Maximum Likelihood Adaptation for Distant Speech Recognition of Stationary and Moving Speakers in Reverberant Environments

Nokas George, Dermatas Evangelos, and Kokkinakis George

Wire Communications Laboratory, Electrical & Computer Engineering Department,
University of Patras, 26100 Patras, Hellas.
E-mail: nokas@george.wcl2.ee.upatras.gr

Abstract

In this paper, a feature transformation method is presented for distant speech recognition in reverberant and noisy environments. In the Maximum Likelihood framework the optimum bias parameters are obtained on-line, using a small number of successive speech frames. The stochastic matching is achieved by assuming a mixture of Gaussian pdfs for the clean speech features. The proposed method was evaluated on the Mel-scaled Frequency Cepstral Coefficient (MFCC) features as well as on MFCC after cepstral mean subtraction and after RASTA filtering. The experiments, carried out in several adverse conditions including room acoustics and additive factory noise for stationary and moving speakers, have shown significant improvement of the classification rate in isolated word speech recognition applications. The proposed method improves the recognition rate of a standing speaker by more than 50%, when SNR is higher than 10db. In the case of a moving speaker the improvement is 8.6% using MFCC while the recognition rate reaches 91.05% using RASTA features.

1. Introduction

1.1. General

After years of research, a great deal of progress has been achieved in building automatic speech recognition (ASR) systems with high recognition rate. Now the challenge is to increase the robustness of these systems by facing adverse interference such as noise, reverberations, microphone transducer distortions, etc, which lead to acoustic mismatches between the training and the testing conditions of the recognition systems. In this paper we investigate the most adverse case where a single, hands-free microphone is used, the speaker is moving and the mismatch effect is created by the room's impulse response and a noise source.

Various methods have been proposed to eliminate the mismatch effect between the original speech and the signal received in the microphone. Among them, the direct estimation of the room's impulse response and the deconvolution of the signal by inverse filtering is a quite difficult procedure, because most typical rooms have unstable inverses. So alternatively, more efficient solutions have been proposed.

Multi-microphone methods have been successfully used for signal dereverberation, employing signal-preprocessing modules in the time domain [1]. In this category of methods, the hypothesis that the impulse responses of the speech source and the microphones transfer function are uncorrelated at different locations is used to suppress the reverberant speech tails by delay-and-sum beam forming.

In the cepstral domain, the most simple and popular method employed for dereverberation is the cepstral mean subtraction (CMS). The concept of this method is to subtract from each cepstral vector the cepstral mean, by assuming that the average of the speech vectors is zero.

Assuming linear propagation of acoustic signals through the air, the microphone signal is the convolution of the original signal, with the impulse response of the surrounding environment. If we assume linear impulse response $h(t)$, then the contaminated signal can be modeled as: $x(t)=y(t)*h(t)+n(t)$, where * denotes convolution, $y(t)$ is the original signal and $n(t)$ is the additive noise. Ignoring the noise component in the cepstral domain the speech features in the microphone become $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{h}$ where $\boldsymbol{y}$ and $\boldsymbol{x}$ are cepstal vectors and $\boldsymbol{h}$ is the additive shift representing the room impulse response.

Another approach is based on the RASTA technique [2]. In this method an IIR filter suppresses the continuous components of the cepstral vectors. Other techniques eliminate the reverberation in an adaptive way based on the blind deconvolution framework [3,4].

1.2. Maximum Likelihood and the EM Algorithm

Recent solutions are based on the stochastic matching framework and the Maximum Likelihood (ML) approach. Acero [5] proposes a codeword-dependent technique that computes a correction vector via maximum likelihood, for environmental adaptation of different microphones. The maximization is achieved via the Expectation-Maximization (EM) algorithm. Rahim and Hwang [6] propose an iterative estimation of $\boldsymbol{h}$ for speech recognition over the telephone. The purpose is to estimate the $\boldsymbol{h}$ that maximizes a likelihood of the form: $p(X|h, \Lambda) = \prod_t \max_i p(x_t - h|\lambda_i)$, where $X=\{\boldsymbol{x_1},\boldsymbol{x_2},\ldots \boldsymbol{x_t},\ldots \boldsymbol{x_T}\}$ is an observation sequence and $\Lambda=\{\lambda_i,\ i = 1,2,\ldots,M\}$ is a Markov model for a speech unit i. The maximization over $\boldsymbol{h}$ is based on an updated nearest neighbor search method. Sankar et al. [7] propose an EM solution for the maximization of the likelihood

$$\nu' = \arg\max_{\nu} p(X \mid \nu)$$

where $\boldsymbol{v}$ is a function of $\boldsymbol{h}$.

Another ML approach [8], estimates the channel, additive and convoluted noise distortions in the frequency domain for robust speaker-independent continuous speech recognition using the EM algorithm. The experimental results showed significant improvement in word recognition rates.

Recently [9], an HMM composition and separation method has been used to eliminate the presence of both acoustic

transfer function and additive noise by assuming a known stochastic model for the noise signal. In a double stochastic optimization method (EM algorithm) the HMM separation is applied in the linear-cepstral domain to remove noise components and after a non-linear transformation, the HMM separation is applied again to estimate the acoustic transfer function.

1.3. Maximum Likelihood and Gradient maximization

Gradient methods and scoring algorithms are competing numerical procedures to the EM algorithm. Compared with EM, gradient methods are simpler in implementation and have impressive numerical stability. In addition, convergence of the EM algorithm is often painfully slow using a large amount of training data.

In our case, where the speaker is moving and the acoustic transfer function varies relatively fast, the ML solution is achieved using the gradient ascent method, performing one update for each incoming frame. Such a solution can be employed in any time-varying application and yields the time-varying local maximum of the likelihood function. The method is data driven, independent of the actual feature extraction and recognition method, and can be implemented as front-end process.

The structure of this paper is as follows. In the following section, a detailed description of the proposed method is given. The speech recognition system and the speech databases are presented in section 3 while in section 4 the experimental results are given. A short discussion in section 5 concludes this work.

2. The Stochastic Matching Framework

The aim of our approach is to derive the bias parameter $\boldsymbol{h}$ that maximizes the likelihood

$$L(X|\Theta) = \ln \prod_t p(x_t - h|\Theta)$$

by assuming a mixture of Gaussian densities with parameters Θ for the clean speech.

Assuming known model parameters for the clean speech, we derive a maximum likelihood estimation of the bias interference as follows:

$$h' = \arg\max_h \ln \prod_{i=0}^{T} p(x_{t-i} - h|\Theta)$$

where $\boldsymbol{x_t}$ is the feature vector of the speech frame received in the microphone, and T is a small number of subsequent frames. The maximization is achieved by the gradient ascent iterative method; the shift difference Δh is set in the direction of the log-likelihood first derivative. This procedure can be expressed as a function of the clean speech model parameters and the distorted data, as follows:

$$\Delta h_k^{(t)} = h_k^{(t)} - h_k^{(t-1)} = a \cdot \frac{\partial}{\partial h_k} \left(\ln \prod_{i=0}^{T} p(x_{t-i} - h;\Theta) \right) \Big|_{h_k \to h_k^{(t-1)}}$$

$$= a \sum_{i=0}^{T} \frac{1}{p(x_{t-i} - h;\Theta)} \cdot \frac{\partial}{\partial h_k} p(x_{t-i} - h;\Theta) \Big|_{h_k \to h_k^{(t-1)}}$$

where t is the frame (iteration) number, k is the coefficient index, and a is the learning rate.

Assuming that the probability density function (pdf) of the clean speech vector y, can be approximated by a mixture of Gaussians, its compact form is given by the following relation:

$$p(y;\Theta) = \sum_{i=1}^{M} c_i . N(y, m_i, \sigma_i)$$

where M is the number of mixtures, c_i is the probability associated to the i mixture component, $\boldsymbol{m_i}$ is the mean vector and $N(.)$ is the Gaussian distribution. For simplicity reasons we assume common variance σ_i for each dimension of the features vector. Thus, the adaptation equation for the case of a mixture of Gaussians becomes:

$$\Delta h_k = a \cdot \sum_{i=0}^{T} G(i) \cdot \sum_{j=1}^{M} c_j \cdot N(x_{t-i} - h, m_j, \sigma_j) \cdot \frac{(m_{j,k} - x_{t-i,k} + h_k)}{\sigma_{j,k}^2}$$

where, $G(i) = \left(\sum_{q=1}^{M} c_q \cdot N(x_{t-i} - h, m_q, \sigma_q) \right)^{-1}$

In the above equation, the bias h is adapted in the direction of maximizing MLT, given the clean speech statistics Θ and T feature vectors. The number of iterations can vary in different applications. When fast adaptation is required, the algorithm may have a non-block form, performing one bias adaptation for each new incoming vector. In the block implementation, the bias adaptations are computed continuously until reaching an *a priori* defined threshold for each incoming feature vector.

In the case of HMM based recognition, the clean speech pdf can be obtained directly from the models of the recognizer [6,7]. In our approach, although we also use HMM based recognition, we suggest an independent from the recognizer solution.

3. The Recognition System and the Speech Databases

The evaluation of the proposed method (Maximum Likelihood Transformation, MLT) was performed in multiple experiments using a speech database (DBclean), which consists of 13 command words and the ten digits of the Greek language recorded by 76 speakers in an anechoic chamber. The recordings of 35 randomly selected speakers composed the training set and the remaining ones were used for testing purposes. A loudspeaker reproduced the DBclean recordings in a real room (6.5x7.2x3.1m). Acquisitions were carried out with sampling frequency of 16 kHz and 16 bit accuracy. Thus, in the real room speech corpus (DB1), the clean speech is affected by the room transfer function, the loudspeaker's impulse response, and the presence of computer fan noise. In our experiments the distance between the microphone and the loudspeaker was 2 m.

Additional experiments were carried out to measure the method's performance in linear distortions by using a simulated speech database. The DBclean was convoluted with a real room's transfer function producing an artificial database (DB2).

A speaker-independent isolated word recognition system based on a whole word CDHMM was used for the classification experiments. Each word model is a five states left to right CDHMM with no state skip. The feature vector consists of 13 mel frequency cepstral coefficients for each frame. The frame was set to 32 ms (512 samples) with a frame

shift of 20 ms. The output distribution probabilities were modeled by means of a Gaussian component with diagonal covariance matrix. The segmental k-means training algorithm was used to estimate the HMM parameters from multiple feature vectors.

4. Experimental Results

Three sets of features were tested; the MFCC, the MFCC after cepstral mean subtraction and the MFCC after RASTA filtering. In these experiments the bias *h* was initialized in the beginning of each spoken word, and was adapted for each incoming frame. The number of mixtures in the pdf estimation was set to 3 and we used the last 50 frames in the MLT implementation. In all experiments the speech model parameters were estimated using 60 sec from the DBclean recordings and the EM algorithm.

In table 1, the recognition rate degradation under matched and mismatched environmental conditions is shown with (second raw) and without (first raw) bias subtraction. The reverberation factor and the impulse response of the loudspeaker , which was used for the reproduction of the database, decrease the recognition accuracy from the rate of 96% to 42%. The proposed method increases the rate to 67%, which means an improvement of 59.1%.

In figure 1, the recognition rate using the MFCC features with and without MLT in the DB1 corpus under additive colored noise (factory noise from NOISEX-92) at different SNR is shown. The best learning rate for the MLT algorithm was found experimentally to be a=5x10^{-5}. The same experiments were carried out with the DB2 corpus. At high SNR the proposed algorithm almost eliminates the distortions. At SNR$\rightarrow\infty$ (clean speech) a recognition rate of 94% is achieved while the anechoic testing gives 96.1% (table 1). Nevertheless below 5dB only small improvements are measured.

MFCC	ANECHOIC RECORDINGS	REVERBERANT RECORDINGS
Anechoic Training	96.1	42.1
Anechoic Training, MLT	96.5	67.0

Table 1. Percent recognition rate with the DB1 speech corpus

The second set of experiments deals with the CMS deconvolution method. In the training process, both HMM and stochastic model parameters were estimated with DBclean recordings and CMS processing: from each incoming speech vector the mean value of the last 50 frames was subtracted. Figure 2 shows the results for the DB1 and DB2 recordings. For this kind of features the best learning rate was found to be a=10^{-7}. For DB1 the CMS technique gave a recognition rate of 73,2% for the clean speech (SNR$\rightarrow\infty$), which is significantly greater than the 42.1%, obtained with the pure MFCC features (table 1). The CMS robustness in the presence of channel and noise can be studied by comparing the contours MFCC_DB1 CMS_DB1 and MFCC_DB2 CMS_DB2 in figures 2,3. Applying the MLT only a slight improvement (1.5%) in the recognition accuracy is acieved giving a rate of 74,5%. For the artificial recordings (DB2) and for clean recordings, the CMS method gives 95,38% almost eliminating the channel distortion. Finally, the MLT method does not improve significantly the recognition rates.

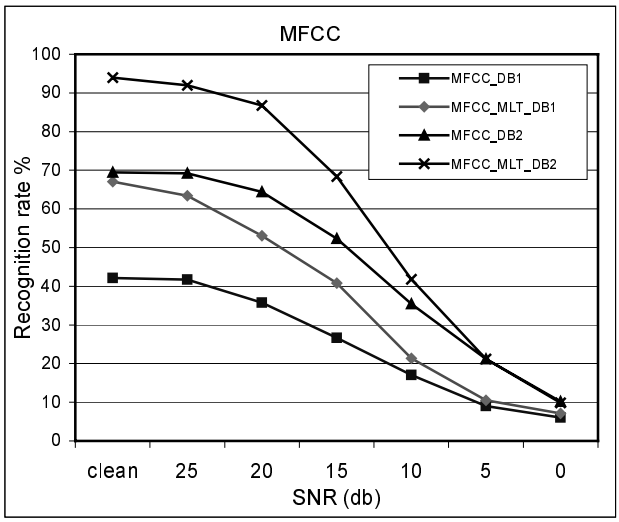

Fig. 1. Recognition experiments with MFCC features.

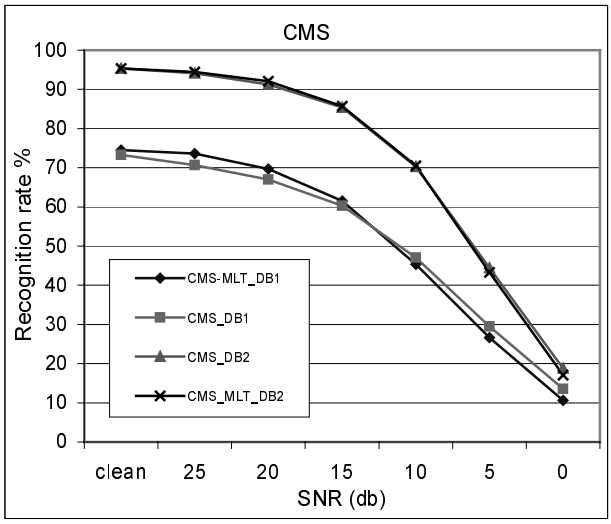

Fig. 2. Recognition experiments using the CMS method.

The next experiment deals with RASTA filtering of the MFCC features Fig.3. The best learning rate for the MLT algorithm was found to be a=10^{-5}. Compared with CMS at SNR$\rightarrow\infty$, the RASTA features of DB1 give a slightly lower recognition rate of 71,05%. On the contrary, when noise distorts the speech signal, the RASTA filtering produces robust MFCC features.

After MLT the recognition rate for the clean recordings is increased from the rate of 71,05% to 80,06%.This improvement is noticeable up to 5 dB SNR. For the artificial recordings and for the clean recordings, the RASTA filtering gives 92,865%, a little worse rate compared with the CMS method. As with the DB1 recordings, the RASTA features are more robust under noise giving higher recognition rates.

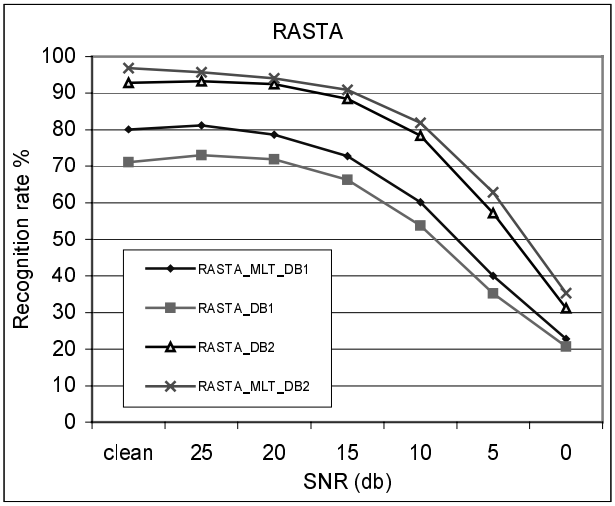

Fig. 3. Recognition experiments after RASTA filtering.

Furthermore, we have investigated the adaptive capability of our algorithm, in the case of a moving speaker. The scenario of our experiment is illustrated in figure 4, where a continuous circular movement of the speaker is assumed. All distances are measured in meters. In this case the channel interference varies continuously. For comparison reasons we also give the recognition rate for a standing speaker in a position inside the area of the circular movement with coordinates (6.6,6.6,1). For the experiments concerning the above scenario the movement of a speaker is simulated using DBclean.

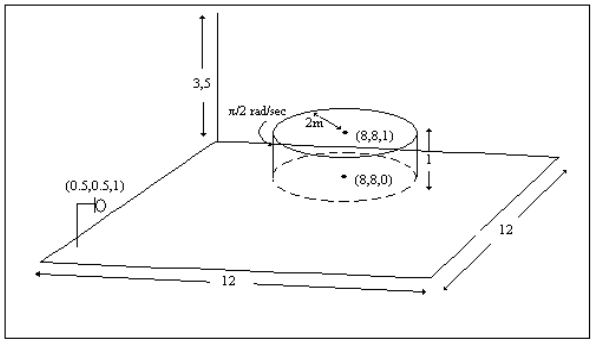

Fig. 4. Scenario of the moving speaker.

In table 2, the experimental results for the standing and moving speaker obtained with the MFCC features, CMS processing and RASTA filtering are given. For comparison reasons a third column containing the results achieved with anechoic recordings (DBclean) is added. In all cases the implementation of our algorithm improves the recognition rate. A significant improvement, approaching the recognition rate of the standing speaker, is achieved by the of RASTA filtering and the proposed method. In the case of the CMS features the improvement is unnoticeable.

5. Conclusion

A new adaptive feature transformation method is proposed for robust speech recognition in reverberant and noisy environments. The on-line adaptive algorithm provides the bias parameters in the feature space using the maximum likelihood criterion. Experiments with MFCC features, MFCC after CMS processing and MFCC after RASTA filtering have shown significant improvements in the recognition rate even in the presence of colored noise. Our method was also tested in a changing environment produced by a moving speaker. The results are very encouraging especially when RASTA features are used.

	STANDING	MOVING	ANECHOIC
MFCC	74.620	70.234	96.140
MFCC_MLT	80.460	76.080	96.550
CMS	89.240	84.620	97.544
CMS_MLT	89.350	84.910	97.250
RASTA	90.877	87.953	98.363
RASTA_MLT	94.730	91.050	98.180

T*a*ble 2. Percent recognition rate for the scenario of figure 4.

6. References

[1] J.L Flanagan, J.D Johnston, R. Zahn and G.W. Elko: ”Computer-steered mocrophone arrays for sound transduction in large rooms”, J. Acoust. Soc. Amer., Vol .78, No. 5, pp. 1508-1518, 1985.

[2] H. Hermansky, N. Morgan, A. Bayya, P. Kohn: “Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)”, in Proc. EUROSPEECH'91, pp. 453-485 1991.

[3] C. Mokbel, D, Jouvet, J. Monne: ”Deconvolution of line effects for speech recognition”, Speech Communication 19, pp. 185-196, 1996.

[4] G. Nokas, E. Dermatas: ”Speech recognition in noisy reverberant rooms using a frequency domain blind deconvolution method”, Eurospeech’99, pp 2853-2856, 1999.

[5] A. Acero: “Acoustical and Environmental Robustness in Automatic Speech Recognition”, Kluwer Academic Publishers, 1991.

[6] M. Rahim and Biing-Hwang Juang: “Signal Bias removal by maximum likelihood Estimation for Robust Telephone Speech Recognition”, IEEE Trans. On Speech And Audio Processing, Vol 4, No 1, pp.19-30,1996.

[7] A. Sankar, Chin-Hui Lee: “A Maximum-Likelihood Aproach to Stochastic Matching for Robust Speech Recognition”, IEEE Trans. On Speech And Audio Processing, Vol. 4, No 3, pp.190-202, 1996.

[8] Yunxin Zhao: "Frequency-Domain Maximum Likelihood Estimation for automatic Speech Recognition in Additive and Convolutive Noises", IEEE Trans. Speech Audio Processing, Vol. 8, pp. 255-267, 2000.

[9] Takiguchi T., Nakamura S., Shikano K: “Speech Recognition for a distant moving Speaker based on HMM composition and Separation”, ICASSP-2000, pp 1587-1590, 2000.