ROBUST INCREMENTAL ADAPTATION OF GMM IN SPEAKER RECOGNITION

JaeYeol Rheem [1], JongJoo Lee [2], and KiYong Lee [2]

[1]School of Information Technology, Korea University of Technology and Education, Korea
[2]School of Electronic Engineering, Soongsil University, Korea

ABSTRACT: Speaker model in speaker recognition system is to be trained from a large data set uttered in multiple sessions. The large data set requires larger amount of memory and computation, and practically it's hard to make users utter a large amount of data in several sessions. Recently proposed incremental adaptation methods cover the problems. However, the data set uttered from multiple sessions is vulnerable to outliers from irregular utterance variation and presence of noise, which results in inaccurate speaker model. In this paper, we propose a robust incremental adaptation method to minimize the influence of outliers on speaker model using Gaussian Mixture Model. The robust adaptation is obtained from an incremental version of M-estimation. Speaker model is initially trained from small amount of data and it is adapted recursively according to new data available. Experimental results from the data set gathered over seven months show that the proposed method is robust against outliers.

INTRODUCTION

Speaker model in speaker recognition system is to be trained using a large data set uttered in multiple sessions(S. Furui, 1981). The large data set requires huge amount of memory and calculation for training the speaker model. In practical system, it is impossible to make users utter a large amount of data in several sessions. Recently, the incremental adaptation methods are proposed to cover the problems. In the incremental adaptation methods, speaker model is initially trained from small amount of data uttered usually in one session and then it is updated incrementally with the new data from the different sessions. However, when the data set uttered from session to session contains outliers, the speaker model obtained from conventional method becomes inaccurate. The outliers can occur from irregular utterance variation and presence of noise.

In this paper, we proposed a robust incremental adaptation method to minimize the influence of outliers on the accuracy of speaker model using GMM (Gaussian Mixture Model) (Reynolds and Rose, 1995). The robust adaptation method for GMM is obtained from an incremental version of M-estimation. The initial speaker model is trained from small amount of data. Whenever new data is available, the parameters of GMM are adapted recursively. Simulation results indicate that the proposed method is robust against outliers.

ROBUST GMM USING M-ESTIMATION

Let $Y^N = \{Y_n, n = 1, \cdots, N\}$, $Y_n = \{y_n(t), t = 1, \cdots, T_n\}$, be a set of *N* training speech sequences with length T_n. Let $y_n(t) \in R^L$ be an *L*-dimensional vector. When outlier exists in Y^N, the parameter estimation procedures of the conventional GMM share the problem of high sensitivity to outliers. Therefore, in order to obtain reliable estimates of GMM, we now present a robust estimation method based on the M-estimation method as

$$J = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \rho\left[\log p\left(y_n(t)|\theta\right)\right] \quad (1)$$

where $\rho[\cdot]$ is a loss function, which is to reduce the effect of outliers. In (1), a Gaussian mixture density $p(y_n(t)|\theta)$ is a weighted sum of M multivariate Gaussian functions

$$p(y_n(t)|\theta) = \sum_{i=1}^{M} p_i b_i(y_n(t)) \tag{2}$$

where $b_i(y_n(t)) = \frac{1}{(2\pi)^{\frac{L}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(y_n(t)-\mu_i)^T \Sigma_i^{-1}(y_n(t)-\mu_i)\right\}$ with mean μ_i and variance matrix Σ_i.

The robust GMM for speaker model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are represented by the notation

$\theta = \{p_i, \mu_i, \Sigma_i\}, i = 1, \cdots, M$.

We can find the re-estimation formulas for θ by minimizing *J* of (1) with respect to each of $p_i, \mu_i,$ and Σ_i, $i = 1, \cdots, M,$ respectively. By $\frac{\partial J}{\partial p_i} = 0$, $\frac{\partial J}{\partial \mu_i} = 0$, and $\frac{\partial J}{\partial \Sigma_i} = 0$, the following re-estimation formulas for robust GMM are obtained:

- Mixture Weights, $$p_i^{\,N} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t) p(i|y_n(t),\theta)}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t)} \tag{3.a}$$

- Means, $$\mu_i^{\,N} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t) p(i|y_n(t),\theta) y_n(t)}{\sum_{n=1}^{N}\sum_{t=1}^{T} w_n(t) p(i|y_n(t),\theta)} \tag{3.b}$$

- Variances, $$\Sigma_i^{\,N} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t) p(i|y_n(t),\theta)(y_n(t)-\mu_i)(y_n(t)-\mu_i)^T}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} w_n(t) p(i|y_n(t),\theta)} \tag{3.c}$$

where $p(i|y_n(t),\theta)$ is the *a posteriori probability* for acoustic class *i* as $p(i|y_n(t),\theta) = \frac{p_i b_i(y_n(t))}{\sum_{j=1}^{M} p_j b_j(y_n(t))}$ and $w_n(t)$ is a weight function defined as $w_n(t) = \frac{\partial \rho[z_n(t)]}{\partial z_n(t)}$, where $z_n(t) = \log p(y_n(t)|\theta)$.

We used the Cauchy's weight function given by $w_n(t) = 1/(1 + z_n(t)/\beta)$, where β is the scale parameter. Since the data with a large $z_n(t)$ have a small $w_n(t)$, the influence of outlier can be reduced in (3).

ROBUST INCREMENTAL ADAPTATION FOR GMM

If a model parameter θ^N is already trained with an initial set of speech data Y^N and new data $Y_{N+1} = \{y_{N+1}(1), \cdots, y_{N+1}(T_{N+1})\}$ is given, the (*N*+1)-th recursive re-estimation equations can be easily obtained from (3) as:

- Mixture Weights, $$p_i^{N+1} = \frac{p_i^{N+1} W(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p\left(i \middle| y_{N+1}(t), \theta^N\right)}{W(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t)} \quad (4.a)$$

- Means, $$\mu_i^{N+1} = \frac{\mu_i^N W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p\left(i \middle| y_{N+1}(t), \theta^N\right) y_{N+1}(t)}{W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p\left(i \middle| y_{N+1}(t), \theta^N\right)} \quad (4.b)$$

- Variances, $$\Sigma_i^{N+1} = \frac{\Sigma_i^N W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p\left(i \middle| y_{N+1}(t), \theta^N\right)\left(y_{N+1}(t) - \mu_i^N\right)\left(y_{N+1}(t) - \mu_i^N\right)^T}{W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p\left(i \middle| y_{N+1}(t), \theta^N\right)} \quad (4.c)$$

where $W(N) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} w_n(t)$ and $W_p(N) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} w_n(t) p\left(i \middle| y_n(t), \theta^N\right)$.

When $Y_{N+1} = \{y_{N+1}(1), \cdots, y_{N+1}(T_{N+1})\}$ contain outliers, the influence of outlier in (4) is reduced by a small $w_{N+1}(t)$. Given the Y_{N+2} data set, $W(N+1)$ and $W_p(N+1)$ in (4) can be obtained recursively, as

$$W(N+1) = W(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) \quad (5.a)$$

$$W_p(N+1) = W_p(N) + \sum_{t=1}^{T_{N+1}} w_{N+1}(t) p\left(i \middle| y_{N+1}(t), \theta^{N+1}\right) . \quad (5.b)$$

EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, text-dependent speaker verification experiments have been performed using both the proposed and the conventional GMM based methods. The database consisted of the sentences from ten male and ten female speakers: five male and five female speakers served as customers and the others served as impostors. The utterances were recorded in seven sessions ($T_0 - T_6$) over seven months. In each session, the sentences are uttered five times by each speaker. For analysis, 15 LPC cepstral coefficients with log-energy and their first derivatives were used. The initial speaker model for adaptation is trained with five sentences from session T_0. Utterances from session T_1 to T_i were used to adapt the model and the utterances from session T_{i+1} to T_6 are used for testing. The speaker adaptation was performed in supervised mode where the data used to adapt a speaker model is certified as belonging to the correct speaker.

Figure 1 shows the EER (Equal Error Rate) of the proposed method, the conventional GMM with adaptation(C. Fredouille, 2000), and the GMM with no adaptation over sessions. At session T_i, speaker models with adaptation are adapted incrementally with one additional training session at a time and speaker model with no adaptation is trained using all the data from T_0 to T_i. The performance of proposed method is equal or good to conventional method. Although the required memory and computational load for proposed method are much smaller than for conventional method, the proposed method performed almost as well as the GMM with no adaptation. These results indicate that the proposed method is robust to session dependent utterance variations and effective for updating the speaker models incrementally.

Figure 2 shows the EER of each method when *K* outliers exist in the training and the testing data, where *K* is a proportion of outliers in the data. The outlier was generated by multivariate normal density function with zero mean vector and covariance matrix with five time unit variance. Then, when outliers exist, we see that EER of the conventional methods are seriously degraded, but the proposed method keeps its performance almost the same.

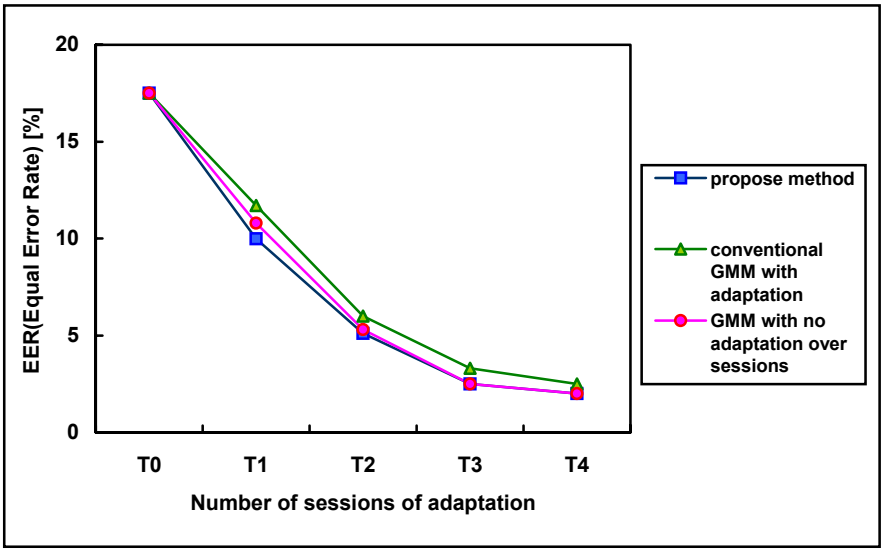

Figure 1. Equal Error Rate at *K*=0%

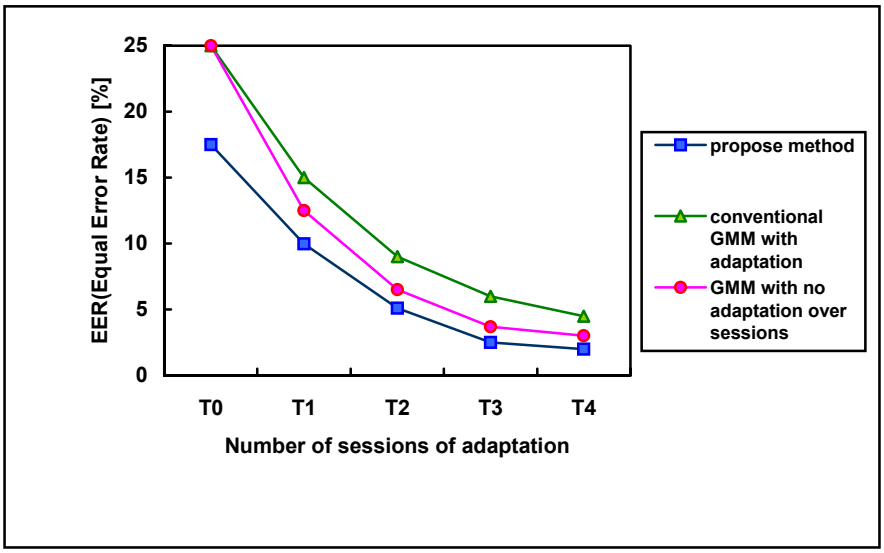

Figure 2. Equal Error Rate at *K*=5%

CONCLUSIONS

In this paper, we proposed a robust incremental adaptation method to minimize the influence of outliers on the accuracy of speaker model using GMM (Gaussian Mixture Model) (Reynolds and Rose, 1995). The performance of proposed method is equal or good to the conventional method. These results indicate that the proposed method is robust to session dependent utterance variations and effective for updating the speaker models incrementally. when outliers exist, we see that EER of the conventional methods are seriously degraded, but the proposed method keeps its performance almost the same.

REFERENCES

C. Fredouille and J. Mariethoz(2000), *Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification,* IEEE int. Conf, Acoustics, Speech, and Signal Processing vol.2 pp.1197-1200

D.A. Reynolds and R.C. Rose, (1995), *Robust text-independent speaker identification using Gaussian mixture speaker models*, IEEE Trans Speech Audio Process., vol.3, no.1, pp.72-83, Jan.

J.L. Gauvain and C.H. Lee, (1994), *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*, IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291-298, Mar.

S. Ahn and H. Ko, (2000), *Speaker adaptation in sparse training data for improved speaker verification*, Electronics Letters, vol. 36, n0.4, pp.371-373, Feb.

S. Furui(1981), *Cepstral analysis technique for automatic speaker verification*, IEEE Trans. ASSP-29, vol 2, pp.254-272

Y.S. Choi and R. Krishnapuram, (1996), *Fuzzy and robust formulations of maximum likelihood based Gaussian mixture decomposition*, IEEE Int. Conf. On Fuzzy Systems, pp.1899-1902,