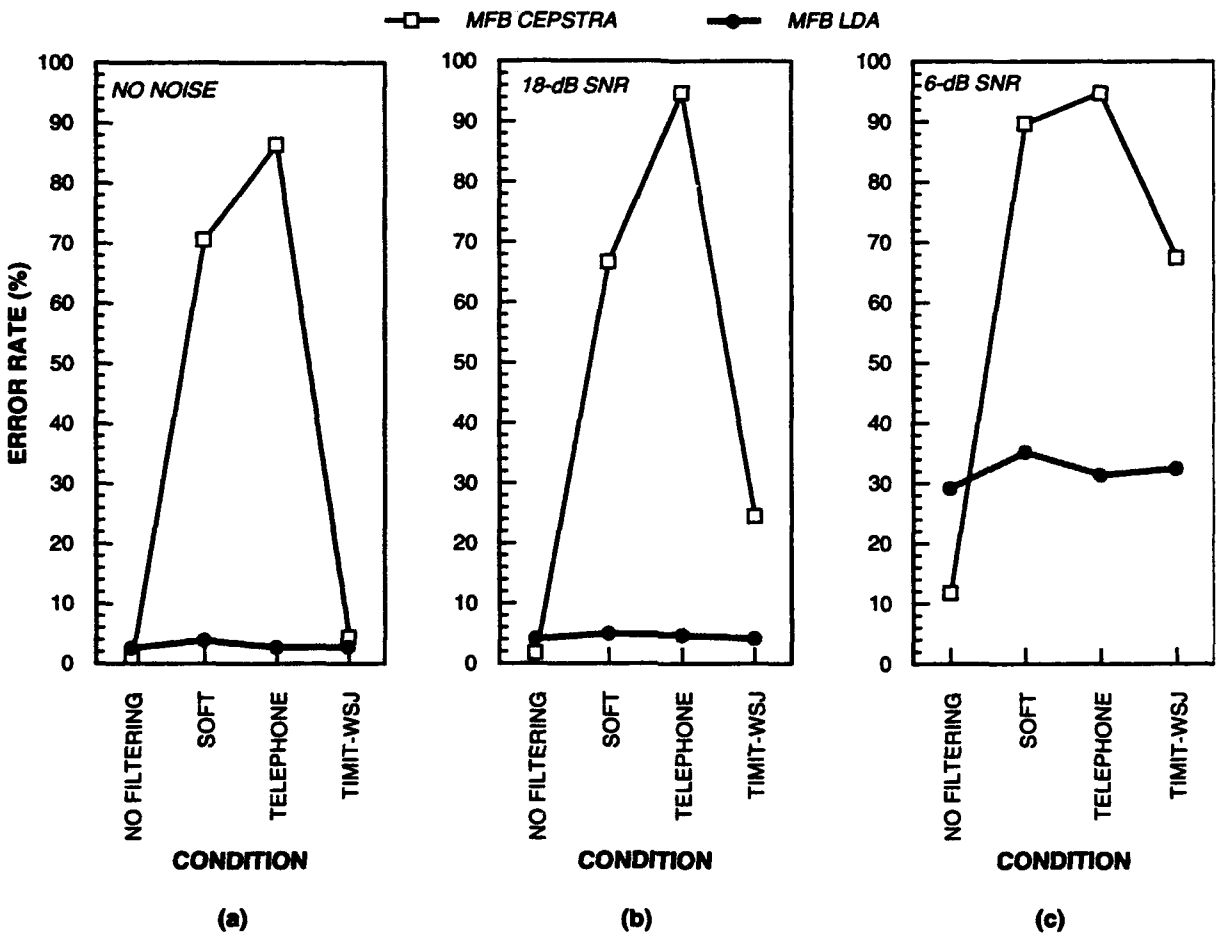

Figure 12. Error rate of MFB cepstra and LDA on mel filter bank energies with noise and spectral variability. No noise in panel (a), 18-dB SNR in panel (b), and 6-dB SNR in panel (c).

Lincoln's results suggest very small performance improvement using the auditory models, not the significant differences found elsewhere. There could be several reasons for the differences. Referring to Ghitza's and CMU's results:

- Both studies used a linear predictive coding (LPC)-based cepstral front end as the control representation; the CMU front end also uses a bilinear transform to approximate a "mel-like" frequency warping [20]. LPC-based front ends do not, however, typically perform well in noisy environments. This study uses a filter bank-based cepstral front end instead.

- Baseline error rates for both studies are significantly higher than here. It has been shown that as the baseline error rate increases, the differences between the traditional front end and auditory models increase. Also, it has been noted that at such high baseline error rates, a recognition system's usefulness is questionable.
- Ghitza and CMU added artificial noise, while Lincoln used recorded speech babble. It is unsure what the effect of this difference might be.
- Lincoln uses a continuous observation HMM system as the speech recognizer. Ghitza used a system based on dynamic time warping (DTW), while CMU used a discrete observation HMM system. It is unsure how the different recognition systems might affect performance.

To evaluate the difference in control front ends (LPC-based cepstra as opposed to FFT filter bank-based cepstra), software was acquired for the CMU front end to compare Lincoln's MFB cepstral front end to the CMU LPC-based scheme. The same data base and recognizer as used for all other isolated-word experiments were used here and tested with both speech babble and artificially generated white noise to study the effect of the noise type. Three different LPC orders (equal to the number of predictor coefficients and the number of vocal tract poles) were used: 8, 14, and 18.

Figure 13 shows the results of the evaluation. The performance of the MFB cepstral front end in both white noise and speech babble is compared to the *best*-performing LPC-based representation. Clearly, the MFB cepstral front end significantly outperforms LPC, especially in noise. This is not terribly surprising, because finding the poles of a vocal tract system function using LPC with noisy speech can be interpreted as finding the roots of a moderate order polynomial with noisy coefficients. These roots are known to have high variability. Curiously, performance was better with the speech babble than with the white noise.

It is believed that these results show the major reason that other studies have found much more performance improvement with auditory models than shown here. Given the same auditory model performance, a study with a poorer performing control front end (such as LPC) will show greater *relative* benefit of auditory models over traditional techniques. It is easy to see how an auditory model might perform better than LPC, as other studies have shown, but perform very similarly to or only slightly better than a better-performing control such as MFB cepstra, which is shown here.

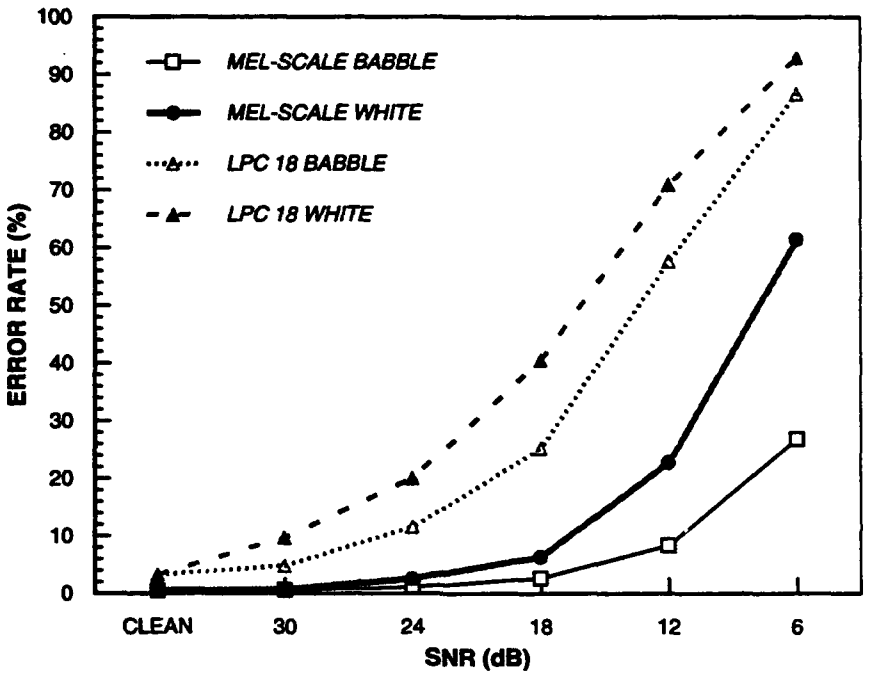

Figure 13. Error rates for MFB cepstra in two types of noise and the best performing LPC representation.

4. CONCLUSION

Experiments were conducted to test the effectiveness of several signal-processing schemes as front ends to automatic speech recognition systems. Front ends were evaluated in clean speech, speech degraded by "babble" type noise, and speech processed with linear filters simulating "real world" acoustic transformations.

For an isolated-word recognition task:

- With additive speech "babble" as noise, auditory models perform very similarly to MFB cepstra but can slightly reduce recognition error rate. With MFB cepstral error rates ranging from 0.5% to 26.9% depending on SNR, auditory models could perform as high as four percentage points better. Large changes are seen at quite high baseline error rates, but here the usability of the recognizer is questionable.
- With speech degraded by linear filtering, where MFB cepstra showed error rates ranging from 0.5% to 3.1%, the EIH and Seneff synchrony auditory outputs could slightly improve performance by as much as 0.4% for conditions with high baseline error rates.
- Multistyle training was quite effective with both noise and spectral variability for all front ends.
- Principal components analysis (PCA) did not improve performance in noise and offered only slight improvement (as much as 0.4 percentage points, with baseline error rates ranging from 0.6% to 2.6%) with spectral variability.
- Linear discriminant analysis (LDA) on mel-scale filter bank outputs, using clustered HMM states as class, performed worse than MFB cepstra in noise but better than MFB cepstra for the more difficult spectral variability conditions.
- LDA provides dramatic improvement in error rate with speech degraded by a combination of noise and spectral variability when LDA is trained on a combination of noisy and filtered speech. With MFB cepstral error rates as high as 95%, LDA could reduce error by as much as 90 percentage points. This is consistent with results found elsewhere [7].
- The current parameters of the MFB cepstral front end—number of filters and number of cepstra—are a reasonable choice for good performance across noise and spectral variability conditions.
- Cepstral coefficients derived from a linear predictive coding (LPC) model perform significantly worse than MFB cepstra when a recognizer is trained in clean speech and tested in noise. This is theoretically expected, so the same result would be expected for continuous speech systems. This result is likely the primary reason that other studies show much more performance improvement with auditory models than is shown here; two other studies use an LPC-based front end as the control representation.

Preliminary experiments were also performed with some of the front ends using a continuous-speech task as opposed to an isolated-word task. A continuous-speech HMM recognition system [11] was used with the ARPA Resource Management (RM) corpus [14]. It was found that:

- PCA could improve recognition performance relative to MFB cepstra for high SNRs (above 18 dB). For lower SNRs, MFB cepstra outperformed PCA.
- Attempts to obtain reasonable continuous-speech performance with Seneff's mean rate and synchrony outputs were not successful. Error rates were significantly higher than with MFB cepstra. Filtering Seneff's outputs and changing important time constants within Seneff's model provided better "looking" psuedo-spectrograms, but recognition performance was still unacceptable. It is unknown why the isolated-word results with Seneff's model in the continuous speech domain were not able to be duplicated.

More work clearly needs to be done in this area; current ASR systems focus primarily on continuous speech tasks.

In summary, front ends based on the human auditory system perform comparably to, and can slightly improve the performance of, an MFB cepstral-based ASR system for isolated words with noise and some spectral variability conditions. The magnitude of the reduction in error rate is small relative to the required increase in computation time. Data reduction techniques such as principal components analysis (PCA) and linear discriminant analysis (LDA) improve ASR performance for some spectral variability conditions, but LDA was especially useful in the case of combined noise and spectral variability. Auditory models are thus most appropriate in situations where computational cost is not an issue. For selected conditions, PCA or LDA alone can provide performance improvement at significantly less cost. As computer hardware continues to improve in performance at an appreciable rate, these issues may become less important.

The importance of having a suitable control representation when testing auditory front ends has also been shown. When training with clean speech and testing with noisy speech, LPC cepstra are not suitable; they perform much worse than MFB cepstra. Much of the difference that others have found between the performance of auditory models and more "standard" representations can be explained by the lack of a high-performing control front end.

One could interpret these results by arguing that speech-processing based on the human auditory system cannot provide appreciable speech recognition performance improvement. One might also suggest that traditional linear techniques are sufficient to code the relevant information necessary for high-performance speech recognition, and thus any further improvement in overall system performance must come from the later stages of the system. Lincoln does not believe that this is necessarily true. The human auditory system is sufficiently complex that thousands of person-years of intensive research in the areas of auditory physiology and perception have left even some fairly basic questions of auditory function unanswered. The models used here are gross simplifications, with components chosen not only to model phenomena that are *currently* believed to be relevant for speech recognition but also to minimize computational complexity. The methods used for converting auditory front end outputs into feature vectors were largely dictated by the structure of the ASR system. For a "fair" test of the effectiveness of auditory

models, researchers must do more work to discover better ways of incorporating features from auditory models into speech recognizers. In the long run, as auditory function and speech perception are better understood, more parameters that are important for speech recognition should be uncovered that have not yet been considered.

REFERENCES

1. S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing* 28, 357–366 (1980).

2. B. Delgutte and N.Y.S. Kiang, "Speech coding in the auditory nerve," *Journal of the Acoustical Society of America* 75, 866–919 (1984).

3. W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA speech recognition database: specifications and status," *Proceedings of DARPA Workshop on Speech and Natural Language*, DARPA ISTO, Palo Alto, California (1986), 93–99.

4. J. Flanagan, "Analog measurements of sound radiation from the mouth," *Journal of the Acoustical Society of America* 32, 1613–1620 (1960).

5. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, New York: Academic Press (1972).

6. O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and Language* 1, 109–130 (1986).

7. M.J. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Glasgow, Scotland (1989), 262–265.

8. R.P. Lippmann, E.A. Martin, and D.B. Paul, "Multistyle training for robust isolated-word recognition," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Dallas, Texas (1987), 705–708.

9. D.B. Paul, "The Lincoln robust continuous speech recognizer," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Glasgow, Scotland (1989), 449–451.

10. D.B. Paul, "Speech recognition using hidden Markov models," *The Lincoln Laboratory Journal* 3, 41–62 (1990).

11. D.B. Paul, "The Lincoln tied-mixture HMM continuous speech recognizer," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Toronto, Ontario (1991), 329–332.

12. D.B. Paul and J.M. Baker, "The design for the *Wall Street Journal*-based CSR corpus," *Proceedings of DARPA Speech and Natural Language Workshop*, DARPA ISTO, Harriman, New York (1992), 357–362.

13. L.C.W. Pols, "Spectral analysis and identification of Dutch vowel in monosyllabic words," Doctoral Dissertation, Free University, Amsterdam, The Netherlands, 1966.

REFERENCES (Continued)

14. P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, "The DARPA 1000-word resource management data base for continuous speech recognition," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* IEEE, New York, New York (1988), 651–654.

15. L. Rabiner and R. Schafer, *Digital Processing of Speech Signals,* Englewood Cliffs, New Jersey: Prentice-Hall (1979).

16. L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE* 77, 257–286 (1989).

17. P.J. Rajesekaran, G.R. Doddington, and J.W. Picone, "Recognition of speech under stress and in noise," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* IEEE, Tokyo, Japan (1986), 733–736.

18. S. Seneff, "A computational model for the peripheral auditory system: application to speech recognition research," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Tokyo, Japan (1986), 1983–1986.

19. H.J.M. Steeneken and F.W.M. Geurtsen, "Description of the RSG-10 noise database," Soesterberg, The Netherlands: Institute for Perception-TNO (1990).

20. K. Shikano, "Evaluation of LPC spectral matching measures for phonetic unit recognition," Technical Report, Computer Science Department, Carnegie Mellon University (1986).

21. R. Stern, F. Lue, Y. Ohshima, T. Sullivan, and A. Acero, "Multiple approaches to robust speech recognition," *Proceedings of DARPA Speech and Natural Language Workshop*, DARPA ISTO, Harriman, New York (1992), 274–279.

22. G. Strang, *Linear Algebra and Its Applications,* San Diego: Harcourt Brace Jovanovich (1988).

23. H.D. Vo, "A comparative study of two front-end auditory models for speech recognition," S.M. thesis, Massachusetts Institute of Technology, August 1992.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	18 July 1994	Technical Report

4. TITLE AND SUBTITLE

A Comparison of Signal-Processing Front Ends for Automatic Speech Recognition

5. FUNDING NUMBERS

C — F19628-90-C-0002
PR — 337
PE — 62301E

6. AUTHOR(S)

Charles R. Jankowski, H-D.H. Vo, and Richard P. Lippmann

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
P.O. Box 73
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-1002

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

ARPA/SISTO
3701 No. Fairfax Dr.
Arlington, VA 22203-1714

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ESC-TR-94-081

11. SUPPLEMENTARY NOTES

None

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

The first stage of any system for automatic speech recognition (ASR) is a signal processing "front end" that converts a sampled speech waveform into a more suitable representation for later processing. Several front ends are compared, three of which are based on knowledge about the human auditory system. The performance of an ASR system with these front ends was compared to a control mel filter bank (MFB)-based cepstral representation in clean speech and with speech degraded by noise and spectral variability. Using the TI-105 isolated word data base, it was found that auditory front ends performed comparably to MFB cepstra, sometimes slightly better in noise. With MFB cepstral recognition error rates ranging from 0.5 to 26.9%, depending on signal-to-noise ratio (SNR), auditory models could perform as high as four percentage points better. With speech degraded by linear filtering, where MFB cepstra showed error rates ranging from 0.5 to 3.1%, auditory outputs could improve performance by as much as 0.4% for con ditions with high baseline error rates. This performance gain comes at a significant computational expense–approximately one-third real time for MFB cepstra as opposed to as much as over 100 times real time for auditory models. These results disagree with previous studies that suggest considerably more improvement with auditory models. However, these earlier studies used a linear predictive coding (LPC)-based control front end, which is shown to perform significantly worse than MFB cepstra under noise conditions (e.g., 2.7% error rate with mel-cepstra vs. 25.3% with LPC at 18-dB SNR). Data-reduction techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) were also evaluated. PCA provided no gain in noise and slight gain with spectral variability. LDA on MFB energies improved performance for more difficult spectral variability conditions. LDA provided significant performance improvement (as much as 4.7% word error with LDA compared with 94.8% for mel-cepstra) with speech degraded by both noise and spectral variability when the LDA is trained on examples of corrupted speech.

14. SUBJECT TERMS			15. NUMBER OF PAGES
auditory model	automatic speech recognition	ensemble interval histogram (EIH)	38
cepstra	linear discriminant analysis (LDA)	linear predictive coding (LPC)	
front end	mel filter bank (MFB)	principal components analysis (PCA)	16. PRICE CODE
synchrony	spectral variability		

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Same as Report