Modified GMM Training for Inexact Observation and Its Application to Speaker Identification

Jin Young Kim* · So Hee Min* · Seung You Na*
Hong Sub Choi** · Seung Ho Choi***

ABSTRACT

All observation has uncertainty due to noise or channel characteristics. This uncertainty should be counted in the modeling of observation. In this paper we propose a modified optimization object function of a GMM training considering inexact observation. The object function is modified by introducing the concept of observation confidence as a weighting factor of probabilities. The optimization of the proposed criterion is solved using a common EM algorithm. To verify the proposed method we apply it to the speaker recognition domain. The experimental results of text-independent speaker identification with VidTimit DB show that the error rate is reduced from 14.8% to 11.7% by the modified GMM training.

Keywords: GMM, speaker identification, optimization

I. Introduction

Observation of signals is often corrupted or distorted by system noise, external noise, and channel characteristics. To cope with corrupted observation problems, there are two approaches. One is an enhancement technique in signal or feature spaces [1-4]. The other system is model adaptation to the system environments [5, 6]. Especially, in speaker recognition, the first method is preferred as in cepstrum mean subtraction (CMS) [2] and RASTA-filtering [1]. Of course, there has been some research applying model adaptation.

On the other hand, Gaussian mixture model (GMM) is widely used in signal processing and pattern recognition problems [7], for the model is very simple and is successfully operated in many subject domains such as speaker recognition, image segmentation, medical diagnosis and so on [8-11]. According to the previous research, confidence of observation is not considered in a GMM training aspect. The problems of noise or distortion have been dealt with in pre-processing or post-processing.

In this paper we propose a modified GMM training by considering observation confidence. Under the assumption that confi-

* Dept. of Electronics and Computer Eng., Chonnam National Univeristy
** Dept. of Electronics Eng., Daejin University
*** Dept. of Multimedia Eng., Dongshin University

dence values of observations are given, we suggest a modified optimization criterion for GMM training by introducing a confidence factor of observations. The modified optimization problem can be solved by a general EM algorithm.

To verify our method we apply it to speaker recognition. In speaker recognition problems observation confidence can be defined as a function of signal-to-noise ratio (SNR). We evaluate our proposed method using the VidTimit database in the speaker identification domain.

2. Modified GMM Criterion and Training

In this section we propose a modified criterion for GMM optimization problems. Then we deduce a training algorithm based on an EM algorithm such as common GMM training.

2.1 Modified GMM Criterion

<Figure 1> is a general model of signal generation. As shown in <Figure 1>, an observed signal is corrupted and distorted by additive noise and the transmission channel. Thus it is impossible to measure the output exactly.

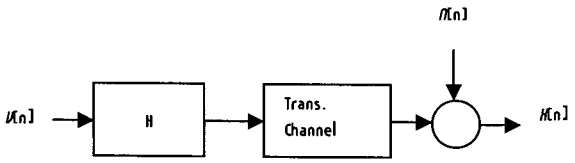


Figure 1. General model of signal generation

There are many techniques for modeling system H. One of them is the Gaussian mixture model (GMM), which is widely used in speech processing, image processing, and any other pattern recognition problems. GMM is to model a system with Gaussian mixtures as Eq.1.

$$p_x(X) = \sum_{k=1}^{C} p(k \mid \theta) p(X \mid k, \theta) = \sum_{k} \alpha_k p_G(\mathrm{M}_k, \Sigma_k), \quad \text{(Eq.1)}$$

where $\sum_{k=1}^{C} p(k \mid \theta) = \sum_{k=1}^{C} \alpha_k = 1$, C is the number of components, $p_G(\cdot)$ is the Gaussian probability density function, α_k, Σ_k and M_k are the weighting factor, covariance matrix and mean vector of k-th component respectively. If we have an observation sequence of n observations, the object function for calculating GMM parameters is given by Eq. 2.

$$O(\alpha_k, M_k, \Sigma_k) = \prod_{n=1}^{N} p_x(X_n) \quad \text{(Eq.2)}$$

Then the objection function is optimized with respect to α_k, M_k and Σ_k for a sequence $\{X_n\}$ given. In other words, GMM problem is

$$\underset{w_k, \mathrm{M}_k, \Sigma_k}{Max} \prod_{n} p_x(X_n).$$

In the object function of Eq. 1, all the observation vectors are evenly treated. The contribution of each observation vector is equal. This means that Eq.1 does not carry noise corruption. Or it assumes that there is no corruption or distortion. We think, however, each vector should be treated differently, for the corruption rate by noise is not fixed to all the observation vectors. For example, the segmental SNR of the speech signal may vary in each frame.

Until now, especially in speech processing, most researchers have struggled to acquire clean speech in a training aspect. They have also tried to develop noise rejection methods and model adaptation techniques. However, we cannot completely eliminate the corruption effect from observations, although we adopt all the possible approaches. Thus it is reasonable to introduce observation confidence in a training stage by modifying the object function. Let's consider the system of <Figure 2>, modified from the system shown in <Figure 1>. In <Figure 2> the system of Hinderer disturbs $H_{T,n}$ into H_n so that a true observation can not be measured. Then we can define observation confidence as how much H_n is close to $H_{T,n}$. d_n is the disturbance quantity to $H_{T,n}$ at time n.

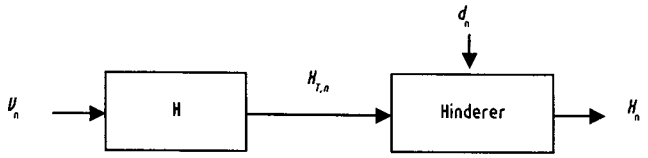


Figure 2. Hinderer as an obstacle making exact observance impossible

Let's assume we have an appropriate confidence value of each observation vector. And let it be $\{m_n\}$. That is, m_n is a confidence value of the n-th observation or membership value representing the possibility of true observation. m_n has the value of between 0 and 1. It may be 1- d_n. Now we have an observation pair of $\{X_n, m_n\}$. Then we can modify the GMM object function as Eq. 3 considering observance confidence.

$$O_M(\alpha_k, M_k, \Sigma_k) = \prod_{n=1}^{N} \left(p_x(X_n)\right)^{m_n} \quad \text{(Eq.3)}$$

According to Eq. 3 the contribution of an observation having low confidence gets depressed as $\left(p_x(X_n)\right)^{m_n}$ becomes closer to 1. Now the GMM training problem is represented as

$$\underset{\alpha_k, M_k, \Sigma_k}{Max} \prod_n \left(p_x(X_n)\right)^{m_n}.$$

The maximization problem above can be solved iteratively using EM algorithm. In the next section we describe the solution of the modified GMM training.

2.2 Optimization Algorithm based on EM algorithm

An EM algorithm is an iterative optimization method to estimate some unknown parameters Θ, given measurement data. However, there are some hidden nuisance variables, which need to be integrated out.

In our problem, the pseudo-likelihood function is defined as

$$L(X;\Theta) = \prod_{n=1}^{N} p_x^{m_n}(X_n;\Theta).$$

Now, introduce the nuisance variable of binary vector $Y_n = \{y_{n,1}, \ldots, y_{n,C}\}$, where $y_{n,c} = 1$, if the sample was produced by the c-th component of C Gaussians. Then the likelihood function can be re-written as

$$L(X,Y;\Theta) = \prod_{n=1}^{N} \left(\prod_{k=1}^{C} \left(\alpha_k p(X_n \mid k, \Theta)\right)^{y_{n,k}} \right)^{m_n} = \prod_{n=1}^{N} \prod_{k=1}^{C} \alpha_k^{y_{n,k} m_n} p(X_n \mid k, \Theta)^{y_{n,k} m_n} \quad \text{(Eq.4)}$$

The log likelihood of Eq. 4 is described as Eq. 5.

$$l(X,Y;\Theta)=\sum_{n=1}^{N}m_n\sum_{k=1}^{C}y_{n,k}\ln\alpha_k+\sum_{n=1}^{N}m_n\sum_{k=1}^{C}y_{n,k}\ln p_x(X_n\mid k,\Theta) \quad \text{(Eq.5)}$$

Let us now write the corresponding auxiliary function as

$$\begin{aligned}Q(\Theta,\Theta^i)&=E_Y\left[\ln L(X,Y\mid\Theta)\mid X,\Theta^i\right]\\&=\sum_{n=1}^{N}m_n\left(\sum_{k=1}^{C}E_Y[y_{n,k}\mid X,\Theta^i]\ln\alpha_k+E_Y[y_{n,k}\mid X,\Theta^i]\ln p_x(X_n\mid k,\Theta)\right).\end{aligned} \quad \text{(Eq.6)}$$

Hence the E-step is to compute the conditional expectation of the complete data log-likelihood, Q-function, given data X and the current estimate Θ^i. That is,

$$E_Y[y_{n,k}\mid X,\Theta^i]=p[y_{n,k}=1\mid x_n,\Theta^i].$$

And the M-step finds the parameters Θ that maximize Q. The derivation process is similar to the conventional GMM optimization process. That is, parameter set, Θ, can be determined by searching for

$$\frac{\partial Q}{\partial\Theta}=0$$

for each parameter ($\alpha_k,\mathrm{M}_k,\Sigma_k$) under the condition of $\sum_{k=1}^{C}\alpha_k=1$. Because the derivation is very similar to the common GMM problem, we skip the derivation details. The EM solution for our proposed criterion is shown in <Table 1>. As shown in <Table 1>, the optimization algorithm is very similar to that of the original GMM method. If confidence of each observation is 1, the algorithm is exactly the same as the conventional GMM training.

Table 1. EM algorithm for GMM training with confidence factor.

E-step	$w_{n,k}$	$w_{n,k}=\dfrac{\alpha_k^i p(X_n\mid k,\Theta^i)}{\sum_{j=1}^{C}\alpha_j^i p(X_n\mid j,\Theta^i)}$ (Eq.7)
M-step	M_k	$M_k^{i+1}=\dfrac{\sum_{n=1}^{N}X_n m_n w_{n,k}}{\sum_{n=1}^{N}m_n w_{n,k}}$ (Eq.8)

	Σ_k	$\Sigma_k^{i+1} = \frac{\sum_{n=1}^{N} m_n w_{n,k} \left(X_n - M_k^{i+1}\right)\left(X_n - M_k^{i+1}\right)^T}{\sum_{n=1}^{N} m_n w_{n,k}}$ (Eq.9)
	α_k	$\alpha_k^{i+1} = \frac{\sum_{n=1}^{N} m_n w_{n,k}}{\sum_{k=1}^{C}\sum_{n=1}^{N} m_n w_{n,k}}$ (Eq.10)

3. Speaker Identification Based on Modified GMM Training

One of the difficult problems in speech processing is to cope with noisy speech, especially if the performance of speech and speaker recognition is highly degraded in noisy environment. Thus, the most important thing is to acquire as clean a speech as possible for training and real services. It is relatively easy to get clean speech in a training stage compared with a real service situation. However, as mobile service and broadcast applications grow rapidly, it is hard to get clean speech even in a training stage [12, 13].

In this paper we apply the modified GMM training to speaker identification in order to verify the proposed method. The flowchart of our method is shown in <Figure 3>. In the figure, solid lines represent a conventional baseline system and dashed lines show our modules added to the conventional approach. In sections 3.1 and 3.2 we describe the baseline system and our proposed system of speaker identification.

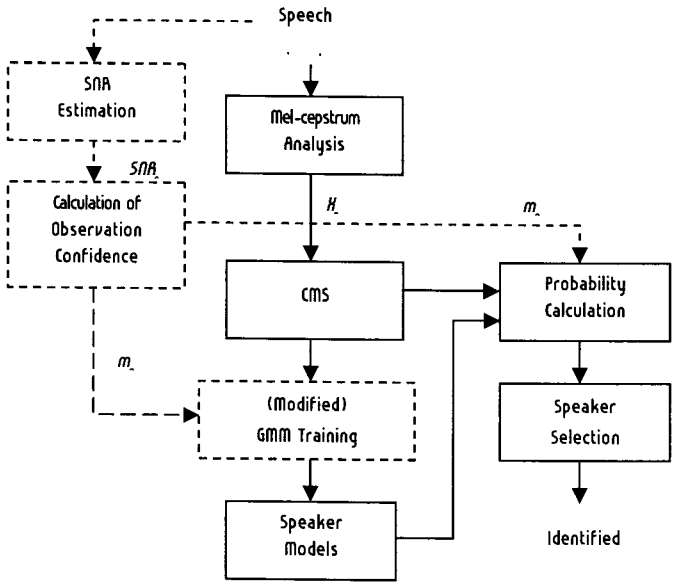


Figure 3. Proposed system for speaker identification

3.1 Baseline Speaker Identification System

The baseline system is composed of common speaker identification processes. In the flowchart of the baseline system (solid lines in <Figure 3>), we apply a mel-cepstrum analysis as a feature extraction method. Mel-cepstrum is one of the most successful features used in speech and speaker recognition. Then the cepstrum mean subtraction (CMS) is adopted to reject the effects of additive noise and the transmission channel. CMS is used to subtract the mean vectors from the feature matrix of a given utterance. So CMS can reject the stationary noise with a computation burden. GMM is used for modeling the speaker's voice. The training algorithm is a common EM based iterative algorithm.

The identification process is very simple. After calculating all the probabilities of each speaker, the speaker having the maximum probability is selected as the identified speaker.

3.2 Modified GMM Training for Speaker Identification

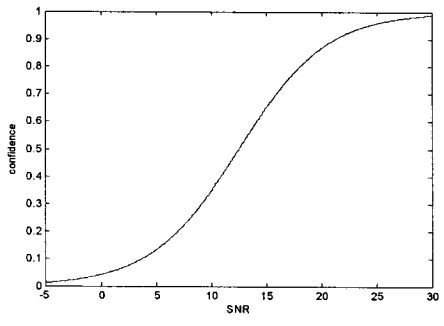
As explained in section 2, a modified GMM training can be applied only when observation confidence is known. That is, we have to measure an observation pair of $\{X_n, m_n\}$. In speaker recognition problems, the only information which can be estimated is signal-to-noise raio (SNR). SNR is easily estimated when additive noise is stable. So, in this paper, we regard noise as the only hinderer of observation. In other words, the observation confidence is a function of SNR; $m_n = f(SNR_n)$. This idea is very reasonable. With noisy training utterances we can easily show that features having high SNR are more discriminative in speaker identification. This is discussed in section 4.

Now, the problem is to we devise the function $f()$. We adopt a simple sigmoid function as the transfomation function from SNR to observation confidence.

$$m_n = \frac{1}{1+e^{-a(SNR_n - b)}} \qquad \text{(Eq.11)}$$

where a and b are control parameters and determined heuristically. <Figure 4> shows an example of the transformation function. As shown in <Figure 4>, SNR is transformed into observation confidence having the value between 0 and 1.

Figure 4. Example of observation confidence (a= -0.25 and b= 12.5)



Then we can apply the modified GMM training with the observance of $\{ X_n, m_n \}$. In a testing stage the function of observation confidence is similarly used for weighting frame level probabilities as shown below;

$$L(X \mid \Theta_p) = \prod_{n=1}^{N} p_x^{m_n}(X_n \mid \Theta_p),$$

where Θ_p is p-th person's model. And the identified person is determined by

$$P^* = \arg\max_p L(X \mid \Theta_p).$$

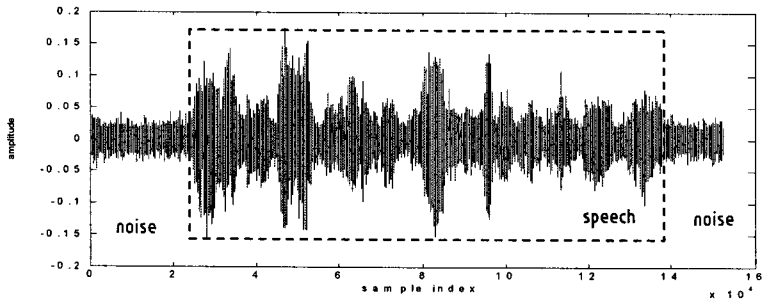
4. Experimental Results and Discussion

We performed experiments of speaker identification with the VidTimit database. The VidTimit database consists of video and corresponding audio recordings of 43 people (19 females and 24 males), reciting short sentences selected from the NTIMIT corpus. The data were recorded in three sessions, with a mean delay of seven days between Sessions 1 and 2, and of six days between Sessions 2 and 3. We tested the proposed method to all the utterances in VidTimit database. <Table 2> shows the specs of our experiments.

Table 2. Specifications of speaker identification experiments

Number of person	43
Number of training sentence per person	7
Number of test sentence per person	3
Sampling & quantization	32 kHz sampling 16 bit quantization
Speech feature	17 mel-cepstrum and energy
Number of GMM components	10
GMM training	EM based iterative training - full covariance - initialization : fuzzy c-means clustering
Speaker identication	Text-independent

The audio quality of the VidTimit database is not good. Utterances were recorded in the real office environment, so the SNR of each utterance is low. <Figure 5> shows the waveform of a sample file. As observed in <Figure 5>, speech is highly corrupted by



noise. The average SNR of the utterances is about 13.8 dB, and the standard deviation is 3.35 dB.

Figure 5. Waveform example of VidTimit utterance.

On the other hand, in section 3.2 we propose observation confidence as a function of SNR. This means that acoustic features with a high SNR have high distinctivity. High distinctivity implies that the feature is well observed in some sense. The distinctivity is defined by

$$Distinctivity(X) = \frac{p_{p\_true}(X)}{\sum_{p=1}^{N_{speaker}} p_p(X)}$$

where $p_p(X)$ is the probability of p-th speaker, $p_{p\_true}(X)$ is the probability of the true speaker and N_{speaker} is the number of speakers. Fig. 6 shows the correlation values between SNR and distinctivity. The average correlation is 0.32. This fact means that the SNR is correlated with distinctivity. So, using SNR for the calculation of observation confidence is reasonable. In this paper, as explained in section 3.2, sigmoid function is adopted as a transformation from SNR to observation confidence.

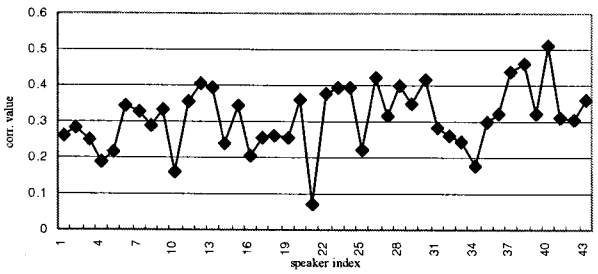


Figure 6. Correlation values between SNR and distinctivity.

In our experiments the control parameter of a is set to -0.25 and the b parameter is 13 dB. Fig. 7 shows identification performance. With the common GMM training, the identification rate is 85.3 %. When the modified GMM training was applied, the rate was increased by as much as 3 % into 88.3 %. This result shows that the proposed approach is one of the methods for overcoming a noise problem.

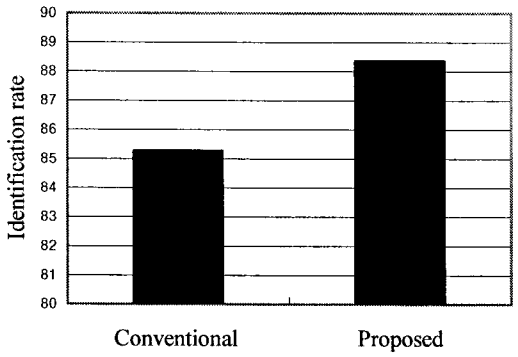


Figure 7. Comparison of Speaker identification performances

5. Conclusion

In this study we proposed a modified GMM training for modeling inexact observations. For this we modified the object function of the optimization, which was optimized with an iterative EM algorithm. The method was verified in the application domain of text-independent speaker identification with the VidTimt database.

The proposed algorithm is so general that it can be applied to any problem having inexact observation with observation confidence. That is, it could be adopted in any pattern recognition such as speech recognition, automatic lip reading, face recognition and so on.

In the future we will adopt our proposed method for automatic lip reading and multi-modal speaker recognition. Further research needs to be conducted to solve how to decide observation confidence. It is problem specific, so it is necessary to devise a proper measure of observation confidence for each given application domain.

Acknowledgement

This paper is supported by the 2005 sabbatical grants of Chonnam National University and Daejin University.

References

[1] Zhen Bin, Wu Kihong, Liu Zhimin, CHI Huisheng. 2000. "An Enhanced RASTA processing for speaker identification." *Proc of 2000 ICSLP*, 251-254.

[2] Rosenberg, A. et al. 1994. "Cepstral channel normalization techniques for HMM-based speaker verification." *Proc. ICSLP-94*, 1835-1838.

[3] Mammone, R. J., Zhang, X. & Ramachandran, R. P. 1996. "Robust Speaker Recognition, A feature-based Approach." *IEEE Signal Processing Magazine* 13(5), 58-71.

[4] Stephane Dupont & Christophe Ris. 2003. "Robust feature extraction and acoustic modeling at Multitel : experiments on the Aurora databases." *Proc of EuroSpeech-2003*, 1789-1792.

[5] Mengusoglu, E. 2003. "Confidence Measure based Model Adaptation for Speaker Verification." *Proc. of the 2nd IASTED International Conference on Communications, Internet and Information Technology.*

[6] Chin-Hung Sit, Man-Wai Mak, & Sun-Yuan Kung. 2004. "Maximum Likelihood and Maximum A Posteriori Adaptation for Distributed Speaker Recognition Systems." *Proc of 1st Int. Conf. on Biometric Authentication.*

[7] Geoffrey McLachlan, David Peel: Finite Mixture Models. 2000. "Wiley Series in Probability and Statistics".

[8] Nakagawa, S. & Zhang, W. 2003. "Text-independent speaker recognition by speaker-specific GMM and speaker adapted syllable-based HMM." *Proc. Eurospeech*, 3017_3020.

[9] Shiri Gordon, Gali Zimmerman, Hayit Greenspan. 2004. "Image Segmentation of Uterine Cervix Images for Indexing in PACS". *Proc. of 17th IEEE Symposium on Computer-Based Medical System*, pp. 298-301.

[10] 2006. "IMAGE SEGMENTATION USING GAUSSIAN MIXTURE MODELS." *Proc. of Twenty sixth International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering.*

[11] David Gibson, Neill Campbell, Barry Thomas. 2002. "Visual Abstraction of Wildlife Footage Using Gaussian Mixture Models and the Minimum Description Length Criterion." *Proc. of 16th International Conference on Pattern Recognition (ICPR'02)* 2, 20814-20817.

[12] Alberto Albiol, Luis Torres. 2003. "The Indexing Of Persons In News Sequences Using Audio-Visual Data." *Proc. of ICASSP-03.*

[13] Woo, R. H. 2005. "Exploration of Small Enrollment Speaker Verification on Handheld Devices." M.S. thesis of MIT.

received: January 28, 2007
accepted: March 9, 2007

▲ Jin Young Kim, So Hee Min, Seung You Na
Dept. of Electronics and Computer Eng., Chonnam National Univeristy
300 Youngbong-Dong, Buk-Gu, Gwangju, 500-757, South Korea
{beyondi, minsh, syna}@chonnam.ac.kr

▲ Hong Sub Choi
Dept. of Electronics Eng., Daejin University

Pocheon, Gyunggi-Do 487-711, South Korea

hschoi@daejin.ac.kr

▲ Seung Ho Choi

Dept. of Multimedia Eng., Dongshin University

Naju, Jollanam-Do, 520-714, South Korea

shchoi@dongshinu.ac.kr