ROBUST SPEECH RECOGNITION IN ADDITIVE AND CHANNEL NOISE ENVIRONMENTS USING GMM AND EM ALGORITHM

Masakiyo Fujimoto [1,2]

[1] ATR Spoken Language Translation Research Laboratories, Kyoto, Japan
[2] Department of Electronics and Informatics, Ryukoku University, Otsu, JAPAN
masa@arikilab.elec.ryukoku.ac.jp

Yasuo Ariki [3]

[3] Research Center for Urban Safety and Security & Department of Computer and System Engineering, Kobe University
Nada, Kobe, 657-8501, JAPAN
ariki@kobe-u.ac.jp

ABSTRACT

In this paper, we evaluated the speech recognition in real driving car environments by using a GMM based speech estimation method and an EM algorithm based channel noise estimation method. The GMM based speech estimation method proposed by Segura et al was not robust for channel noise such as an acoustic transfer function, a microphone characteristic and so on. To cope with this problem, we propose a channel noise estimation method based on the EM algorithm. Furthermore, we estimate the speech signal more accurately by using a speech GMM and a silence GMM instead of the GMM trained without speech/silence discrimination. Our proposed method has been evaluated on the AURORA3 tasks. In the evaluation results, the proposed method showed the significant improvement in the high-mismatched condition test of AURORA3 tasks.

1. INTRODUCTION

In recent years, many types of speech recognition systems have been proposed and developed toward the practical use in a real world. However, most of the works recognize clean speech collected in quiet environments. In practical use it is required for recognition systems to be robust to interfering noises.

Robust speech recognition systems are classified into two types. One adapts itself to any kinds of noises based on model adaptation techniques[1]-[4]. The other reduces the noise component from noisy speech based on noise reduction techniques[5, 6].

Parallel model combination(PMC) method[1] has been proposed which adapts the speech recognition system to any kinds of noises. To improve the recognition accuracy under non-stationary noisy environments, a compensation method for time varying residual noise has been proposed[3]. However, PMC has a problem that it needs a huge quantity of computation, if it is applied to the acoustic model for a large number of phonemes with mixture distributions like a triphone model HMM.

On the other hand, spectral subtraction(SS) method[5] has been proposed as a conventional noise reduction method. However, SS has a problem that it degrades the recognition rate due to spectral distortion caused by over or under subtraction. In addition, the SS does not consider the time varying property of noise spectra, because it estimates the noise spectra as mean spectra within the time section assumed to be noise (usually, beginning of utterance).

For the above mentioned problems, J.C.Segura et al proposed a Gaussian mixture model(GMM) based speech estimation method[6]. It estimates the expectation of the mismatch factor between clean speech and noisy speech at each frame by using GMM of clean speech and mean vector of noise, and showed the significant improvements in recognition accuracy. However, the Segura's method considered only the additive noise environments and it did not consider about the channel noise problem such as an acoustic transfer function, a microphone characteristic and so on. For example, in car speech recognition with a distant(hands-free) microphone, it is necessary to cope with not only the additive noise but also the acoustic transfer function(channel noise). From this viewpoint, in this study, we propose a channel noise estimation method based on an EM algorithm, in the same spirit of CDCN[7] and VTS[8] method. Furthermore, we estimate the speech signal more accurately by using a speech GMM and a silence GMM instead of the GMM trained without speech/silence discrimination, because the differences of the feature parameters are large between speech sections and silence sections. By using the proposed additive and channel noise estimation method, the speech signal corrupted by additive and channel noise is recognized accurately.

Our proposed method has been evaluated on the AURORA3 tasks[9]. In the evaluation results, the proposed method showed the significant improvement in the high-mismatched condition test of AURORA3 tasks.

2. GMM BASED SPEECH ESTIMATION

2.1. Signal model

At the ith frame, the logarithmic output energy of Mel filter bank of observed noisy speech is represented as follows[6]:

$$\begin{aligned}\mathbf{X}(i) &= \log\left[\exp(\mathbf{S}(i)) + \exp(\mathbf{N}(i))\right] \\ &= \mathbf{S}(i) + \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \\ &= \mathbf{S}(i) + \mathbf{G}(i) \qquad (1)\end{aligned}$$

$$\mathbf{G}(i) = \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \qquad (2)$$

where $\mathbf{X}(i)$, $\mathbf{S}(i)$ and $\mathbf{N}(i)$ denote the vectors which have logarithmic output energy of Mel filter bank of observed noisy speech, clean speech and noise, respectively.

In Eq.(1), $\mathbf{G}(i)$ is equivalent to the mismatch factor between $\mathbf{X}(i)$ and $\mathbf{S}(i)$.

2.2. GMM based mismatch factor estimation

At first, it is supposed that $\mathbf{S}(i)$ can be modeled by GMM with K mixture distributions as follows:

$$p(\mathbf{S}(i)) = \sum_{k=1}^{K} P_k p_k\left(\mathbf{S}(i)|\mu_{S,k}, \mathbf{\Sigma}_{S,k}\right) \quad (3)$$

$$p_k\left(\mathbf{S}(i)|\mu_{S,k}, \Sigma_{S,k}\right) = \mathcal{N}\left(\mathbf{S}(i), \mu_{S,k}, \mathbf{\Sigma}_{S,k}\right) \quad (4)$$

where $p(\mathbf{S}(i))$ denotes the output probability of $\mathbf{S}(i)$, and P_k, $\mu_{S,k}$ and $\mathbf{\Sigma}_{S,k}$ denote the mixture weight, mean vector and diagonal covariance matrix of the kth Gaussian distribution, respectively.

Next, it is supposed that $\mathbf{X}(i)$ can be modeled by GMM with K mixture distributions as well as $\mathbf{S}(i)$ as follows:

$$p(\mathbf{X}(i)) = \sum_{k=1}^{K} P_k p_k\left(\mathbf{X}(i)|\mu_{X,k}, \mathbf{\Sigma}_{X,k}\right). \quad (5)$$

Here, in most case, the GMM of $\mathbf{X}(i)$ cannot be given as a prior information. However, when GMM of $\mathbf{S}(i)$ is given, GMM of $\mathbf{X}(i)$ can be obtained approximately by using log-add compensation method[2] in the following way.

Let μ_N denotes the mean vector of $\mathbf{N}(i)$ which is estimated using the first 10 frames of the observed noisy speech $\mathbf{X}(i)$. Then the mean vector of $\mathbf{X}(i)$ at the kth Gaussian distribution is estimated as follows based on Eq.(1) and (2).

$$\begin{aligned} \mu_{X,k} &\simeq \mu_{S,k} + \log\left[1 + \exp(\mu_N - \mu_{S,k})\right] \\ &= \mu_{S,k} + \mu_{G,k} \end{aligned} \quad (6)$$

On the other hand, the covariance matrix of $\mathbf{X}(i)$ is not modified, because the estimation accuracy of the covariance matrix of $\mathbf{N}(i)$ estimated from the first 10 frames of the observed noisy speech $\mathbf{X}(i)$ is very poor. Therefore, the covariance matrix $\mathbf{\Sigma}_{S,k}$ is used as the covariance matrix $\mathbf{\Sigma}_{X,k}$ as shown in Eq.(7).

$$\mathbf{\Sigma}_{X,k} \simeq \mathbf{\Sigma}_{S,k} \quad (7)$$

In Eq.(6), $\mu_{G,k}$ corresponds to the mean vector of the mismatch factor at kth Gaussian distribution. Therefore, the expectation of $\mathbf{G}(i)$ is estimated as weighted average of $\mu_{G,k}$ by using a posterior probability $P_{i,k}$ as follows[6]:

$$\hat{\mathbf{G}}(i) = \sum_{k=1}^{K} P_{i,k}\mu_{G,k} \quad (8)$$

$$P_{i,k} = \frac{P_k p_k\left(\mathbf{X}(i)|\mu_{X,k}, \mathbf{\Sigma}_{X,k}\right)}{\sum_{k'=1}^{K} P_{k'} p_{k'}\left(\mathbf{X}(i)|\mu_{X,k'}, \mathbf{\Sigma}_{X,k'}\right)} \quad (9)$$

From the procedure described above, the clean speech $\hat{\mathbf{S}}(i)$ is estimated by subtracting $\hat{\mathbf{G}}(i)$ from $\mathbf{X}(i)$ as follows[6]:

$$\hat{\mathbf{S}}(i) = \mathbf{X}(i) - \hat{\mathbf{G}}(i). \quad (10)$$

3. SPEECH ESTIMATION IN ADDITIVE AND CHANNEL NOISE ENVIRONMENT

3.1. Signal model with additive and channel noise

In the GMM based speech estimation method described in Sec.2, the signal model represented by Eq.(1) is constructed under the assumption that the additive noise exists alone. However, in car speech recognition with distant(hands-free) microphone, it is necessary to cope with not only the additive noise but also the acoustic transfer function(channel noise).

Let $\mathbf{H}$ denotes the mean vector of channel noise, the logarithmic output energy of Mel filter bank of the observed signal in additive and channel noise environment is represented as follows:

$$\begin{aligned} \mathbf{X_H}(i) &= \log\left[\exp(\mathbf{H})\left(\exp(\mathbf{S}(i)) + \exp(\mathbf{N}(i))\right)\right] \\ &= \mathbf{S}(i) + \mathbf{G}(i) + \mathbf{H} \\ &= \mathbf{X}(i) + \mathbf{H} \end{aligned} \quad (11)$$

where $\mathbf{X_H}(i)$ denotes the observed signal with channel noise and $\mathbf{H}$ is assumed to be a time constant parameter.

By using above signal model, the GMM of $\mathbf{X_H}(i)$ can be represented by

$$p(\mathbf{X_H}(i)) = \sum_{k=1}^{K} P_k p_k(\mathbf{X_H}(i)|\mu_{X_H,k}, \mathbf{\Sigma}_{X_H,k}). \quad (12)$$

Since $\mathbf{H}$ is a time constant parameter, the mean vector of $\mathbf{H}$ can be represented as $\mu_H = \mathbf{H}$. Furthermore, μ_H is assumed to be a common parameter in each Gaussian distribution of the GMM of $\mathbf{X_H}(i)$. From these assumptions, in kth Gaussian distribution, the mean vector of $\mathbf{X_H}(i)$ is represented by

$$\begin{aligned} \mu_{X_H,k} &= \mu_{S,k} + \mu_{G,k} + \mu_H \\ &= \mu_{X,k} + \mu_H. \end{aligned} \quad (13)$$

On the other hand, the diagonal covariance matrix of $\mathbf{X_H}(i)$ is represented as follows:

$$\begin{aligned} \mathbf{\Sigma}_{X_H,k} &= E\left[\left(\mathbf{X_H}(i) - \mu_{X_H,k}\right)\left(\mathbf{X_H}(i) - \mu_{X_H,k}\right)^T\right] \\ &= E\left[\left(\mathbf{X}(i) + \mathbf{H} - \mu_{X,k} - \mu_H\right)\right. \\ &\quad \left.\times\left(\mathbf{X}(i) + \mathbf{H} - \mu_{X,k} - \mu_H\right)^T\right], \end{aligned} \quad (14)$$

finally, from the assumption of $\mu_H = \mathbf{H}$, $\mathbf{\Sigma}_{X_H,k}$ can be represented as

$$\begin{aligned} \mathbf{\Sigma}_{X_H,k} &= E\left[\left(\mathbf{X}(i) - \mu_{X,k}\right)\left(\mathbf{X}(i) - \mu_{X,k}\right)^T\right] \\ &= \mathbf{\Sigma}_{X,k} \left(= \mathbf{\Sigma}_{S,k}\right). \end{aligned} \quad (15)$$

3.2. Discrimination of speech section and silence section

In Sec.2.2, the mismatch factor $\mu_{G,k}$ is estimated by using the GMM which is trained without discrimination of speech sections and silence sections in training materials. However, it is necessary to estimate the mismatch factor by using the speech GMM and the silence GMM which are both trained after discrimination of speech sections and silence sections in training materials, because the differences of the feature parameters are large between speech sections and silence sections. Therefore, in this study, we propose a mismatch factor estimation method using the speech GMM and the silence GMM.

At first, we trained the speech GMM and the silence GMM by using clean speech training materials and composed the GMM for $\mathbf{X_H}(i)$ by log-add compensation method[2] using the speech GMM and the silence GMM as follows:.

$$p^{(s)}(\mathbf{X_H}(i)) = \sum_{k=1}^{K} P_k^{(s)} p_k\left(\mathbf{X_H}(i)|\mu_{X_H,k}^{(s)}, \mathbf{\Sigma}_{X_H,k}^{(s)}\right) \quad (16)$$

where if $s = 0$, the parameters of GMM are composed using the silence GMM. Otherwise, they are composed using the speech GMM.

By using Eq.(13) and Eq.(15), the parameters of each composed GMM are represented as follows:

$$\begin{aligned} \mu_{X_H,k}^{(s)} &= \mu_{S,k}^{(s)} + \mu_{G,k}^{(s)} + \mu_H \\ &= \mu_{X,k}^{(s)} + \mu_H \end{aligned} \quad (17)$$

$$\mathbf{\Sigma}^{(s)}_{X_H,k} = \mathbf{\Sigma}^{(s)}_{X,k} \ \left(= \mathbf{\Sigma}^{(s)}_{S,k}\right) \tag{18}$$

where μ_H is assumed to be a common parameter to two types of composed GMMs.

3.3. Channel noise estimation based on the EM algorithm

In Sec.3.1, the parameter μ_H is introduced. However, μ_H is a prior unknown parameter, when $\mathbf{X_H}(i)$ is observed. To solve this problem, we estimated the μ_H by maximizing the GMMs with likelihood of $\mathbf{X_H}(i)$ based on an EM algorithm.

When the incomplete data(observable data) $\mathbf{x} = (\mathbf{X_H})$ and the unobservable data $\mathbf{y}$ are given, the complete data of the EM algorithm consists of $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Here, the model parameters consist of $\lambda = (\mathbf{P}, \mu_X, \mathbf{\Sigma}_X, \mu_H)$.

$$\mathbf{X_H} = (\mathbf{X_H}(0), \ldots, \mathbf{X_H}(i), \ldots, \mathbf{X_H}(N-1)) \tag{19}$$

$$\begin{aligned} \mathbf{P} &= \left(P^{(0)}_1, \ldots, P^{(0)}_k, \ldots, P^{(0)}_K, \right. \\ &\quad \left. P^{(1)}_1, \ldots, P^{(1)}_k, \ldots, P^{(1)}_K\right) \end{aligned} \tag{20}$$

$$\begin{aligned} \mu_X &= \left(\mu^{(0)}_{X,1}, \ldots, \mu^{(0)}_{X,k}, \ldots, \mu^{(0)}_{X,K}, \right. \\ &\quad \left. \mu^{(1)}_{X,1}, \ldots, \mu^{(1)}_{X,k}, \ldots, \mu^{(1)}_{X,K}\right) \end{aligned} \tag{21}$$

$$\begin{aligned} \mathbf{\Sigma}_X &= \left(\mathbf{\Sigma}^{(0)}_{X,1}, \ldots, \mathbf{\Sigma}^{(0)}_{X,k}, \ldots, \mathbf{\Sigma}^{(0)}_{X,K}, \right. \\ &\quad \left. \mathbf{\Sigma}^{(1)}_{X,1}, \ldots, \mathbf{\Sigma}^{(1)}_{X,k}, \ldots, \mathbf{\Sigma}^{(1)}_{X,K}\right) \end{aligned} \tag{22}$$

where N is the number of frame.

Here, the mean vector $\mu^{(s)}_{X_H,k}$ and the diagonal covariance matrix $\mathbf{\Sigma}^{(s)}_{X_H,k}$ of GMMs for $\mathbf{X_H}(i)$ are given by Eq.(17) and Eq.(18). In Eq.(17) and Eq.(18), $\mu^{(s)}_{X,k}$ and $\mathbf{\Sigma}^{(s)}_{X,k}$ which are given by Eq.(6) and Eq.(7) are the prior known parameters to the EM algorithm based estimation. Furthermore, the mixture weight $P^{(s)}_k$ is also a prior known parameter. From these facts, the unknown model parameter to the EM algorithm based estimation is only μ_H.

If complete data and incomplete data are defined, the expectation of auxiliary function $Q(\mathbf{x}, \mu_H, \hat{\mu}_H)$ is given as follows:

$$Q(\mathbf{x}, \mu_H, \hat{\mu}_H) = \int \log p(\mathbf{x}, \mathbf{y}|\lambda)\, p\left(\mathbf{y}|\mathbf{x}, \hat{\lambda}\right) d\mathbf{y} \tag{23}$$

where $\hat{\mu}_H$ is the estimate of μ_H.

Finally, $\hat{\mu}_H$ is estimated as the parameter which maximizes the above auxiliary function with iteration of the E-step(expectation) and the M-step(maximization) described bellow.

(E-step)

At first, by using the estimate $\hat{\mu}^{(l)}_H$ at the lth iteration, the speech GMM and the silence GMM defined in Eq.(16) are composed. Then, the speech sections and the silence sections of the observed signal are discriminated by comparing the output probability from each composed GMM as follows:

$$id(i,l) = \begin{cases} 0 & if\ \ p^{(0)}(\mathbf{X_H}(i)) > p^{(1)}(\mathbf{X_H}(i)) \\ 1 & if\ \ p^{(0)}(\mathbf{X_H}(i)) \leq p^{(1)}(\mathbf{X_H}(i)) \end{cases} \tag{24}$$

where if $id(i,l) = 0$, the ith frame is identified as the silence section. Otherwise, it is identified as the speech section.

Furthermore, to cope with time varying noise, the noise mean vector μ_N is updated as follows:

$$\mu_N(i) = \begin{cases} \rho\mu_N(i-1) + (1-\rho)\mathbf{X_H}(i) & if\ \ id(i,l) = 0 \\ \mu_N(i-1) & if\ \ id(i,l) = 1 \end{cases} \tag{25}$$

where ρ was set to 0.97 in this study. The initial value $\mu_N(0)$ is given by

$$\mu_N(0) = \frac{1}{10}\sum_{i=0}^{9} \mathbf{X_H}(i). \tag{26}$$

Finally, the expectation of auxiliary function $Q(\mathbf{x}, \mu_H, \hat{\mu}_H)$ is given as follows:

$$\begin{aligned} Q\left(\mathbf{x}, \mu_H, \hat{\mu}^{(l)}_H\right) &= \sum_i^{N-1}\sum_k^K P^{(l),(id(i,l))}_{i,k} \\ &\times \left(\log P^{(id(i,l))}_k + \log p_k\left(\mathbf{X_H}(i)|\mu^{(id(i,l))}_{X,k}, \Sigma^{(id(i,l))}_{X,k}, \mu_H\right)\right) \end{aligned} \tag{27}$$

$$\begin{aligned} &P^{(l),(id(i,l))}_{i,k} = \\ &\frac{P^{(id(i,l))}_k p_k\left(\mathbf{X_H}(i)|\mu^{(id(i,l))}_{X,k}, \Sigma^{(id(i,l))}_{X,k}, \hat{\mu}^{(l)}_H\right)}{\sum_{k'=1}^K P^{(id(i,l))}_{k'} p_{k'}\left(\mathbf{X_H}(i)|\mu^{(id(i,l))}_{X,k'}, \Sigma^{(id(i,l))}_{X,k'}, \hat{\mu}^{(l)}_H\right)} \end{aligned} \tag{28}$$

$$\begin{aligned} &p_k\left(\mathbf{X_H}(i)|\mu^{(id(i,l))}_{X,k}, \Sigma^{(id(i,l))}_{X,k}, \mu_H\right) = \\ &\mathcal{N}\left(\mathbf{X_H}(i), \mu^{(id(i,l))}_{X,k} + \mu_H, \mathbf{\Sigma}^{(id(i,l))}_{X,k}\right). \end{aligned} \tag{29}$$

(M-step)

By solving $\partial Q\left(\mathbf{x}, \mu_H, \hat{\mu}^{(l)}_H\right)/\partial\mu_H = 0$, μ_H is obtained as the parameter which maximizes the expectation of $Q\left(\mathbf{x}, \mu_H, \hat{\mu}^{(l)}_H\right)$.

$$\begin{aligned} &\frac{\partial Q\left(\mathbf{x}, \mu_H, \hat{\mu}^{(l)}_H\right)}{\partial\mu_H} = \sum_i^{N-1}\sum_k^K P^{(l),(id(i,l))}_{i,k} \\ &\times \frac{\partial \log p_k\left(\mathbf{X_H}(i)|\mu^{(id(i,l))}_{X,k}, \Sigma^{(id(i,l))}_{X,k}, \mu_H\right)}{\partial\mu_H} = 0 \end{aligned} \tag{30}$$

$$\begin{aligned} \mu_H &= \left(\sum_i^{N-1}\sum_k^K P^{(l),(id(i,l))}_{i,k}\left(\Sigma^{(id(i,l))}_{X,k}\right)^{-1}\right)^{-1} \\ &\times \sum_i^{N-1}\sum_k^K P^{(l),(id(i,l))}_{i,k}\left(\Sigma^{(id(i,l))}_{X,k}\right)^{-1} \\ &\times \left(\mathbf{X_H}(i) - \mu^{(id(i,l))}_{X,k}\right) \end{aligned} \tag{31}$$

By iterating the E-step and the M-step until convergence, the optimum estimation $\hat{\mu}_H$ is obtained. The convergence condition in this study is $\left|Q\left(\mathbf{x}, \mu_H, \hat{\mu}^{(l)}_H\right) - Q\left(\mathbf{x}, \mu_H, \hat{\mu}^{(l-1)}_H\right)\right| \leq 0.001$ and the initial value is $\hat{\mu}^{(0)}_H = \mathbf{0}$.

3.4. Estimation of $\mathbf{S}(i)$

In this section, the estimation method of $\mathbf{S}(i)$ by using the speech GMM and the silence GMM is described.

At first, the speech sections and the silence sections of the observed signal is discriminated based on Eq.(24) by using the composed GMMs with the estimated $\hat{\mu}_H$. Then, by using the discrimination results, the mismatch factor $\mathbf{G}(i)$ is estimated as follows:

$$\hat{\mathbf{G}}(i) = \sum_{k=1}^K P^{(id(i))}_{i,k}\mu^{(id(i))}_{G,k} \tag{32}$$

$$P^{(id(i))}_{i,k} = \frac{P^{(id(i))}_k p_k\left(\mathbf{X_H}(i)|\mu^{(id(i))}_{X,k} + \hat{\mu}_H, \mathbf{\Sigma}^{(id(i))}_{X,k}\right)}{\sum_{k'=1}^K P^{(id(i))}_{k'} p_{k'}\left(\mathbf{X_H}(i)|\mu^{(id(i))}_{X,k'} + \hat{\mu}_H, \mathbf{\Sigma}^{(id(i))}_{X,k'}\right)} \tag{33}$$

where $id(i)$ is discrimination result.

Finally, the estimate of $\mathbf{S}(i)$ is given by

$$\hat{\mathbf{S}}(i) = \mathbf{X_H}(i) - \hat{\mathbf{G}}(i) - \hat{\mu}_H. \quad (34)$$

4. EXPERIMENTS

4.1. Experimental set up

The proposed method has been evaluated on the AURORA3 database which is provided by ELRA[9]. The AURORA3 database contains the speech data(digit speech) spoken in 4 European languages which are Danish, Finnish, German and Spanish. The speech data of the AURORA3 is recorded in the several real driving car environments with a close-talking(CT) microphone and a hands-free(HF) microphone. The driving conditions are quiet(idling), low speed and high speed condition.

The 3 test conditions are defined in the AURORA3 database; namely, well-matched(WM), moderate-mismatched(MM) and high-mismatched(HM) condition. Table 1 and 2 show the speech data used for training and testing(◯ indicates the used data).

Table 1. The used training data

Condition	WM		MM		HM	
Microphone	CT	HF	CT	HF	CT	HF
Quiet	◯	◯		◯	◯	
Low speed	◯	◯		◯	◯	
High speed	◯	◯			◯	

Table 2. The used test data

Condition	WM		MM		HM	
Microphone	CT	HF	CT	HF	CT	HF
Quiet	◯	◯				
Low speed	◯	◯				◯
High speed	◯	◯		◯		◯

The acoustic features used in this evaluations are composed of 39 parameters with 13 MFCCs(with zero-th MFCC) and their first and second order derivatives as shown in Table3. The zero-th MFCC was used for energy coefficient instead of standard Log-energy. The AURORA3 standard whole word HMMs(16states, 3 mixture distributions per state) are used for all the evaluations.

Table 3. Feature extraction conditions

Pre-emphasis	$1 - 0.97z^{-1}$
Feature parameter	13th order MFCC(with zero-th) + Δ + $\Delta\Delta$
Frame length	25ms
Frame shift	10ms
Window type	Hamming window

The clean speech GMMs are trained by using the training materials which are recorded in quiet condition with a close-talking microphone. The number of mixture distribution of the each GMM is 256.

4.2. Experimental results

In this study, we evaluated the following 4 methods on the HM condition of AURORA3 database, because the HM condition has both additive noise and channel condition mismatch between training data and test data.

Method 1 : Baseline at ICSLP2002[10]

Method 2 : Spectral subtraction + Cepstral mean subtraction

Method 3 : Segura's method + Cepstral mean subtraction

Method 4 : Proposed method

Table 4 and 5 show the recognition results by each method. Compared with the Method 3, the proposed method(Method 4) showed about 4.3% of word error rate improvement and about 9.7% of relative improvement in average. These results indicate that the proposed EM algorithm based channel noise estimation worked effectively, comparing with the conventionally used cepstral mean subtraction.

However, the proposed method assumed that the length of the impulse response of the channel noise is shorter than the analysis window size. If it is longer than the analysis window size such as a room reverberation, the performance of the proposed method will degrade. To solve this problem, the channel noise estimation method to be robust for the channel noise with long impulse response is required in future work.

Table 4. Word error rate(%)

Language	Danish	Finnish	German	Spanish	Average
Method 1	60.63	59.47	26.83	48.45	48.85
Method 2	59.01	43.92	20.44	37.68	40.26
Method 3	38.75	15.48	13.32	23.55	22.78
Method 4	**30.80**	**13.71**	**9.02**	**20.24**	**18.44**

Table 5. Relative improvement(%)

Language	Danish	Finnish	German	Spanish	Average
Method 1	0.00	0.00	0.00	0.00	0.00
Method 2	2.67	26.15	23.82	22.23	18.72
Method 3	36.09	73.97	50.35	51.39	52.95
Method 4	**49.20**	**76.95**	**66.38**	**58.22**	**62.69**

5. CONCLUSIONS

In this paper, we proposed a robust speech recognition method in additive and channel noise environments by using GMM and EM Algorithm. In the evaluation on the high-mismatched condition test of AURORA3, our method showed the significant improvement in word error rate and relative improvement. In future, we are planning to cope with the non-stationary additive noise and more accurate channel noise estimation method.

6. REFERENCES

[1] M.J.F.Gales and S.J.Young: "Robust Continuous Speech Recognition Using Parallel Model Combination", IEEE Trans. Speech and Audio Processing, Vol.4, No.5, pp.352-359(1996).

[2] Y.Gong: "A Comparative Study of Approximations for Parallel Model Combination of Static and Dynamic Parameters," ICSLP'02, Vol.III, pp.1209-1032(2002).

[3] K.Yao, B.E.Shi, P.Fung and Z.Cao: "Residual Noise Compensation for Robust Speech Recognition in Nonstationary Noise", ICASSP'00, Vol.II, pp.1125-1128(2000).

[4] C.L. Leggetter and P.C. Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol.9, pp.171-185(1995).

[5] S.F.Boll: "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoustic Speech Signal Processing, Vol.27, No.2, pp.113-120(1979).

[6] J.C.Segura, A.de la Torre, M.C.Benitez and A.M.Peinado: "Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks", EuroSpeech'01, Vol.I, pp.221-224(2001).

[7] A.Acero and R.M.Stern: "Environmental Robustness in Automatic Speech Recognition", ICASSP'90, pp.849-852(1990).

[8] P.J.Moreno, B.Raj and R.M.Stern: "A Vector Taylor Series Approach for Environment-Independent Speech Recognition", ICASSP'96, pp.733-736(1996).

[9] ELRA Web site: http://www.elra.info

[10] AURORA2,3 Spread sheet: http://icslp2002.colorado.edu/special_sessions/aurora/