A New Adaptation Approach to High-Level Speaker-Model Creation in Speaker Verification

Shi-Xiong Zhang and Man-Wai Mak

Center for Multimedia Signal Processing,
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
zhang.sx@alumni.polyu.edu.hk, enmwmak@polyu.edu.hk

Abstract

Research has shown that speaker verification based on high-level speaker features requires long enrollment utterances to guarantee low error rate during verification. However, in practical speaker verification, it is common to model speakers based on a limited amount of enrollment data, which will make the speaker models unreliable. This paper proposes four new adaptation methods for creating high-level speaker models to alleviate this undesirable effect. Unlike conventional methods in which only the phoneme-dependent background model is adapted, the proposed adaptation methods also adapts the phoneme-independent speaker model to fully utilize all the information available in the training data. A proportional factor, which is derived from the ratio between the phoneme-dependent background model and the phoneme-independent background model, is used to adjust the phoneme-independent speaker models during adaptation. The proposed method was evaluated under the NIST 2000 and NIST 2002 SRE frameworks. Experimental results show that the proposed adaptation method can alleviate the data-sparseness problem effectively and achieves a better performance when compared with traditional MAP adaptation.

Key words: speaker verification, high-level features, model adaptation, maximum-a-posterior (MAP) adaptation

1 Introduction and Motivation

In most text-independent speaker verification systems, short-term spectra of speech signals are extracted to train speaker-dependent Gaussian mixture models (GMMs). To enhance the discrimination between the client (target) speakers and impostors, the distribution of impostors' speech is represented by a GMM-based background model [1]; verification decisions are then based on likelihood-ratio hypothesis tests in which the client and background GMMs represent the distribution of the null and alternative hypotheses, respectively. The background model can be trained using the speech of non-target speakers from large speech corpora. Therefore, collecting sufficient amount of speech for training a background model is not a problem. However, it is difficult to request a user to provide a large amount of speech for enrollment, because this will impose too much burden on the user.

To address this problem, adaptation techniques such as maximum a posteriori (MAP) [2], maximum-likelihood linear regression (MLLR) [3, 4], and speaker clustering [5] have been proposed for creating low-level acoustic speaker models from a moderate amount of client data [1, 6]. When client utterances are extremely short (e.g., a few seconds) and verification is text-dependent, it is possible to create phoneme-dependent HMMs for each client speaker by adapting a universal phoneme-dependent HMM [7]. When the verification task is text-independent, it has been shown that creating speaker models by linearly combining several reference models in an eigenvoice (EV) [8] space can achieve good performance [9]. The EV adaptation has been extended to the eigenspace MLLR (EMLLR) [10]. In EMLLR, an eigenspace is derived from the MLLR transformations of a set of speaker-dependent (SD) models; a speaker is then represented by a point in the speaker space spanned by the leading eigenvectors of the MLLR-eigenspace. To introduce nonlinearity to the adaptation, EMLLR has been extended to kernel eigen-space MLLR (KEMLLR) [11] and its fast version called embedded KEMLLR (eKEMLLR) [12]. The idea is to replace linear PCA in EMLLR adaptation by kernel PCA in a way analogous to kernel eigenvoice (KEV) adaptation [11].

It has been shown that KEMLLR outperforms other adaptation methods when the amount of enrollment data is extremely limited (e.g., 2s enrollment utterances for speaker-dependent GMMs) and that when a small amount of enrollment data is available (e.g., 32s enrollment utterances for speaker-dependent GMMs), MAP is a better candidate for creating speaker models [13]. Comparison studies in [6] also show that MAP is the best adaptation method for the NIST99 database. Therefore, in this paper we will compare our new adaptation method with MAP.

Recently, to improve the robustness of speaker verification systems, researchers have started to investigate the possibility of using long-term, high-level features to characterize speakers. The idea is based on the observation that humans rely not only on the low-level acoustic information but also on some high-level information to recognize speakers. There is convincing evidence supporting this idea. For example, studies in speech prosody have shown that individual speakers exhibit substantial differences in voluntary speaking behaviors such as lexicon, prosody, intonation, pitch range, and pronunciation [14, 15]. Studies in linguistics have shown that speaking styles (e.g., read speech versus spontaneous speech) have significant effect on pronunciation patterns [16]. Kuehn and Moll [17] measured the velocity and displacement of the tongue during speech production and found appreciable variation of these two measurements among different speakers. Shaiman et al. [18] used X-ray to capture the movement of the upper lip and jaw and found substantial speaker-dependent patterns in the articulator coordination.

The use of long-term or high-level features for automatic speaker recognition was pioneered by Doddington [19] and the SuperSID project [20]. These works have led to extensive investigations into high-level features, in which prosodic features [18,21–25], pronunciation features [26–30], and idiolect features [19,31] were proposed and combined with acoustic features. The results show that there is significant benefit of fusing high- and low-level features for speaker verification. Among the high-level features investigated, the conditional pronunciation modeling (CPM) technique [30] that extracts multilingual phone sequences from utterances achieves the best performance [20]. One limitation of the CPM in [30] is that it requires multi-lingual corpora to build speaker and background models. To overcome this limitation, Leung et al. [32] proposed using articulatory feature (AF) streams to construct CPM and called the resulting models AFCPM. It was found in [32] that AFCPM can reduce the error rate of conventional CPM by 25%. The state-of-the-art high-level features for speaker verification and their modeling methods are summarized in Table 1.

One problem of using high-level features is that a large amount of speech data is required to create reliable speaker models. As a result, data-sparseness can cause serious problems in high-level speaker verification. Unlike the low-level acoustic GMM speaker models where plenty of adaptation methods have been proposed and evaluated, adaptation of high-level speaker models has largely remained unexplored. The closest method is the MAP adaptation of phonetic N-gram speaker models in [34] and language models in [35]. Leung et al. [32] have shown in their articulatory feature-based pronunciation model (AFCPM) that high-level speaker models can be created by using MAP adaptation. However, the client models that they created are essentially a linear weighted sum of enrollment data's distribution and background models. It was found that the modeling capability of the AFCPMs drops rapidly when the amount of enrollment data decreases [36, 37].

To alleviate the above problem, this paper proposes to adapt not only the phoneme-dependent background models but also the phoneme-independent speaker models to create client speaker models. A scaling factor, which is derived from the ratio between the phoneme-dependent background model and the phoneme-independent background model, will be used to adjust the phoneme-independent speaker models during adaptation. The results show that the proposed adaptation method, which uses as much information as possible from the training data, significantly outperforms the classical MAP adaptation method. It was also found that the new adaptation approach can effectively alleviate the data sparseness problem in phoneme-dependent AFCPMs, resulting in a significantly lower error rate.

The paper is organized as follows. In Section 2, a high-level speaker verification system that is based on the phoneme-dependent articulatory feature-based conditional pronunciation models (AFCPMs) is introduced. Then, four new adaptation methods for creating AFCPM speaker models are proposed and discussed in Section 3. Section 4 outlines the scoring procedure during verification sessions. In Section 5, experimental evaluations on all proposed adaptation methods are presented and compared.

2 Phoneme-dependent AFCPM

Articulatory features (AFs) are representations describing the movements or positions of different articulators during speech production. In [32,36], manner and place of articulation were used for pronunciation modeling. The manner and place properties are shown in Table 2. AFs can be automatically determined from speech signals using AF-based multilayer perceptrons (MLPs) shown in Figure 1.

Specifically, for each articulatory property, an AF-MLP takes 9 consecutive frames of MFCCs X_t as input to determine the output classes at frame t [32, 36]:

$$\begin{aligned} l_t^{\mathrm{M}} &= \arg\max_{m \in \mathcal{M}} P(\text{Manner} = m | X_t) \\ l_t^{\mathrm{P}} &= \arg\max_{p \in \mathcal{P}} P(\text{Place} = p | X_t). \end{aligned} \tag{1}$$

The two AF streams—one from the manner MLP and another from the place MLP—for creating the conditional pronunciation models are formed by concatenating l_t^{M}'s and l_t^{P}'s for $t = 1, \ldots, T$, where T is the total number of frames in the utterance. See [32] for a detailed description of the AFCPM approach.

In phoneme-dependent AFCPMs, N phoneme-dependent universal background models (UBMs) are trained from the AF and phoneme streams of a large number of speakers to represent the speaker independent pronunciation characteristics. Each UBM comprises the joint probabilities of the manner and place classes conditioned on a phoneme. The training procedure begins with aligning two AF streams (l_t^{M} and l_t^{P}) obtained from the AF-MLPs and a phoneme sequence q_t obtained from a null-grammar recognizer. The joint probabilities corresponding to a particular phoneme q is given by

$$\begin{aligned} P_b(m,p|q) &= P_b(\text{Manner}=m, \text{Place}=p|\text{Phoneme}=q, \text{Background}) \\ &= \frac{\#((m,p,q) \text{ in the data of all background speakers})}{\#((*,*,q) \text{ in the data of all background speakers})}, \end{aligned} \tag{2}$$

where $m \in \mathcal{M}, p \in \mathcal{P}, (m,p,q)$ denotes the condition for which Manner $=$ m, Place $= p$, and Phoneme $= q$, $*$ represents all possible members in that class, and $\#()$ represents the total number of frames with phoneme labels and AF labels fulfill the description inside the parentheses.

The unadapted speaker models $P_s(m,p|q)$ are created in the same way:

$$\begin{aligned} P_s(m,p|q) &= P_s(\text{Manner}=m, \text{Place}=p|\text{Phoneme}=q, \text{speaker}=s) \\ &= \frac{\#((m,p,q) \text{ in the enrollment utterance of speaker } s)}{\#((*,*,q) \text{ in the enrollment utterance of speaker } s)}. \end{aligned} \tag{3}$$

We can see for each phoneme, a total of 60 probabilities can be obtained. These probabilities are the products of 6 manner classes and 10 place classes. The procedure of creating a phoneme-dependent AFCPM speaker model is illustrated in Figure 1. However, this naive approach can result in many zero entries in the probability mass functions, primarily because of the data sparseness problem. To overcome this problem, this paper proposes several new adaptation and model creation methods.

3 Adaptation Methods for AFCPMs

Here, we firstly review the classical MAP adaptation and then propose four MAP-based adaptation methods that use as much information from training data as possible.

The four adaptation methods investigated in this paper are summarized as follows:

Method A: *Classical MAP.* Adapted from phoneme-dependent background models, $P_b(m,p|q)$. This is based on the classical MAP used in [32].

Method B: *Phoneme-independent adaptation (PIA).* Adapted from phoneme-dependent speaker models $P_s(m,p|q)$ and phoneme-independent speaker models $P_s(m,p|*)$.

Method C: *Scaled phoneme-independent adaptation (SPI).* Adapted from phoneme-independent speaker models $P_s(m,p|*)$ with a phoneme-dependent scaling factor that depends on both the phoneme-dependent and phoneme-independent background models.

Method D: *Mixed phoneme-dependent and scaled phoneme-independent adaptation (MSPI).* Adapted from phoneme-dependent background models $P_b(m, p|q)$ and phoneme-independent speaker models $P_s(m, p|*)$ with a phoneme-dependent scaling factor that depends on both the phoneme-dependent and phoneme-independent background models. This method is a combination of Methods A and C.

Method E: *Mixed phoneme-independent and scaled phoneme-dependent adaptation (MSPD).* Adapted from phoneme-independent speaker models $P_s(m, p|*)$ and phoneme-dependent background models $p_b(m, p|q)$ with a speaker-dependent scaling factor that depends on both the phoneme-independent speaker model and background models. This method is a combination of Methods B and C.

Figure 2 illustrates how these five adaptation methods use the available information from training data. Note that Method A is treated as the baseline, and Methods B to E are the four proposed methods.

3.1 Classical MAP (Method A) and Its Limitations

For discrete probability models, MAP adaptation can be viewed as count merging or model interpolation [38]. Following the N-gram language model adaptation in [35], we assume that the prior distribution of the model parameters is the Dirichlet density. This assumption leads to the adaptation formula:

$$\widehat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) P_b(m, p|q) \tag{4}$$

where $\beta_s^q \in [0, 1]$ is a phoneme-dependent adaptation coefficient controlling the contribution of the enrollment data and the background models (Eq. 2) on the MAP-adapted model. It is obtained by

$$\beta_s^q = \frac{\#((*, *, q) \text{ in the enrollment utterances of speaker } s)}{\#((*, *, q) \text{ in the enrollment utterances of speaker } s) + r_\beta} \tag{5}$$

where r_β is a fixed relevance factor common to all phonetic classes and speakers. Figure 3 illustrates the procedure of applying MAP adaptation (Method A) for speaker-model creation.

The relationship between the adapted, unadapted, and background models is illustrated in Figure 4. These models are projected onto the first two principal axes [39] in the model space. When enrollment data is sufficient, MAP adaptation can create client models that capture the phoneme-dependent characteristics of speakers. However, when the amount of enrollment data is limited, this speaker-model creation method may have three fundamental problems:

Problem 1: The method will make the client models of the same phoneme very close to the background model of that phoneme (see Figure 4), even though the clients may have very different pronunciation characteristics. This will cause the client models fail to discriminate the true speakers from the imposters.

Problem 2: The method does not fully utilize the information available in the training data.

Problem 3: The method imposes too much constraint on the adaptation.

Problem 1 is further exemplified in Figure 5, which shows that the adapted models (Figs. 5(d) and 5(e)) of two speakers look very similar because they are very similar to the background model. This will make the speaker models fail to discriminate the true speakers from impostors.

For Problem 2, Method A only uses two out of four possible unadapted speaker and background models for adaptation. Figure 2 shows the four possible models from which the target models can be adapted. Method A uses the phoneme-dependent models only and ignores the fact that the phoneme-independent models ($P_b(m, p|*)$ and $P_s(m, p|*)$) can also be used to create target speaker models.

For Problem 3, Method A uses all of the background speakers' data to train phoneme-dependent background models from which phoneme-dependent target speaker models are created by MAP adaptation. Creating a phoneme-dependent speaker model from the corresponding phoneme-dependent background model means that the resulting speaker model is constrained by the articulatory properties of a single phoneme. In other words, the method does not allow cross-phoneme adaptation. Note that the classical MAP adaptation for acoustic GMMs does not have such a hard constraint. Instead, a soft constraint is implicitly imposed by the posterior probabilities of the mixture components.

3.2 New Adaptation Methods for AFCPMs

Our new adaptation methods attempt to utilize all of the available information from the training data. To relax the constraint imposed by classical MAP adaptation (see Problem 3 in Section 3.1), we introduce phoneme-independent models for target speakers and background speakers as follows:

$$
\begin{aligned}
P_b(m,p|*) &= \Pr(\text{Manner} = m, \text{Place} = p|\text{Background}) \\
&= \frac{\#((m,p,*) \text{ in the data of all background speakers})}{\#((*,*,*) \text{ in the data of all background speakers})}, \qquad (6)\\
P_s(m,p|*) &= \Pr(\text{Manner} = m, \text{Place} = p|\text{speaker} = s) \\
&= \frac{\#((m,p,*) \text{ in the enrollment utterance of speaker } s)}{\#((*,*,*) \text{ in the enrollment utterance of speaker } s)}, \qquad (7)
\end{aligned}
$$

where $m \in \mathcal{M}, p \in \mathcal{P}$ are defined in Section 2, and $(m,p,*)$ denotes the condition for which Manner $= m$ and Place $= p$. Based on the definition of $P_s(m,p|*), P_b(m,p|*)$ and $P_b(m,p|q)$, we can further derive (see Appendix):

$$
\begin{aligned}
P_s(m,p|*) &= \sum_{i=1}^{46} P_s(m,p|q^{(i)})P_s(q^{(i)}), \\
P_b(m,p|q) &= \sum_{k=1}^{M} P_{b_k}(m,p|q)P(b_k|q), \qquad (8)\\
P_b(m,p|*) &= \sum_{i=1}^{46} P_b(m,p|q^{(i)})P_b(q^{(i)}),
\end{aligned}
$$

where M is the number of background speakers, b_k is one of these background speakers, $q^{(i)}$ represents one of the 46 phonemes in English, $P(b_k|q)$ is the conditional probability:

$$
P(b_k|q) = \frac{\#((*,*,q) \text{ in the utterances of background speaker } b_k)}{\#((*,*,q) \text{ in the utterances of all background speakers})},
$$

and $P_s(q^{(i)})$ is the probability of phoneme $q^{(i)}$:

$$
P_s(q^{(i)}) = \frac{\#((*,*,q) \text{ in the utterances of speaker } s)}{\#((*,*,*) \text{ in the utterances of speaker } s)}.
$$

Figure 6 illustrates how the phoneme-independent models are used for creating speaker models, which will be discussed next.

3.2.1 Method B: Phoneme-independent adaptation (PIA)

Instead of adapting from the phoneme-dependent UBM, we can create the speaker model $\widehat{P}_s(m,p|q)$ by adapting the phoneme-independent speaker model $P_s(m,p|*)$, i.e.,

$$
\widehat{P}_s(m,p|q) = \beta_s^q P_s(m,p|q) + (1-\beta_s^q)P_s(m,p|*). \qquad (9)
$$

Figure 7 illustrates the relationship (based on real data) between the unadapted and adapted speaker models created by this method. During the adaptation, the unadapted phoneme-dependent speaker models $P_s(m, p|q)$ (represented by “blue cross” and “black plus” in Figure 7) will move towards their corresponding phoneme-independent speaker models $P_s(m, p|*)$ (represent by “blue circle” and “black circle”). As a result, the adapted phoneme-dependent speaker models (represented by “green square” and “red diamond”) will be created according to how much phoneme-dependent data the speaker possesses. The advantage of this method is that all of the unadapted phoneme-dependent models $P_s(m, p|q)$ will move towards their respective phoneme-independent models (“blue circle” and “black circle”) instead of towards a single background model as in MAP method. Therefore, for a given phoneme, the adapted speaker models of different speakers created by Method B will not concentrate in one place of the model space.

Figure 12(b) shows all of the 46 adapted phoneme-dependent and phoneme-independent speaker models of speaker 1018 and 1042. Evidently, because the speaker models were adapted from different phoneme-independent speaker models, the adapted models belonging to the two speakers are well separated.

While this method can help solve Problems 1 and 3 mentioned in Section 3.1, it does have its own problem. The problem is that for a particular client, all of his/her phoneme-dependent models are adapted from the same phoneme-independent model, causing loss of phoneme-dependence in the client model. In fact, the method uses enrollment data only, as illustrated in Figure 2. This loss of phoneme-dependence, however, violates the requirement of the scoring procedure (see Section 4) where the speaker and background models are assumed to be phoneme-dependent. Fortunately, the phoneme-dependence in the client models can be easily retained by introducing a phoneme-dependent scaling factor in the adaption equation. This is to be discussed next.

3.2.2 Method C: Scaled phoneme-independent adaptation (SPI)

In this method, a phoneme-dependent scaling factor $\frac{P_b(m,p|q)}{P_b(m,p|*)}$ is added to the adaptation formula in Eq. 9, yielding

$$\widehat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) \left[\frac{P_b(m, p|q)}{P_b(m, p|*)} P_s(m, p|*) \right], \qquad (10)$$

where $P_b(m, p|*)$ represents the phoneme-independent background model. With this factor, the model to be adapted becomes

$$f_s^q = \frac{P_b(m, p|q)}{P_b(m, p|*)} P_s(m, p|*). \qquad (11)$$

Therefore, the resulting target model $\widehat{P}_s(m,p|q)$ is now adapted from a model with certain degree of phoneme-dependence instead of adapting from a purely phoneme-independent model $(P_s(m,p|*))$.

Note that f_s^q in Eq. 11 can also be written as $\frac{P_s(m,p|*)}{P_b(m,p|*)}P_b(m,p|q)$. In that case, we can interpret $\frac{P_s(m,p|*)}{P_b(m,p|*)}$ as a phoneme-independent scaling factor for the classical MAP adaptation in Eq. 4. This factor can help alleviates Problems 2 and 3 in classical MAP mentioned earlier, because it implicitly incorporates the speaker-dependent articulatory properties of other phonemes into the adaptation equation.

More interestingly, using Eq. 8, f_s^q in Eq. 11 can be written as:

$$\frac{P_b(m,p|q)}{P_b(m,p|*)}P_s(m,p|*) = \frac{\left[\sum_{k=1}^{M} P_{b_k}(m,p|q)P(b_k|q)\right]\cdot\left[\sum_{i=1}^{46} P_s(m,p|q^{(i)})P_s(q^{(i)})\right]}{\sum_{k=1}^{M}\sum_{i=1}^{46} P_{b_k}(m,p|q^{(i)})P(b_k|q)P_b(q^{(i)})} \tag{12}$$

where M is the number of background speakers, b_k is one of these background speakers, and $q^{(i)}$ represents one of the 46 phonemes in English. If we assume $P_b(q^{(1)}) = \cdots = P_b(q^{(46)}) = P_s(q^{(1)}) = \cdots = P_s(q^{(46)}) =$ constant and $P(s_1|q) = \cdots = P(s_M|q) = P(s_1|q^{(i)}) = \cdots = P(s_M|q^{(i)}) =$ constant $\forall\ i$, then we have

$$\frac{P_b(m,p|q)}{P_b(m,p|*)}P_s(m,p|*) = \frac{\left[\sum_{k=1}^{M} P_{b_k}(m,p|q)\right]\cdot\left[\sum_{i=1}^{46} P_s(m,p|q^{(i)})\right]}{\sum_{k=1}^{M}\sum_{i=1}^{46} P_{b_k}(m,p|q^{(i)})}. \tag{13}$$

Eqs. 12 and 13 suggest that all of the available information have been harnessed during the adaptation process.

Figure 8 illustrates the projection of the unadapted and adapted speaker models on the first two principal axes. During this adaptation, we firstly used the phoneme-independent speaker models $P_s(m,p|*)$ (red dot), the phoneme-independent background model $P_b(m,p|*)$ (pink circle), and the phoneme-dependent background models $P_b(m,p|q)$ (purple circles) to generate the speaker-dependent phoneme-dependent term f_s^q (orange dashed circles). Then, the adapted phoneme-dependent speaker models (green square) will be produced by linearly combining $P_s(m,p|q)$ (blue cross) and f_s^q (yellow dashed circles). The advantage of this method is that each of the unadapted phoneme-dependent speaker models $P_s(m,p|q)$ will move towards a different position which is dependent on the position of $P_s(m,p|*)$, $P_b(m,p|*)$, and $P_b(m,p|q)$. Therefore, using Method C, not only do the adapted models of different speakers become well separated, the phoneme-dependence can also be maintained. This argument is supported by Figure 12(c), which shows that the adapted phoneme-dependent models of speakers 1018 and 1042 do not overlap with each other.

3.2.3 Method D: Mixed phoneme-dependent and scaled phoneme-independent adaptation (MSPI)

It becomes clear that Method A is likely to impose too much constraint on the adaptation. Method B aims to relax such constraint by introducing a phoneme-independent model in its adaptation equation. However, the relaxation may be overdone so that the phoneme-dependent scaling factor in Method C is necessary to limit the loss of phoneme-dependence. Nevertheless, the target models created by Method C depend implicitly on the phoneme-dependent background models $P_b(m, p|q)$ through the scaling factor. To strengthen the dependence of these background models while allowing certain degree of phoneme-independence, we may combine Methods A and C, which results in Method D:

$$\widehat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) \left[\alpha_b^q P_b(m, p|q) + (1 - \alpha_b^q) \frac{P_b(m, p|q)}{P_b(m, p|*)} P_s(m, p|*) \right] \tag{14}$$

where, $\alpha_b^q \in [0, 1]$ is a phoneme-dependent adaptation coefficient. It is obtained by

$$\alpha_b^q = \frac{\#((*, *, q) \text{ in the utterances of all background speakers})}{\#((*, *, q) \text{ in the utterances of all background speakers}) + r_\alpha} \tag{15}$$

where r_α is a fixed relevance factor.

Figure 9 illustrates the relationship between different models in Method D, and Figure 10 explains why this method is better than Method A via an illustrative example. During adaptation, we firstly used the phoneme-independent speaker models $P_s(m, p|*)$ (red dot in Figure 9), the phoneme-independent background model $P_b(m, p|*)$ (red circle) and the phoneme-dependent background models $P_b(m, p|q)$ (purple circles) to generate the speaker-dependent phoneme-dependent term f_s^q in Eq. 11 (orange dashed circles). Then unlike Method C, the adapted phoneme-dependent speaker models $\widehat{P}_s(m, p|q)$ (green square) in this method was produced by double adaptation to further enhance the phoneme-dependence. During the first adaptation f_s^q (yellow dashed circles) and $P_b(m, p|q)$ (blue circles) were linearly combined to generate a new point ($\star$). Then, during the second adaptation, the new point and $P_s(m, p|q)$ (blue cross) were linearly combined. Therefore, in Method D, the adapted models of different speakers will not only be well separated but also keep the phoneme-dependence, which results in higher discriminative power.

Comparing Figures 5 and 10 reveals that the Euclidean distance and dissimilarity between the AFCPM models of speakers 1018 and 1042 become larger (the distance increases from 4.39 to 14.17 and the correlation coefficient reduces from 0.9966 to 0.8013). Therefore the two speakers will be better discriminated if Method D is used to create their model.

3.2.4 Method E: Mixed phoneme-independent and scaled phoneme-dependent adaptation (MSPD)

Using the same idea in Method D, the phoneme-independent speaker model and phoneme-dependent UBMs can be linearly combined first. The contribution of the latter is controlled by another scaling factor. The method is described mathematically as follows:

$$\widehat{P}_s(m,p|q) = \beta_s^q P_s(m,p|q) + (1-\beta_s^q)\left[\alpha_b^q P_b(m,p|q)\frac{P_s(m,p|*)}{P_b(m,p|*)} + (1-\alpha_b^q)P_s(m,p|*)\right] \tag{16}$$

where $\frac{P_s(m,p|*)}{P_b(m,p|*)}$ is a phoneme-independent scaling factor used for incorporating speaker-dependency into the phoneme-dependent UBM. The relationship between the unadapted and adapted models created by Method E is illustrated in Figure 11.

3.3 An Illustrative Example

Figure 12 shows the relationship between the phoneme-dependent background and adapted models (corresponding to 46 phonemes) of two speakers for Methods A to D. Apparently, Problem 1 in Method A mentioned in Section 3.1 does not appear in Method D.

4 Scoring Method

Following the scoring method in [1], the verification score of a test utterance $X = \{X_1, \ldots, X_t, \ldots, X_T\}$ is defined as:

$$S_{\mathrm{AF}}(X) = \sum_{t=1}^{T} \left(\log \widehat{p}_s(X_t) - \log p_b(X_t)\right), \tag{17}$$

where the speaker models $\widehat{P}_s(m,p|q)$ and background models $P_b(m,p|q)$ created by using different adaptation methods discussed in Section 3 are used to compute the scores:

$$\widehat{p}_s(X_t) = \widehat{P}_s(l_t^{\mathrm{M}}, l_t^{\mathrm{P}}|q_t) = \widehat{P}_s(\text{Manner} = l_t^{\mathrm{M}}, \text{Place} = l_t^{\mathrm{P}}|\text{Phoneme} = q_t, \text{Speaker} = s) \tag{18}$$

$$p_b(X_t) = P_b(l_t^{\mathrm{M}}, l_t^{\mathrm{P}}|q_t) = P_b(\text{Manner} = l_t^{\mathrm{M}}, \text{Place} = l_t^{\mathrm{P}}|\text{Phoneme} = q_t, \text{Background}). \tag{19}$$

In Eqs. 18 and 19, q_t is the phoneme of frame t in the test utterance recognized by a null-grammar phoneme recognizer, and l_t^{M} and l_t^{P} are the AF labels determined by the AF-MLPs [32].

For the acoustic GMM-UBM system [1], we applied several channel compensation techniques, including feature warping [40], Z-norm [41], short-time Gaussianization (STG) [42] and fast blind stochastic feature transformation (fBSFT) [43]. Acoustic scores $S_{\text{GMM-UBM}}$ were computed based on the log-likelihood ratio:

$$S_{\text{GMM-UBM}}(X) = \sum_{t=1}^{T} \left[\log p(\mathbf{x}_t|\Lambda_s) - \log p(\mathbf{x}_t|\Lambda_b)\right] \tag{20}$$

where Λ_s and Λ_b are the acoustic GMM of speaker s and the acoustic UBM, respectively.

To demonstrate the state-of-the-art acoustic speaker verification system can still be improved by high-level features, we also fused the scores obtained from AFCPMs and GMM-SVM [44]. For the GMM-SVM system, acoustic scores $S_{\text{GMM-SVM}}$ were computed based on the SVM framework [44]:

$$S_{\text{GMM-SVM}}(\text{utt}_c) = \alpha_0 K\left(\text{utt}_c, \text{utt}_s\right) - \sum_{i=1}^{M} \alpha_i K\left(\text{utt}_c, \text{utt}_{b_i}\right) + d, \tag{21}$$

where

$$K\left(\text{utt}_c, \text{utt}_s\right) = \sum_{i=1}^{N} \left(\sqrt{\lambda_i}\Sigma_i^{-\frac{1}{2}}\mathbf{m}_i^c\right)^{\mathrm{T}} \left(\sqrt{\lambda_i}\Sigma_i^{-\frac{1}{2}}\mathbf{m}_i^s\right) \tag{22}$$

is the GMM-supervector kernel [44]. λ_i and Σ_i are the mixture weights and covariances of UBM Gaussians, respectively. $\mathbf{m}_i^s$ and $\mathbf{m}_i^c$ are the mean of the i-th Gaussian belonging to speaker s and claimant c, respectively. utt_s represents the utterance pronounced by speaker s. α_0 is the lagrange multiplier corresponding to the target speaker, [1] and α_i $(i = 1, \ldots, M)$ are Lagrange multipliers (some of them may be zero) corresponding to the background speakers. M is the number of background speakers.

[1] Assuming one enrollment utterance per target speaker, which is the case in NIST00 and NIST02.

5 Experiments and Results

5.1 Speech Data

NIST99, NIST00, NIST01, NIST02, SPIDRE [45], and HTIMIT [46] were used in the experiments. [2] NIST99 and NIST01 were used for creating the background models, and NIST00 and NIST02 were used for creating speaker models and for performance evaluation. 3,794 utterances selected from HTIMIT were used to train the manner and place MLPs (see [32] for the architecture), and utterances from SPIDRE were used to train a null-grammar phoneme recognizer with 46 context-independent phoneme models (HMMs with 3 states, 16 mixtures per state). For NIST00 evaluation, the training part of NIST99 was used for creating phoneme-dependent AF-based UBMs. For NIST02 evaluation, the training part of NIST01 was used for creating the AF-based UBMs. The purposes of the databases used in this work are summarized in Table 3.

NIST00 contains landline telephone speech extracted from the SwitchBoard-II, Phase 1 and Phase 4 Corpus. The evaluation set comprises 457 male and 546 female target speakers, each with approximately 2 minutes of enrollment speech, and after silence removal, approximately 1 minute of speech remains. There are 3,026 female and 3,026 male verification utterances. Each verification utterance has length not exceeding 60 seconds and is evaluated against 11 hypothesized speakers of the same gender as the speaker of the verification utterance. This amounts to 6,096 speaker trials and 60,476 impostor attempts.

NIST02 contains cellular telephone speech. The evaluation set comprises 139 male and 191 female target speakers, each with approximately 2 minutes of speech for enrollment. There are 2,983 speaker trials and 36,287 impostor attempts.

5.2 Low-Level Features and Models

The phone recognizer uses standard 39-D input vectors comprising MFCCs, energy, and their derivatives. The inputs to the manner and place MLP comprise 9 frames of 26-D acoustic vectors: 12 MFCCs, log-energy, and their first derivatives. For the NIST00 evaluation, the acoustic vectors for the GMM-UBM comprise 19 MFCCs plus their first derivative. For the NIST02 evaluation, the acoustic vectors comprise 12 MFCCs and 12 delta-MFCCs. [3]

[2] See http://www.nist.gov/speech/tests/sre/ for NIST SRE plans.

[3] We have also tried using 38-D acoustic vector for NIST02 evaluation, but the performance is inferior to that using 24-D vectors.

For the GMM-UBM, gender-dependent UBMs with 1,024 Gaussians were used. The GMMs of target speakers were adapted from the UBMs using MAP adaptation [1]. Each supervector in the GMM-SVM comprises the means of a MAP-adapted GMM, each with 256 Gaussians. [4] The SVM of each target speaker in NIST02 was trained by using his/her training utterance as the positive sample and the training utterances of the same gender in NIST01 as negative training samples. This amounts to 112 male and 122 female negative samples for each SVM. SVMlight was used for training the SVMs. The penalty factor (-c) and cost factor (-f) were set to 5,000 and 100, respectively.

5.3 Score Fusion of AFCPMs and Acoustic GMMs

Research has shown that features and classifiers of different types may complement each other, and thus improvement in classification performance can be obtained by fusing them [20, 47]. The AFCPMs and the acoustic GMMs characterize speakers at two different levels. The former represents the pronunciation behaviors of individual speakers, whereas the latter focuses on their vocal-tract characteristics. Therefore, fusing their scores is expected to improve speaker verification performance. In this work, the scores from AFCPMs and the acoustic GMMs were linearly combined to obtain the fused scores.

Because high-level AFCPMs and low-level GMM produce scores with different dynamic range, score normalization should be applied before fusion:

$$S_{\mathrm{F}}(X) \;=\; \alpha_u \frac{S_{\mathrm{AFCPM}}(X) - \mu_{\mathrm{AFCPM}}}{\sigma_{\mathrm{AFCPM}}} + (1-\alpha_u)\frac{S_{\mathrm{GMM}}(X) - \mu_{\mathrm{GMM}}}{\sigma_{\mathrm{GMM}}} \qquad (23)$$

where μ and σ are the mean and standard deviation of scores.

Figure 13 shows that normalizing the scores before fusion can make the EER less sensitive to the fusion weight α_u. Another advantage of score normalization is that the value of α_u can suggest which set of scores is more reliable. For example, in Figure 13, the scores produced by the GMM-UBM system are more reliable because the best fusion weight is about 0.4.

5.4 Choice of Relevance Factors

All the adaptation methods mentioned in Section 3 use relevance factors to control the dependence of the adapted model on target speaker's data. The discriminative power of the resulting speaker models depends on the amount

[4] We have also tried using 1,024 Gaussians, but the performance is poorer than that using 256 Gaussians.

of adaptation, which in turn depends on the relevance factors (Eq. 5 and Eq. 15). To investigate the sensitivity of the adapted models with respect to the relevance factors, we used NIST02 data and varied the relevance factors r_β in Eq. 5 and r_α in Eq. 15. The EER performance is shown in Table 4 and Table 5.

Clearly, the performance is very stable across a wide range of r_β, suggesting that the relevance factor is very robust. Nevertheless, the relevance factor cannot be too large or too small; otherwise, the speaker models will either be identical to the background models or depend purely on the adaptation data. Both scenarios are undesirable. In this work, we set r_β to 180 and r_α to 9.5×10^4 in an attempt to avoid these extreme scenarios.

5.5 Effect of Phone Recognition Errors

We have tried replacing the null-grammar recognizer with a full-blown speech recognizer equipped with a good language model.[5] However, the results turn out to be slightly worse. We conjecture that this is mainly because a good language model will help the recognizer to “correct” the pronunciation mistakes made by a speaker; therefore, the performance of AFCPMs may degrade if the langauge model is too good.

5.6 Verification Performance

Table 6 shows the equal error rate (EER) and p-values [48] (with respect to Method A) achieved by different adaptation methods. It shows that Methods C, D, and E achieve a lower error rate as compared to the classical MAP adaption. This confirms our earlier argument that better speaker models can be obtained by adapting the phoneme-independent models in addition to the phoneme-dependent models.

We have also compared our methods with the adaptation method for acoustic GMMs proposed by Hansen et al. [49]. Applying the idea in [49], the adaptation equation for AFCPM can be written as:

$$\widehat{P}_s(m,p|q) = \beta_s^q P_s(m,p|q) + (1-\beta_s^q)\left[\alpha_b^q P_b(m,p|q) + (1-\alpha_b^q)P_s(m,p|*)\right], \quad (24)$$

which can be considered as a combination of Methods A and B. The EER is 25.65%. Evidently, its performance is better than that of Methods A and B but is worse than Method D. Because the performance of Method B is worse

[5] We thank M.H. Siu for providing the phone sequences.

than that of Method C, combining Methods A and Method B is unlikely to give better result than combining Methods A and C.

The DET plots corresponding to Tables 6 are shown in Figure 14. Evidently, Method D achieves the best performance across a wide range of decision threshold. It was found that the proposed adaptation approaches can effectively alleviate the data sparseness problem, resulting in a significantly lower error rate. Apparently, Problems 2 and 3 in Method A have also been alleviated by Method D.

Figure 15 shows the DET performance when the low-level GMM scores and high-level AFCPM scores are fused. It demonstrates that the AFCPMs are complementary to the acoustic GMMs, leading to a slightly better performance when the scores of the two types of models are combined.

5.7 Conclusion

To minimize the undesirable effect of insufficient enrollment data on system performance, this paper proposes four new adaptation methods for creating speaker models based on high-level features. The best performing method is the one that adapts not only the phoneme-dependent background model but also the phoneme-independent speaker model. The amount of adaptation in the latter is adjusted by a proportional factor derived from the phoneme-independent background models. The proposed method was compared with traditional MAP adaptation under the NIST2000 and NIST2002 SRE frameworks. Experimental results show that the proposed method can alleviate the data-sparseness problem effectively and achieves a better performance when compared with traditional MAP adaptation.

6 Acknowledgment

This work was supported by the Research Grant Council of the Hong Kong SAR Project No. PolyU5241/07E and HKPolyU Project No. 1-BB9W.

References

[1] 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.

[2] O. Siohan, C. Chesta, and C. H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 417–428, 2001.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[4] O. Kimball, M. Schmidt, H. Gish, and J.Waterman, "Speaker verification with limited enrollment data," in *Eurospeech'97*, 1997, pp. 967–970.

[5] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering," *J. Comput. Speech Lang.*, vol. 10, pp. 54–77, 1996.

[6] J. Mariéthoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," in *ICSLP*, 2002, pp. 581–584.

[7] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Proc. ICASSP*, 1993, vol. 1, pp. 391–394.

[8] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 695–707, 2000.

[9] O. Thyes, R. Kuhn, P. Nguyen, and JC Junqua, "Speaker identification and verification using eigenvoices," in *Proc. ICSLP*, 2000, vol. 2, pp. 242–245.

[10] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum-likelihood linear regression," in *Proc. ICSLP*, 2000, vol. 3, p. 742.

[11] B. Mak, S. Ho, R. Hsiao, and J. T. Kwok, "kernel eigenvoice speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 984–992, Spe. 2005.

[12] B. Mak, S. Ho, R. Hsiao, and J. T. Kwok, "Embedded kernel eigenvoice speaker adaptation and its implication to reference speaker weighting," *IEEE Transactions on Speech and Audio Processing*, vol. 14, pp. 1267– 1280, 2006.

[13] M. W. Mak, R. Hsiao, and B. Mak, "A comparison of various adaptation methods for speaker verification with limited enrollment data," in *ICASSP*, 2006, pp. 929–932.

[14] E. Blaauw, "The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech," *Speech Communication*, vol. 14, pp. 359–375, 1994.

[15] D. Dahan and J. M. Bernard, "Interspeaker variability in emphatic accent production in French," *Language and Speech*, vol. 39, no. 4, pp. 341–374, 1996.

[16] J. Sussman, E. Dalston, and S. Gumbert, "The effect of speaking style on a locus equation characterization of stop place articulation," *Phonetica*, vol. 55, no. 4, pp. 204–255, 1998.

[17] D. P. Kuehn and K.L. Moll, "A cineradiographic study of VC and CV articulatory velocities," *J. Phonetics*, vol. 23, no. 4, pp. 303–320, 1976.

[18] E. Shriberg, et al., "Modeling prosodic sequences for speaker recognition," *Speech Communication*, vol. 4, pp. 455–472, 2005.

[19] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eurospeech*, Aalborg, Sept. 2001, pp. 2521–2524.

[20] D. Reynolds, et. al., "The superSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP*, Hong Kong, April 2003, vol. 4, pp. 784–787.

[21] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP*, 2003, vol. 4, pp. 788–791.

[22] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *ICSLP*, 1998, vol. 4, pp. 3189–3192.

[23] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *Proc. ICASSP*, 2002, vol. 1, pp. 141–144.

[24] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusáček, D. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," in *Proc. ICASSP*, 2003, vol. 4, pp. 792–795.

[25] D. Chappell and J. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. ICASSP*, 1998, vol. 1, pp. 885–888.

[26] W. Andrews, et al., "Gender-dependent phonetic refraction for speaker recognition," in *Proc. ICASSP*, 2002.

[27] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proc. Eurospeech*, 2003, pp. 2665–2668.

[28] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, "Phonetic speaker recognition using maximum likelihood binary decision tree models," in *Proc. ICASSP*, 2003, vol. 4, pp. 796–799.

[29] Q. Jin, et al., "Combining cross-stream and time dimensions in phonetic speaker recognition," in *Proc. ICASSP*, 2003.

[30] D. Klusacek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," in *Proc. ICASSP*, 2003, vol. 4, pp. 804–807.

[31] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "Speech recognition performance comparison between DSR and AMR transcoded speech," in *Proc. ICASSP*, 2005, vol. 1, pp. 173– 176.

[32] K. Y. Leung, M. W. Mak, and S. Y. Kung, “Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification,” *Speech Communication*, vol. 48, no. 1, pp. 71–84, 2006.

[33] Elizabeth Shriberg, “Higher-level features in speaker recognition,” *Speaker Classification*, vol. 1, pp. 241–259, 2007.

[34] B. Baker, B. Vogt, M. Mason, and S. Sridharan, “Improved phonetic and lexical speaker recognition through MAP adaptation,” in *Proc. ODYSSEY*, May 31-June 3, 2004, pp. 91–96.

[35] M. Federico, “Bayesian estimation methods for n-gram language model adaptation,” in *ICSLP*, 1996, pp. 279–282.

[36] S. X. Zhang, M. W. Mak, and H. M. Meng, “Speaker verification via high-level feature based phonetic-class pronunciation modeling,” *IEEE Trans. on Computers*, vol. 56, no. 9, pp. 1189–1198, 2007.

[37] S. X. Zhang and M. W. Mak, “A new adaptation method for speaker-model creation in high-level speaker verification,” in *Advances in Multimedia Information Processing (PCM)*, Hong Kong, Springer LNCS 4810, 2007, pp. 325–335.

[38] J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[39] J. Shlens, “A tutorial on principal component analysis,” *http://www.snl.salk.edu/˜shlens/pub/notes/pca.pdf*, 2005, December.

[40] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. Speaker Odyssey*, 2001, pp. 213–218.

[41] D. A. Reynolds, “Comparison of background normalization methods for text-independent speaker verification,” in *Proc. Eurospeech’97*, 1997, pp. 963–966.

[42] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, “Short-time Gaussianization for robust speaker verification,” in *Proc. ICASSP*, 2002, vol. 1, pp. 681–684.

[43] M. W. Mak, K. K. Yiu, and S. Y. Kung, “Probabilistic feature-based transformation for speaker verification over telephone networks,” *Neurocomputing, special issue on Neural Networks for Speech and Audio Processing*, vol. 71, pp. 137–146, 2007.

[44] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[45] J. P. Campbell and D. A. Reynolds, “Corpora for the evaluation of speaker recognition systems,” in *Proc. ICASSP 1999*, 1999, vol. 2, pp. 829–832.

[46] D. A. Reynolds, "HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects," in *Proc. ICASSP*, 1997, vol. 2, pp. 1535–1538.

[47] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[48] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.

[49] E. G. Hansen, E. Raymond, and T. R. Anderson, "Speaker recognition using phoneme-specific GMMs," in *Proc. ODYSSEY*, May 31-June 3, 2004, pp. 179–184.

7 Appendix

Denote

$$\begin{aligned} a_i &= \#\Big((m,p,q^{(i)}) \text{ in the utterances of all backgroud speakers}\Big) \\ b_i &= \#\Big((*,*,q^{(i)}) \text{ in the utterances of all backgroud speakers}\Big) \end{aligned}$$

where $q^{(i)}$ represents one of the 46 phonemes (including silence) in English. We have

$$\begin{aligned} P_b(m,p|q^{(i)}) &= \frac{\#((m,p,q^{(i)}) \text{ in the utterances of all backgroud speakers })}{\#((*,*,q^{(i)}) \text{ in the utterances of all backgroud speakers})} = \frac{a_i}{b_i} \\ P_b(m,p|*) &= \frac{\#((m,p,*) \text{ in the utterances of all backgroud speakers}}{\#((*,*,*) \text{ in the utterances of all backgroud speakers}} = \frac{\sum\limits_{i=1}^{46} a_i}{\sum\limits_{i=1}^{46} b_i}. \end{aligned} \tag{25}$$

Assume that there exist constants A_i that satisfy

$$P_b(m,p|*) = \sum_{i=1}^{46} A_i P_b(m,p|q^{(i)}). \tag{26}$$

Substituting Eq. 25 into Eq. 26, we obtain

$$\frac{\sum\limits_{i=1}^{46} a_i}{\sum\limits_{i=1}^{46} b_i} = \sum_{i=1}^{46}\left(A_i \frac{a_i}{b_i}\right) \tag{27}$$

$$\Longrightarrow \sum_{i=1}^{46} a_i = \sum_{i=1}^{46} \left[\left(A_i \frac{\left(\sum_{j=1}^{46} b_j\right)}{b_i} \right) a_i \right] \tag{28}$$

$$\Longrightarrow A_i = \frac{b_i}{\sum_{j=1}^{46} b_j} = \frac{\#((*,*,q^{(i)}) \text{ in the utterances of all backgroud speakers})}{\#((*,*,*) \text{ in the utterances of all backgroud speakers})} = P_b(q^{(i)}) \tag{29}$$

which suggests that

$$P_b(m,p|*) = \sum_{i=1}^{46} P_b(m,p|q^{(i)}) P_b(q^{(i)}). \tag{30}$$

Other equations in Eq. 8 can be derived similarly.

Feature Category	Feature Description	Feature Extractor	Feature Time Span	System Models
Pronunciations (Place of birth, education, socioeconomic status, etc.)	Multilingual phone streams	Language-dependent phone ASR	Several frames	N-gram [26]; Binary tree [28]
	Multilingual phone cross-streams	Language-dependent phone ASR	Several frames	N-gram [29]; CPM [30]
	Articulatory features	MLP and phone ASR	Several frames	AFCPM [32]
Idiolect (Education, socioeconomic status, etc.)	Word streams	Word ASR	Several frames	N-gram [19]; SVM [31]
Prosodic (Personality type, parental influence, etc.)	F0 & Energy distribution	Energy estimator	One frame	GMM [21]
	Pitch contour	F0 estimator & word ASR	Several frames	DTW [25]
	F0 & energy contour & duration dynamics	F0 & energy estimator & phone ASR	Several frames	N-gram [21]
	Prosodic statistics from F0 & duration	F0 & energy estimator & word ASR	Several frames	KNN [24]
Acoustic (Physical structure of vocal organs)	MFCC & its time derivatives	MFCC extractor	One/Several frames	GMM [1]

Table 1
A summary of high-level features in speaker verification. The level of features decreases from top to bottom. (After [33]).

Articulatory Properties	**Classes**	**Number of Classes**
Manner ($\mathcal{M}$)	Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral	6
Place ($\mathcal{P}$)	Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal	10

Table 2
The manner and place properties in AFCPMs. The products of 6 manner and 10 place classes produce 60 probabilities.

Database	**Purpose**
SPIDRE	To train the null-grammar phone recognizer
HTIMIT	To train the manner and place MLPs
NIST99	To create the background models for NIST00 evaluation
NIST01	To create the background models for NIST02 evaluation
NIST00 & NIST02	To create speaker models and evaluate their performance

Table 3
The purposes of the databases used in this study.

Relevance factor r_β	60	120	180	240	300
EER (%)	25.58	25.22	24.93	25.18	25.36

Table 4
The effect of varying the relevance factor r_β in Eq. 5 on the system performance. Results based on the female part of NIST02. Classical MAP (Eq. 4) was used in the adaptation.

Relevance factor r_α ($\times 10^4$)	8	9	9.5	10	11
EER (%)	24.16	23.79	23.50	23.76	24.21

Table 5
The effect of varying the relevance factor r_α in Eq. 15 on the system performance. Results based on the female part of NIST02, and $r_\beta = 180$. Method D (Eq. 14) was used in the adaptation.

Adaptation Method	**EER** (%)	***p*-values**	**H-L Fusion**
Method A (MAP)	25.89	—	15.89
Method B (PIA)	26.18	< 0.00001	16.05
Method C (SPI)	24.63	0.0042	15.78
Method D (MSPI)	23.91	< 0.00001	15.56
Method E (MSPD)	24.86	< 0.00001	15.72
Score Fusion (A+D)	23.67	< 0.00001	13.19

(a)

Adaptation Method	**EER** (%)	***p*-values**	**H-L Fusion**
Method A (MAP)	24.87	—	8.42
Method B (PIA)	25.76	< 0.00001	8.51
Method C (SPI)	24.14	0.0018	8.14
Method D (MSPI)	23.46	< 0.00001	8.10
Method E (MSPD)	24.22	0.0127	8.26
Score Fusion (A+B+C+D+E)	23.18	< 0.00001	8.01

(b)

Table 6

Results based on (a) NIST00 and (b) NIST02. The EERs were obtained by phoneme-dependent AFCPMs created by the methods described in Section 3. The p-values between the classical MAP and the proposed adaptation methods are listed in the third column ($p < 0.01$ implies that the difference between the two EERs is statistically significant). The H-L (high- and low-level) Fusion means linearly combining of the scores of AFCPM and acoustic GMM systems (GMM-UBM + GMM-SVM). The EER of the GMM systems for (a) NIST00 is 13.88 and that for (b) NIST02 is 8.60.

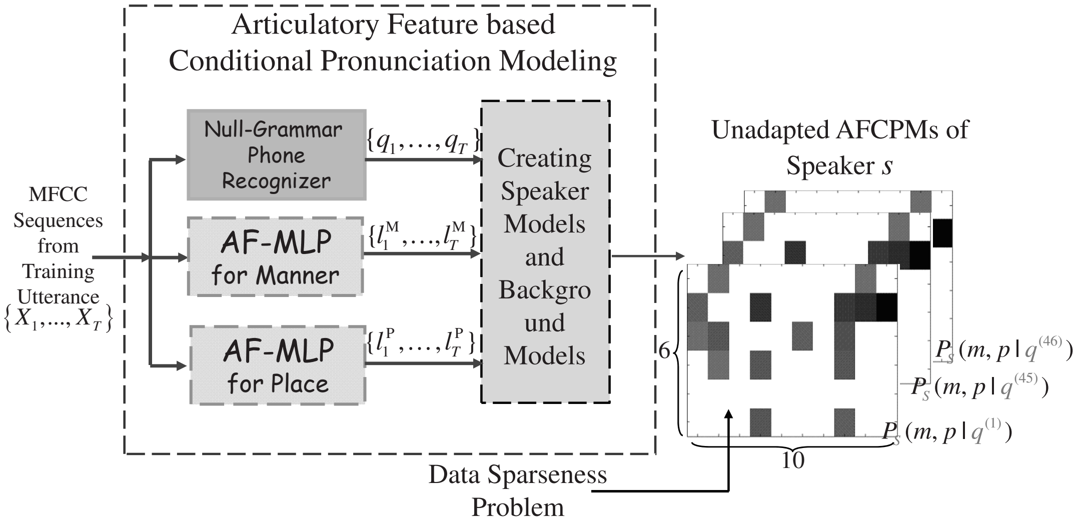

Articulatory Feature based
Conditional Pronunciation Modeling
MFCC Sequences from Training Utterance $\{X_1,\ldots,X_T\}$
Null-Grammar Phone Recognizer
AF-MLP for Manner
AF-MLP for Place
$\{q_1,\ldots,q_T\}$
$\{l_1^M,\ldots,l_T^M\}$
$\{l_1^P,\ldots,l_T^P\}$
Creating Speaker Models and Backgro und Models
Unadapted AFCPMs of Speaker s
$P_S(m,p\,|\,q^{(46)})$
$P_S(m,p\,|\,q^{(45)})$
$P_S(m,p\,|\,q^{(1)})$
6
10
Data Sparseness Problem

Fig. 1. Training of unadapted phoneme-dependent AFCPM speaker models and the data-sparseness problem they may encounter. Note that there are 46 AFCPMs for each client speaker because there are 46 phones (including silence) in English.

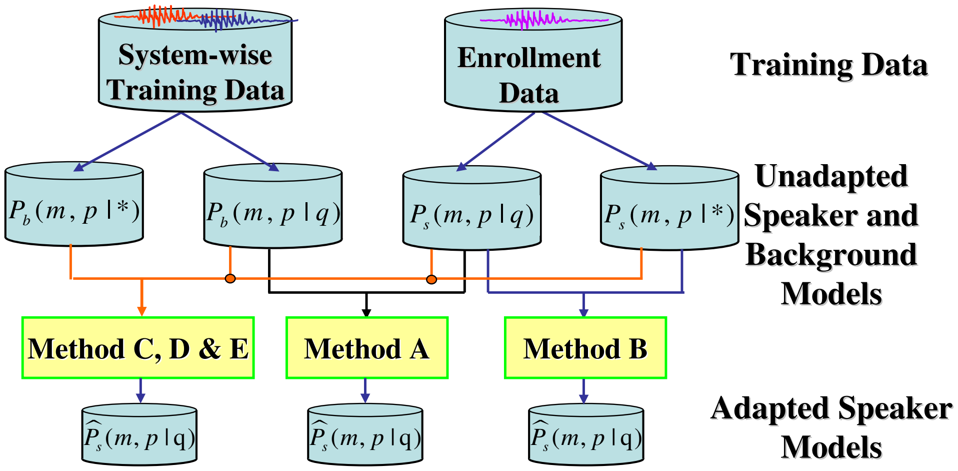
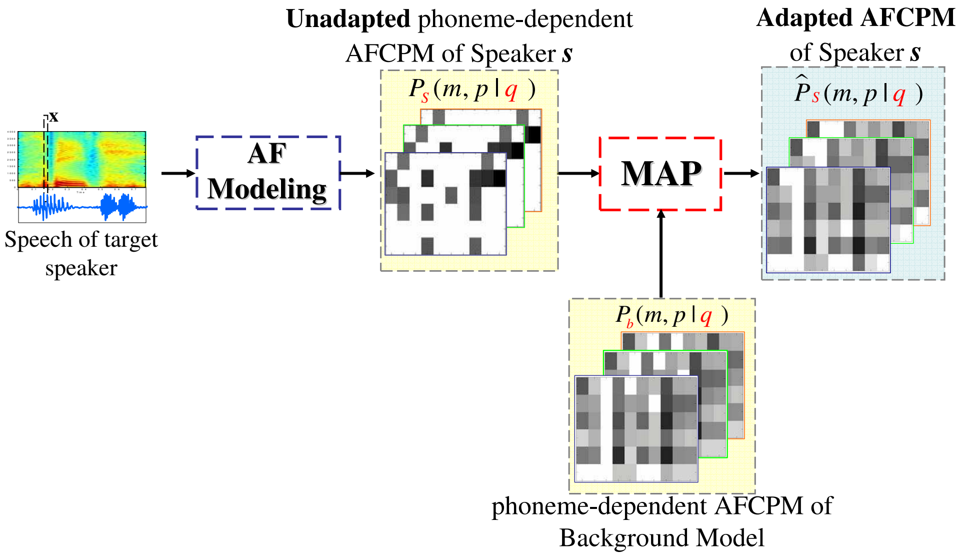


Fig. 2. The use of available information from training data. Four different types of unadapted models can be directly derived from training data using Eqs. 2, 3, 6 and 7, and the adapted speaker models can be derived from different combinations of these unadapted models. Method A only uses part of the available information via phoneme-dependent background models and phoneme-dependent speaker models. Similar situation occurs in Method B. Methods C, D and E fully utilize all of the possible information (via all types of unadapted models) that can be obtained from the training data. A model with an '$*$' means that it is phoneme-independent, whereas a model with a density function conditioned on q means that it depends on phoneme q.



x
Speech of target speaker
AF Modeling
Unadapted phoneme-dependent AFCPM of Speaker *s*
$P_s(m,p|q)$
MAP
Adapted AFCPM of Speaker *s*
$\hat{P}_s(m,p|q)$
$P_b(m,p|q)$
phoneme-dependent AFCPM of Background Model

Fig. 3. The procedure of applying MAP adaptation (Method A) to create a phoneme-dependent AFCPM for a target speaker.

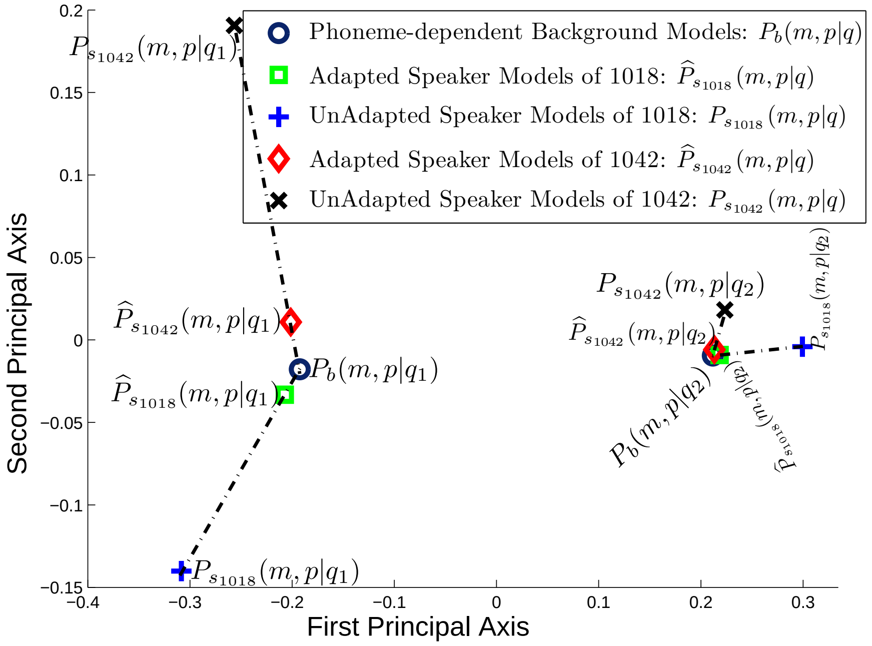


Fig. 4. *Method A*. Relationship (based on real data, q_1 = /jh/ and q_2 = /uw/) between the background, unadapted, and adapted AFCPMs in classical MAP. The linear combination in Eq. 4 suggests that the adapted model will lie along the straight line passing through the unadapted model and the background model. Note that the adapted models (◇ and □) are close to the background model (◯).

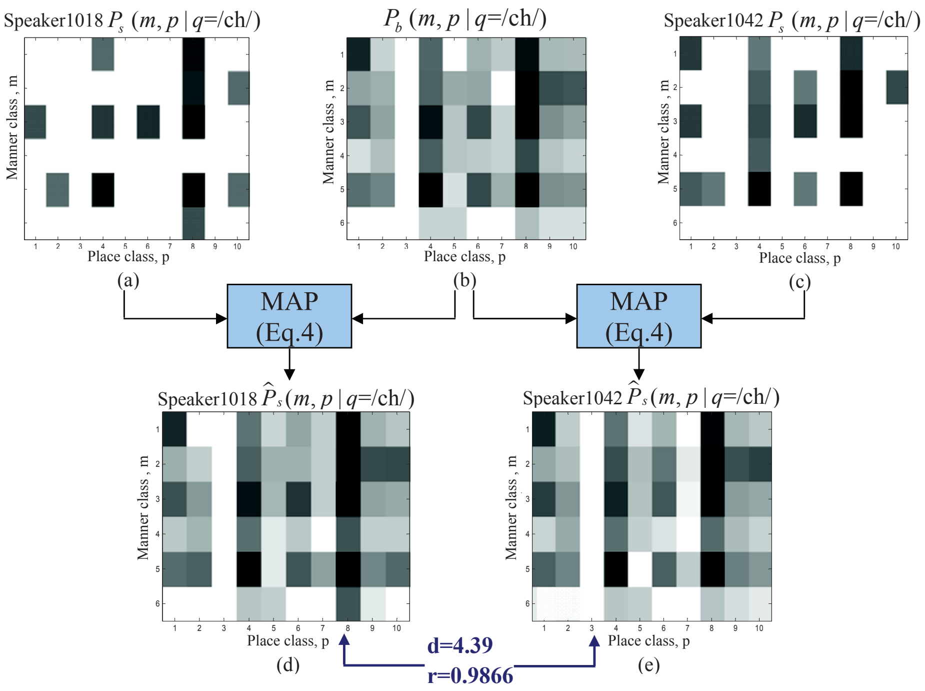


Fig. 5. Phoneme-dependent AFCPMs correspond to phoneme /ch/ of (a) speaker 1018 from NIST00, (b) background speakers from NIST99, and (c) speaker 1042 from NIST00. (d) and (e): Phoneme-dependent speaker models of the two speakers adapted from (b) using the traditional MAP adaptation (see Method A in section 3.1). d and r represent the Euclidean distance and the correlation coefficient between the models pointed to by arrows. The 60 discrete probabilities corresponding to the combinations of the 6 manner and 10 place classes are nonlinearly quantized to 256 gray levels using log-scale, where white represents 0 and black represents 1.

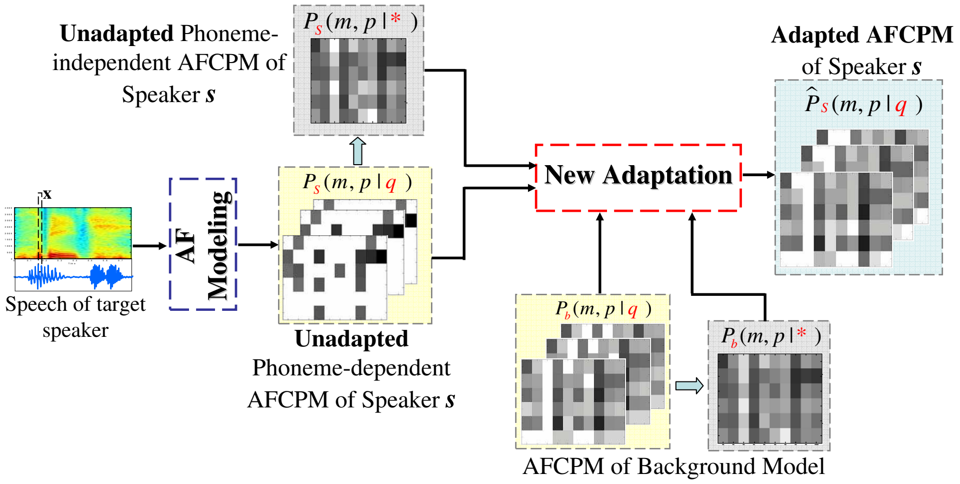

Unadapted Phoneme-independent AFCPM of Speaker s
$P_s(m, p\,|\,*)$
Adapted AFCPM of Speaker s
$\hat{P}_s(m, p\,|\,q)$
x
Speech of target speaker
AF Modeling
$P_s(m, p\,|\,q)$
New Adaptation
Unadapted Phoneme-dependent AFCPM of Speaker s
$P_b(m, p\,|\,q)$
$P_b(m, p\,|\,*)$
AFCPM of Background Model

Fig. 6. Procedures of creating an adapted speaker model using Methods B to E.

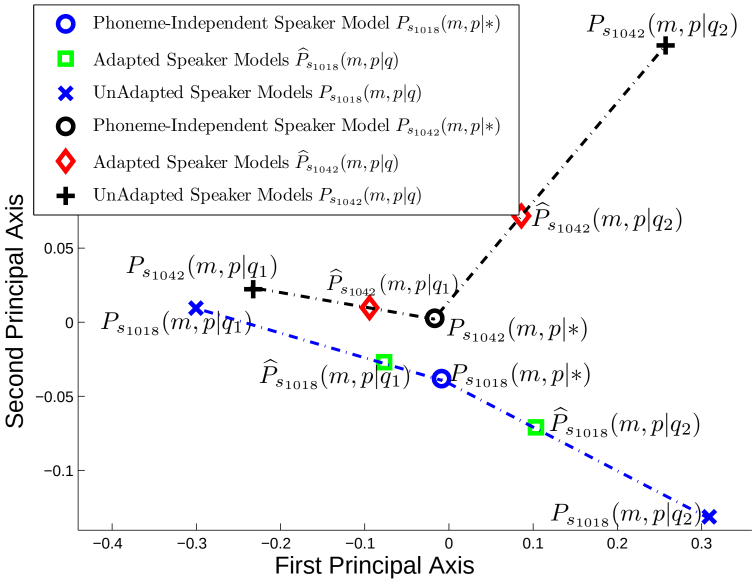


Fig. 7. *Method B.* Relationship between the phoneme-independent speaker model, unadapted speaker models, and adapted speaker models for speakers 1018 and 1042 (q_1 = /jh/, q_2 = /uw/).

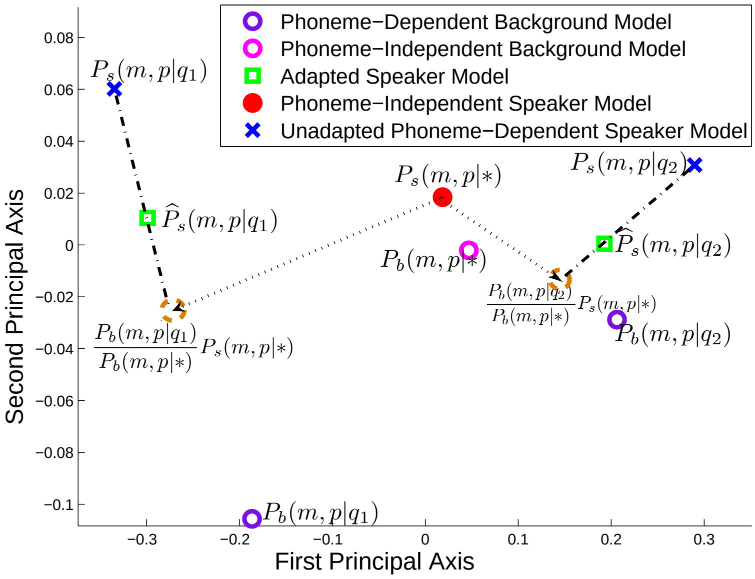


Fig. 8. *Method C.* Relationship between the phoneme-independent speaker model, unadapted speaker models, and adapted speaker models for speaker 1018. (q_1=/jh/ and q_2=/uw/).

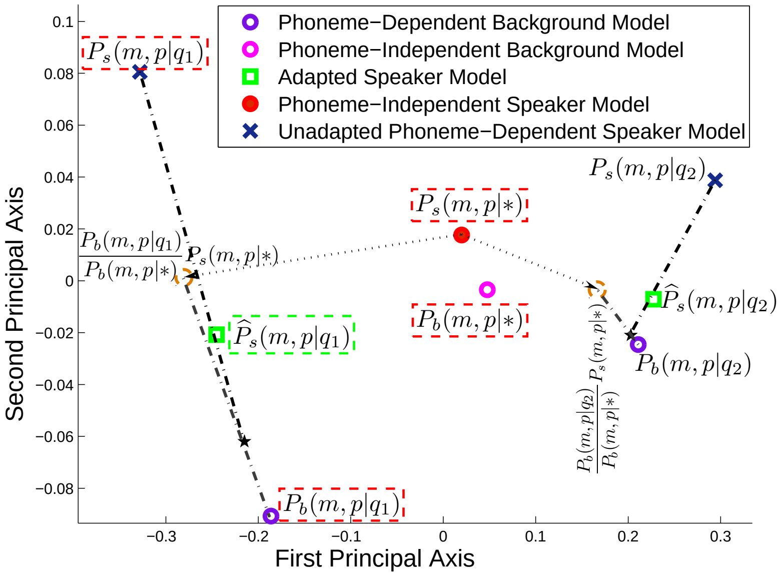


Fig. 9. *Method D*. Relationship between the phoneme-independent speaker model, unadapted speaker models, and adapted speaker models for speaker 1018. (q_1 = /jh/ and q_2 = /uw/; the marker '★' represents the term inside the square brackets in Eq. 14.)

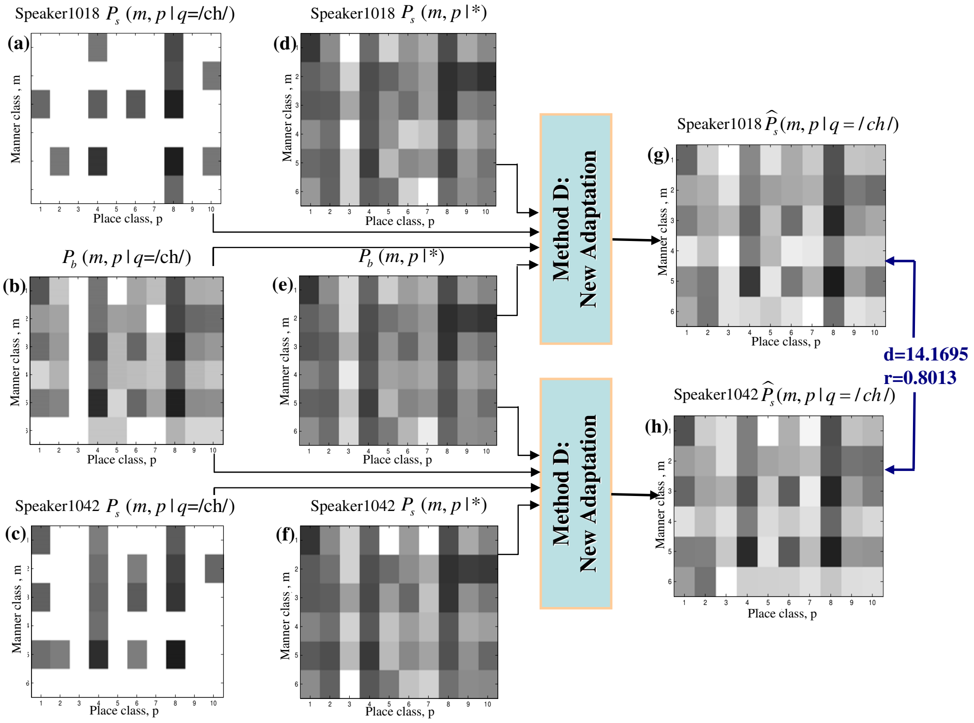

Speaker1018 P_s (m, p | q=/ch/)
(a)
Speaker1018 P_s (m, p | *)
(d)
P_b (m, p | q=/ch/)
(b)
P_b (m, p | *)
(e)
Speaker1042 P_s (m, p | q=/ch/)
(c)
Speaker1042 P_s (m, p | *)
(f)
Manner class , m
Place class, p
Method D: New Adaptation
Method D: New Adaptation
Speaker1018 $\widehat{P}_s(m, p \mid q = /ch/)$
(g)
Speaker1042 $\widehat{P}_s(m, p \mid q = /ch/)$
(h)
d=14.1695
r=0.8013

Fig. 10. Phoneme-dependent AFCPMs ((g) and (h)) of speakers 1018 and 1042 created by Method D. (a) and (c): Unadapted speaker models. (b) Phoneme-dependent background model. (d) and (f): Phoneme-independent speaker models. (e) Phoneme-independent background model. d and r represent the Euclidean distance and correlation coefficient between the adapted models pointed to by arrows.

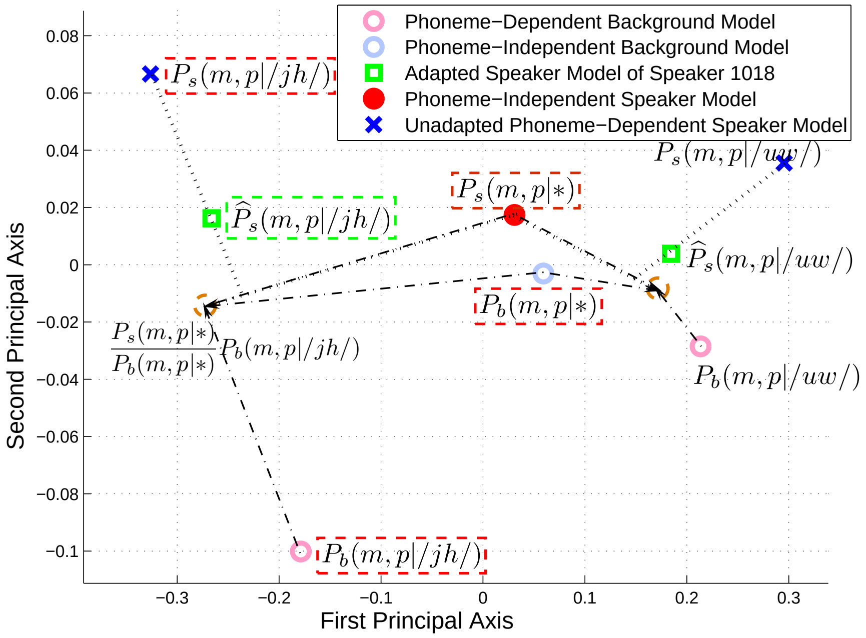


Fig. 11. *Method E.* Relationship between the unadapted, adapted phoneme-dependent and phoneme-independent speaker models for speaker 1018 in Method E.

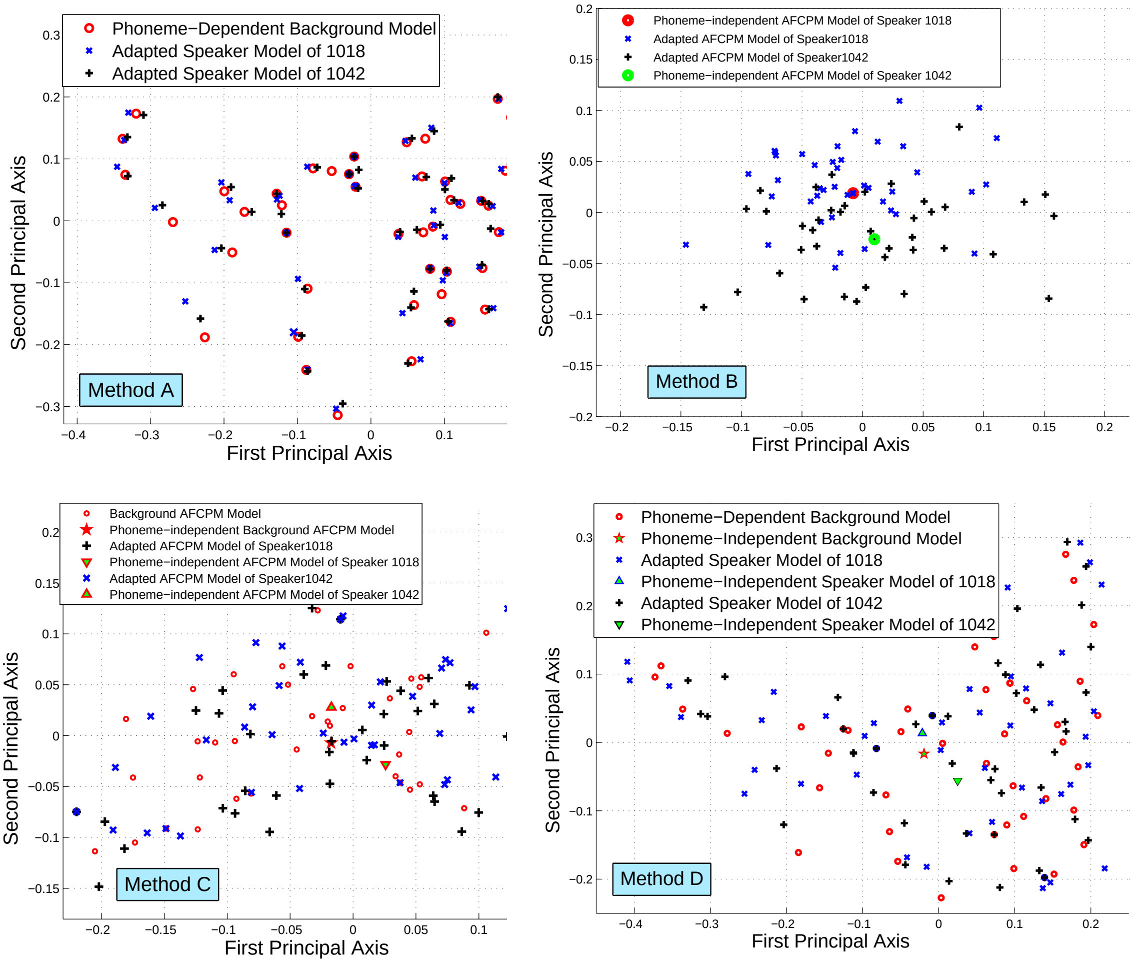


Fig. 12. The projection of adapted phoneme-dependent speaker models and phoneme-dependent background models on the first two principal axes for speakers 1018 and 1042 based on Methods A to D.

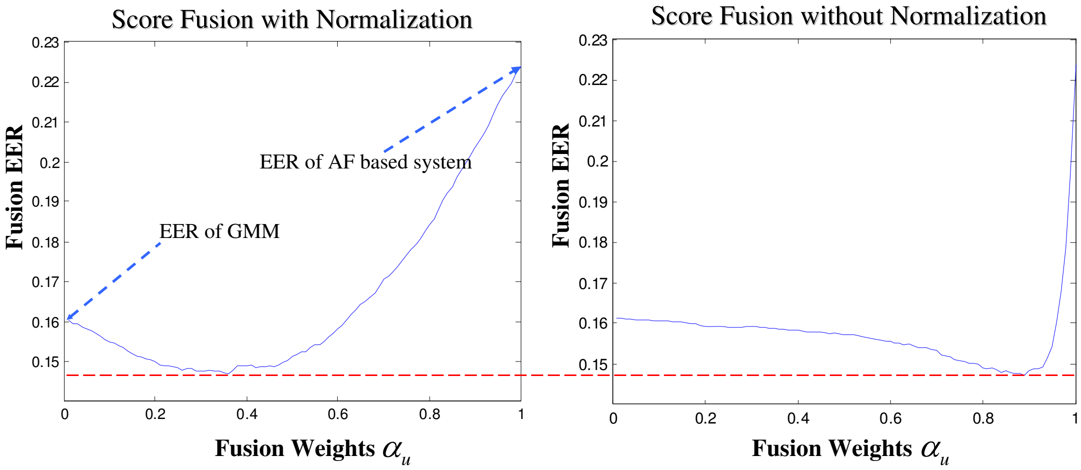
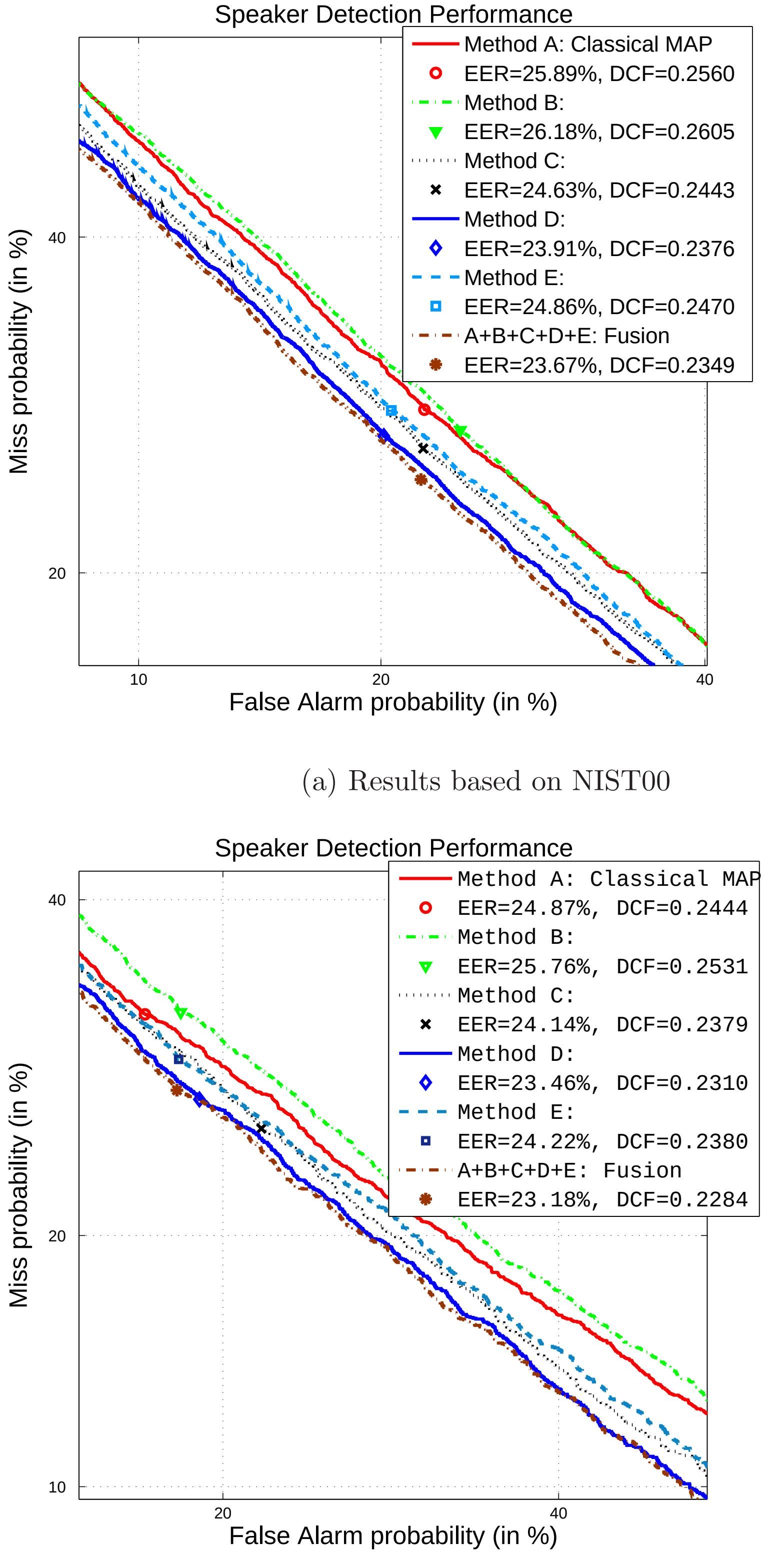


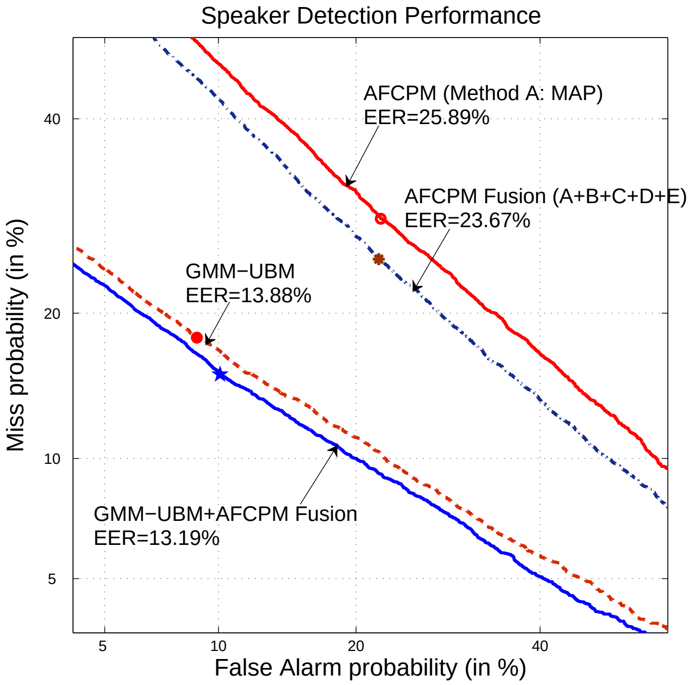
Fig. 13. Score fusion with (left) and without (right) normalization.



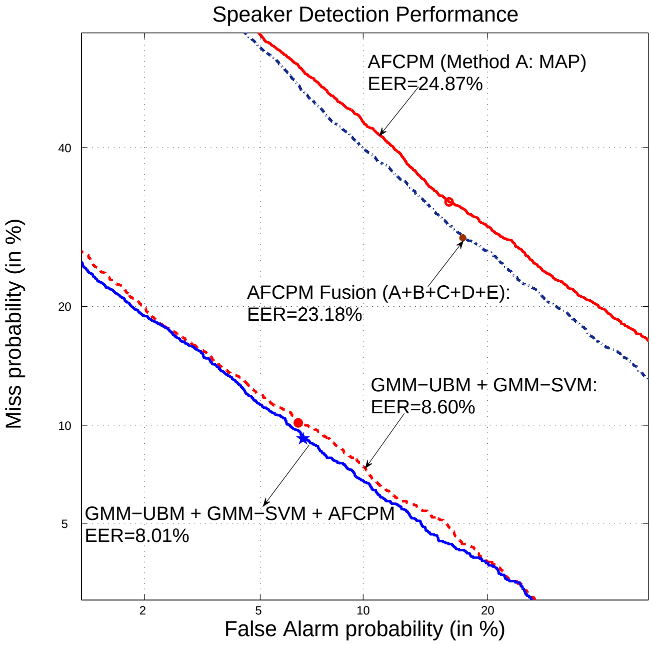
(a) Results based on NIST00

(b) Results based NIST02

Fig. 14. DET performance of AFCPM speaker verification systems using different adaptation methods. (a) NIST00 results. (b) NIST02 results.



(a) Results based on NIST00



(b) Results based NIST02

Fig. 15. DET performance of AFCPM, GMM-UBM, GMM-SVM, and their fusions. For (a), short-time Gaussianization (STG) and fast blind stochastic feature transformation (fBSFT) [43] were applied to the low-level features, and for (b) feature warping was applied.