A MAXIMUM LIKELIHOOD EQUALIZATION TECHNIQUE FOR ROBUST SPEECH RECOGNITION IN ADVERSE ENVIRONMENTS

K.K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia

ABSTRACT

In this paper, we study the problem of robust speech recognition in adverse environments. We focus our attention to the following two types of distortions: 1) the additive noise distortion, 2) the channel mismatch distortion. The maximum likelihood (ML) equalization technique is used to compensate for these distortions. Performance of the ML technique is compared with the following channel equalization techniques: the global mean subtraction (GMS) technique, the local mean subtraction (LMS) technique, the finite impulse response (FIR) highpass filtering technique, the infinite impulse response (IIR) highpass filtering technique, the RASTA (bandpass) filtering technique, and the masking-based filtering technique. These techniques have been recently proposed in the literature and are computationally much simpler than the ML equalization technique. It is shown that the ML equalization technique does not offer any significant advantage over the other channel equalization techniques in terms of recognition performance.

1. INTRODUCTION

Though the currently-available speech recognizers work reasonably well in normal environments, their performance deteriorates drastically when they are deployed in adverse environments. These environments introduce noise, channel mismatch and other types of distortions in the speech signal. Since a speech recognizer is designed for a given acoustic environment, any mismatch in environmental conditions degrades its performance. For example, an isolated word recognizer that can recognize 10 English digits perfectly when spoken in laboratory environment, recognizes only 30% of the spoken digits when white noise is added to the signal with 10 dB signal-to-noise ratio (SNR) [1]. Similar degradations in recognition performance are observed due to channel mismatch. The recognition accuracy of the SPHINX speech recognition system on a speaker-independent alphanumeric task dropped from 85% correct to 20% correct when the close-talking Sennheiser microphone used in training was replaced by the omnidirectional Crown desktop microphone [2].

In the present paper, our aim is to study the problem of robust speech recognition in adverse environments. We focus our attention to the following two types of distortions: 1) the additive noise distortion, 2) the channel mismatch distortion. We propose to use a maximum likelihood (ML) equalization technique to compensate for these distortions. In the ML equalization technique, the distortion within an input utterance is assumed to be stationary. The ML equalization technique transforms the feature vectors (representing the distorted input speech utterance) in such a way that it maximizes the log-likelihood function. We use here a hidden Markov model (HMM) based speech recognition framework. The ML equalization technique is consistent with the HMM-based speech recognizer as likelihood is maximized for the transformation of the feature vectors in the former case and for the recognition of the input utterance in the later case. See references [2, 3, 4, 5, 6] for related work reported in the literature. In this paper, we compare the performance of the ML technique with the following channel equalization techniques: the global mean subtraction (GMS) technique [7], the local mean subtraction (LMS) technique [6], the finite impulse response (FIR) highpass filtering technique [8], the infinite impulse response (IIR) highpass filtering technique [9], the RASTA (bandpass) filtering technique [10], and the masking-based filtering technique [11]. These techniques have been recently proposed in the literature and are computationally much simpler than the ML equalization technique.

This paper is organized as follows. The maximum likelihood equalization technique is developed in Section 2. Section 3 describes the data base used for evaluating the channel equalization techniques for robust speech recognition under adverse environments. Speech recognition experiments and results are described in Section 4. Section 5 summarizes the paper.

2. MAXIMUM LIKELIHOOD EQUALIZATION TECHNIQUE

Here, we develop the maximum likelihood equalization procedure for isolated word recognition. However, it is general enough and can be used for continuous speech recognition employing sub-word units. We use here the cepstral coefficients derived through linear prediction analysis as recognition features.

Let the input utterance to be recognized be represented by a sequence of observation (cepstral) vectors, $Y = \{Y_1, Y_2, \ldots, Y_T\}$, where T is the number of frames in the input utterance. Since this utterance is spoken under adverse conditions, it is distorted. Our aim here is to clean up this distortion. For this, we transform each vector of this utterance such that the likelihood function is minimized. Let F_η denote the transformation (parameterized in terms of a parameter vector η), and $X = \{X_1, X_2, \ldots, X_T\}$ the observation sequence after transformation. Then

$$X_t = F_\eta(Y_t), \qquad \text{for } 1 \leq t \leq T. \tag{1}$$

Our goal is to find this transformation such that it maximizes the likelihood function under the HMM framework. For finding this transformation, consider a con-

tinuous density HMM $\lambda = [N, \pi, A, B]$, where N = the number of states in the model, $\pi = \{\pi_i, 1 \leq i \leq N\}$, the initial state probability vector (π_i is the probability that the model is in state i initially), $A = \{a_{ij}, 1 \leq i, j \leq N\}$, the transition matrix of underlying Markov chain (a_{ij} is the probability of transition from state i to state j), and $B = \{b_j(X_t), 1 \leq j \leq N\}$, the output probability matrix. Here $b_j(X_t)$ is the probability of outputting the vector X_t when the model is in state j. In our study, we represent $b_j(X_t)$ as a mixture of M normal probability density functions (PDFs); i.e.,

$$b_j(X_t) = \sum_{k=1}^{M} C_{jk} \mathbf{N}(X_t, \mu_{jk}, \sigma_{jk}) \quad (2)$$

$$= \sum_{k=1}^{M} \frac{C_{jk}}{[(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_{jki}]} \exp[-\sum_{i=1}^{d} (X_{ti} - \mu_{jki})^2 / 2\sigma_{jki}^2)], \quad (3)$$

where d is the number of features in an observation (cepstral) vector, C_{jk} is the mixture weight of k-th mixture of j-th state, and μ_{jk} and σ_{jk} are the mean and standard deviation vectors, respectively, of j-th state and k-th mixture. Note that we use here only diagonal covariance matrices (i.e., we assume off-diagonal elements to be zero).

The transformation F_η is estimated by the maximum likelihood formulation in two steps: segmentation and maximization. In the segmentation stage, the model λ is assumed to be known and the Viterbi algorithm is used to segment the observation sequence into states. Let the state sequence be given by,

$$q_1^T = \{q_1, q_2, \ldots, q_T\}.$$

In the maximization stage, the transformation is obtained by maximizing the likelihood function which is expressed as the probability of the sequence X given the model and state sequence and is written as

$$P(X \mid q_1^T, \lambda) = \prod_{t=1}^{T} b_{q_t}(X_t) \quad (4)$$

Let us denote q_t by j. Then, the log-likelihood of the sequence X is

$$L(X \mid q_1^T, \lambda) = \log P(X \mid q_1^T, \lambda) \quad (5)$$

$$= \sum_{t=1}^{T} \log(b_j(X_t)) \quad (6)$$

By substituting the value of X_t from Eq. (1) into Eq. (6), we get the likelihood function in terms of the transformation F_η. In order to solve for this transformation, we consider two cases. In case 1, we assume that the functional form of the transformation is known and it is represented in terms of a few parameters. For example, we know that the additive noise introduces multiplicative distortion in the cepstral vector. This means that $X_t = \alpha Y_t$, where α is a constant for a given utterance whose value depends on the amount of additive noise distortion. Similarly, we know that the channel mismatch distortion becomes additive in the cepstral domain. This means that $X_t = Y_t - B$, where B is the bias vector which remains constant for the input utterance. The parameters of this transformation can be computed by minimizing the likelihood function (Eq. (6)). In case 2, we do not know the functional form of the transformation. In this case, we use a multilayer perceptron to approximate this transformation. Note that the multilayer perceptron can provide nonlinear transformation. The connection weights of the multilayer perceptron can be estimated by the back-propagation algorithm using the likelihood function (Eq. (6)) as the cost function.

Note that in this procedure we compute the transformation for each input utterance we recognize. This may be computationally expensive in some applications. For this, we suggest an alternate way where we provide a small amount of calibration speech to the recognizer before its use to compute the transformation for a given adverse environment. Once we have learnt this transformation, we can use the recognizer with this transformation as long as our environmental condition does not change.

In the present paper, we assume this transformation to be linear and it is characterized by an additive bias vector B in the cepstral domain; i.e.,

$$X_t = Y_t - B. \quad (7)$$

The bias B is obtained by maximizing the following log likelihood function:

$$L(Y \mid B, q_1^T, \lambda) = \sum_{t=1}^{T} \log[\sum_{m=1}^{M} C_{jm} N(Y_t - B, \mu_{jm}, \sigma_{jm})], \quad (8)$$

where state q_t is denoted by j, and observation probability from this state is given by weighted sum of M Gaussian mixtures. The k-th component of bias vector B, obtained from this maximization, is given by

$$B_k = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M} \{\gamma_{jmt}(Y_{tk} - \mu_{jmk})/\sigma_{jmk}^2\}}{\sum_{t=1}^{T} \sum_{m=1}^{M} \{\gamma_{jmt}/\sigma_{jmk}^2\}}, \quad (9)$$

where

$$\gamma_{jmt} = \frac{C_{jm} N(Y_t, \mu_{jm}, \sigma_{jm})}{\sum_{l=1}^{M} C_{jl} N(Y_t, \mu_{jl}, \sigma_{jl})}. \quad (10)$$

3. SPEECH DATA BASE

In this paper, we study the problem of robust speech recognition under adverse environments using a speaker independent isolated word recognition system as a test bed. For this, we use the ISOLET spoken letter database from Oregon Graduate Institute[12]. Here, the vocabulary consists of 26 English alphabets (A-Z). From this data base, we use 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-sample the speech utterances from this data base to 8 kHz using a lowpass filter with cutoff frequency of 3.5 kHz. The speech signal is analyzed every 15 ms with a frame width of 45 ms (with Hamming window and preemphasis), and each frame is represented in terms of 10 cepstral features. LP analysis is done through the autocorrelation method with predictor order of 10.

For studying the speech recognition performance 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 is studied here as a function of SNR.

In order to evaluate the recognition performance 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 design a finite impulse response (FIR) digital filter using the windowing technique which approximates the parametric function representing the channel mismatch distortion. The input (unknown) utterance (from the test data) is filtered by this FIR filter to introduce the channel mismatch distortion. We measure the distortion by the value of this amplitude in decibels and evaluate the recognition performance for different dB values. Fig. 1 illustrates the frequency response of the FIR filter used to introduce a 12 dB channel mismatch distortion at 8 kHz sampling rate.

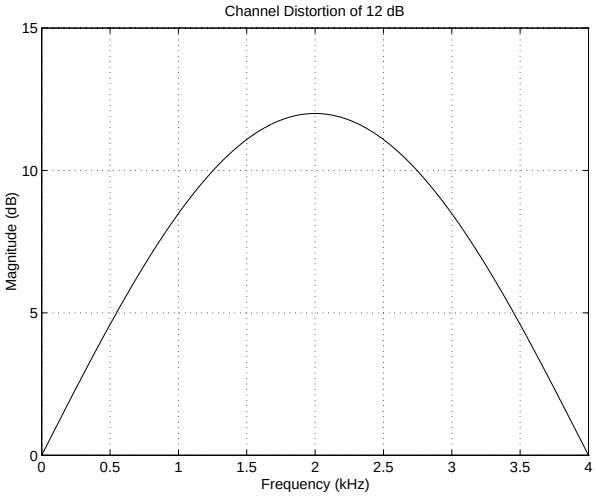

Figure 1: Frequency response of the FIR filter used to generate 12 dB channel mismatch distortion.

4. RECOGNITION EXPERIMENTS AND RESULTS

In order to see the effectiveness of the equalization technique for recognizing speech in adverse environments, we use an isolated-word speech recognition system as a test bed. We study its recognition performance in speaker independent mode. The recognition system uses a multimixture continuous density HMM framework. We use a 5-state continuous density HMM recognizer with probability density functions approximated by a mixture of 5 multivariate normal distributions with diagonal covariance matrices.

In order to put the ML equalization technique in a proper perspective, we compare its performance with other channel equalization techniques recently proposed in the literature. The channel equalization techniques used in our comparison are as follows: the global mean subtraction (GMS) technique [7], the local mean subtraction (LMS) technique [6], the finite impulse response (FIR) highpass filtering technique [8], the infinite impulse response (IIR) highpass filtering technique [9], the RASTA (bandpass) filtering technique [10], and the masking-based filtering technique [11].

We use 10 cepstral coefficients as recognition features and evaluate the recognition performance for different noise and channel mismatch conditions. Here we provide results for clean speech, 15 dB noisy speech and 12 dB channel distorted speech. Speech recognition results for different channel equalization techniques are shown in Table 1. Note that the ML results are obtained by modeling the distortion as a constant additive factor (or, bias) in the input utterance. It can be seen from this table that all the channel equalization techniques (except for the masking-based filtering technique) give better results than the baseline case where no processing is done for channel equalization. Among the channel equalization techniques reported in this table, only the ML and GMS techniques assume the distortion within the input utterance to be stationary, others assume it to be slowly varying function of time. The ML equalization technique performs slightly better than the GMS technique, but its performance is comparable with other techniques. Inclusion of delta-cepstrum and delta-delta-cepstrum in the feature set gives similar results as shown in Tables 2 and 3, respectively.

Table 1: Recognition performance of different channel equalization techniques using 10 cepstrum features.

Type of Processing	Recognition accuracy		
	Clean	Noisy	Channel
No processing	74.94	57.63	63.91
GMS	82.37	72.88	80.32
LMS	85.06	74.68	82.50
FIR HPF	84.81	72.50	82.50
IIR HPF	84.10	75.00	82.88
RASTA BPF	84.36	74.94	82.37
Mask	76.22	62.50	66.15
ML	84.88	74.10	82.58

Table 2: Recognition performance of different channel equalization techniques using 10 cepstrum and 10 delta cepstrum features.

Type of Processing	Recognition accuracy		
	Clean	Noisy	Channel
No processing	86.92	73.01	81.22
GMS	88.46	79.87	87.12
LMS	87.69	78.40	86.60
FIR HPF	88.46	79.81	87.56
IIR HPF	89.29	80.45	88.53
RASTA BPF	88.53	79.94	88.08
Mask	86.28	77.18	82.88
ML	88.72	80.15	88.15

We have also studied another version of the ML equalization technique, where effect of distortion on cepstral feature vectors is assumed to be a multiplicative factor. This version gives improved recognition results for speech corrupted by additive white noise. These results are consistent with our earlier results obtained by introducing

Table 3: Recognition performance of different channel equalization techniques using 10 cepstrum, 10 delta cepstrum and 10 delta-delta cepstrum features.

Type of Processing	Recognition accuracy		
	Clean	Noisy	Channel
No processing	88.97	76.86	86.41
GMS	89.29	80.64	88.53
LMS	87.11	78.91	85.13
FIR HPF	89.36	81.73	88.27
IIR HPF	88.91	80.32	88.65
RASTA BPF	88.85	81.86	88.21
Mask	88.97	78.85	84.81
ML	88.91	80.75	88.45

first-order equalization in the HMM framework [1]. These results will be reported later on in a separate paper.

5. SUMMARY

In this paper, speech recognition in presence of additive noise distortion and channel mismatch distortion is studied. The ML equalization technique is used to compensate for these distortions. Performance of the ML technique is compared with the following channel equalization techniques: the global mean subtraction (GMS) technique, the local mean subtraction (LMS) technique, the finite impulse response (FIR) highpass filtering technique, the infinite impulse response (IIR) highpass filtering technique, the RASTA (bandpass) filtering technique, and the masking-based filtering technique. These techniques are computationally much simpler than the ML equalization technique. It is shown that the ML equalization technique does not offer any significant advantage over the other channel equalization techniques in terms of recognition performance.

REFERENCES

[1] B.H. Juang and K.K. Paliwal, "Hidden Markov models with first-order equalization for noisy speech recognition", *IEEE Trans. Signal Processing*, Vol. 40, pp. 2136-2143, Sept. 1992.

[2] A. Acero and R.M. Stern, "Environmental robustness in automatic speech recognition", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Albuquerque, NM), Apr. 1990, pp. 849-952.

[3] S.J. Young, "Cepstral mean compensation for HMM recognition in noise", in *Proc. Workshop on Speech Processing in Adverse Environments*, (Cannes, France), Nov. 1992.

[4] Y. Zhao, "A new speaker adaptation technique using very short calibration speech" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* 1993, Vol. II, pp. 562-565.

[5] M.G. Rahim and B.H. Juang, "Signal bias removal for robust telephone based speech recognition in adverse environments", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* 1994, Vol. I, pp. 445-448.

[6] A.E. Rosenberg, C.H. Lee and F.K. Soong, "Cepstral channel normalization techniques for HMM-based speaker recognition", in *Proc. Int. Conf. Spoken Language Processing*, 1994, pp. 1835-1838.

[7] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Amer.*, Vol. 55, pp. 1304-1312, June 1974.

[8] D. Geller, R.H. Umbach and H. Ney, "Improvements in speech recognition for voice dialing in car environment", in *Proc. ECSA Workshop on Speech Processing in Adverse Conditions*, 1992, pp. 203-206.

[9] H. Murveit, J. Butzberger and M. Weintraub, "Reduced channel dependence for speech recognition", in *Proc. Speech and Natural Language Workshop (DARPA)*, 1992, pp. 280-284.

[10] H. Hermansky and N. Morgan, "RASTA processing of speech", *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589, Oct. 1994.

[11] K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* 1993, Vol. 2, pp. 668-671.

[12] 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.