A New Adaptation Method for Speaker-Model Creation in High-Level Speaker Verification

Shi-Xiong Zhang and Man-Wai Mak⋆

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong
{zhang.sx,enmwmak}@polyu.edu.hk

Abstract. Research has shown that speaker verification based on high-level speaker features requires long enrollment utterances to be reliable. However, in practical speaker verification, it is common to model speakers based a limited amount of enrollment data. To minimize the undesirable effect of insufficient enrollment data on system performance, this paper proposes a new adaptation method for creating speaker models based on high-level features. Different from conventional methods, the proposed adaptation method not only adapts 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 SRE framework. Experimental results show that the proposed method can solve the data-spareness problem effectively and achieves a better performance when compare with traditional MAP adaptation.

1 Introduction

Text-independent speaker verification systems typically extract speaker-dependent features from short-term spectra of speech signals to build speaker-dependent Gaussian mixture models (GMMs) [1]. To increase the ability to discriminate between client (target) speakers and impostors, a GMM-based background model is used to represent the characteristics of impostors. The background model can be trained using the speech of non-target background speakers from large speech corpora. Therefore, finding enough speech to train the background model is usually not too difficult. However, obtaining a large number of client utterances is difficult and impractical because most clients are not willing to spend a long time for enrollment. To address this problem, various adaptation methods, such as maximum a posteriori (MAP) [1], maximum-likelihood linear regression (MLLR) [2], kernel eigen-space MLLR (KEMLLR) [3], and adaptation of phoneme-independent speaker models [4] have been proposed for creating low-level acoustic speaker models from a small amount of client data. It has been shown that KEMLLR outperforms other adaptation methods when the amount of enrollment data is very limited and that when a large amount of enrollment data is available, MAP is a better candidate for creating speaker models [5].

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 [6]. One problem of using high-level features is that it requires a large amount of speech data for creating reliable speaker models. Although Leung et al. [7] have shown in their articulatory feature-based pronunciation model (AFCPM) that this problem can be tackled by classical MAP adaptation, 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 [8].

To alleviate this problem, we propose 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 also 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.

2 Phoneme-dependent AFCPM

Articulatory features (AFs) are representations describing the movements or positions of different articulators during speech production. In Leung et al. [7], manner and place of articulation were used for pronunciation modeling. The manner property has 6 classes, $\mathcal{M}$ ={Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral}, and the place property has 10 classes, $\mathcal{P}$ ={Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal}. The AFs were automatically determined from speech signals using AF-based multilayer perceptrons (MLPs). See [7] for detail description of AFCPM approach.

In phoneme-dependent AFCPM, 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

$$P_b(m,p|q) = P_b(L^{\mathrm{M}} = m, L^{\mathrm{P}} = p|\text{Phoneme} = q, \text{Background})$$

$$= \frac{\#((m,p,q) \text{ in the data of all background speakers})}{\#((*,*,q) \text{ in the data of all background speakers})}, \quad (1)$$

where $m \in \mathcal{M}, p \in \mathcal{P}, (m,p,q)$ denotes the condition for which $L^{\mathrm{M}} = m, L^{\mathrm{P}} = p$, 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:
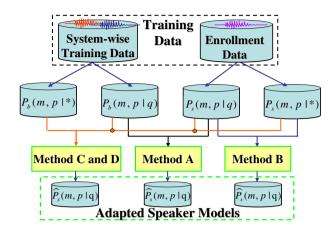
Fig. 1: Data-set utilization in different adaptation methods. Methods A and B only use part of available models. Methods C and D fully utilize all of the possible models that can be obtained from training data. '*' means that the corresponding model is phoneme-independent.

$$P_s(m, p|q) = P_s(L^{\mathrm{M}} = m, L^{\mathrm{P}} = p|\mathrm{Phoneme} = q, \mathrm{speaker} = s)$$
$$= \frac{\#((m, p, q) \ \ \mathrm{in\ the\ enrollment\ utterrence\ of\ speaker\ } s)}{\#((*, *, q) \ \ \mathrm{in\ the\ enrollment\ utterrence\ of\ speaker\ } s)}. \quad (2)$$

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.

3 Adaptation Methods for AFCPMs

Here, we review the classical MAP adaptation and propose three MAP-based adaptation methods that use as much information obtainable from training data as possible (see Fig. 1).

Method A: Adapted from phoneme-dependent background models (classical MAP used in [7]).

Method B: Adapted from phoneme-independent speaker models.

Method C: Adapted from phoneme-independent speaker models with a phoneme-dependent scaling factor.

Method D: Adapted from phoneme-dependent background models and phoneme-independent speaker models with a phoneme-dependent scaling factor.

Method A: In [7], MAP adaptation is applied as follows:

$$\widehat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) P_b(m, p|q) \quad (3)$$

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. 1) on the
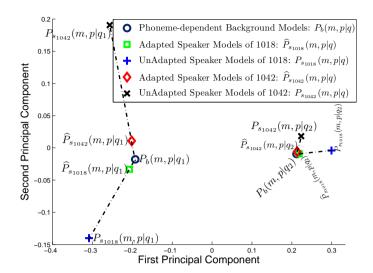
Fig. 2: *Method A*. Relationship (based on real data) between the background, unadapted, and adapted AFCPMs in classical MAP (q_1=/jh/, q_2=/uw/). The linear combination in Eq. 3 suggests that the adapted model will lie along the straight line passing through the unadapted model and the background model.

MAP-adapted model. It is obtained by

$$\beta_s^q = \frac{\#((*,*,q) \text{ in the enrollment utterances of speaker } s)}{\#((*,*,q) \text{ in the enrollment utterances of speaker } s) + r} \qquad (4)$$

where r is a fixed relevance factor common to all phonetic classes and speakers. The relationship between the adapted, unadapted and background models is illustrated in Fig. 2. 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 too close to the background model of the corresponding phoneme, 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 exemplified in Fig. 3, where the adapted models of two speakers are very similar because they are very close to the background model. Comparison between Figs. 3(d) and 3(e) reveals that the model of speaker 1018 are very similar to that of speaker 1042. This will make the speaker models fail to discriminate the true speakers from impostors. For Problem 2, the method only uses two out of four possible models for adaptation. Fig. 1 shows the possible models from which the target models can be adapted. Method A uses
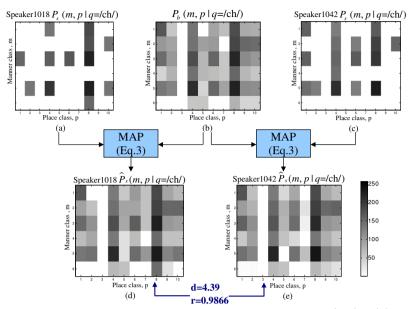
Fig. 3: Phoneme-dependent AFCPMs correspond to phoneme /ch/ of (a) speaker 1018 from NIST00, (b) background speakers from NIST99, and (c) speaker 1042 from NIST00. (d) and (e): Phoneme-dependent speaker models of the two speakers adapted from (b) using the traditional MAP adaptation (see Method A in section 3). 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 non-linearly quantized to 256 gray levels using log-scale, where white represents 0 and black represents 1.

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, the method 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.

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

$$\widehat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) P_s(m, p|*). \tag{5}$$

While this method can help solve Problems 1 and 3 mentioned in Method A, it does have its own problem. The problem is that for a particular client, all
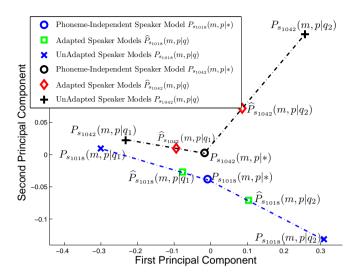
Fig. 4: *Method B*. Relationship between the phoneme-independent speaker model, unadapted speaker models, and adapted speaker models for speakers 1018 and 1042 (q_1=/jh/, q_2=/uw/).

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 Fig. 1. 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.

Method C: In this method, a phoneme-dependent scaling factor is added to the adaptation formula in Eq. 5:

$$\widehat{P}_s(m,p|q) = \beta_s^q P_s(m,p|q) + (1 - \beta_s^q) \cdot \left[\frac{P_b(m,p|q)}{P_b(m,p|*)} \cdot P_s(m,p|*) \right] \qquad (6)$$

where $P_b(m,p|*)$ represents the phoneme-independent background model and $\frac{P_b(m,p|q)}{P_b(m,p|*)}$ is the scaling factor. With this factor, the model to be adapted becomes $\frac{P_b(m,p|q)}{P_b(m,p|*)} P_s(m,p|*)$. 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 $\frac{P_b(m,p|q)}{P_b(m,p|*)} P_s(m,p|*)$ in Eq. 6 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. 3. 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.
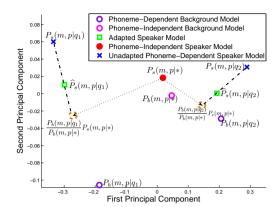
Fig. 5: *Method C.* Relationship between the phoneme-independent speaker model, unadapted speaker models, and adapted speaker models for speaker 1018 (q_1=/jh/, q_2=/uw/).
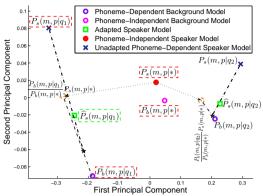

Fig. 6: *Method D.* Relationship between the phoneme-independent speaker model, unadapted speaker models, and adapted speaker models for speaker 1018. (q_1=/jh/, q_2=/uw/ and the marker '★' represents the term inside the square brackets in Eq. 7.)

Method D: 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 too far 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. We refer to the resulting adaptation as Method D whose adaptation equation is written as:

$$\widehat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) \left[\alpha_b^q P_b(m, p|q) + (1 - \alpha_b^q) \frac{P_b(m, p|q)}{P_b(m, p|*)} P_s(m, p|*) \right] \tag{7}$$

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

$$\alpha_b^q = \frac{\#((*, *, q) \text{ in the utterances of all background speakers})}{\#((*, *, q) \text{ in the utterances of all background speakers}) + r_\alpha} \tag{8}$$

where r_α is also a fixed relevance factor.

Fig. 6 illustrates the relationship between different models in Method C, and Fig. 7 explains why this method is better than Method A via an illustrative example.

Comparing Figs. 3 and 7 reveals that the Euclidean distance and dissimilarity between the AFCPM models of speakers 1018 and 1042 become larger
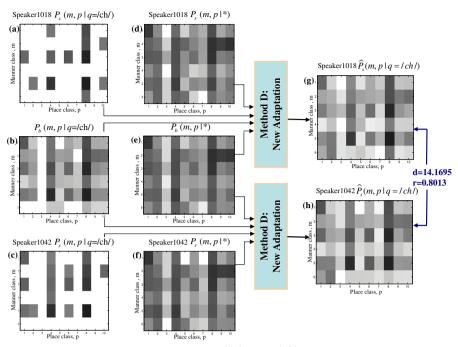
Fig. 7: Phoneme-dependent AFCPMs ((g) and (h)) of speakers 1018 and 1042 created by Method D. (a) and (c): Unadapted speaker models. (b) Phoneme-dependent background model. (d) and (f): Phoneme-independent speaker models. (e) Phoneme-independent background model. d and r represent the Euclidean distance and the correlation coefficient between the adapted models pointed to by arrows.

(the distance increases from 4.39 to 14.17 and the correlation coefficient reduces from 0.9966 to 0.8013). Therefore Method D makes the speaker models easier to discriminate speakers.

4 Scoring

We follow the scoring method in [1]. Specifically, we define the verification score of a test utterance $X = \{X_1, \ldots, X_t, \ldots, X_T\}$ as:

$$S(X) = \sum_{t=1}^{T} \left(\log \widehat{p}_s(X_t) - \log p_b(X_t) \right) \tag{9}$$

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:

$$\widehat{p}_s(X_t) = \widehat{P}_s(l_t^{\mathrm{M}}, l_t^{\mathrm{P}}|q_t) = \widehat{P}_s(L^{\mathrm{M}} = l_t^{\mathrm{M}}, L^{\mathrm{P}} = l_t^{\mathrm{P}}|\text{Phoneme} = q_t, \text{Speaker} = s) \tag{10}$$

$$p_b(X_t) = P_b(l_t^{\mathrm{M}}, l_t^{\mathrm{P}}|q_t) = P_b(L^{\mathrm{M}} = l_t^{\mathrm{M}}, L^{\mathrm{P}} = l_t^{\mathrm{P}}|\text{Phoneme} = q_t, \text{Background}), \tag{11}$$

In Eqs. 10 and 11, q_t is the phoneme of frame t in the test utterance recognized by a null gram phoneme recognizer, and l_t^{M} and l_t^{P} are the AF labels determined by the AF-MLPs [7].

5 Experiments and Results

5.1 Procedures

NIST99, NIST00, SPIDRE [9], and HTIMIT [10] were used in the experiments. NIST99 was used for creating the background models, and the female part of NIST00 was used for creating speaker models and for performance evaluation. HTIMIT and SPIDRE were used for training the AF-MLPs and the null-grammar phone recognizer, respectively.

The phone recognizer uses standard 39-D vectors comprising MFCCs, energy, and their derivatives. The training part of NIST99 was used for creating phoneme-dependent AF-based UBMs. We followed the evaluation protocol of NIST00. Specifically, for each female client speaker in NIST00, her phoneme-dependent speaker models were created using Methods A to D.

5.2 Results and Discussion

Fig. 8 shows the relationship between the phoneme-dependent background and adapted models (corresponding to 46 phonemes) of two speakers for Methods A and D. Apparently, Problem 1 in Method A (left figure) mentioned in Section 3 does not appear in Method D (right figure).
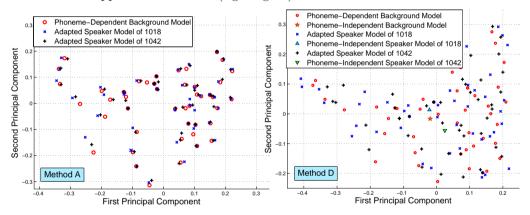

Fig. 8: The distribution of all adapted phoneme-dependent speaker models and phoneme-dependent background models in principal component space for speaker 1018 and 1042 based on Method A (left) and Method D (right).

Table 1 shows the equal error rate (EER) and p-values [11] (with respect to Method A) achieved by different adaptation methods. It shows that Methods C and D achieve a lower error rate as compare 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.The DET plots corresponding to Table 1 are shown in Fig. 9. Evidently, Method D achieves the best performance across a wide range of decision threshold. It was found that the proposed adaptation approaches can effectively solve the data sparseness problem, resulting in a significantly lower error rate. Apparently, Problem 2 and 3 in Method A have also been overcome by method D.
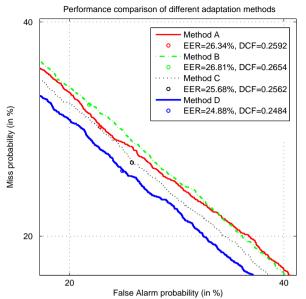
Performance comparison of different adaptation methods

	Method A
	EER=26.34%, DCF=0.2592
	Method B
	EER=26.81%, DCF=0.2654
	Method C
	EER=25.68%, DCF=0.2562
	Method D
	EER=24.88%, DCF=0.2484

Fig. 9: DET performance of AFCPM-based speaker verification systems using different adaptation methods.

Table 1: EERs obtained by phoneme-dependent AFCPMs created by MAP-based adaptation methods described in Section 3. The p-values between the classical MAP and the new adaptation methods are listed in the last column.

Adaptation Method	EER (%)	p-values
Method A	26.34	–
Method B	26.81	0.04560
Method C	25.68	0.00008
Method D	24.88	0.00000

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. 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.
3. 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.
4. T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Proc. ICASSP 1993*, 1993, vol. 1, pp. 391–394.
5. M.W. Mak, R. Hsiao, and B. Mak, "A comparison of various adaptation methods for speaker verification with limited enrollment data," in *ICASSP'06*, 2006, pp. 929–932.
6. D. Reynolds, et. al., "The superSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. International Conference on Audio, Speech, and Signal Processing*, Hong Kong, April 2003, vol. 4, pp. 784–787.
7. 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.
8. S. X. Zhang, M. W. Mak, and Helen H. Meng, "Speaker verification via high-level feature based phonetic-class pronunciation modeling," *IEEE Trans. on Computers*, 2007, *to appear*.
9. 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.
10. D. A. Reynolds, "HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects," in *Proc. ICASSP'97*, 1997, vol. 2, pp. 1535–1538.
11. L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP'89*, 1989, pp. 532–535.