On Speaker Adaptations with Sparse Training Data for Improved Speaker Verification*

Sungjoo Ahn, Sunmee Kang*** and Hanseok Ko****

ABSTRACT

This paper concerns effective speaker adaptation methods to solve the over-training problem in speaker verification, which frequently occurs when modeling a speaker with sparse training data. While various speaker adaptations have already been applied to speech recognition, these methods have not yet been formally considered in speaker verification. This paper proposes speaker adaptation methods using a combination of MAP and MLLR adaptations, which are successfully used in speech recognition, and applies to speaker verification. Experimental results show that the speaker verification system using a weighted MAP and MLLR adaptation outperforms that of the conventional speaker models without adaptation by a factor of up to 5 times. From these results, we show that the speaker adaptation method achieves significantly better performance even when only small training data is available for speaker verification.

Keywords: speaker verification, speaker adaptation, training

I. INTRODUCTION

Reliable speaker verification provides an added security measure to online services and is rapidly positioning itself as the key to success in the online service industry. As the first step to enable the speaker verification task in these online service systems, a new subscriber is required to register by enrolling his/her voice imprints and creating users personal speech model. Each prospective subscriber must make voice utterances to construct a personal speech model. In general, at least 8 to 10 utterances per speaker are required to make a faithful speaker model. However, in practical online services, it is desirable to contain the number of utterances required of a prospective subscriber to minimize inconveniences at initial enrollment. As a result, we are confronted with only a few enrollment data available for training. To avoid the cost of compromising user convenience, the desirable number of enrollment data

* This work was supported by a grant from the Interdisciplinary Research Program (Contract No. 1999-2-302-016-5) of the KOSEF.

** Dept. of Electronics Engineering, Korea University

*** Dept. of Computer Science, Seokyeong University

should be no more than 2 to 3 utterances for speaker verifications applied in electronic commerce or tele-banking service. The consequence of using small enrollment data when making speaker model is visibly noted that the performance of speaker verification system decreases drastically. Thus, an effective speaker adaptation method is required to cope with the problem of very small enrollment data for constructing faithful speaker model [1]. To date, various speaker adaptation methods proposed have been applied primarily to improving speech recognition with some good results but not in speaker verification.

In this paper, we propose an effective speaker adaptation for speaker verification by considering the key elements of an adaptation scheme successfully used in speech recognition such as Maximum A Posteriori (MAP) [2] and Maximum Likelihood Linear Regression (MLLR) [4].

II. SPEAKER ADAPTATION METHODS

The premise of speaker adaptation is such that a speaker-independent model is adapted to a new speaker by means of speaker specific adaptation data from the speaker without going through a full length training. We propose several candidate speaker adaptation methods employing a variation and/or combination of MAP and MLLR aimed at constructing an effective speaker model. In the proposed adaptation models, only Gaussian mean vectors are adapted while the Gaussian covariance matrix is disregarded since speakers can be characterized primarily by the estimates of the Gaussian means. The impact of Gaussian variance is minimal.

(1) MAP Adaptation Method: MAP adaptation [2][3] provides a way of incorporating prior information of a previously trained system into the adaptation model using newly acquired speaker-specific data. Among the various parameters for consideration, in the experiment for the performance analysis, only Gaussian mean vectors are made to adaptation while the rest of parameters are disregarded.

(2) MLLR Adaptation Method: In the MLLR approach [4][5], the model is adapted to a new speaker by transformation of the model parameters. In the experiment, only Gaussian mean vectors are adapted and the Gaussian covariance matrix is disregarded. Also, tying of mixtures of all states is used.

(3) Updated Weight MLLR Adaptation Method: In general, the MLLR adaptation method adapts the Gaussian means and covariance but not Gaussian weights. No adaptation of weight can be restricted to make a more efficient model. Thus in this paper, we adapt the Gaussian weights using Equation (1) as a way to construct more efficient speaker model.

$$\hat{w}_{ik} = \frac{\tau + \sum_{t=1}^{T} c_{ikt}}{K \times \tau + \sum_{k=1}^{K} \sum_{t=1}^{T} c_{ikt}} \quad (1)$$

where c_{ikt} is the probability of being in state i with the mixture component label k at time t given that the model generates observation, and K is the number of mixtures, and τ is the weight of new observations as compared to the prior information.

(4) Combination of MAP and MLLR Adaptation: Since the adaptation by MAP or MLLR alone is anticipated to show limited effectiveness, the performance may be improved by combining the strengths of the two methods.

◆ MAP→MLLR, MLLR→MAP Approach: When adaptation data is made available, MAP (MLLR) adaptation is executed first and then MLLR (MAP) adaptation is followed using the adapted parameters. This procedure is iterated recursively. In this configuration, the weaknesses of each method may be compensated by cascading the two adaptations.

◆ A Hybrid Approach: With only limited data available for training, the construction of an effective speaker model is difficult using MAP and MLLR adaptation alone. With a hybrid speaker adaptation by a weighted sum of MAP and MLLR, it is anticipated that their individual strength can be combined. We achieve the combination of MAP and MLLR by a weighted sum of individual mean as shown in Equation (2).

$$\hat{\mu}_{New} = \alpha \times \hat{\mu}_{MAP} + (1-\alpha) \times \hat{\mu}_{MLLR} \quad (2)$$

where α is the weight.

III. EXPERIMENTS

To construct an effective speaker model with limited training data, text-dependent speaker verification experiments were performed using the speaker adaptation methods discussed above. The speech database used in the experiments contains isolated Korean words uttered by 46 speakers (23 male and 23 female) and collected in 5 sessions over 2 months in order to capture the variations in speakers' voice over time. The speech samples were recorded using a microphone in an office environment and sampled at 8 kHz. In the experiments, each speech signal was parameterized using 12 MFCC (Mel-Frequency Cepstral Coefficient) plus log-energy, and their first derivatives. The analysis window size was 25 ms with 10 ms overlap. Acoustic

models for the speaker produced left-right continuous density Hidden Markov Model (HMM) of five-state. The likelihood was normalized with world model for decision. 10 utterances of each customer were used to test the false rejection and 3 utterances of imposters were used to test false acceptance. We fixed the number of adaptation data to 3 in this experiment. The performance of the speaker verification is evaluated with the EER (Equal Error Rate), meaning that the False Rejection Rate (FRR) is equal to the False Acceptance Rate (FAR). While EER indicates the error rate committed by the verification system, the score measures how accurately the model is constructed.

As the baseline experiment, speaker verification without adaptation was conducted. From Figure 1, it is shown that if the training data is as small as 2 or 3, the speaker model can not be estimated accurately and as a result, the verification performance falls drastically.

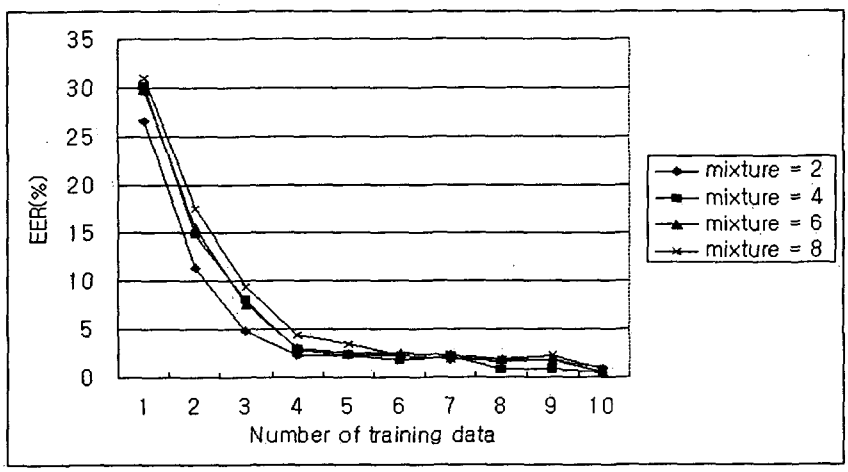

Figure 1. The performance against the number of training data with various mixtures

We then applied MAP and MLLR adaptation to speaker verification. As can be seen in Figure 2, MAP is superior to MLLR when same mixtures are used. Comparing the performance of MAP alone in terms of mixture size, the EER of mixture 8 is lower than that of mixture 4. The result is identical to that of the MLLR case. This is because the greater number of mixtures present, the more accurate model can be made. When comparing the Figure 2 with Figure 1 (the number of training data = 3), it can be seen that the performance of the speaker verification increases by applying the adaptation. Also, from Figure 2, the EER of updated weight MLLR is improved more higher than that of means-only adapted MLLR. This result shows that by updating the weights, the speaker model improves in terms of accuracy.

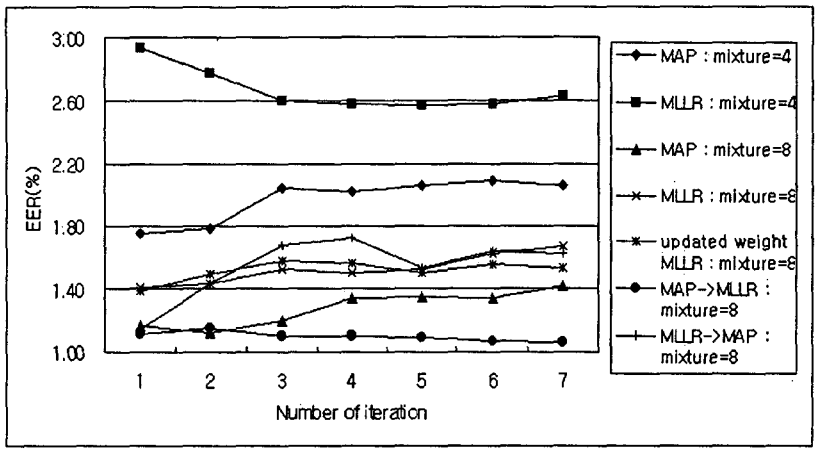

Figure 2. Comparison of EER with various adaptation methods when the number of training data is 3

Figure 2 shows the performance of MAP→MLLR adaptation and MLLR→MAP adaptation compared against MAP and MLLR adaptation alone. It shows that the combining methods in the form of MAP→MLLR and MLLR→MAP do not improve the model adaptation over that of MAP alone.

Finally, Figure 3 shows speaker verification EER using the hybrid approach with a variation of combination parameter α. The number of mixtures per states was set at 8. From the result, it shows that the hybrid approach attains the performance improvement of speaker verification in comparison to the case when α=1(i.e., MAP alone). Figure 3 also shows that the weighting parameter, α, should be set at 0.6 to produce the best performance.

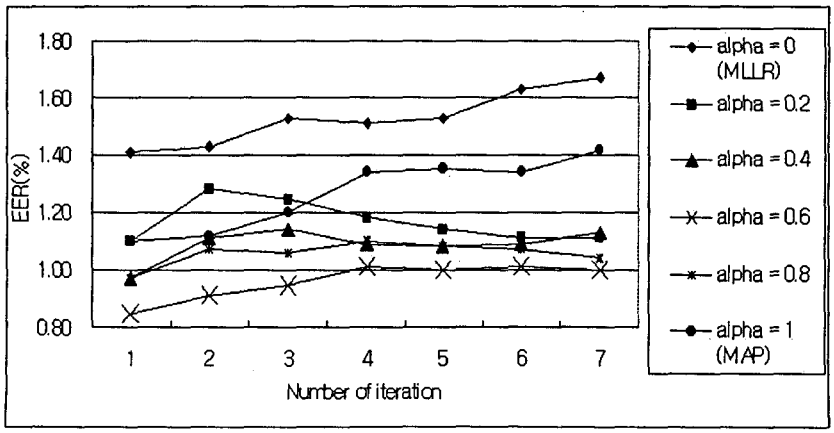

Figure 3. Speaker verification performance using a hybrid approach when the number of training data is 3

Table 1 shows the overall experimental results. The best performance of each adaptation method is used to fill the table. In case of no adaptation, the number of mixture is set at 2 while it is set at 8 for other adaptation methods.

Table 1. Performance comparison of various adaptation methods

method	**without Adaptation**	MAP	MLLR	weighted MLLR	MAP→ MLLR	MLLR→ MAP	**Hybrid**
EER(%)	**4.88**	1.12	1.41	1.39	1.06	1.14	**0.85**

From the above results, a hybrid approach method achieves better than other adaptation methods and the best performance is obtained with the weighting parameter α set at 0.6. These results demonstrate that the speaker verification system using a hybrid adaptation method outperforms that of the conventional speaker models without adaptation (the number of training data = 3) by a factor of up to 5 times.

IV. CONCLUSIONS

In this paper, we presented various candidate speaker adaptation methods to solve the over-training problem in speaker verification when modeling a speaker with sparse training data. In particular, we compared MAP and MLLR and their variations as candidate adaptation models for speaker verification. The experimental results confirmed that the speaker adaptation methods indeed improve the speaker verification performance under sparse training data by a factor of up to 5 times.

REFERENCES

[1] Lee. C. H., Soong. F. K. and Pliwal. K. K., Automatic Speech and Speaker Recognition Advanced Topics, *Kluwer Academic Publishers*, Second Printing, pp. 31-56, 1997.

[2] Gauvain. J. L. and Lee. C. H., Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, April 1994.

[3] C. H. Lee, C. H. Lin, and B. H. Juang, A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models, *IEEE Trans. on Signal Processing*, Vol. 39, No. 4, pp. 806-814, April 1991.
[4] Gales. M. J. F. and Woodland. P. C., Mean and variance adaptation within the MLLR framework, *Computer Speech and Language*, Vol. 10, pp. 249-264, 1996.
[5] C. J. Leggetter and P. C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density HMMs, *Computer Speech and Language*, Vol. 9, pp. 171-185, 1995.

Received : Jan. 10, 2000
Accepted : Feb. 28, 2000

▲ Sungjoo Ahn
Dept. of Electronics Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku
Seoul, 136-701, KOREA
Tel: +82-2-2927-6115 (O)
Fax: +82-2-3291-2450 H/P: 019-375-6709
e-mail: sjahn@ispl.korea.ac.kr

▲ Sunmee Kang
Dept. of Computer Science, Seokyeong University
16Ka-1, Chongnung-Dong, Sungbuk-ku,
Seoul, 136-704, KOREA
Tel: +82-2-940-7144 (O),
Fax: +82-2-919-0345 H/P: 011-9760-7144
e-mail: smkang@bukak.seokyeong.ac.kr

▲ Hanseok Ko
Dept. of Electronics Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku
Seoul, 136-701, KOREA
Tel: +82-2-3290-3239 (O)
Fax: +82-2-3291-2450 H/P: 011-9001-3239
e-mail: hsko@ispl.korea.ac.kr