論文 / 著書情報
Article / Book Information

Title	Robust Methods of Updating Model and a Priori Threshold in Speaker Verification
Author	Tomoko Matsui, Takashi Nishitani, Sadaoki Furui
Journal/Book name	IEEE ICASSP 96, Vol. 1, No. , pp. 97-100
発行日 / Issue date	1996, 5
権利情報 / Copyright	(c)1996 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

ROBUST METHODS OF UPDATING MODEL AND A PRIORI THRESHOLD IN SPEAKER VERIFICATION

Tomoko Matsui† *Takashi Nishitani*‡ *Sadaoki Furui*†‡

†NTT Human Interface Laboratories, 3-9-11, Midori-cho, Musashino-shi, Tokyo, Japan
‡Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

ABSTRACT

We describe a method of updating a hidden Markov model (HMM) for speaker verification using a small amount of new data for each speaker. The HMM is updated by adapting the model parameters to the new data by maximum a posteriori (MAP) estimation. The initial values of the a priori parameters in MAP estimation are set using training speech used for first creating a speaker HMM. We also present a method of resetting the a priori threshold as the updating of the model proceeds. Evaluation of the performance of the two methods using 10 male speakers showed that the verification error rate was about 42% of that without updating.

1. INTRODUCTION

Speaker verification systems are supposed to be used to judge the identities of individual speakers many times over a long period. The reference model of each speaker that is made using a small amount of data uttered in one session, however, is not robust against utterance variations. Such variations include session-to-session variation and text-dependent variation in text-independent recognition. Still, it is impractical to make users utter a large amount of samples in several sessions before using the system. Therefore, we need a scheme for updating the reference model on the basis of a small amount of new data in different sessions.

In template-based verification, Furui [1] reported a method of updating a reference template by bringing the latest utterances into time registration and averaging them. There have been only a few studies on updating reference models in HMM-based verification. Ideally, a large amount of previous data for each speaker would be saved and the reference model recreated using the large data set containing utterance variations. Setlur et al. [2] reported a method of recreating the reference model allowing for increased model complexity to better capture utterance variations. However, while such methods would be quite effective, the required memory and amount of calculation would be huge. Therefore, we investigate a method, in which the reference model is updated by adapting the model parameters to the new data using Bayes estimation in HMM-based speaker verification. Such adaptation can be asymptotically considered as recreation of the reference model by maximum likelihood (ML) estimation using a large amount of previous data. In Bayes estimation, since the a priori probability density function is introduced, HMM parameters are estimated stably when only a small amount of data is available.

Bayes estimation, well-known as a maximum a posteriori (MAP) estimation, was formulated by Lee et al. [3]$^-$[5] for use in speaker adaptation for HMM-based speech recognition. We set the initial values of the a priori parameters using training speech used for first creating a speaker HMM.

Since the likelihood value between the input speech and the reference model varies, how to set the a priori threshold for robustness against utterance variations is crucially important, especially when training data is collected over few sessions. In this paper, we investigate a method of resetting the threshold a priori as the updating of the reference model proceeds.

2. MAP ESTIMATION

In MAP estimation [3]$^-$[5], an HMM parameter vector θ is estimated so that $f(X|\theta)g(\theta)$ is locally maximized, where X is an observation sample, $f(\cdot|\theta)$ is the likelihood function obtained by the Viterbi algorithm, and $g(\theta)$ is the a priori density function.

The HMM parameters (mean vector μ_{sm} and weighting factor w_{sm} of mixture component m in state s) are re-estimated using equations (1) and (2). (Here, covariance matrices of the HMM parameters are fixed to the initial values.)

$$\tilde{\mu}_{sm} = \frac{\tau_{sm}\mu_{sm} + \sum_{t=1}^{T_U} c_{smt}x_t}{\tau_{sm} + \sum_{t=1}^{T_U} c_{smt}} \tag{1}$$

$$\tilde{w}_{sm} = \frac{\nu_{sm} - 1 + \sum_{t=1}^{T_U} c_{smt}}{\sum_{m=1}^{M} \nu_{sm} - M + \sum_{m=1}^{M} \sum_{t=1}^{T_U} c_{smt}} \tag{2}$$

Here, c_{smt} is the probability of being in state s with mixture component m at time t $(1 \leq t \leq T_U)$, given that the HMM with θ generates the observation vector x_t; ν_{sm} is calculated using w_{sm} and τ_{sm} according to equation (3). T_U is the length of speech for updating the speaker HMM.

$$\nu_{sm} = w_{sm} \sum_{m=1}^{M} \tau_{sm} \tag{3}$$

Originally, τ_{sm} was defined in the relation in which the conditional distribution of the mean vector of mixture component m in state s when the precision matrix is r_{sm} is a multivariate normal distribution with mean vector μ_{sm} and precision matrix $\tau_{sm}r_{sm}$ such that $\tau_{sm} > 0$ [6].

Initial τ_{sm} is set to $\sum_{t=1}^{T_I} c_{smt}$ with the length of training speech T_I used for first creating a speaker HMM, and

0-7803-3192-3/96 $5.00©1996 IEEE

renewed every updating of the speaker HMM according to

$$\tau'_{sm} = \tau_{sm} + \sum_{t=1}^{T_U} c_{smt}, \tag{4}$$

Here, since estimation of τ_{sm} values is poor when using only a small amount of speech, the values are averaged over all states and mixtures for each speaker HMM.

3. A PRIORI THRESHOLD

Setting the threshold a priori causes two kinds of errors, false acceptance (FA) and false rejection (FR). In some applications, the FA rate may be more important than the FR rate or vice versa. However, when the purpose is unknown, the optimal threshold should be set at the equal error rate threshold (e.e.t.) according to the Bayes rule. Therefore, our objective is to set the a priori threshold to a value close to the e.e.t.

We encounter two problems in setting the a priori threshold. First, the likelihood values for open-set samples of the genuine speaker become much smaller than those for the training samples, and the FR rate for open-set samples using the e.e.t. criterion for the training samples becomes high. This is particularly true when the reference model is not robust against utterance variations. Second, the FR rate is very difficult to estimate when the number of training samples is small.

To cope with these problems, the a priori threshold should be set at a value with higher FA rate (i.e., a lower FR rate) than the equal error rate for the training samples, or should even be set using the FA rate alone. Furui [1] reported a method of setting the a priori threshold using the mean and the standard deviation for the distribution of inter-speaker distance, which considers only the FA rate.

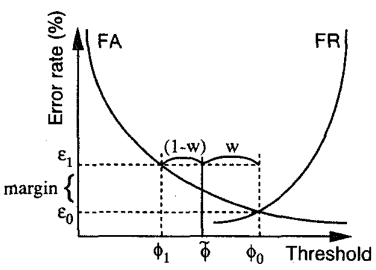

Figure 1. Method of updating the a priori threshold.

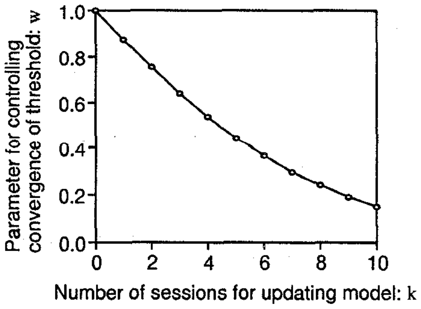

Figure 2. Parameter for controlling the convergence of the threshold: w.

Here, we propose a new method in which the a priori threshold is reset using the samples for updating the reference model. For each speaker, the FR rate is calculated using the likelihood values between one of the training samples and a model that is made by adapting the reference model to the set of samples excluding that sample; this procedure is repeated, rotating through all the samples.

In this method, the a priori threshold $\tilde{\phi}$ is set according to equation (5). In this method, the threshold for open-set samples converges from a threshold value that has a higher FA rate to the e.e.t. value for the samples for updating the model as the updating of the reference model proceeds.

$$\tilde{\phi} = w\phi_1 + (1-w)\phi_0 \tag{5}$$

$$w = \frac{2}{1+\exp(a \cdot k)} \tag{6}$$

Here, ϕ_0 is the e.e.t. for the samples for updating the model; ϕ_1 is the threshold which determines the upper bound of the FA rate (Figure 1). In our experiment, ϕ_1 was set to where the FA rate was 1% higher than the error rate at e.e.t. The w is a parameter for controlling the convergence of the threshold, and for instance, defined as (6) with k being the number of sessions for updating the model and a being an experimental parameter (0.25 in our case, Figure 2).

4. EXPERIMENTAL CONDITIONS

The proposed methods were evaluated by text-independent speaker verification experiments. The database consisted of sentence data uttered by 20 male speakers; 10 male speakers served as customers and the remainder served as impostors. The speech was recorded in six sessions (T0-5) over fifteen months. The cepstral coefficients were calculated by LPC analysis with an order of 16, a frame period of 8 ms, and a frame length of 32 ms. Each set of ten sentences from session T0 to T3 was used for training; five sentences from sessions T4 or T5 were used individually for testing. The training involved two cases: in case A, ten sentences from session T0 were used for creating each speaker model and sentences from T1 and T2 were used to update the models; in case B, ten sentences from session T1 were used for creating each speaker model and sentences from T2 and T3 were used to update the models. The texts of half of the ten training sentences were the same for all customers and all sessions, while the texts of the other half differed from customer to customer and from session to session. The sentences for testing were different from the sentences for training, but were the same for all customers and impostors and all recording sessions.

Two hundred utterances (20 people × 5 sentences × 2 sessions) were used for evaluation. The average duration of each sentence was 4.2 s. The continuous HMM (1-state, 64-mixture, diagonal) was used as the model for each registered speaker. When first creating the speaker HMMs, the Baum-Welch algorithm was used to estimate the HMM parameters.

5. RESULTS

5.1. Effect of updating reference model

The effect of updating the reference models was evaluated in terms of verification error rates using a threshold that was set the a posteriori e.e.t. Here, two types of updating were examined: a batch type using 10 consecutive sentences at once and an incremental type using each sentence successively.

Figure 3 shows verification error rates for when the mean vectors and the weighting factors of the mixture components in the reference models were updated by MAP estimation by the batch type. "Recalculation" in this figure means that the mean and variance values and the weighting factors of the mixture components in the reference models were recalculated by the Baum-Welch algorithm using all the data including the previous data. Although the required memory and computational load for MAP are much smaller than for Recalculation, MAP performed almost as well as Recalculation. The average verification error rate for cases A and B using our method of updating the models with 20 of the total number of sentences was roughly $0.2/1.5 \approx 13\%$ of that without the updating ($k = 0$). These results indicate that MAP is effective for updating the reference models.

Figure 4 shows verification error rates for when the reference models were updated by the incremental type. Like in the batch type, MAP performed almost as well as Recalculation. The average verification error rate for cases A and B using our method of updating the models with 20 of the total number of sentences was roughly $0.3/1.5 \approx 20\%$ of that without the updating. For case B in particular, the error rates for MAP varied less than those for Recalculation. In MAP estimation, since the distribution of the mean vector of mixture component m in state s is assumed to be a multivariate normal distribution and the deviation of the mean vector is restricted, the mean vector is estimated stably even when only a small amount of data is available.

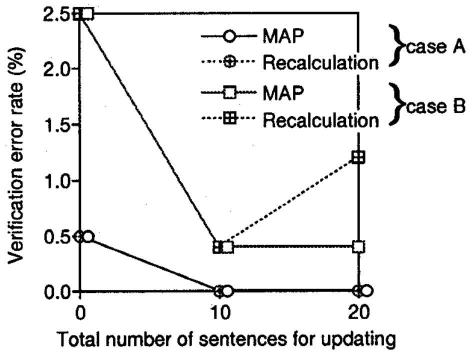

Figure 3. Verification error rates when updating the reference models using 10 consecutive sentences (batch type).

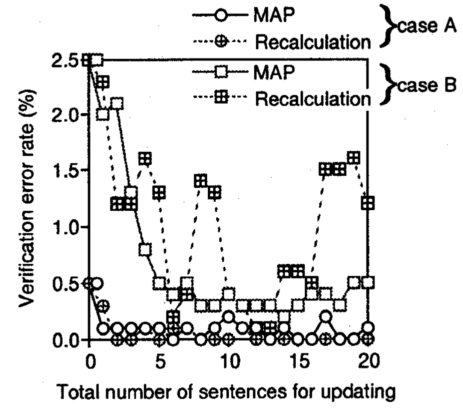

Figure 4. Verification error rates when updating the reference models successively using each sentence (incremental type).

5.2. Effect of resetting a priori threshold

The effect of resetting the a priori threshold was evaluated in terms of verification error rates using a method of updating the reference models by MAP estimation with the 10 consecutive sentences in the batch type. Figure 5 shows the FA and FR rates when using our method of resetting the a priori threshold. Figure 6 shows the FA and FR rates when setting/resetting the a priori threshold to where the FA rate is 1% higher than the error rate at e.e.t. for the samples for creating/updating the model. Figure 7 shows the FA and FR rates when first setting the a priori threshold to where the FA rate is 1% higher than the error rate

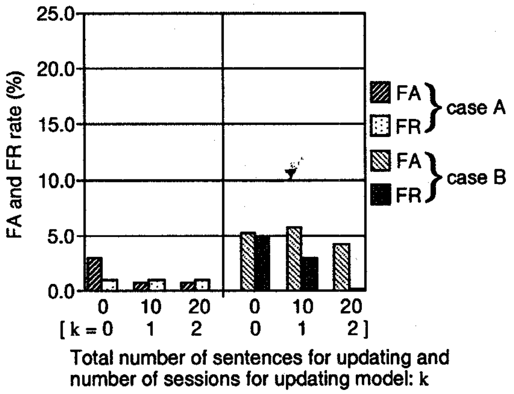

Figure 5. FA and FR rates when using our method of resetting the a priori threshold.

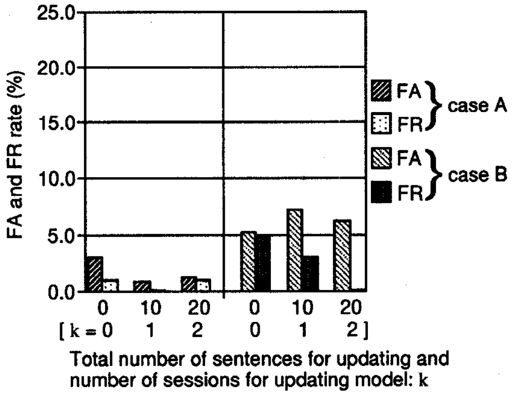

Figure 6. FA and FR rates when resetting the a priori threshold to where the FA rate is 1% higher than the error rate at e.e.t.

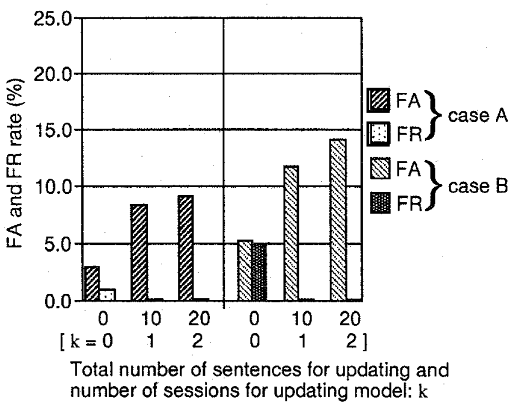

Figure 7. FA and FR rates without resetting the a priori threshold.

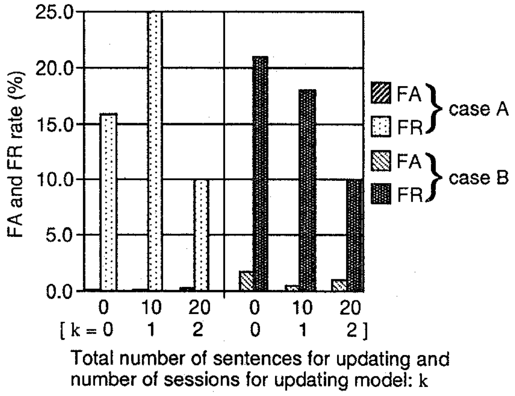

Figure 8. FA and FR rates when resetting the a priori threshold to the e.e.t.

at e.e.t. for the samples for creating the model, and after that without resetting. Figure 8 shows the FA and FR rates when setting/resetting the a priori threshold to the e.e.t. for the samples for creating/updating the model.

As shown in Figure 5, the average FA and FR rate for cases A and B when using our method of resetting the a priori threshold with 20 of the total number of sentences was roughly 1.5/3.6 $\approx$ 42% of that without the resetting ($k = 0$). When comparing the average FA and FR rates for cases A and B with 20 of the total number of sentences in Figures 5, 6, 7 and 8, the average rate for our method in Figure 5 was roughly 1.5/2.2 $\approx$ 68% of that in Figure 6, 1.5/5.3 $\approx$ 28% of that in Figure 7, and 1.5/5.8 $\approx$ 26% of that in Figure 8. These results indicate that our proposed method of resetting the a priori threshold is effective.

For Figures 5 and 6, the differences between the FA and FR rates in Figure 5 were smaller than those in Figure 6. This result supports the appropriateness of our method, in which the threshold for open-set samples converges from a threshold value that has a higher FA rate to the e.e.t. value for the samples for updating the model as the updating of the reference model proceeds.

In Figure 7, the error rates when updating the reference models were higher than those without updating. This result indicates that the a priori threshold needs to be reset as the updating of the reference model proceeds.

6. CONCLUSIONS

We presented a method of updating the reference model for each speaker using a small amount of new data based on MAP estimation and a method of resetting the a priori threshold as the updating of the model proceeds. Our method of updating the reference model achieved almost the same performance as when the reference model was recalculated by ML estimation using all the data, including the previous data, yet the required memory and computational load was much smaller for our method. The verification error rates (by a threshold set a posteriori to equalize the FR and FA rate) using our method of updating the reference model were roughly 13% of the values for when the updating was not done. Moreover, the verification error rates also using our method of resetting the a priori threshold were roughly 26% of those with setting the a priori threshold to the e.e.t. for the training samples, and roughly 42% of those without updating of the reference model and with the a priori threshold fixed to where the FA rate was 1% higher than the error rate at e.e.t. for the samples for first creating the model.

Further study will include confirming the performance of our method using data recorded in a larger number of sessions and investigating a method for estimating the covariance matrices of the mixture components in the reference model for MAP estimation. We will also examine our method in text-prompted speaker verification.

7. ACKNOWLEDGMENT

We wish to thank the members of the Furui Research Laboratory of NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

[1] S. Furui, *Cepstral analysis technique for automatic speaker verification*, IEEE Trans. ASSP-29, 2, pp.254-272, 1981.

[2] A. Setlur and T. Jacobs, *Results of a speaker verification service trial using HMM models*, Proc. Eurospeech, pp. I-53-56, 1995.

[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 ASSP, April, 1991.

[4] C.-H. Lee and J.L. Gauvain, *Speaker adaptation based on MAP estimation of HMM parameters*, Proc. ICASSP, Minneapolis, pp. II-558-561, 1993.

[5] J.L. Gauvain and C.-H. Lee, *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp.291-298, 1994.

[6] M.H. DeGroot, *Optimal statistical decisions*, McGraw-Hill, 1970.