DISCRIMINATIVE FEEDBACK ADAPTATION FOR GMM-UBM SPEAKER VERIFICATION

Yi-Hsiang Chao1,2, Wei-Ho Tsai3, and Hsin-Min Wang1

1 Institute of Information Science, Academia Sinica, Taipei
2 Department of Computer Science, National Chiao Tung University, Hsinchu
3 Department of Electronic Engineering, National Taipei University of Technology, Taipei

ABSTRACT

The GMM-UBM system is the current state-of-the-art approach for text-independent speaker verification. The advantage of the approach is that both target speaker model and 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 impostor’s 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 preserves the generalization ability of the GMM-UBM approach. This is done by adapting the UBM to a target-speaker-dependent anti-model based on a minimum verification squared-error criterion, rather than estimating 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.

Index Terms—Discriminative feedback adaptation, log-likelihood ratio, minimum verification squared-error linear regression, 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:

\[ H_0: \text{U was spoken by the target speaker,} \]
\[ H_1: \text{U was not spoken by the target speaker.} \]

The LLR test can be expressed as

\[ L(U) = \log p(U | \lambda) - \log p(U | \bar{\lambda}) \geq \theta \quad \text{accept} \ H_0 \]
\[ L(U) < \theta \quad \text{accept} \ H_1 , \]

where \( \theta \) is a decision threshold; \( \lambda \) is the target speaker model; and \( \bar{\lambda} \) is the so-called anti-model or impostor model. Both \( \lambda \) and \( \bar{\lambda} \) 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 \( \lambda \) and \( \bar{\lambda} \). This approach pools all speech data from a large number of background speakers to form a universal background model (UBM) [2] as \( \bar{\lambda} \) via the expectation-maximization (EM) algorithm [3], and then adapts the UBM to \( \lambda \) via the maximum a posteriori (MAP) estimation [2] technique. GMM-UBM is effective because its generalization ability allows \( \lambda \) to handle “unseen” acoustic patterns, even when the amount of training data from the target speaker is very limited. However, since \( \lambda \) and \( \bar{\lambda} \) are trained according to separate criteria, the optimization procedure can not distinguish a target speaker from background speakers optimally. In particular, GMM-UBM uses a common UBM \( \bar{\lambda} \) for all target speakers, it tends to be weak in rejecting impostors’ voices that are similar to the target speaker’s voice. Moreover, as \( \lambda \) is derived from \( \bar{\lambda} \), both models may correspond to a similar probability distribution. To improve the GMM-UBM approach, we propose a discriminative feedback adaptation (DFA) framework that allows generalization and discrimination to be considered jointly.

Although several discriminative training methods [4], such as the minimum classification error (MCE) training method [5], the minimum verification error (MVE) training method [6], and the maximum mutual information (MMI) training method [7], have been proposed, they tend to over-train a model if the amount of training data is insufficient. In contrast, the DFA framework is based on the minimum verification squared-error (MVSE) adaptation strategy, which is modified from the MVE training method. The DFA framework regards the target speaker model and the UBM, both of which are 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, it does not destroy the generalization ability of the two models even if they are updated iteratively until convergence. Because a small number of mis-verified training samples may not be able to adapt a large number of model parameters, we propose the minimum verification squared-error linear regression (MVSELR) adaptation method to implement DFA.

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed DFA framework. Section 3 describes the proposed MVSELR adaptation technique used to implement DFA. Section 4 details the experiment results. Then, in Section 5, we summarize our conclusions.

2. DISCRIMINATIVE FEEDBACK ADAPTATION

Fig. 1 shows the block diagram of the proposed discriminative feedback adaptation (DFA) framework, which can be divided into two phases. The first phase, indicated by the dotted 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, 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 resolve this problem, DFA tries to generate a discriminative anti-model exclusively for each target speaker by using the negative samples from the cohort \( \lambda \) of each target speaker to adapt both \( \lambda \) and \( \overline{\lambda} \). Since the models may affect each other, the DFA framework also uses the positive samples to avoid increasing the miss probability while reducing the false alarm probability. The resulting \( \lambda \) and \( \overline{\lambda} \) are then updated iteratively. Second, since the DFA framework only uses mis-verified training samples as adaptation data in each iteration, it actually fine-tunes the model parameters according to 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 develop the following algorithms.

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

The minimum verification error (MVE) training method [6] is modified to fit our requirement that only mis-verified training samples should be considered. We call it the minimum verification squared-error (MVSE) adaptation strategy. The goal of DFA is to minimize the overall expected loss \( D \) defined as

\[
D = x_0 \ell_0 + x_1 \ell_1,
\]

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

\[
\ell_i = \frac{1}{N_i} \sum_{U \in \Omega_i} s(d(U)),
\]

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

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

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

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

\[
s(d(U)) = \begin{cases} a(d(U) - b)^2 & \text{if } d(U) > b \\ 0 & \text{if } d(U) \leq b, \end{cases}
\]

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 \( d(U) > b \), and \( s(d(U)) \) is a response squared-error value. Fig. 2 contrasts the curve of the \( s \) function with that of the well-known sigmoid function. If \( d(U) \leq b \), 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 parameters. In contrast, as the value of \( d(U) \) increases, the sigmoid function used in MVE [6] will become flat, and the obtained gradient will approximate zero. As a result, the mis-verified utterance \( U \) will not contribute to model adaptation. In addition, the proposed DFA framework differs from the conventional MVE training method because the latter always updates the model parameters if the value of the sigmoid function is not 0 or 1.

![Fig. 2](image_url)

**2.2. Fast scoring for DFA**

To speed up DFA, we use a fast scoring approach [2] to compute the LLR. Given an utterance \( U = \{a_1, \ldots, a_T\} \), the computation of LLR can be written as

\[
L(U) = \frac{1}{T} \sum_{t=1}^{T} \log \sum_{m=1}^{M} a_m p(o_t | g_m) - \log \sum_{m=1}^{M} a_m p(o_t | \overline{g}_m)
\]

\[
eq \frac{1}{T} \sum_{i=1}^{C} \log \sum_{m=1}^{M} a_{C(i)} p(o_t | g_{C(i)}) - \log \sum_{m=1}^{M} a_{C(i)} p(o_t | \overline{g}_{C(i)})
\]

where \( g_m \) and \( \overline{g}_m \) are the \( m \)-th Gaussian mixture components of the target speaker model and the anti-model, respectively; and \( a_m \) is a mixture weight, \( m = 1, \ldots, M \). 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, \( C(i), i = 1, \ldots, C \), in the UBM; 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. MINIMUM VERIFICATION SQUARED-ERROR LINEAR REGRESSION (MVSELR) ADAPTATION**

Because 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 the minimum verification squared-error linear regression (MVSELR) adaptation method to implement DFA. MVSELR is motivated by the minimum classification error linear regression (MCELIR) techniques [9-11].
that have been used in speech recognition applications. We assume the initial target speaker model \( \hat{\lambda}^{(0)} \) and the anti-model \( \overline{\lambda}^{(0)} \) have \( M \) Gaussian mixtures \( \hat{x}_m^{(0)} \sim N(\mu_m^{(0)}, \Sigma_m) \) and \( \overline{x}_m^{(0)} \sim N(\overline{\mu}_m^{(0)}, \overline{\Sigma}_m) \), respectively, where \( \mu_m^{(0)} \) and \( \overline{\mu}_m^{(0)} \) are \( r \)-dimensional mean vectors obtained with the GMM-UBM method; and \( \Sigma_m \) is an \( r \times r \) covariance matrix of the UBM. We assume MVSELR, the new mean vectors of the target speaker model and the anti-model take the following respective forms:

\[
\mu_m = \mathbf{W} \hat{x}_m^{(0)}
\]

and

\[
\overline{\mu}_m = \overline{\mathbf{W}} \overline{x}_m^{(0)},
\]

where \( \mathbf{W} \) and \( \overline{\mathbf{W}} \) are \( r \times (r+1) \) transformation matrices; and

\[
\hat{x}_m^{(0)} = [1 \, \mathbf{W}_m^{(0)T}]^T \quad \text{and} \quad \overline{x}_m^{(0)} = [1 \, \overline{\mathbf{W}}_m^{(0)T}]^T.
\]

Given initial transformation matrices \( \mathbf{W}_m^{(0)} = \overline{\mathbf{W}}_m^{(0)} = [0 \, I] \), where \( I \) is an \( r \times 1 \) zero vector and \( I \) is an \( r \times r \) identity matrix, the parameters \( \mathbf{W} \) and \( \overline{\mathbf{W}} \) can be iteratively optimized using

\[
\mathbf{W}^{(k+1)} = \mathbf{W}^{(k)} - \eta \frac{\partial D}{\partial \mathbf{W}^{(k)}}
\]

and

\[
\overline{\mathbf{W}}^{(k+1)} = \overline{\mathbf{W}}^{(k)} - \eta \frac{\partial D}{\partial \overline{\mathbf{W}}^{(k)}},
\]

respectively, where \( \eta \) is the step size; and

\[
\frac{\partial D}{\partial \mathbf{W}^{(k)}} = x_0 \frac{\partial \ell_0}{\partial \mathbf{W}^{(k)}} + \frac{\partial \ell_1}{\partial \mathbf{W}^{(k)}} + \frac{\partial \ell_2}{\partial \mathbf{W}^{(k)}} + \frac{\partial \ell_3}{\partial \mathbf{W}^{(k)}}
\]

\[
= x_0 \frac{1}{N_1} \sum_{i=1}^{N_1} \left( \frac{2a_i \cdot (-L(U) - b)}{\ell_0} \right) \frac{\partial \ell_0}{\partial \mathbf{W}^{(k)}}
\]

\[
+ x_1 \frac{1}{N_1} \sum_{i=1}^{N_1} \left( \frac{2a_i \cdot (L(U) - b)}{\ell_1} \right) \frac{\partial \ell_1}{\partial \mathbf{W}^{(k)}}
\]

\[
+ x_2 \frac{1}{N_1} \sum_{i=1}^{N_1} \left( \frac{2a_i \cdot (L(U) - b)}{\ell_2} \right) \frac{\partial \ell_2}{\partial \mathbf{W}^{(k)}}
\]

\[
+ x_3 \frac{1}{N_1} \sum_{i=1}^{N_1} \left( \frac{2a_i \cdot (L(U) - b)}{\ell_3} \right) \frac{\partial \ell_3}{\partial \mathbf{W}^{(k)}}
\]

for the \( i \)-th speaker’s utterances, \( U_i \), and the \( j \)-th speaker’s utterances, \( U_j \), respectively. For each cohort speaker, we extracted six 3-second speech segments from his/her training utterances as negative samples. Thus, each target speaker had 60 negative samples in total. All the 3-second segments extracted from each target speaker’s training utterances served as positive samples for MVSELR adaptation. To remove silence-noise frames, we processed all the speech data with a Voice Activity Detector (VAD) [13]. 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 (MFCC) [3] and their first time derivatives. To compensate for channel mismatch effects, we applied feature warping [14] 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. The MVSELR adaptation procedure was trained until it 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 \( C_{\text{Min}} \times P_{\text{Target}} \) and \( C_{\text{FalseAlarm}} \times (1 - P_{\text{Target}}) \), respectively, according to the NIST Detection Cost Function (DCF) [12]:

\[
C_{\text{DET}} = C_{\text{Min}} \times P_{\text{Min}} \times P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{\text{FalseAlarm}} \times (1 - P_{\text{Target}}),
\]

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

### 4.1. Experiment results

To evaluate the performance of the DFA framework, we use the Detection Error Tradeoff (DET) curve [15] and the NIST DCF, which reflects the performance at a single operating point on the DET curve. For the performance comparison, we use the GMM-UBM method (denoted as “MAP”) as our baseline. In addition to the proposed MVSELR method (denoted as “MAP+MVSELR”), we also implement the conventional MVE training method with the sigmoid function (denoted as “MAP+MVE”). Both methods use the target speaker GMM and the UBM obtained from the GMM-UBM method as initial models.

Fig. 3 shows the results of experiments conducted on 2,038 target speaker trials and 20,380 impostor trials in the evaluation set. From the figure, we observe that “MAP+MVSELR” outperforms both “MAP” and “MAP+MVE”. Note that the performance of “MAP+MVE” is not absolutely better than that of “MAP” because it tends to over-train the well-trained models obtained from the GMM-UBM method with the correctly-verified training utterances. Moreover, it is difficult to select the optimal stopping point in MVE training. Table 1 summarizes the minimum DCF of each system. Significantly, “MAP+MVSELR” achieves a 14.35% relative improvement over the baseline GMM-UBM system (“MAP”) and a 9.22% relative improvement over the “MAP+MVE” method.

### 5. CONCLUSIONS

We have proposed a discriminative feedback adaptation (DFA) framework to improve the state-of-the-art GMM-UBM approach for speaker verification. The proposed framework not only preserves the generalization ability of the GMM-UBM approach, but also reinforces the discrimination between \( H_0 \) and \( H_1 \). Our approach is based on the minimum verification squared-error (MVSE) adaptation strategy, which is modified from the MVE training method. Because a small number of mis-verified training samples may not be able to adapt a large number of model parameters, we propose the minimum verification squared-error linear regression (MVSELR) adaptation method to implement DFA. In addition, we use a fast LLR scoring approach to speed up 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.

### 6. REFERENCES


