Improving GMM-UBM Speaker Verification Using Discriminative Feedback Adaptation

Yi-Hsiang Chao[1,2], Wei-Ho Tsai[3], and Hsin-Min Wang[1]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan

[2] Department of Applied Geomatics, Ching Yun University, Taoyuan, Taiwan

[3] Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan

yschao@iis.sinica.edu.tw, whtsai@ntut.edu.tw, whm@iis.sinica.edu.tw

Abstract

The Gaussian Mixture Model - Universal Background Model (GMM-UBM) system is one of the predominant approaches for text-independent speaker verification, because both the target speaker model and the impostor model (UBM) have generalization ability to handle "unseen" acoustic patterns. However, since GMM-UBM uses a common anti-model, namely UBM, for all target speakers, it tends to be weak in rejecting impostors' voices that are similar to the target speaker's voice. To overcome this limitation, we propose a discriminative feedback adaptation (DFA) framework that reinforces the discriminability between the target speaker model and the anti-model, while preserving the generalization ability of the GMM-UBM approach. This is achieved by adapting the UBM to a target speaker dependent anti-model based on a minimum verification squared-error criterion, rather than estimating the model from scratch by applying the conventional discriminative training schemes. The results of experiments conducted on the NIST2001-SRE database show that DFA substantially improves the performance of the conventional GMM-UBM approach.

Keywords: discriminative feedback adaptation, log-likelihood ratio, minimum verification squared-error, speaker verification

1. Introduction

In essence, speaker verification is a hypothesis testing problem that can be solved by using a log-likelihood ratio (LLR) test [1]. Given an input utterance *U*, the goal is to determine whether or not *U* was spoken by the target speaker. Let us consider the following two hypotheses:

H_0: *U* was spoken by the target speaker,

H_1: *U* was not spoken by the target speaker.

The LLR test can be expressed as

$$L(U) = \log p(U \mid \lambda_0) - \log p(U \mid \lambda_1) \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{accept } H_1 \end{cases}, \tag{1}$$

where θ is a decision threshold; λ_0 is the target speaker model; and λ_1 is the so-called anti-model or impostor model. Both λ_0 and λ_1 are usually represented by Gaussian mixture models (GMMs) [1]. The current state of the art GMM-UBM approach for text-independent speaker verification uses the UBM-MAP technique [2] to generate λ_0 and λ_1. This approach pools all speech data from a large number of background speakers to form a universal background model (UBM) [2] as λ_1 via the expectation-maximization (EM) algorithm [3]. It then adapts the UBM to λ_0 via the maximum a posteriori (MAP) estimation [4] technique. GMM-UBM is effective because its generalization ability allows λ_0 to handle acoustic patterns not covered by the limited training data of the target speaker. However, since λ_0 and λ_1 are trained according to separate criteria, the optimization procedure can not distinguish a target speaker from background speakers optimally. In particular, since GMM-UBM uses a common UBM λ_1 for all target speakers, it tends to be weak in rejecting impostors' voices that are similar to the target speaker's voice. Moreover, as λ_0 is derived from λ_1, both models may correspond to a similar probability distribution.

One possible way to improve the performance of GMM-UBM is to use discriminative training methods, such as the minimum classification error (MCE) method [5] and the maximum mutual information (MMI) method [6]. In [7], a minimum verification error (MVE) training method is developed by adapting MCE training to the binary classification problem, in which the parameters of λ_0 and λ_1 are estimated using the generalized probabilistic descent (GPD) approach [8]. However, as the MVE training method requires a large number of positive and negative samples to estimate a model's parameters, it tends to over-train the model if the amount of training data is insufficient. In addition, it is difficult to select the optimal stopping point in GPD-based training.

To resolve the limitation of MVE training, we propose a framework called discriminative feedback adaptation (DFA), which improves the discrimination ability of GMM-UBM while preserving its generalization ability. The rationale behind DFA is that only mis-verified training samples are considered in the discriminative training process, rather than all the training samples used in the conventional MVE method. More specifically, DFA regards the UBM and the target speaker model obtained by the GMM-UBM approach as initial models, and then reinforces the discriminability between the models by using the mis-verified training samples. Since the reinforcement is based on model adaptation rather than training from scratch, it does not destroy the generalization ability of the two models, even if they are updated iteratively until convergence. However, recognizing that a small number of mis-verified training samples may not be able to adapt a large number of model parameters, to implement DFA, we propose two adaptation techniques: a linear regression-based minimum verification squared-error (LR-MVSE) adaptation method and an eigenspace-based minimum verification squared-error (E-MVSE) adaptation method. LR-MVSE is motivated by the minimum classification error linear regression (MCELR) techniques [9-12], which have been studied in the context of automatic speech recognition; while

E-MVSE is motivated by the MCE/eigenvoice technique [13], which has been studied in the context of speaker identification.

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed DFA framework. Sections 3 and 4 describe, respectively, the proposed LR-MVSE and E-MVSE adaptation techniques used to implement DFA. Section 5 presents simplified versions of LR-MVSE and E-MVSE. Section 6 details the experimental results. Then, in Section 7, we summarize our conclusions.

2. Discriminative Feedback Adaptation

Fig. 1 shows a block diagram of the proposed discriminative feedback adaptation (DFA) framework, which is divided into two phases. The first phase, indicated by the dashed line, utilizes the conventional GMM-UBM approach. The initial target speaker model and the UBM obtained in the first phase serve as the initial models for DFA in the second phase. The basic strategy of DFA is to reinforce the discriminability between the initial target speaker model and the UBM for ambiguous data that is mis-verified by the GMM-UBM approach. The reinforcement strategy is based on two concepts. First, since the GMM-UBM approach uses a single anti-model, UBM, for all target speakers, it tends to be weak in rejecting impostors' voices that are similar to the target speaker's voice. To resolve this problem, DFA tries to generate a discriminative anti-model exclusively for each target speaker by using the negative training samples from the cohort [14] of each target speaker to adapt both λ_0 and λ_1. Since the models may affect each other, the DFA framework also uses the positive training samples from the target speaker to avoid increasing the miss probability while reducing the false alarm probability. The resulting λ_0 and λ_1 are then updated iteratively. Second, since the DFA framework only uses mis-verified positive and negative training samples as adaptation data in each iteration, it actually fine-tunes the parameters of both λ_0 and λ_1 based on a small amount of adaptation data. It thus preserves the generalization ability of

the GMM-UBM approach while reinforcing the discrimination between H_0 and H_1. To implement the above concepts, we developed the following algorithms.

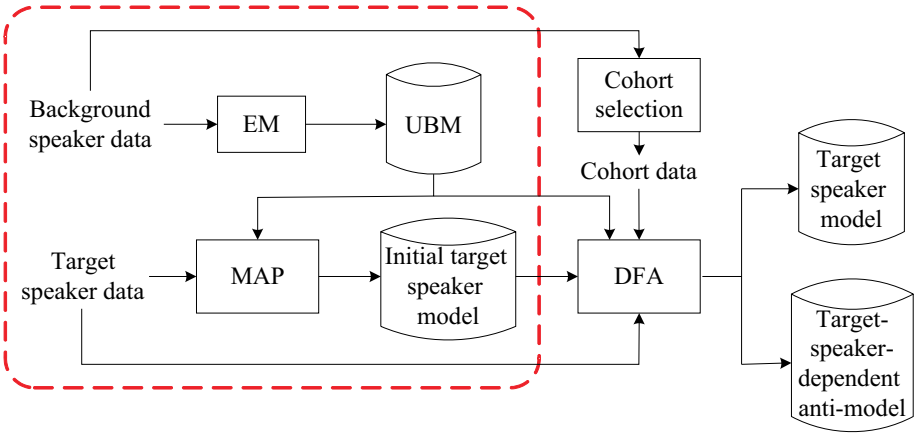


Figure 1: The proposed discriminative feedback adaptation framework, where the dashed line area utilizes the standard GMM-UBM training approach.

2.1. Minimum verification squared-error (MVSE) adaptation strategy

We modify the minimum verification error (MVE) training method [7] to fit our requirement that only mis-verified training samples should be considered. This is called the minimum verification squared-error (MVSE) adaptation strategy. The goal of DFA is to minimize the overall expected loss D, defined as

$$D = x_0 \ell_0 + x_1 \ell_1, \tag{2}$$

where x_0 and x_1 reflect which type of error is of more concern in a practical application; and ℓ_i is a loss function that describes the average false rejection loss ($i = 0$) or false acceptance loss ($i = 1$), defined as

$$\ell_i = \frac{1}{N_i} \sum_{U \in H_i} s(d(U)), \tag{3}$$

where N_0 and N_1 are the numbers of training utterances from the target speaker and the cohort, respectively; and $d(U)$ is a mis-verification measure defined as

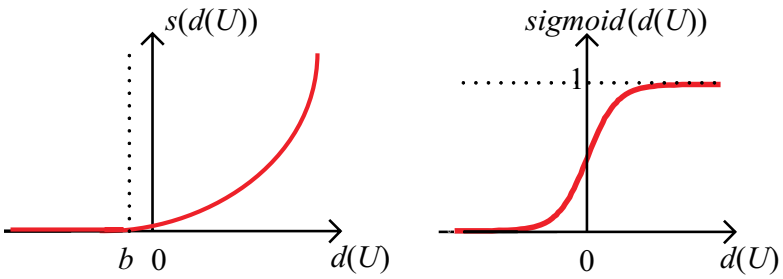
$$d(U) = \begin{cases} -L(U) & \text{if } U \in H_0 \\ L(U) & \text{if } U \in H_1 \end{cases}, \tag{4}$$

where $L(U)$ is the LLR defined in Eq. (1).

To reflect the requirement that only mis-verified training utterances should be considered, we define a function $s(\cdot)$ to represent the verification error as an adjustable quantity as follows:

$$s(d(U)) = \begin{cases} a(d(U)-b)^2 & \text{if } d(U) > b \\ 0 & \text{if } d(U) \le b, \end{cases} \tag{5}$$

where a is a scalar and b is a bias for controlling the convergence speed of DFA. The input utterance U is considered incorrectly verified if $d(U) > b$. Therefore, $s(d(U))$ is a response squared-error value. Fig. 2 contrasts the curve of the s function with that of the well-known sigmoid function. If $d(U) \le b$, the response value $s(d(U)) = 0$, i.e., the utterance U is verified correctly; hence, it will not be used for model adaptation. If $d(U) > b$, the steeper slope of the s function for a larger value of $d(U)$ results in a larger gradient to update the model's parameters. In contrast, as the value of $d(U)$ increases, the sigmoid function used in MVE [7] will become flat, and the obtained gradient will approximate zero. As a result, the mis-verified utterance U will not contribute to model adaptation. Another difference between the proposed DFA framework and the conventional MVE training method is that the latter always updates the model's parameters if the value of the sigmoid function is not 0 or 1; thus, it may over-train the well-trained models obtained from the GMM-UBM method with the correctly-verified training utterances.



(a) *s* function (b) sigmoid function

Figure 2: The *s* function compared to the sigmoid function.

2.2. Fast scoring for DFA

To speed up DFA, we use a fast scoring approach [2] to compute the LLR. Given an utterance $U = \{o_1, \ldots, o_T\}$, the computation of LLR for a GMM with M Gaussian mixture components can be written as

$$\begin{aligned} L(U) &= \frac{1}{T}\sum_{t=1}^{T}\left(\log\sum_{m=1}^{M}\alpha_m p(o_t \mid \mathbf{g}_{0,m}) - \log\sum_{m=1}^{M}\alpha_m p(o_t \mid \mathbf{g}_{1,m})\right) \\ &\approx \frac{1}{T}\sum_{t=1}^{T}\left(\log\sum_{i=1}^{C}\alpha_{C_i(t)} p(o_t \mid \mathbf{g}_{0,C_i(t)}) - \log\sum_{i=1}^{C}\alpha_{C_i(t)} p(o_t \mid \mathbf{g}_{1,C_i(t)})\right), \end{aligned} \tag{6}$$

where $\mathbf{g}_{0,m}$ and $\mathbf{g}_{1,m}$ are the m-th Gaussian mixture components of the target speaker model and the anti-model, respectively; and α_m is the mixture weight, m = 1,…, M. Note that the target speaker model has the same mixture weights as the anti-model. For each frame o_t, we determine the top C scoring mixture indices, $C_i(t)$, i = 1,…, C, in the UBM, where $C << M$; hence, it requires $M + C$ Gaussian computations in the first iteration, and $2C$ Gaussian computations per iteration thereafter. In this study, the value of C is set at 5 [2].

3. Linear Regression-based MVSE (LR-MVSE) Adaptation

Recognizing that a small amount of adaptation data selected from the mis-verified training samples may not be able to adapt a large number of model parameters, we propose using a linear regression method to implement MVSE adaptation. We call it linear regression-based MVSE (LR-MVSE) adaptation. Our strategy is motivated by the minimum classification error linear regression (MCELR) techniques [9-12], which have been studied in the context of automatic speech recognition. We assume that the initial target speaker model $\lambda_0^{(0)}$ and anti-model $\lambda_1^{(0)}$ have M Gaussian mixtures $\mathbf{g}_{0,m}^{(0)} \sim N(\boldsymbol{\mu}_{0,m}^{(0)}, \boldsymbol{\Sigma}_m)$ and $\mathbf{g}_{1,m}^{(0)} \sim N(\boldsymbol{\mu}_{1,m}^{(0)}, \boldsymbol{\Sigma}_m)$, respectively, where $\boldsymbol{\mu}_{0,m}^{(0)}$ and $\boldsymbol{\mu}_{1,m}^{(0)}$ are r–dimensional mean vectors obtained with the GMM-UBM method; and $\boldsymbol{\Sigma}_m$ is an $r \times r$ covariance matrix of the UBM, $m = 1,\ldots, M$. Note that, in this study, we only adapt the mean vectors of GMMs. After adaptation, the new mean vectors of the target speaker model or the anti-model take the following form:

$$\boldsymbol{\mu}_{j,m} = \mathbf{W}_j \boldsymbol{\xi}_{j,m}^{(0)}, \tag{7}$$

where $\mathbf{W}_j$, $j = 0$ (for the target speaker model) or 1 (for the anti-model), is an $r \times (r+1)$ transformation matrix; and $\boldsymbol{\xi}_{j,m}^{(0)} = [1 \ \boldsymbol{\mu}_{j,m}^{(0)\prime}]'$. Given initial transformation matrices $\mathbf{W}_0^{(0)} = \mathbf{W}_1^{(0)} = [\mathbf{0} \ \mathbf{I}]$, where $\mathbf{0}$ is an $r \times 1$ zero vector and $\mathbf{I}$ is an $r \times r$ identity matrix, the parameter $\mathbf{W}_j$ can be iteratively optimized using

$$\mathbf{W}_j^{(k+1)} = \mathbf{W}_j^{(k)} - \delta \frac{\partial D}{\partial \mathbf{W}_j^{(k)}}, \tag{8}$$

where the superscript "(k)" denotes the k-th iteration, and δ is the step size. In addition,

$$\begin{aligned} \frac{\partial D}{\partial \mathbf{W}_j^{(k)}} &= x_0 \frac{\partial \ell_0}{\partial s(d(U))} \cdot \frac{\partial s(d(U))}{\partial d(U)} \cdot \frac{\partial d(U)}{\partial L(U)} \cdot \frac{\partial L(U)}{\partial \mathbf{W}_j^{(k)}} + x_1 \frac{\partial \ell_1}{\partial s(d(U))} \cdot \frac{\partial s(d(U))}{\partial d(U)} \cdot \frac{\partial d(U)}{\partial L(U)} \cdot \frac{\partial L(U)}{\partial \mathbf{W}_j^{(k)}} \\ &= x_0 \cdot \frac{1}{N_0} \sum_{U \in H_0, -L(U) > b} \left\{ 2a \cdot (-L(U) - b) \cdot \left(-\frac{\partial L(U)}{\partial \mathbf{W}_j^{(k)}} \right) \right\} + x_1 \cdot \frac{1}{N_1} \sum_{U \in H_1, L(U) > b} \left\{ 2a \cdot (L(U) - b) \cdot \frac{\partial L(U)}{\partial \mathbf{W}_j^{(k)}} \right\}, \end{aligned} \tag{9}$$

where

$$\frac{\partial L(U)}{\partial \mathbf{W}_j^{(k)}} = \frac{1}{T}\sum_{t=1}^{T}\frac{(-1)^j}{p(o_t \mid \lambda_j^{(k)})}\left(\sum_{i=1}^{C}\alpha_{C_i(t)}\frac{\partial p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)})}{\partial \mathbf{W}_j^{(k)}}\right), \tag{10}$$

where the target speaker model $\lambda_0^{(k)}$ with mixtures $\mathbf{g}_{0,m}^{(k)}$ and the anti-model $\lambda_1^{(k)}$ with mixtures $\mathbf{g}_{1,m}^{(k)}$, $m = 1,\ldots, M$, are obtained by LR-MVSE adaptation in k iterations, and

$$\frac{\partial p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)})}{\partial \mathbf{W}_j^{(k)}} = p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)})\Sigma_{C_i(t)}^{-1}\left(o_t - \mathbf{W}_j^{(k)}\xi_{j,C_i(t)}^{(0)}\right)\xi_{j,C_i(t)}^{(0)\prime}. \tag{11}$$

If we assume that all covariance matrices $\mathbf{\Sigma}_m$ of the UBM, $m = 1,\ldots, M$, are diagonal, Eq. (11) can be rewritten as

$$\frac{\partial p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)})}{\partial \mathbf{W}_j^{(k)}(r_1, r_2)} = \frac{p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)})}{\sigma_{C_i(t)}^2(r_1)}\left(o_t(r_1) - \sum_{s=1}^{r+1}\mathbf{W}_j^{(k)}(r_1, s)\xi_{j,C_i(t)}^{(0)}(s)\right)\xi_{j,C_i(t)}^{(0)}(r_2), \tag{12}$$

where $\sigma_m^2(r_1)$ is the r_1-th diagonal element of $\mathbf{\Sigma}_m$; $o_t(r_1)$ is the r_1-th element of o_t; $\xi_{j,m}^{(0)}(r_2)$ is the r_2-th element of $\xi_{j,m}^{(0)}$; and $\mathbf{W}_j^{(k)}(r_1, r_2)$ is the r_1-th row and r_2-th column element of $\mathbf{W}_j^{(k)}$, $r_1 = 1,\ldots, r$, and $r_2 = 1,\ldots, (r+1)$.

4. Eigenspace-based MVSE (E-MVSE) Adaptation

Alternatively, we can use the eigenspace method to implement MVSE adaptation. We call it eigenspace-based MVSE (E-MVSE) adaptation. E-MVSE is motivated by the MCE/eigenvoice technique [13], which has been studied in the context of speaker identification. In this case, we also assume that only the mean vectors of GMMs are adapted. Let $\mathbf{u}_0^{(0)}$ and $\mathbf{u}_1^{(0)}$ be $(rM)\times 1$ supervectors [15, 16] obtained by concatenating all the mean vectors of the initial target speaker model $\lambda_0^{(0)}$ and anti-model (a clone of the UBM) $\lambda_1^{(0)}$, where

$$\mathbf{u}_j^{(0)} = [\boldsymbol{\mu}_{j,1}^{(0)\prime}\ \boldsymbol{\mu}_{j,2}^{(0)\prime}\ldots\boldsymbol{\mu}_{j,M}^{(0)\prime}]', \tag{13}$$

j = 0 or 1. Following the eigenvoice approach, we use the principal component analysis (PCA) technique [17] to construct a speaker eigenspace $\mathbf{E}$ = span$\{\mathbf{e}_1, \mathbf{e}_2,\ldots, \mathbf{e}_Z\}$ based on R supervectors derived from R pre-trained background speaker GMMs, where $Z \le R-1$. According to the orthogonality principle [18], we can decompose $\mathbf{u}_j^{(0)}$ into

$$\mathbf{u}_j^{(0)} = \boldsymbol{\eta} + \sum_{z=1}^{Z} w_{j,z}^{(0)} \mathbf{e}_z + w_{j,Z+1}^{(0)} \mathbf{e}_j^{\perp}, \tag{14}$$

where $\boldsymbol{\eta}$ is the sample mean vector of R supervectors. The second term in Eq. (14) represents the results of projecting $(\mathbf{u}_j^{(0)} - \boldsymbol{\eta})$ onto the eigenspace $\mathbf{E}$. Note that, in most cases, $(\mathbf{u}_j^{(0)} - \boldsymbol{\eta}) \notin \mathbf{E}$, since the initial target speaker model and anti-model are not included in the background speaker model set. The coordinates, $w_{j,z}^{(0)}$, z = 1,.., Z, are computed by

$$w_{j,z}^{(0)} = \mathbf{e}_z'(\mathbf{u}_j^{(0)} - \boldsymbol{\eta}). \tag{15}$$

The third term in Eq. (14) represents the residual after the projection. If the residual is not zero, we can define $w_{j,Z+1}^{(0)}$ as

$$w_{j,Z+1}^{(0)} = \left\| \mathbf{u}_j^{(0)} - \boldsymbol{\eta} - \sum_{z=1}^{Z} w_{j,z}^{(0)} \mathbf{e}_z \right\|, \tag{16}$$

and define $\mathbf{e}_j^{\perp}$ as

$$\mathbf{e}_j^{\perp} = \frac{\mathbf{u}_j^{(0)} - \boldsymbol{\eta} - \sum_{z=1}^{Z} w_{j,z}^{(0)} \mathbf{e}_z}{\left\| \mathbf{u}_j^{(0)} - \boldsymbol{\eta} - \sum_{z=1}^{Z} w_{j,z}^{(0)} \mathbf{e}_z \right\|}. \tag{17}$$

Since both $\mathbf{e}_0^{\perp}$ and $\mathbf{e}_1^{\perp}$ are orthogonal to $\mathbf{E}$, $\mathbf{u}_0^{(0)}$ and $\mathbf{u}_1^{(0)}$ can be represented, respectively, by the initial coordinates $[w_{0,1}^{(0)}\ w_{0,2}^{(0)} \ldots w_{0,Z}^{(0)}\ w_{0,Z+1}^{(0)}]'$ and $[w_{1,1}^{(0)}\ w_{1,2}^{(0)} \ldots w_{1,Z}^{(0)}\ w_{1,Z+1}^{(0)}]'$ in a target speaker space $\mathbf{E}_{\lambda_0}$ with an orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2,\ldots, \mathbf{e}_Z, \mathbf{e}_0^{\perp}\}$ and an anti-model space $\mathbf{E}_{\lambda_1}$ with an orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2,\ldots, \mathbf{e}_Z, \mathbf{e}_1^{\perp}\}$. If $w_{j,Z+1}^{(0)} = 0$, $\mathbf{e}_j^{\perp}$ is a zero vector, and the (Z+1)-th

coordinate is not included in a coordinate vector $\in \mathbf{E}_{\lambda_j} = \mathbf{E}$ with a basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_Z\}$, $j = 0$ or 1. Our goal is to find the best coordinates $[w_{0,1}\ w_{0,2} \ldots w_{0,Z}\ w_{0,Z+1}]'$ in $\mathbf{E}_{\lambda_0}$ and $[w_{1,1}\ w_{1,2} \ldots w_{1,Z}\ w_{1,Z+1}]'$ in $\mathbf{E}_{\lambda_1}$ such that the reconstructed models can optimally distinguish the target speaker's voice from the non-target speakers' voices. The reconstructed mean vectors of the target speaker model or the anti-model take the following form:

$$\boldsymbol{\mu}_{j,m} = \boldsymbol{\eta}_m + \sum_{z=1}^{Z} w_{j,z}\mathbf{e}_{z,m} + w_{j,Z+1}\mathbf{e}^{\perp}_{j,m}, \tag{18}$$

where $\boldsymbol{\eta}_m$, $\mathbf{e}_{z,m}$, and $\mathbf{e}^{\perp}_{j,m}$ represent the m-th subvectors of $\boldsymbol{\eta}$, $\mathbf{e}_z$, and $\mathbf{e}^{\perp}_j$, respectively, and correspond to the mean vector of the m-th Gaussian mixture component of the target speaker model ($j = 0$) and the anti-model ($j = 1$), $m = 1,\ldots, M$. The coordinates, $w_{j,z}$, $j = 0, 1$, $z = 1,.., Z+1$, can be iteratively optimized using

$$w^{(k+1)}_{j,z} = w^{(k)}_{j,z} - \delta \frac{\partial D}{\partial w^{(k)}_{j,z}}, \tag{19}$$

where δ is the step size. In addition,

$$\begin{aligned}\frac{\partial D}{\partial w^{(k)}_{j,z}} &= x_0 \frac{\partial \ell_0}{\partial s(d(U))} \cdot \frac{\partial s(d(U))}{\partial d(U)} \cdot \frac{\partial d(U)}{\partial L(U)} \cdot \frac{\partial L(U)}{\partial w^{(k)}_{j,z}} + x_1 \frac{\partial \ell_1}{\partial s(d(U))} \cdot \frac{\partial s(d(U))}{\partial d(U)} \cdot \frac{\partial d(U)}{\partial L(U)} \cdot \frac{\partial L(U)}{\partial w^{(k)}_{j,z}} \\ &= x_0 \cdot \frac{1}{N_0} \sum_{U \in H_0, -L(U) > b} \left\{ 2a \cdot (-L(U) - b) \cdot \left(-\frac{\partial L(U)}{\partial w^{(k)}_{j,z}} \right) \right\} + x_1 \cdot \frac{1}{N_1} \sum_{U \in H_1, L(U) > b} \left\{ 2a \cdot (L(U) - b) \cdot \frac{\partial L(U)}{\partial w^{(k)}_{j,z}} \right\},\end{aligned} \tag{20}$$

where

$$\frac{\partial L(U)}{\partial w^{(k)}_{j,z}} = \frac{1}{T} \sum_{t=1}^{T} \frac{(-1)^j}{p(o_t \mid \lambda^{(k)}_j)} \left(\sum_{i=1}^{C} \alpha_{C_i(t)} \frac{\partial p(o_t \mid \mathbf{g}^{(k)}_{j,C_i(t)})}{\partial w^{(k)}_{j,z}} \right), \tag{21}$$

where the Gaussian mixture components, $\mathbf{g}^{(k)}_{j,m}$, $m = 1,\ldots, M$, of $\lambda^{(k)}_j$ are the results of the k-th iteration, and

$$\frac{\partial p(o_t \mid \mathbf{g}^{(k)}_{j,C_i(t)})}{\partial w^{(k)}_{j,z}} = \begin{cases} p(o_t \mid \mathbf{g}^{(k)}_{j,C_i(t)}) \left(o_t - \boldsymbol{\mu}^{(k)}_{j,C_i(t)} \right)' \Sigma^{-1}_{C_i(t)} \mathbf{e}^{\perp}_{j,C_i(t)} & \text{if } z = Z+1 \\ p(o_t \mid \mathbf{g}^{(k)}_{j,C_i(t)}) \left(o_t - \boldsymbol{\mu}^{(k)}_{j,C_i(t)} \right)' \Sigma^{-1}_{C_i(t)} \mathbf{e}_{z,C_i(t)} & \text{otherwise,} \end{cases} \tag{22}$$

where

$$\boldsymbol{\mu}_{j,C_i(t)}^{(k)} = \boldsymbol{\eta}_{C_i(t)} + \sum_{z=1}^{Z} w_{j,z}^{(k)} \mathbf{e}_{z,C_i(t)} + w_{j,Z+1}^{(k)} \mathbf{e}_{j,C_i(t)}^{\perp}. \tag{23}$$

If we assume that all covariance matrices $\boldsymbol{\Sigma}_m$ of the UBM, $m = 1,\ldots, M$, are diagonal, Eqs. (22) and (23) can be rewritten, respectively, as

$$\frac{\partial p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)})}{\partial w_{j,z}^{(k)}} = \begin{cases} p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)}) \sum_{r_1=1}^{r} \dfrac{\left(o_t(r_1) - \boldsymbol{\mu}_{j,C_i(t)}^{(k)}(r_1)\right) \mathbf{e}_{j,C_i(t)}^{\perp}(r_1)}{\sigma_{C_i(t)}^2(r_1)} & \text{if } z = Z+1 \\ p(o_t \mid \mathbf{g}_{j,C_i(t)}^{(k)}) \sum_{r_1=1}^{r} \dfrac{\left(o_t(r_1) - \boldsymbol{\mu}_{j,C_i(t)}^{(k)}(r_1)\right) \mathbf{e}_{z,C_i(t)}(r_1)}{\sigma_{C_i(t)}^2(r_1)} & \text{otherwise,} \end{cases} \tag{24}$$

and

$$\boldsymbol{\mu}_{j,C_i(t)}^{(k)}(r_1) = \boldsymbol{\eta}_{C_i(t)}(r_1) + \sum_{z=1}^{Z} w_{j,z}^{(k)} \mathbf{e}_{z,C_i(t)}(r_1) + w_{j,Z+1}^{(k)} \mathbf{e}_{j,C_i(t)}^{\perp}(r_1). \tag{25}$$

where $\boldsymbol{\eta}_m(r_1)$, $\mathbf{e}_{z,m}(r_1)$, and $\mathbf{e}_{j,m}^{\perp}(r_1)$, $m = 1,\ldots, M$, $r_1 = 1,\ldots, r$, represent the r_1-th elements of the m-th subvectors $\boldsymbol{\eta}_m$, $\mathbf{e}_{z,m}$, and $\mathbf{e}_{j,m}^{\perp}$, respectively.

5. Simplified Versions of LR-MVSE and E-MVSE

As far as reliability is concerned, a target speaker model trained with the GMM-UBM approach may be effective in characterizing the target speaker's voice. In contrast, a UBM generated from a number of background speakers may not be able to represent the imposters with respect to each specific target speaker. In other words, it may not be able to distinguish between imposters and the target speaker. Thus, it is more important to reinforce discriminability in the UBM than in the target speaker model. Moreover, in our experience, the training samples of target speakers are seldom mis-verified; i.e., nearly all the mis-verified training samples are from the cohort. Accordingly, to adapt the UBM to the target speaker dependent anti-model, it might be sufficient to use only negative training samples in our DFA framework. In this case, the training goal can be

simplified to one of minimizing the average false acceptance (false alarm) loss ℓ_1 defined in Eq. (3). For LR-MVSE adaptation, the parameter $\mathbf{W}_1$ is iteratively optimized using

$$\mathbf{W}_1^{(k+1)} = \mathbf{W}_1^{(k)} - \delta \frac{\partial \ell_1}{\partial \mathbf{W}_1^{(k)}}, \tag{26}$$

where

$$\frac{\partial \ell_1}{\partial \mathbf{W}_1^{(k)}} = \frac{\partial \ell_1}{\partial s(d(U))} \cdot \frac{\partial s(d(U))}{\partial d(U)} \cdot \frac{\partial d(U)}{\partial L(U)} \cdot \frac{\partial L(U)}{\partial \mathbf{W}_1^{(k)}} = \frac{1}{N_1} \sum_{U \in H_1,\, L(U) > b} \left\{ 2a \cdot (L(U) - b) \cdot \frac{\partial L(U)}{\partial \mathbf{W}_1^{(k)}} \right\}, \tag{27}$$

and $\frac{\partial L(U)}{\partial \mathbf{W}_1^{(k)}}$ is computed by Eq. (10). For E-MVSE adaptation, the coordinates, $w_{1,z}$, $z = 1,.., Z+1$, are iteratively optimized using

$$w_{1,z}^{(k+1)} = w_{1,z}^{(k)} - \delta \frac{\partial \ell_1}{\partial w_{1,z}^{(k)}}, \tag{28}$$

where

$$\frac{\partial \ell_1}{\partial w_{1,z}^{(k)}} = \frac{\partial \ell_1}{\partial s(d(U))} \cdot \frac{\partial s(d(U))}{\partial d(U)} \cdot \frac{\partial d(U)}{\partial L(U)} \cdot \frac{\partial L(U)}{\partial w_{1,z}^{(k)}} = \frac{1}{N_1} \sum_{U \in H_1,\, L(U) > b} \left\{ 2a \cdot (L(U) - b) \cdot \frac{\partial L(U)}{\partial w_{1,z}^{(k)}} \right\}, \tag{29}$$

and $\frac{\partial L(U)}{\partial w_{1,z}^{(k)}}$ is computed by Eq. (21). When $N_0 \approx N_1$, the training times of the simplified versions of LR-MVSE and E-MVSE are about one-quarter of the training times of the respective original versions.

6. Experiments

6.1. Experimental setup

In our experiments, we used the NIST 2001 cellular speaker recognition evaluation (NIST2001-SRE) database [19], and divided it into two subsets: an evaluation set and a development set. The evaluation set contained 74 male and 100 female speakers. On average, each

speaker had approximately 2 minutes of training utterances and 10 test segments. The development set contained 38 males and 22 females as background speakers that did not overlap with the speakers in the evaluation set. To scale up the number of background speakers, we also included 139 male and 191 female speakers extracted from the NIST2002-SRE corpus [19]. Thus, we collected the training utterances of 177 male and 213 female background speakers to build two gender-dependent UBMs, each containing 1,024 mixture components. To train each target speaker's GMM, we only adapted the mean vectors from the speaker's corresponding gender-dependent UBM in the GMM-UBM method. Then, for each male or female target speaker, we chose the B closest speakers from the 177 male or 213 female background speakers, respectively, as a cohort based on the degree of closeness measured in terms of the pairwise distance defined by [1]

$$\mathrm{dist}(\lambda_i, \lambda_j) = \log \frac{p(U_i \mid \lambda_i)}{p(U_i \mid \lambda_j)} + \log \frac{p(U_j \mid \lambda_j)}{p(U_j \mid \lambda_i)}, \tag{30}$$

where λ_i and λ_j are speaker GMMs trained using the i-th speaker's utterances, U_i, and the j-th speaker's utterances, U_j, respectively. For each cohort speaker, we extracted J 3-second speech segments from his/her training utterances as negative training samples of a target speaker. Thus, each target speaker had $J \times B$ negative training samples in total. All the 3-second segments extracted from each target speaker's training utterances served as positive training samples in LR-MVSE or E-MVSE adaptation.

To remove silence/noise frames, we processed all the speech data with a Voice Activity Detector (VAD) [20]. Then, using a 32-ms Hamming-windowed frame with 10-ms shifts, we converted each utterance into a stream of 30-dimensional feature vectors, each consisting of 15 Mel-frequency cepstral coefficients (MFCCs) [3] and their first time derivatives. To compensate for channel mismatch effects, we applied feature warping [21] after MFCC extraction.

In the experiments, a and b in the s function defined in Eq. (5) were set at 3 and 0.01, respectively. For E-MVSE adaptation, we generated two gender-dependent Z-dimensional eigenspaces using the GMMs of the 177 male and 213 female background speakers, respectively, with Z set to 70 or 140. The LR-MVSE and E-MVSE adaptation procedures were trained until they almost converged, i.e., until the number of mis-verified training samples approximated zero. For the overall expected loss D defined in Eq. (2), x_0 and x_1 were set as $C_{Miss} \times P_{Target}$ and $C_{FalseAlarm} \times (1 - P_{Target})$, respectively, according to the NIST Detection Cost Function (DCF) [19]:

$$C_{DET} = C_{Miss} \times P_{Miss} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm} \times (1 - P_{Target}), \tag{31}$$

where P_{Miss} and $P_{FalseAlarm}$ are the miss (false rejection) probability and the false alarm (false acceptance) probability, respectively; C_{Miss} and $C_{FalseAlarm}$ are the respective relative costs of the detection errors; and P_{Target} is the *a priori* probability of the target speaker. Following the NIST2001-SRE protocol, C_{Miss}, $C_{FalseAlarm}$, and P_{Target} were set at 10, 1, and 0.01, respectively.

6.2. Experimental results

To evaluate the performance of the DFA framework, we used the Detection Error Tradeoff (DET) curve [22] and the NIST DCF; the latter reflects the performance at a single operating point on the former. We implemented the proposed DFA framework in three ways:

a) LR-MVSE adaptation ("MAP + LR-MVSE"),

b) E-MVSE adaptation with the first 70 eigenvectors ("MAP + E-MVSE70"), and

c) E-MVSE adaptation with the first 140 eigenvectors ("MAP + E-MVSE140").

For the performance comparison, we used two baseline systems:

a) GMM-UBM ("MAP") and

b) conventional MVE (MCE) training with the sigmoid function ("MAP + MVE").

The target speaker GMM and the UBM obtained from the GMM-UBM method served as the initial models for the proposed DFA-related methods and the conventional MVE method.

Fig. 3 plots the minimum DCFs against the total number of negative training samples per target speaker for each adaptation method. The experiments involved 2,025 target speaker trials and 20,250 impostor trials of the evaluation set. We considered different numbers of negative training samples, but not different numbers of positive training samples because the same target speaker data had been used to train the initial target speaker model in the GMM-UBM method. From the figure, we observe that "MAP + E-MVSE70" achieves the lowest minDCF in cases where the adaptation data only includes 6 or 12 negative training samples per target speaker; while "MAP + LR-MVSE" achieves the lowest minDCF in cases where the adaptation data includes 36 or 60 negative training samples per target speaker. As expected, a small amount of adaptation data favors the methods in which a smaller number of model parameters must be estimated. Note that the larger the number of negative training samples used, the lower the minDCF that can be achieved.

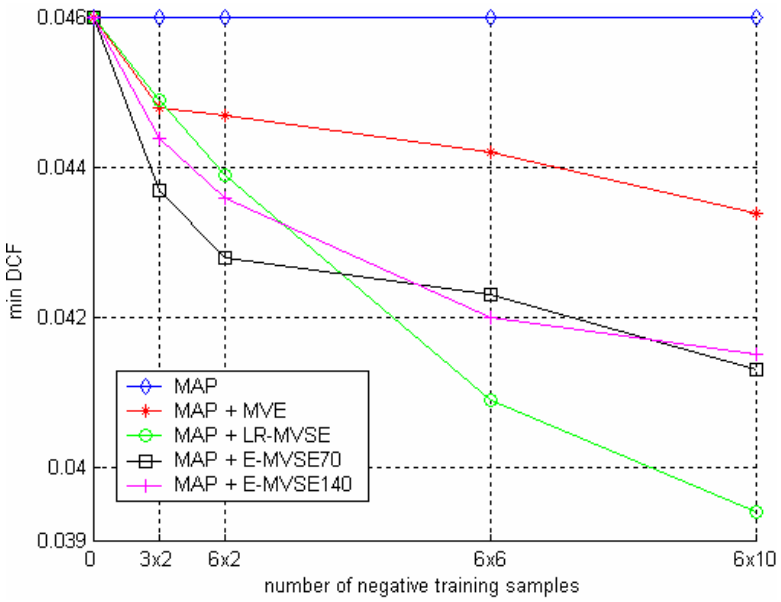


Figure 3: The minimum DCFs versus the number ($J \times B$) of 3-second negative training samples per target speaker.

Fig. 4 shows the DET curves obtained by evaluating the above systems for the case with 60 negative training samples per target speaker. It is clear that the performances of the three proposed methods, "MAP + LR-MVSE", "MAP + E-MVSE70", and "MAP + E-MVSE140", are comparable; and they all outperform the conventional methods "MAP" and "MAP + MVE". Interestingly, the performance of "MAP + MVE" is not always better than that of "MAP". This is because MVE tends to over-train the models obtained from the GMM-UBM method, and it is difficult to select the optimal stopping point in MVE training.

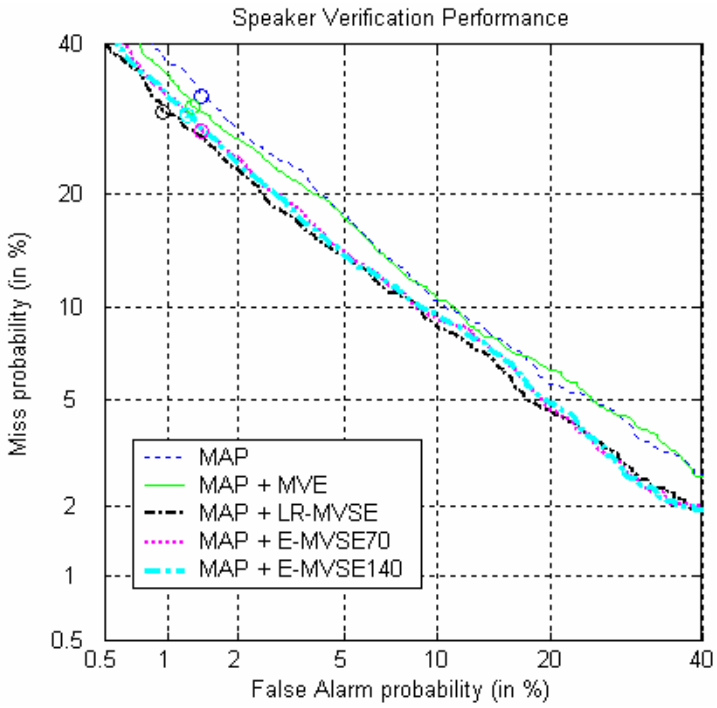


Figure 4: Experiment results in DET curves. The circles indicate the minimum DCFs.

In the above experiments, we found that nearly all the mis-verified training samples in each adaptation iteration were negative training samples. Fig. 5 shows the number of mis-verified training samples versus the number of iterations in LR-MVSE adaptation for an example target speaker (ID number "5609"). Thus, we further compared the simplified versions of the LR-MVSE and E-MVSE methods with the respective original versions. Fig. 6 shows the DET curves for the

case of 60 negative training samples per target speaker. It is clear that the simplified versions perform comparably to the respective original versions. This confirms our assumption that reinforcing the discriminability in the UBM is more beneficial than reinforcing the discriminability in the target speaker model.

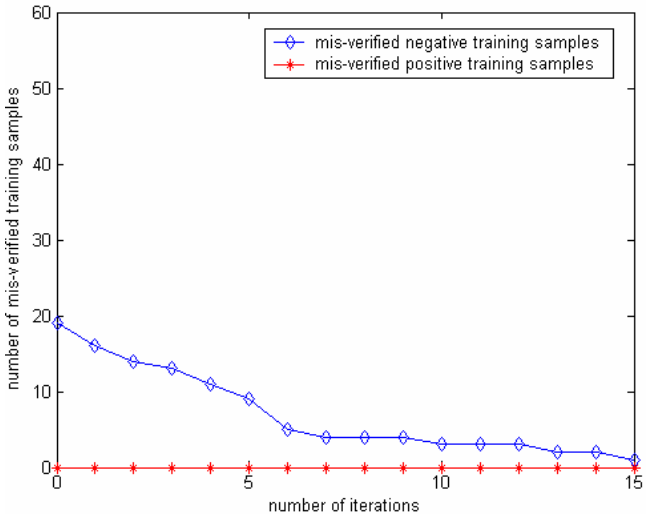
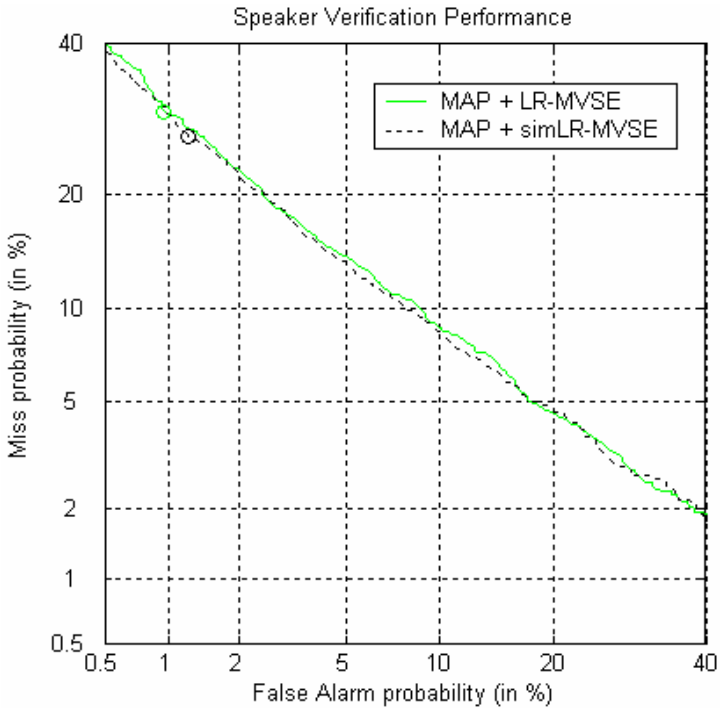


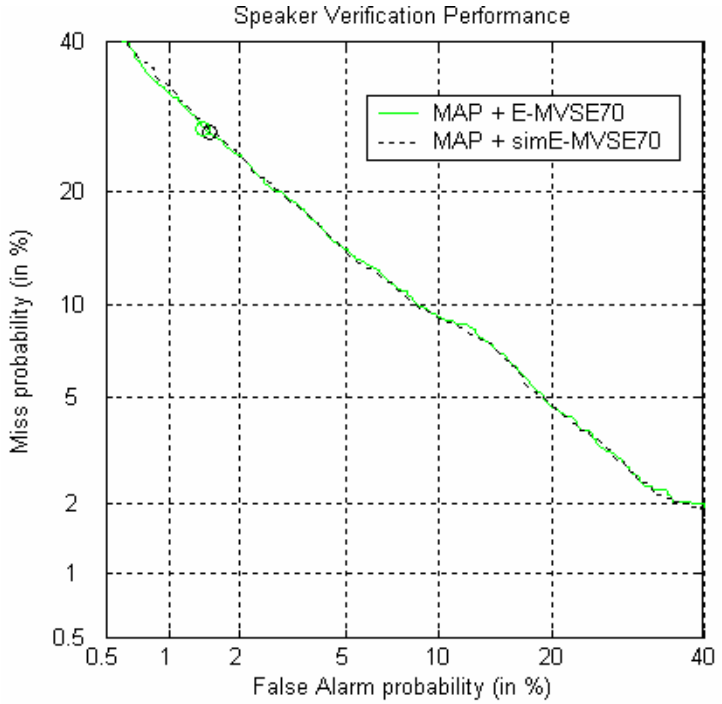
Figure 5: The number of mis-verified training samples versus the number of iterations in LR-MVSE adaptation for an example target speaker (ID number "5609") having 60 negative training samples and 33 positive training samples. The 0-th (initial) iteration represents the result obtained with the standard GMM-UBM method.

Table 1 summarizes the minimum DCFs of each system shown in Figs. 4 and 6. We observe that "MAP + LR-MVSE" achieves a 14.35% relative DCF reduction over the baseline GMM-UBM system ("MAP") and a 9.22% relative DCF reduction over the "MAP + MVE" method. "MAP + simLR-MVSE" even performs slightly better than the original version "MAP + LR-MVSE", but the difference is not statistically significant. Table 2 compares the correlation of correct and incorrect decisions between "MAP" and "MAP + LR-MVSE" for the minimum DCF [23]. Using McNemar's test [24] with a significance level = 0.005, the resulting *P*-value is smaller

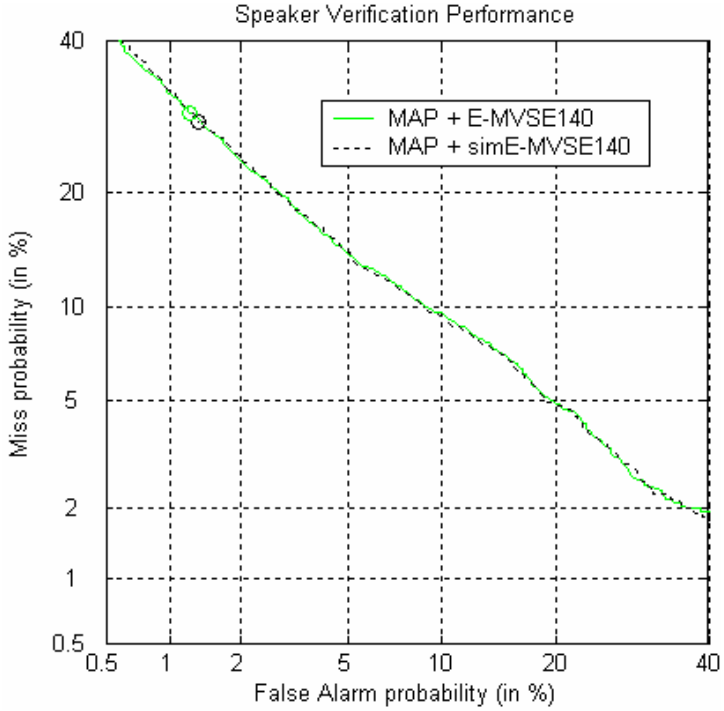
than 0.005; therefore, we conclude that "MAP + LR-MVSE" performs significantly better than "MAP".



(a) LR-MVSE vs. the simplified version of LR-MVSE (simLR-MVSE)



(b) E-MVSE70 vs. the simplified version of E-MVSE70 (simE-MVSE70)



(c) E-MVSE140 vs. the simplified version of E-MVSE140 (simE-MVSE140)

Figure 6: The DET curves of the LR-MVSE and E-MVSE systems and their simplified versions. The circles indicate the minimum DCFs.

Table 1. Summary of the minimum DCFs in Figures 4 and 6.

Methods	minDCF
MAP	0.0460
MAP + MVE	0.0434
MAP + LR-MVSE	0.0394
MAP + E-MVSE70	0.0413
MAP + E-MVSE140	0.0415
MAP + simLR-MVSE	0.0390
MAP + simE-MVSE70	0.0420
MAP + simE-MVSE140	0.0416

Table 2. Correlations of errors made by "MAP + LR-MVSE" and "MAP", where P and N denote the positive (target speaker) trial and the negative (impostor) trial, respectively. There are 2,025 P and 20,250 N in total.

Trials		MAP	
		Correct	Incorrect
MAP + LR-MVSE	Correct	1,296P + 19,918N	119P + 142N
	Incorrect	77P + 50N	533P + 140N

7. Conclusion

We have proposed a discriminative feedback adaptation (DFA) framework to improve the state of the art GMM-UBM speaker verification approach. The framework not only preserves the generalization ability of the GMM-UBM approach, but also reinforces the discrimination between H_0 and H_1. Our method is based on the minimum verification squared-error (MVSE) adaptation strategy, which is modified from the MVE training method so that only mis-verified training utterances are considered. Because a small number of mis-verified training samples may not be able to adapt a large number of model parameters, to implement DFA, we developed two adaptation techniques: the linear regression-based minimum verification squared-error (LR-MVSE) method and the eigenspace-based minimum verification squared-error (E-MVSE) method. In addition, we use a fast LLR scoring approach and the simplified version of LR-MVSE or E-MVSE to improve the efficiency and effectiveness of the DFA framework. The results of experiments conducted on the NIST2001-SRE database show that the proposed DFA framework can substantially improve the performance of the conventional GMM-UBM approach.

8. References

[1] D. A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", *Speech Communication*, vol.17, pp. 91-108, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[3] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, New Jersey, 2001.

[4] J. L. Gauvain and C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291-298, 1994.

[5] B. H. Juang, W. Chou, and C. H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 257-265, 1997.

[6] C. Y. Ma and E. Chang, "Comparison of Discriminative Training Methods for Speaker Verification", in *Proc. ICASSP2003*.

[7] A. E. Rosenberg, O. Siohan, and S. Parthasarathy, "Speaker Verification Using Minimum Verification Error Training", in *Proc. ICASSP1998*.

[8] W. Chou and B. H. Juang, *Pattern Recognition in Speech and Language Processing*, CRC Press, 2003.

[9] R. Chengalvarayan, "Speaker Adaptation Using Discriminative Linear Regression on Time-Varying Mean Parameters in Trended HMM", *IEEE Signal Processing Letters*, vol. 5, no. 3, pp. 63-65, 1998.

[10] J. Wu and Q. Huo, "Supervised Adaptation of MCE-Trained CDHMMs Using Minimum Classification Error Linear Regression", in *Proc. ICASSP2002*.

[11] X. D. He and W. Chou, "Minimum Classification Error Linear Regression for Acoustic Model Adaptation of Continuous Density HMMs", in *Proc. ICASSP2003*.

[12] X. D. He and W. Chou, “Minimum Classification Error (MCE) Model Adaptation of Continuous Density HMMs”, in *Proc. Eurospeech2003*.

[13] F. Valente and C. Wellekens, “Minimum Classification Error/Eigenvoices Training for Speaker Identification”, in *Proc. ICASSP2003*.

[14] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong, “The Use of Cohort Normalized Scores for Speaker Verification”, in *Proc. ICSLP1992*.

[15] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid Speaker Adaptation in Eigenvoice Space”, *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 695-707, 2000.

[16] O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua, “Speaker Identification and Verification Using Eigenvoices” , in *Proc. ICSLP2000*.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd. ed., John Wiley & Sons, New York, 2001.

[18] G. Strang, *Linear Algebra and Its Applications*, 4th. ed., Brooks/Cole, 2005.

[19] http://www.nist.gov/speech/tests/spk/index.htm

[20] The VIMAS speech codec. http://www.vimas.com

[21] J. Pelecanos and S. Sridharan, “Feature Warping for Robust Speaker Verification”, in *Proc. Odyssey2001*.

[22] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET Curve in Assessment of Detection Task Performance”, in *Proc. Eurospeech1997*.

[23] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, “NIST and NFI-TNO Evaluations of Automatic Speaker Recognition”, *Computer Speech and Language*, vol. 20, pp. 128-158, 2006.

[24] L. Gillick and S. J. Cox, “Some Statistical Issues in the Comparison of Speech Recognition Algorithms”, in *Proc. ICASSP1989*.