A COMPARATIVE STUDY OF FEATURE REPRESENTATIONS FOR ROBUST SPEECH RECOGNITION IN ADVERSE ENVIRONMENTS

K.K. Paliwal[1] and B.S. Atal

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

ABSTRACT – In this paper, a number of feature representations are studied as to their recognition performance in presence of additive noise and channel mismatch distortions. It is shown that 1) the linear prediction analysis technique provides more robust cepstral features than the homomorphic analysis technique, 2) the filter-bank power spectrum is capable of generating more robust cepstral features than the power spectrum derived through the fast Fourier transform algorithm, and 3) use of human auditory properties in an acoustic front-end makes it more robust.

1. INTRODUCTION

Objective of speech recognition is to take the speech waveform of an unknown (input) utterance, and classify it as one of a set of spoken words, phrases, or sentences. Typically, this is done in two steps. In the first step, an acoustic front-end is used to perform feature analysis where the speech signal is analysed to extract a set of features or characteristics sequentially in time. Second step deals with pattern classification where the sequence of feature vectors is compared against the machine's knowledge of speech (in the form of acoustics, lexicon, syntax, semantics, etc.) to arrive at a transcription of the input utterance.

Selection of proper acoustic features is perhaps the most important task in the design of a speech recognition system. It directly affects the performance of a speech recognizer. These features should be selected in such a manner that they should contain maximum information necessary for speech recognition and, at the same time, discard irrelevant information such as speaker characteristics, manner of speaking, background noise, channel distortion, etc. Feature selection is a difficult task and a great deal of research has been done to identify these features (see [1] and references given therein for different front-ends).

Once these features are selected, the task of the acoustic front-end is to extract these features from the speech signal. For this, it divides the speech signal into overlapping time frames and computes the values of these features for each frame. The complexity of the acoustic front-end depends on the type of features selected. These features may be as simple as the energy and zero-crossing rate of the waveform during each frame. A better, but more complex, method for feature analysis is based on the source/system model of the speech production system. It is generally considered that the system part of this model represents the vocal tract response and it contains most of the linguistic information necessary for speech recognition. The power spectrum of each speech frame contains information about the source part (in the form of fine structure) and vocal tract system part (in the form smooth spectral envelope). The task of the acoustic front-end is to compute the smooth spectral envelope from the power spectrum by removing the fine structure. Once the smooth spectral envelope is estimated, it can be represented in terms of a few parameters (such as cepstral coefficients). These parameters are used as acoustic features in a speech recognition system

Traditionally, the power spectrum of a speech frame is computed either by fast Fourier transform (FFT) algorithm or through filter-bank analysis technique. The smooth spectral envelope is computed from this power spectrum by using one of the following two signal processing techniques: 1) linear prediction (LP) analysis and 2) homomorphic analysis. In the LP analysis technique, the smooth spectral envelope is modeled by an all-pole filter and parameters of this filter are estimated through a least-squares procedure. In the homomorphic analysis technique, a logarithmic function is used on the power spectrum which makes the source and system components additive in the log power spectrum. This allows a simple linear filter to remove the source component (fine structure) from the log power spectrum. This is done by computing an inverse Fourier transform of the log power spectrum where a first few terms (called cepstral coefficients) represent the smooth spectral envelope.

Most of the speech recognizers reported in the literature use cepstral features which are derived from the FFT power spectrum by using the LP analysis technique. These cepstral features are known to be very sensitive to additive noise and channel mismatch distortions which are very common in practice. As a result, the performance of these recognition systems deteriorates drastically in the presence of these distortions. Human listeners, on the contrary, can recognize speech even in the presence of large amount of noise and channel distortions. Therefore, it is argued that the acoustic front-end can be made more robust to these distortions by utilizing the properties of human auditory system. We call these front-ends as auditory front-ends.

A number of auditory front-ends have been proposed in the literature. These front-ends employ some property of human auditory system to modify the power spectrum and then use either the LP analysis technique or the homomorphic analysis technique to get the smooth spectral envelope which, in turn, is represented in terms of a few cepstral features. Some examples of popular auditory front-ends are Mel filter-bank analysis [2], perceptual linear prediction analysis [3], ensemble interval histogram (EIH) analysis [4], etc. The Mel filter-bank analysis procedure [2] is based on the fact that the frequency sensitivity of the human ear is higher at low frequencies than at higher frequencies. Therefore, this method computes the power spectrum of a given speech frame by using a nonuniform filter bank where filter bandwidth increases logarithmically with filter frequency (according to Mel scale). The cepstral features representing the smooth spectral envelope are computed from the power spectrum using the homomorphic analysis technique. The PLP analysis technique [3] uses more detailed properties of the human auditory system than the Mel filter-bank analysis technique to compute the power spec-

[1]Present address: School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111, Australia

trum. In addition to nonuniform filter-bank (where filters are spaced according to Bark scale), it uses equal loudness curve and the intensity-loudness power law to model the auditory system better. The cepstral features are estimated from the resulting power spectrum by using the LP analysis technique. The EIH analysis technique [4] uses a measure of the spatial (tonotopic) extent of coherent neural activity across the stimulated auditory nerve to compute the power spectrum. The cepstral features are computed from this power spectrum using the LP analysis technique.

In this paper, our aim is to make a comparative evaluation of some of the acoustic and auditory front-ends. We use here a hidden Markov model (HMM) based isolated-word speech recognition system as test-bed. We provide here speech recognition results for both speaker-dependent and independent modes. Speech recognition performance is studied here for clean as well as distorted speech. Two types of distortions are investigated here: additive white noise and channel mismatch. The following four front-ends are studied here:

1. *FFTL front-end:* In this front-end, power spectrum of a speech frame is computed through FFT algorithm and cepstral features are computed by using the LP analysis technique.
2. *FFTH front-end:* In this front end, power spectrum of a speech frame is computed through FFT algorithm and cepstral features are computed by using the homomorphic analysis technique.
3. *UNIH front-end:* In this front-end, power spectrum of a speech frame is computed by using the filter-bank analysis technique. Here, the filters are equally spaced to cover the frequency range of interest (which is from zero to half of the sampling frequency). Though the filter-bank can can be designed by implementing the filters explicitly, we use here the FFT algorithm to derive the power spectrum and compute the filter output by averaging within the filter band. Cepstral features are computed from the filter-bank power spectrum by using the homomorphic analysis technique.
4. *MELH front-end:* This front-end is similar to the UNIH front-end, except that the filter spacing is non-uniform and according to Mel scale [2]. Cepstral features are computed from the Mel filter-bank power spectrum using the the homomorphic analysis technique.

This paper is organized as follows. Section 2 describes the data base used here for evaluating the recognition performance of these front-ends. Speech recognition experiments and results are described in Section 3. Section 4 provides conclusions.

2. SPEECH DATA BASE

We use two different speech data bases for conducting speaker-dependent and speaker-independent isolated-word recognition experiments. In speaker-dependent experiments, a vocabulary of 39 English alpha-digits (26 alphabets (A-Z) + 10 digits (0-9) + 3 command words 'stop', 'error' and 'repeat') is used. The data base consists of speech from 4 talkers (2 males and 2 females). 5 utterances of each word from each of these 4 talkers were used for training and an additional 5 utterances for testing. The training and testing utterances were recorded over the local dialed-up telephone lines, and digitized at a sampling rate of 6.67 kHz. The speech signal was analysed every 15 ms with a frame width of 45 ms (with Hamming window and preemphasis), and each frame was represented as a feature vector having 8 cepstral coefficients as its components. LP analysis was done through the autocorrelation method with predictor order of 8. Endpoints of each utterance were manually determined.

For speaker-independent experiments, the ISOLET spoken letter database from Oregon Graduate Institute was used [5]. Here, the vocabulary consists of 26 English alphabets (A-Z). From this data base, we used 90 utterances for each word from 90 talkers (45 male and 45 female) for training and 30 utterances for each word from 30 talkers (15 male and 15 female, different from training talkers) for testing. In the original data base, these utterances were digitized at 16 kHz sampling rate. We down-sampled the signal to 8 kHz using a lowpass filter with cutoff frequency of 3.5 kHz. The speech signal was analysed every 15 ms with a frame width of 45 ms (with Hamming window and preemphasis), and each frame was represented in terms of 10 cepstral features. LP analysis was done through the autocorrelation method with predictor order of 10.

For studying the speech recognition performance of the acoustic front-ends for noisy speech, we use machine-generated, zero-mean, white Gaussian noise and add it to each test utterance to get the desired signal-to-noise ratio. Speech recognition performance of the acoustic front-ends is studied here as a function of SNR.

In order to evaluate the recognition performance of the front-ends for speech with channel mismatch distortion, we need a procedure to introduce a controlled amount of channel mismatch distortion in the speech signal. However, to the best our knowledge, no such procedure is available in the literature. In order to devise such a procedure, we model the channel mismatch distortion by a parametric function. For simplicity, we assume it to be represented by a half sinusoid cycle, with two ends of the half sinusoid being at zero and $F_s/2$ frequencies and maximum of the half sinusoid being at frequency $F_s/4$. (Here, F_s is the sampling frequency of the speech signal.) The channel mismatch distortion can be controlled by changing the amplitude of the sinusoid. We measure the distortion by the value of this amplitude in decibels and evaluate the recognition performance of the acoustic front-ends for different dB values.

3. RECOGNITION EXPERIMENTS AND RESULTS

In our experiments, an HMM-based speech recognizer is used for the recognition of isolated words. Here, the HMM for each word has five states. Transitions between states are allowed only in left-to-right direction with no skipping of states. Mixtures of multivariate Gaussian functions are used to characterize the probability density functions of cepstral vectors in different states. The covariance matrices in the Gaussian probability density functions are assumed to be diagonal. The number of mixtures is 1 for the speaker-dependent system and 5 for the speaker-independent system. The Viterbi algorithm is used for training as well as for testing the recognizer.

Effect of additive noise distortion on the recognition performance of the four front-ends (FFTL, FFTH, UNIH and MELH) is shown in Table 1 for the speaker-dependent case and in Table 2 for the speaker-independent case. The following observations can be made from these tables:

1. Comparison of FFTL and FFTH results (see columns 2 and 3 in Tables 1 and 2) shows that the FFT homomorphic cepstral features are affected more by the additive noise distortion than the FFT LP cepstral features. A possible explanation for this may be as follows. Lower power regions in the power spectrum are affected more by the additive noise in terms of SNR than the higher power regions (i.e., the lower power regions have smaller SNR than the higher

Table 1: Effect of additive noise distortion on the recognition performance of the speaker-dependent isolated word recognizer for different front-ends.

SNR (in dB)	Recognition accuracy (in %) with			
	FFTL	FFTH	UNIH	MELH
Clean	90.1	88.6	89.1	86.5
30	68.8	59.2	69.6	71.3
25	59.9	46.5	59.6	64.0
20	47.4	32.9	47.8	55.1

Table 2: Effect of additive noise distortion on the recognition performance of the speaker-independent isolated word recognizer for different front-ends.

SNR (in dB)	Recognition accuracy (in %) with			
	FFTL	FFTH	UNIH	MELH
Clean	76.4	76.2	76.3	75.6
30	71.4	39.7	75.8	74.6
25	67.3	25.5	73.1	73.0
20	59.4	12.5	67.3	69.0

power regions). The FFTH front-end uses log power spectrum to compute the cepstral features. The logarithmic operation reduces the dynamic range of the power spectrum and makes the lower SNR regions in the power spectrum comparable to the higher SNR regions. Therefore, the resulting cepstral features are affected by the additive white noise distortion. The FFTL front-end operates on power spectrum itself (i.e., there is no logarithmic operation), the lower SNR regions in the power spectrum are negligible in magnitude than the larger SNR regions. Because of this, the additive noise distortion affects the LP cepstral features less than the homomorphic cepstral features. This explains why the FFTL front-end performs better than the FFTH front-end in presence of additive noise distortion .

2. The filter-bank based front-ends (UNIH and MELH) are slightly inferior to the FFT-based front-ends (FFTL and FFTH) in terms of recognition performance for clean speech, but are more robust to additive noise distortion. This observation can be explained as follows: The filter-bank analysis technique computes the power in a given band by averaging spectral power at frequencies within the band. Since lower SNR regions in the FFT power spectrum are much smaller than the higher SNR regions, the averaging operation neglects the contributions from these smaller SNR regions and the resulting filter-bank power spectrum is dominated by the larger SNR regions. Therefore, the filter-bank power spectrum is more robust to additive noise distortion than the power spectrum derived directly through the FFT algorithm. The UNIH and MELH front-ends use this filter-bank power spectrum and employ homomorphic analysis to smooth it further. This explains why these two front-ends are more robust to additive noise distortion than the FFT-based front-ends (FFTL and FFTH), but on clean speech their performance is slightly inferior. Note that the use of LP analysis on filter-bank power spectrum, as done in PLP analysis [3], may provide more robust cepstral features. However, it has to be investigated.

3. Comparison of the UNIH and MELH results (see columns 4 and 5 in Tables 1 and 2) shows that the MELH front-end is inferior in recognition performance than the UNIH front-end for clean speech, but it is more robust to additive white noise distortion. Thus, use of human auditory properties in an acoustic front-end makes it more robust to additive noise distortion. Similar observation has been made in earlier studies [4].

Effect of channel mismatch distortion on the recognition performance of the four front-ends (FFTL, FFTH, UNIH and MELH) is shown in Table 3 for the speaker-dependent case and in Table 4 for the speaker-independent case. Comparison of these tables with Tables 1 and 2 reveals that degradation in recognition performance due to channel mismatch distortion is less severe than that due to additive noise distortion. In addition, we can make following observations from Tables 3 and 4:

Table 3: Effect of channel mismatch distortion on the recognition performance of the speaker-dependent isolated word recognizer for different front-ends.

Distortion (in dB)	Recognition accuracy (in %) with			
	FFTL	FFTH	UNIH	MELH
Clean	90.1	88.6	89.1	86.5
4	86.7	83.9	85.4	83.9
8	75.7	71.2	74.9	76.5
12	52.2	43.3	56.9	67.1

Table 4: Effect of channel mismatch distortion on the recognition performance of the speaker-independent isolated word recognizer for different front-ends.

Distortion (in dB)	Recognition accuracy (in %) with			
	FFTL	FFTH	UNIH	MELH
Clean	76.4	76.2	76.3	75.6
4	73.8	73.0	74.6	73.1
8	70.5	68.9	72.1	71.4
12	64.1	60.7	67.4	68.1

1. The FFTL front-end performs better than the FFTH front-end in presence of channel mismatch distortion (see columns 2 and 3 of Tables 3 and 4). Thus, the LP analysis technique provides more robust cepstral features than the homomorphic analysis technique.

2. The UNIH front-end performs better than the FFTH front-end in presence of channel mismatch distortion (see columns 3 and 4 of Tables 3 and 4). Thus, the power spectrum derived from a filter-bank provides more robust cepstral features than the FFT power spectrum.

3. The MELH front-end performs better than the UNIH front-end in presence of channel mismatch distortion (see columns 4 and 5 of Tables 3 and 4). Thus, the use of human auditory properties in an acoustic front-end makes it more robust.

Note that the channel mismatch distortion affects the four acoustic front-ends in a similar manner as the additive noise distortion.

4. CONCLUSIONS

In this paper, four acoustic front-ends (FFTL, FFTH, UNIH and MELH) have been studied as to their recognition performance in presence of additive noise and channel mismatch distortions. It has been found that 1) the LP analysis technique provides more robust cepstral features than the homomorphic analysis technique, 2) the power spectrum derived from a filter-bank provides more robust cepstral features than the FFT power spectrum, and 3) use of human auditory properties in an acoustic

front-end makes it more robust.

References

[1] J.W. Picone, "Signal Modeling techniques in speech recognition", *Proc. IEEE*, Vol. 81, No. 9, Sept. 1993.

[2] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, Vol. 87, No. 4, pp. 1738-1752, Apr. 1990.

[4] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environments", *Computer Language and Speech*, Vol. 1, pp. 109-130, 1986.

[5] R. Cole, Y. Muthusamy and M. Fanty, "The ISOLET spoken letter database", Technical Report No. CSE 90-004, Dept. of Computer Science and Engineering, Oregon Institute of Science and Technology, Beaverton, OR, USA, Mar. 1990.