Speaker Verification via High-Level Feature Based Phonetic-Class Pronunciation Modeling

Shi-Xiong Zhang, Man-Wai Mak, *Member* and Helen M. Meng, *Member*

Abstract— **It has been shown recently that the pronunciation characteristics of speakers can be represented by articulatory feature-based conditional pronunciation models (AFCPMs). However, the pronunciation models are phoneme-dependent, which may lead to speaker models with low discriminative power when the amount of enrollment data is limited. This paper proposes to mitigate this problem by grouping similar phonemes into phonetic classes and representing background and speaker models as phonetic-class dependent density functions. Phonemes are grouped by (1) vector quantizing the discrete densities in the phoneme-dependent universal background models, (2) using the phone properties specified in the classical phoneme tree, or (3) combining vector quantization and phone properties. Evaluations based on 2000 NIST SRE show that this phonetic-class approach effectively alleviates the data spareness problem encountered in conventional AFCPM, which results in better performance when fused with acoustic features.**

Index Terms— **Speaker verification, pronunciation modeling, articulatory features, phonetic classes, NIST speaker recognition evaluation.**

I. Introduction

State-of-the-art 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]. One advantage of using short-term spectra is that promising results can be obtained from a limited amount of training data. However, the lack of robustness to mismatched conditions remains a serious problem. Although approaches such as feature transformation [2], [3], model transformation [4], and score normalization [5] have shown promise in reducing the mismatches, these methods have almost reached their limit in terms of error rate reduction.

To further reduce error rate, 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 [6], [7]. Studies in linguistics have shown that speaking styles have significant effect on pronunciation patterns [8]. Kuehn and Moll [9] 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. [10] 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 in 2001 [11]. This work has led to extensive investigations into high-level features in the SuperSID project [12] in which prosodic features [13], [14], phone features [15]–[18], and conversational features [19] were combined and fused 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 [20] that extracts multilingual phone sequences from utterances achieves the best performance [12]. CPM aims to model speaker-specific pronunciations by learning the relationship between what has been said (phonemes) and how speech is pronounced (phones). The rationale behind using CPM for speaker verification is that different speakers have different ways of pronouncing the same phonemes.

One limitation of the CPM in [20] is that it requires multilingual corpora to build speaker and background models. To overcome this limitation, Leung et al. [21] proposed using articulatory feature (AF) streams to construct CPM and called the resulting models AFCPM. AFs are abstract classes describing the movement or positions of different articulators during speech production. The idea hinges upon the linkage between the states of articulation during speech production and the actual phones produced by speakers. In contrast to the conventional speaker recognition systems in which short-term spectral characteristics are represented by Gaussian mixture models (GMM) [1], AFCPM-based systems use discrete probabilistic models to represent two articulatory properties: manner and place of articulation. More specifically, the speaker models are composed of conditional probabilities of articulatory classes in these two properties, and each speaker has N phoneme-dependent discrete probabilistic models, one for each phoneme. It was found in [21] that AFCPM can reduce the error rate of conventional CPM by 25%.

While promising results have been obtained, AFCPM requires a large amount of speech data for training the phoneme-dependent speaker models. Insufficient enrollment data will lead to imprecise speaker models and poor performance. To improve the accuracy of articulatory feature-based models, this paper proposes using phonetic-class based AFCPM. In this method, phonemes with similar manner and place of articulation are grouped together based on the similarity between the AFCPM-based phoneme-dependent universal background models. Then, a discrete density function is computed for each phonetic class. It was found that this phonetic-class AFCPM approach can reduce the side effect caused by the error in the phoneme recognizer and effectively solve the data sparseness problem encountered in conventional AFCPM. Experimental results show that the proposed modification leads to a significantly lower error rate as compared to the

S. X. Zhang and M. W. Mak are with Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University. H. Meng is with Dept. of Systems Engineering and Engineering Management, The Chinese University of Hong Kong.

conventional AFCPM. Results also show that further performance gain can be obtained by fusing the scores derived from AFCPMs and acoustic GMMs.

This paper is organized as follows. Section II introduces articulatory features and explains how they can be extracted from speech signals. Section III outlines the phoneme-dependent AFCPM and discusses the problem that may arise when the amount of training data is limited. Section IV then proposes the phonetic-class dependent AFCPM and three phoneme-to-phonetic class mapping functions to address the problem of insufficient enrollment data. Sections V and VI demonstrate the advantage of the proposed approach via experimental evaluations using the NIST2000 corpus. Finally, conclusions are drawn in Section VII.

II. Articulatory Feature Extraction

Articulatory features (AFs) are the representations of some important phonological properties appeared during speech production. More precisely, AFs are abstract classes describing the movements or positions of different articulators during speech production. Since AFs are closely related to the speech production process, they are suitable for capturing the pronunciation characteristics of speakers.

In Leung et al. [21], the manner and place of articulation as shown in Table I were used for pronunciation modeling. These properties describe the way and location that the air-stream along the vocal tract is constricted by the articulators. Leung et al. [21] adopted the AF extraction approach outlined in [22]. Specifically, the AFs were automatically determined from speech signals using AF-based multilayer perceptrons (MLPs) [23] as shown in Figure 1. For each articulatory property, an AF-MLP takes 9 consecutive frames of 26-dimensional normalized MFCCs X_t (with consecutive frame indexes ranging from $t-4$ to $t+4$) as input to determine the posterior probabilities of the output classes at frame t. For example, given X_t at frame t, the place MLP determines ten posterior probabilities of the output classes, i.e., $P(L^{\mathrm{P}}=p|X_t)$ where $p \in \mathcal{P}$ with $\mathcal{P}$ defined in Table I. Using these probabilities, the manner class label $l_t^{\mathrm{M}} \in \mathcal{M}$ and place class label $l_t^{\mathrm{P}} \in \mathcal{P}$ at frame t are determined by

$$l_t^{\mathrm{M}} = \arg\max_{m\in\mathcal{M}} P(L^{\mathrm{M}}=m|X_t)$$
$$l_t^{\mathrm{P}} = \arg\max_{p\in\mathcal{P}} P(L^{\mathrm{P}}=p|X_t). \tag{1}$$

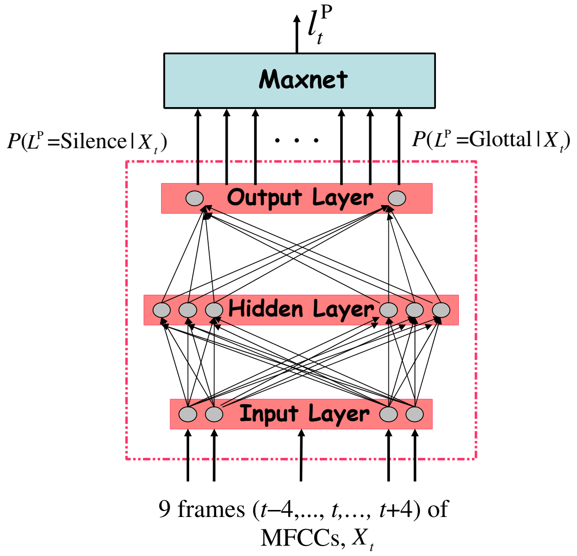
l_t^{P}
Maxnet
$P(L^{\mathrm{P}}$=Silence$|X_t)$
$P(L^{\mathrm{P}}$=Glottal$|X_t)$
Output Layer
Hidden Layer
Input Layer
9 frames (t–4,..., t,..., t+4) of MFCCs, X_t

Fig. 1. Articulatory feature-based multilayer perceptrons (AF-MLP) for the place of articulation. The MLP for the manner of articulation has a similar architecture.

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 I
Articulatory properties and the number of classes in each property.

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.

Interestingly, the AF-MLPs do not need to be very accurate for the purpose of capturing articulatory features.[1] This is mainly because their main purpose is to capture the articulatory features of speakers instead of classifying the articulatory properties. Therefore, as long as the patterns of mistakes made by these MLPs are consistent for the same speaker and different for different speakers, they can still provide valuable speaker information for building the pronunciation models.

III. Phoneme-Dependent AFCPM

A. Phoneme-Dependent UBMs

As illustrated in Figure 3, N phoneme-dependent universal background models (UBMs) are trained from the AF and phoneme streams of a large number of speakers[2] 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 obtained from the AF-MLPs and a phoneme sequence obtained from a null-grammar recognizer [21]. The joint probabilities corresponding to a particular phoneme q is given by

$$\begin{aligned} & P_b^{\mathrm{PD}}(m,p|q) \\ &= P_b^{\mathrm{PD}}(L^{\mathrm{M}}=m, L^{\mathrm{P}}=p|\text{Phoneme}=q, \text{Background}) \\ &= \frac{\#((m,p,q) \text{ in the utterances of all background speakers})}{\#((*,*,q) \text{ in the utterances of all background speakers})} \end{aligned} \tag{2}$$

where $m \in \mathcal{M}, p \in \mathcal{P}, (m,p,q)$ denotes the condition for which $L^{\mathrm{M}}=m, L^{\mathrm{P}}=p$, 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. For each phoneme, a total of 60 probabilities can be obtained (some of them could be zero). These probabilities are the products of 6 manner classes and 10

[1] In our experiments, the manner and place MLPs achieve an average accuracy of 79.49% and 67.69% on the HTIMIT corpus.

[2] In this work, 230 male and 309 female background speakers from NIST99 were used.

place classes. Therefore, a system with N phonemes has $60N$ probabilities in the UBMs. Eq. 2 will be used in Section IV-A to train a mapping function that maps phonemes to phonetic classes.

B. Phoneme-Dependent Speaker Models

A speaker model can be obtained from speaker-dependent data as follows:

$$\begin{aligned} & P_s^{\text{PD}}(m,p|q) \\ & = P_s^{\text{PD}}(L^{\text{M}}=m, L^{\text{P}}=p|\text{Phoneme}=q, \text{Speaker}=s) \\ & = \frac{\#((m,p,q) \text{ in the utterances of speaker } s)}{\#((*,*,q) \text{ in the utterances of speakers} s)}. \end{aligned} \tag{3}$$

However, the accuracy of speaker models obtained by Eq. 3 is limited by the amount of training data available. For some phonemes (e.g., /th/, /sh/, and /v/), the number of occurrences is too small for an accurate estimation of the joint probabilities. To overcome this data-sparseness problem, speaker models can be adapted from the UBMs. Specifically, given the background model corresponding to phoneme q, the joint probabilities $\widehat{P}_s^{\text{PD}}(m,p|q)$ for speaker s are given by

$$\widehat{P}_s^{\text{PD}}(m,p|q) = \beta_q P_s^{\text{PD}}(m,p|q) + (1-\beta_q)P_b^{\text{PD}}(m,p|q)$$

where $\beta_q \in [0,1]$ is a phoneme-dependent adaptation coefficient controlling the contribution of the unadapted speaker model (Eq. 3) and the background model (Eq. 2) on the adapted model. Similar to MAP adaptation of GMM-based systems [1], β_q can be obtained by

$$\beta_q = \frac{\#((*,*,q) \text{ in the utterances of speaker } s)}{\#((*,*,q) \text{ in the utterances of speaker } s) + r_\beta},$$

where r_β is a fixed relevance factor common to all phonemes and speakers. The purpose of r_β is to control the dependence of the adapted model on speaker's data. If the number of occurrences of $(*,*,q)$ is much less than r_β, then β_q will be very close to 0 and the estimation of the new model is less dependent on speaker's data. On the contrary, if the number of occurrences of $(*,*,q)$ is significantly greater than r_β, then β_q will be very close to 1 and the the adapted model will become more dependent on speaker's data.

C. Problem of Phoneme-Dependent Speaker Models

While it has been demonstrated that phoneme-dependent AFCPM can achieve reasonably good performance [21], it has its own limitation. Since the method is phoneme based, it builds phoneme-dependent models regardless of the fact that some phonemes are very similar in terms of articulatory properties. This causes some of the phoneme-based background models to be almost identical. Worse yet, because the speaker models are adapted from the background models, for those "similar" phonemes that rarely occur in the speakers' utterances, the corresponding speakers models will be almost identical to the background models, making the speaker models fail to discriminate the speakers. This situation is exemplified in Figure 2 where the density functions of background and speaker models are illustrated as gray-scale images. Evidently, there is substantial similarity between the two background models (Figures 2(a) and 2(b)). Comparisons between Figures 2(c) and 2(d) and between Figures 2(e) and 2(f) also reveal that the models of speaker 1018 are very similar to those of speaker 3823.

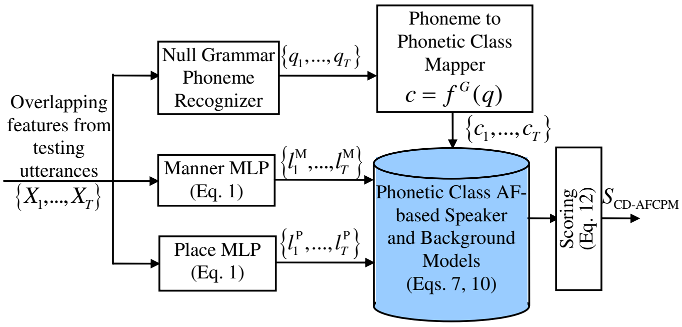
Overlapping features from testing utterances
$\{X_1,...,X_T\}$
Null Grammar Phoneme Recognizer
$\{q_1,...,q_T\}$
Phoneme to Phonetic Class Mapper
$c = f^G(q)$
$\{c_1,...,c_T\}$
Manner MLP (Eq. 1)
$\{l_1^{\text{M}},...,l_T^{\text{M}}\}$
Place MLP (Eq. 1)
$\{l_1^{\text{P}},...,l_T^{\text{P}}\}$
Phonetic Class AF-based Speaker and Background Models (Eqs. 7, 10)
Scoring (Eq. 12)
$S_{\text{CD-AFCPM}}$

Fig. 4. The verification phase of phonetic-class dependent AFCPM. $f^G(q) \in \{f_{\text{VQ}}^G(q), f_{\text{P}}^G(q), f_{\text{P+VQ}}^G(q)\}$.

IV. Phonetic-Class Dependent AFCPM

The limitation of phoneme-dependent AFCPM mentioned earlier can be overcome by grouping the similar AFCPMs into a model set. In other words, each density function can be conditioned on a phonetic-class instead of a single phoneme. Figures 3 and 4 illustrates the training and verification procedures of the phonetic-class dependant AFCPMs, respectively.

A. Mapping Functions

The success of the proposed approach relies on an effective mapping function that groups similar phonemes into a phonetic class. There are several ways of grouping the phonemes: (1) according to the similarity (Euclidean distance) between the AFCPMs, (2) according to the phoneme properties as depicted in the classical phoneme tree [24], and (3) combination of (1) and (2).

1) Method 1: Grouping based on Euclidean distance: The phoneme-dependent UBMs, $P_b^{\text{PD}}(m,p|q)$, are vectorized to N 60-dimensional vectors called AFCPM vectors (see Figure 3):

$$\mathbf{a}_q = \begin{bmatrix} P_b^{\text{PD}}(L^{\text{M}} = \text{‘Vowel’}, L^{\text{P}} = \text{‘High’}|\text{Phoneme} = q) \\ P_b^{\text{PD}}(L^{\text{M}} = \text{‘Vowel’}, L^{\text{P}} = \text{‘Low’}|\text{Phoneme} = q) \\ \cdots \\ P_b^{\text{PD}}(L^{\text{M}} = \text{‘Nasal’}, L^{\text{P}} = \text{‘Glottal’}|\text{Phoneme} = q) \end{bmatrix}$$

where $q \in \{\text{Phoneme } 1, \ldots, \text{Phoneme } N\}$. Then, K-means clustering or VQ can be applied to cluster the N AFCPM vectors into G classes. The mapping from a specific phoneme to its corresponding phonetic class index c is defined as a mapping function:

$$c = f_{\text{VQ}}^G(q), \quad c \in \{1, 2, \ldots, G\}. \tag{4}$$

This function will be used to train the phonetic-class UBMs and speaker models, which is to be detailed in Section IV-B.

2) Method 2: Grouping based on phoneme properties: Because the phoneme grouping in classical phoneme tree [24] is partly based on articulatory properties, we can also use the tree to determine the mapping between phonemes and phonetic classes. This results in the mapping function

$$c = f_{\text{P}}^G(q), \quad c \in \{1, 2, \ldots, G\}. \tag{5}$$

Table II shows the mapping between the phonemes and phonetic classes obtained from the classical phoneme tree [24] for three different values of G.

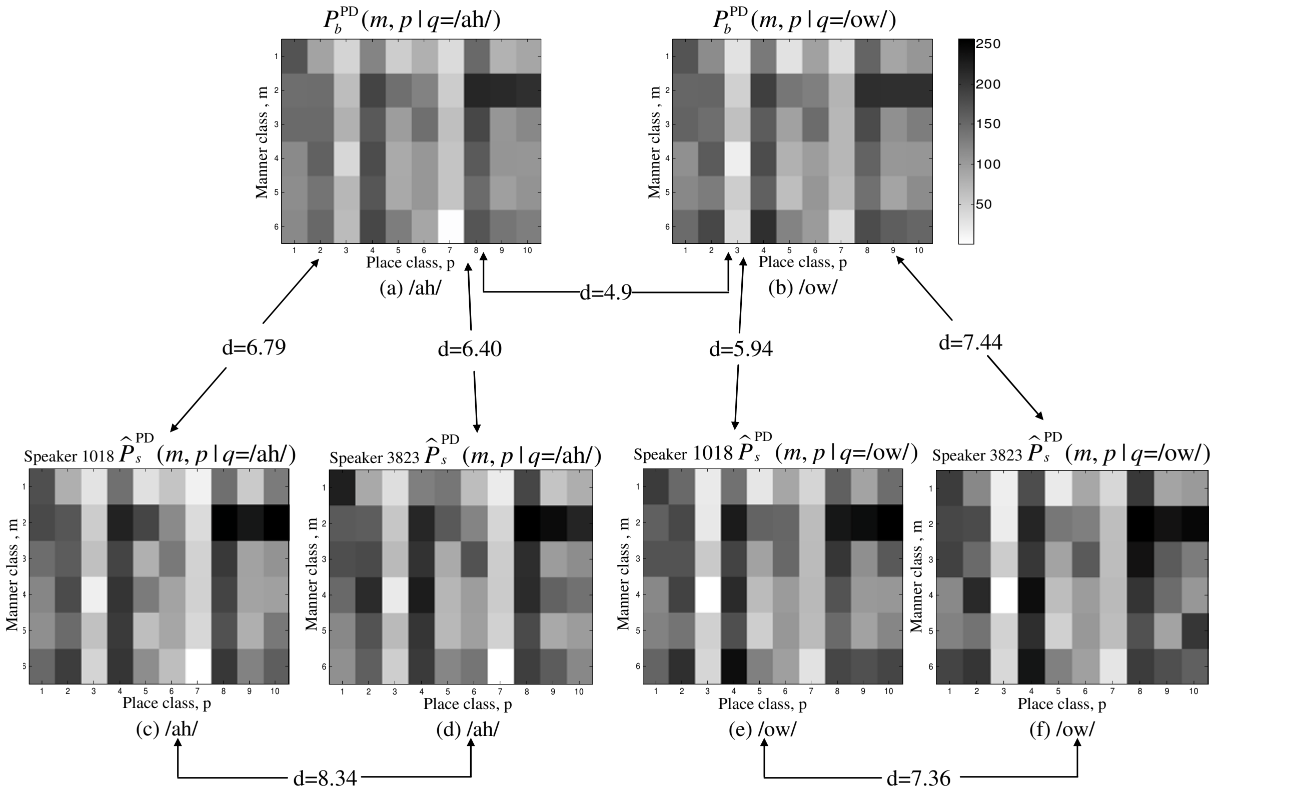

Fig. 2. Phoneme-dependent AFCPM background models correspond to (a) phoneme /ah/ and (b) phoneme /ow/ based on the training utterances in NIST99. (c) to (f): Phoneme-dependent speaker models of two speakers in NIST00 adapted from (a) and (b). d represents the Euclidean distance between the models pointed 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. The 6 manner and 10 places classes in ascending order of the axis labels are: {Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral} and {Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal}, respectively.

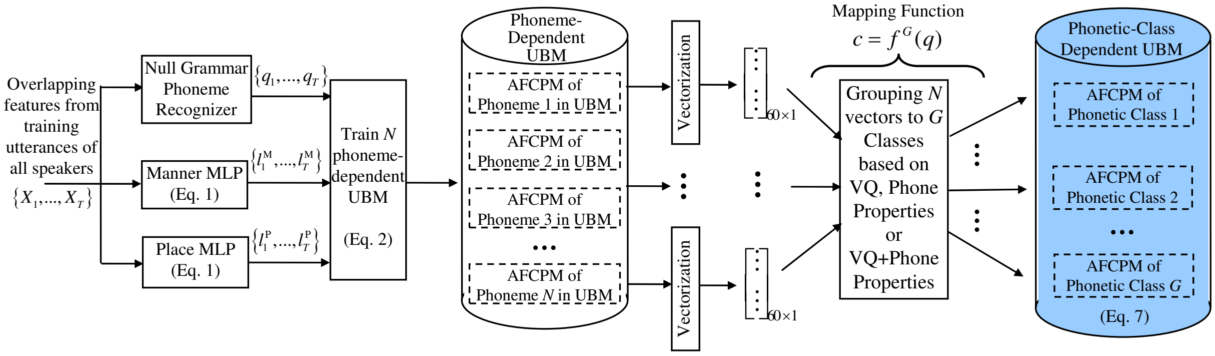
Overlapping features from training utterances of all speakers
$\{X_1,\ldots,X_T\}$
Null Grammar Phoneme Recognizer
$\{q_1,\ldots,q_T\}$
Manner MLP (Eq. 1)
$\{l_1^M,\ldots,l_T^M\}$
Place MLP (Eq. 1)
$\{l_1^P,\ldots,l_T^P\}$
Train N phoneme-dependent UBM
(Eq. 2)
Phoneme-Dependent UBM
AFCPM of Phoneme 1 in UBM
AFCPM of Phoneme 2 in UBM
AFCPM of Phoneme 3 in UBM
AFCPM of Phoneme N in UBM
Vectorization
Vectorization
60×1
60×1
Mapping Function
$c = f^G(q)$
Grouping N vectors to G Classes based on VQ, Phone Properties or VQ+Phone Properties
Phonetic-Class Dependent UBM
AFCPM of Phonetic Class 1
AFCPM of Phonetic Class 2
AFCPM of Phonetic Class G
(Eq. 7)

Fig. 3. The procedure of training the mapping function and creating the UBMs for the phonetic-class dependent AFCPM. $f^G(q) \in \{f^G_{\text{VQ}}(q), f^G_{\text{P}}(q), f^G_{\text{P+VQ}}(q)\}$, $N = 46$.

3) Method 3: Grouping based on Euclidean distance and phoneme properties: Note that Method 1 and Method 2 group phonemes according to different criteria. Specifically, the former is based on the articulatory properties, whereas the latter is based on continuant/noncontinuant properties of phonemes. For example, phonemes are grouped in part by the vertical positions (high, middle, and low) of the tongue via the place of articulation in Method 1, whereas they are grouped by the horizontal tongue positions (front, central, and back) in Method 2. Since these two ways of phoneme classification may complement each other, we propose a hybrid method based on the classical phoneme tree and Euclidean distance between AFCPMs to build the third mapping function:

$$c = f^G_{\text{P+VQ}}(q), \quad c \in \{1, 2, \ldots, G\}. \tag{6}$$

In this method, phonemes are grouped firstly by using phoneme properties. The phonemes in the same group are then further divided into subgroups by VQ. For example, all phonemes belonging to ‘Vowels’ in Table II are grouped together and then divided into 3 subgroups by using VQ. For the groups with a very small number of phonemes, such as ‘Affricates’, no clustering will be applied to the corresponding AFCPM vectors. A similar procedure is also applied to the phoneme groups with members that seldom appear in the training utterances. Table III shows the mapping function $f^G_{\text{P+VQ}}(q)$ used in this work.

Phoneme q	Class label for phoneme q		
	G=8	G=11	G=13
Front Vowels: iy, ih, ey, eh, ae	1	1	1
Mid Vowels: er, ax, ah		2	2
Back Vowels: uw, uh, ow, ao, aa		3	3
Voiced Fricatives: v, dh, z, zh	2	4	4
Unvoiced Fricatives: f, th, s, sh		5	5
Whisper: hh	3	6	6
Affricates: jh, ch	4	7	7
Diphthongs: ay, aw, oy	5	8	8
Liquids: r, l, el	6	9	9
Glides: w, y			10
Voiced Consonants: b, d, g	7	10	11
Unvoiced Consonants: p, t, k			12
Nasals: m, en, n, ng	8	11	13

TABLE II

THE MAPPING BETWEEN THE PHONEMES AND PHONETIC CLASSES BASED ON THE CLASSICAL PHONEME TREE FOR THREE DIFFERENT VALUES OF G. SEE APPENDIX I FOR THE DETAILED RELATIONSHIP BETWEEN THE PHONEMES AND THE PHONETIC CLASSES.

Phonetic Class c	**Phoneme** q	**Obtained by**
1	iy, uw, ih	P+VQ
2	er, uh, ax, ey	P+VQ
3	eh, ah, ow, ae, ao, aa	P+VQ
4	v, f, th, dh	P+VQ
5	z, zh, s, sh	P+VQ
6	hh	P
7	jh, ch	P
8	ay, aw, oy	P
9	r, l, el, w, y	P
10	b, d, p, t	P+VQ
11	g, k	P+VQ
12	m, en, n, ng	P

TABLE III

THE RELATIONSHIP BETWEEN PHONEMES AND PHONETIC CLASSES IN THE MAPPING FUNCTION $f^G_{\text{P+VQ}}(q)$, I.E., EQ. 6. VQ: VECTOR QUANTIZATION; P: PHONEME PROPERTIES. PHONEMES ARE FIRSTLY DIVIDED INTO 8 GROUPS ACCORDING TO THE PHONEME PROPERTIES (SEE TABLE VI IN APPENDIX I). THEN, SOME OF THESE GROUPS ARE FURTHER DIVIDED INTO SUBGROUPS VIA VQ.

B. Phonetic-Class Dependent UBMs

Given the mapping functions, phonetic-class dependent UBMs can be obtained as follows. For a particular phonetic class c, the joint probabilities of the phonetic-class dependent UBMs are determined by:

$$\begin{aligned}&P_b^{\text{CD}}(m,p|c)\\&=P_b^{\text{CD}}(L^{\text{M}}=m,L^{\text{P}}=p|\text{PhoneClass}=c,\text{Background})\\&=\frac{\#((m,p,c)\text{in the untterances of all background speaker }s)}{\#((*,*,c)\text{in the untterances of all background speaker }s)}\end{aligned} \tag{7}$$

where $m\in\mathcal{M},p\in\mathcal{P}$, (m,p,c) denotes the condition for which $L^{\text{M}}=m,L^{\text{P}}=p$, and $\text{PhoneClass}=c$.

Note that the accuracy of the mapping functions and hence the phonetic-class dependent UBMs depends on the amount of data in individual phonetic classes. Therefore, it is necessary to weight the models' density functions according to the amount of data available for training the mapping functions. Here, we propose to compute the weighting coefficients as follows:

$$w_c=\frac{\dfrac{\#\big((*,*,c)\text{ in the untterances of all background speakers}\big)}{\#\big((*,*,c)\text{ in the untterances of all background speakers}\big)+r_w}}{\displaystyle\sum_{c=1}^{G}\frac{\#\left((*,*,c)\text{ in the untterances of all background speakers}\right)}{\#\left((*,*,c)\text{ in the untterances of all background speakers}\right)+r_w}} \tag{8}$$

where $c\in\{1,\ldots,G\}$ and r_w is a relevance factor. These coefficients will be used for weighting the phonetic-class dependent speaker models (see Section IV-C below).

C. Phonetic-Class Dependent Speaker Models

A phonetic-class speaker model can be obtained from speaker-dependent data as follows (see Figure 5):

$$\begin{aligned}&P_s^{\text{CD}}(m,p|c)\\&=P_s^{\text{CD}}(L^{\text{M}}=m,L^{\text{P}}=p|\text{PhoneClass}=c,\text{Speaker}=s)\\&=\frac{\#((m,p,c)\text{ in the utterances of speaker }s)}{\#((*,*,c)\text{ in the utterances of speaker }s)}.\end{aligned} \tag{9}$$

Similar to the phoneme-dependent case, MAP adaptation is applied to obtain the final speaker model:[3]

$$\widehat{P}_s^{\text{CD}}(m,p|c)=\beta_c w_c P_s^{\text{CD}}(m,p|c)+(1-\beta_c)w_c P_b^{\text{CD}}(m,p|c) \tag{10}$$

where, $\beta_c\in[0,1]$ is a phonetic class-dependent adaptation coefficient controlling the contribution of the speaker data and the background models (Eq. 7) on the MAP-adapted model. It is obtained by:

$$\beta_c=\frac{\#((*,*,c)\text{ in the utterances of speaker }s)}{\#((*,*,c)\text{ in the utterances of speaker }s)+r_\beta} \tag{11}$$

where r_β is a fixed relevance factor common to all phonetic classes and speakers. Its purpose is to control the dependence of the adapted model on speaker's data.[4]

For each speaker, the accuracy of his/her phonetic-class models depends on the amount of training data for estimating the mapping functions. Therefore, it is intuitive to weight the density functions by the weighting coefficients w_c in Eq. 10. Alternatively, we may also train an MLP to optimally weight the phonetic classes, as in [25].

Figure 6 shows the background model for phonetic class $c=3$ of which phonemes /ah/ and /ow/ in Figure 2 are members. Also shown are the phonetic-class speaker models of speakers 1018 and 3823 in NIST00. Figures 6(b) and 6(c) show that the two phonetic-class speaker models are more distinctive (therefore more discriminative) than the phoneme-dependent speaker models shown in Fig. 2. The Euclidean distance d between the phonetic-class speaker models (Figures 6(b) and 6(c)) is also 1.3 times that of the phoneme-dependent models (Figures 2(c)–(f)): 11.08 vs. 8.34 and 7.36. Moreover, the distances between the speaker

[3] Although strictly speaking $\widehat{P}_s^{\text{CD}}(m,p|c)$ is not probability because of the weighting factor w_c, we use the symbol $\widehat{P}$ here for readability and consistency.

[4] Our results suggest that the system performance is not sensitive to the relevance factor, see Section VI-C.

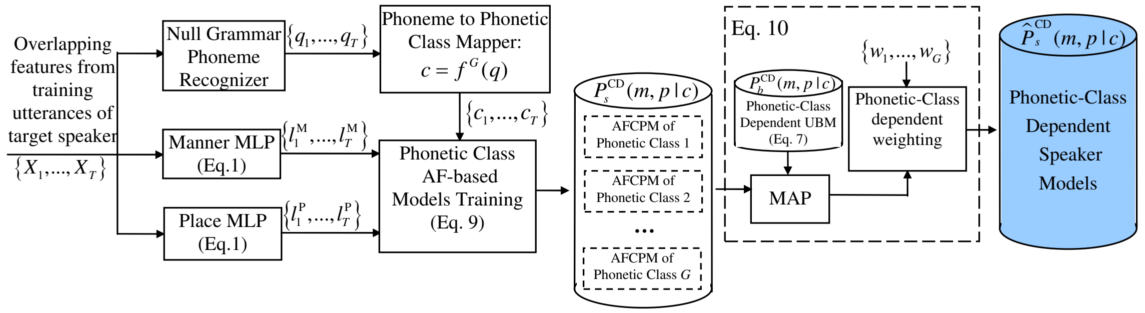
Overlapping features from training utterances of target speaker
$\{X_1,...,X_T\}$
Null Grammar Phoneme Recognizer
$\{q_1,...,q_T\}$
Phoneme to Phonetic Class Mapper: $c = f^G(q)$
$\{c_1,...,c_T\}$
Manner MLP (Eq.1)
$\{l_1^M,...,l_T^M\}$
Place MLP (Eq.1)
$\{l_1^P,...,l_T^P\}$
Phonetic Class AF-based Models Training (Eq. 9)
$P_s^{CD}(m,p|c)$
AFCPM of Phonetic Class 1
AFCPM of Phonetic Class 2
AFCPM of Phonetic Class G
Eq. 10
$P_b^{CD}(m,p|c)$
Phonetic-Class Dependent UBM (Eq. 7)
$\{w_1,...,w_G\}$
Phonetic-Class dependent weighting
MAP
$\hat{P}_s^{CD}(m,p|c)$
Phonetic-Class Dependent Speaker Models

Fig. 5. The procedure of training the phonetic-class dependent speaker models. $f^G(q) \in \{f^G_{\text{VQ}}(q), f^G_{\text{P}}(q), f^G_{\text{P+VQ}}(q)\}$.

models and the background models are also larger in the phonetic-class case, primarily because of more data are available for training the phonetic-class speaker models. All of these results suggest that phonetic-class dependent speaker models are more discriminative.

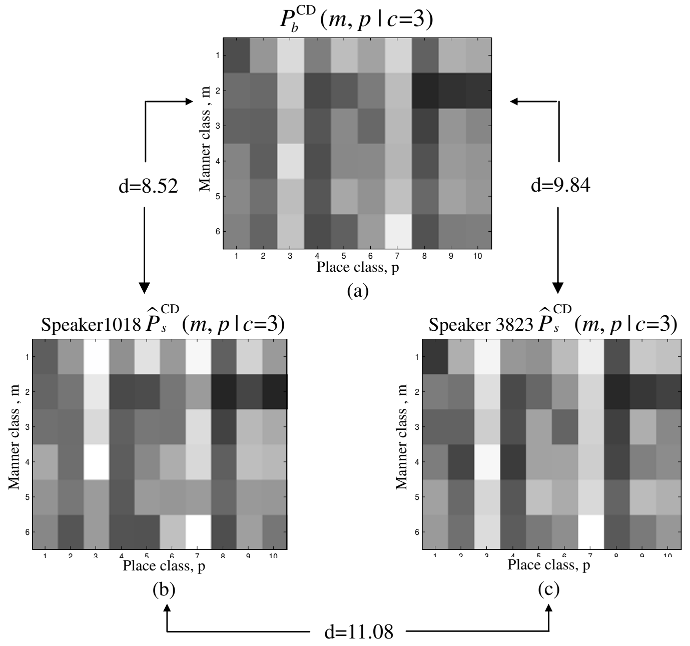

Fig. 6. Phonetic-class dependent models in which the phonemes /ah/ and /ow/ are members of the phonetic class ($c = 3$ in Table III). The speaker models were obtained from the training utterances of speakers 1018 and 3823 in NIST00, using the mapping function $f^G_{\text{P+VQ}}(q)$. d represents the Euclidean distance between the models pointed to by arrows. Refer to Figure 2 for the manner and place class labels.

D. Scoring Method

Following the scoring method in [21], we define the verification score of a test utterance $X = \{X_1, \ldots, X_t, \ldots, X_T\}$ as:

$$S_{\text{CD-AFCPM}}(X) = \sum_{t=1}^{T} \left[\log \hat{p}_s^{\text{CD}}(X_t) - \log p_b^{\text{CD}}(X_t)\right] \qquad (12)$$

where the speaker models (Eq. 10) and background models (Eq. 7) are used to compute the scores

$$\begin{aligned}\hat{p}_s^{\text{CD}}(X_t) &= \hat{P}_s^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|c_t)\\ &= \hat{P}_s^{\text{CD}}(L^{\text{M}} = l_t^{\text{M}}, L^{\text{P}} = l_t^{\text{P}}|\text{PhoneClass} = c_t, \text{Speaker} = s)\end{aligned} \qquad (13)$$

and

$$\begin{aligned}p_b^{\text{CD}}(X_t) &= P_b^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|c_t)\\ &= P_b^{\text{CD}}(L^{\text{M}} = l_t^{\text{M}}, L^{\text{P}} = l_t^{\text{P}}|\text{PhoneClass} = c_t, \text{Background}),\end{aligned} \qquad (14)$$

where $c_t = f^G(q_t)$ is the phonetic class of frame t.

For the acoustic GMM system, we applied feature transformation [26] and short-time Gaussianization [27] to reduce the effect of channel distortion. Then, acoustic scores S_{GMM} were computed based on GMM-UBM framework [1]:

$$S_{\text{GMM}}(X) = \sum_{t=1}^{T} [\log p(X_t|\Lambda_s) - \log p(X_t|\Lambda_b)] \qquad (15)$$

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

V. Experiments

A. Speech Corpora and Speech Features

NIST99 [28], NIST00 [29], SPIDRE [30], and HTIMIT [31] were used in the experiments. NIST99 was used for creating the background models and mapping functions, and 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.

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. For each speaker, approximately 2 minutes of speech is available for enrollment, 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 sex as the speaker of the verification utterance. This amounts to 6,096 speaker trials and 60,476 impostor attempts.

The acoustic features for training the HMMs and speaker models are slightly different. For the HMMs, acoustic vectors of 39 dimensions—each comprising of 12 Mel-frequency cepstral coefficients (MFCCs) [32], the normalized energy, and their first- and second-order derivatives—were used. For the MFCC-based and AFCPM-based speaker models, 19-dimensional MFCCs and their first-order derivatives were computed every 10ms using a Hamming window of 25ms. The MFCCs and delta MFCCs were concatenated to form 38-dimensional feature vectors. Cepstral mean subtraction (CMS), fast blind stochastic features transformation (fBSFT) [26], [3] and short-time Gaussianization (STG) [27] were applied to the MFCCs to remove channel effects.

The feature vectors for the AF-MLPs comprise 12-dimensional MFCCs and energy plus their derivatives. These vectors were extracted from speech signals at 100Hz using a Hamming windows of 25ms.

B. Training and Evaluation Procedures

3,794 utterances selected from HTIMIT were used to train the manner and place MLPs, 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).

The training part of NIST99 was used to create gender-dependent acoustic (MFCC-based) background models with 1024 mixtures. The same set of data was also used to build phoneme-dependent and phonetic-class dependent AF-based UBMs, which were subsequently used for obtaining the gender-dependent mapping functions based on the three methods mentioned in Section IV-A. Then, for each target speaker in NIST00, his/her speaker models were created using Eq. 10 and the 2-minute enrollment speech based on the mapping functions and the phonetic-class dependent UBMs.

We followed the evaluation protocol of NIST00. To ensure statistical significance of our results, we also computed the p-values [33] between the error rates obtained by phoneme-dependent AFCPM and phonetic-class dependent AFCPM.

C. Fusion of MFCC- and AFCPM-Based Systems

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 [12], [34]. The phonetic-class 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 acoustic GMMs were linearly combined to obtain the fused scores.

VI. Results and Discussion

A. Comparing Different Mapping Functions

Table IV shows the equal error rates (EERs) obtained by phoneme-dependent AFCPM (PD-AFCPM) and phonetic-class dependent AFCPM (CD-AFCPM) using the three phoneme-to-phonetic class mapping methods. It shows that the mapping function $f^G_{\text{P+VQ}}(q)$ achieves the lowest error rates in CD-AFCPM. This result suggests that phone properties and Euclidean distance between AF models (VQ) play a complementary role. We conjecture that the phone properties constrain the possible partitioning of phonemes and VQ provides a fine division within the phoneme groups where phone properties alone cannot entirely represent the articulatory properties of speech. In particular, for some large phoneme groups (e.g., vowels), it may be better to partition the groups into subgroups based on the distribution of the AF models than to divide the groups based purely on their phone properties. Completely relying on the distribution of AF models, however, is inappropriate because some constraints are essential for forming the large phoneme groups.

	Phoneme-to-Phonetic Class Mapping Method	No. of Classes G	EER(%)			
			Female		Male	
			Equally weighted	Class weighted	Equally weighted	Class weighted
CD-AFCPM	VQ $c = f^G_{\text{VQ}}(q)$	8	26.72	26.42	23.85	23.74
		10	25.22	24.93	23.70	23.65
		12	25.64	25.36	23.73	23.71
	Phone Properties $c = f^G_{\text{P}}(q)$	8	25.04	24.85	24.32	24.11
		11	24.13	23.92	23.31	23.24
		13	24.48	24.25	23.09	23.10
	Phone Properties+VQ $c = f^G_{\text{P+VQ}}(q)$	12	23.63	23.46	22.89	22.83
			Mix gender: **23.76**			
PD-AFCPM			26.35		24.66	
			Mix gender: **25.91**			
GMM (fBSFT)			Mix gender: **16.11**			
PD-AFCPM + GMM (fBSFT)			Mix gender: **15.91**			
CD-AFCPM + GMM (fBSFT)			Mix gender: **14.87**			
GMM (STG+fBSFT)			Mix gender: **13.81**			
PD-AFCPM + GMM (STG+fBSFT)			Mix gender: **13.71**			
CD-AFCPM + GMM (STG+fBSFT)			Mix gender: **13.16**			

TABLE IV

Equal error rates (EERs) obtained by acoustic GMM, phoneme-dependent AFCPM (PD-AFCPM) and phonetic-class dependent AFCPM (CD-AFCPM) using three different phoneme-to-phonetic class mapping methods. "Equally weighted" means that $w_c = 1$ in Eq. 10 for all c. Note that the fusion of phonetic-class AFCPM and GMM is based on the phonetic-class AFCPM that uses the mapping function $f^G_{\text{P+VQ}}$. The p-values between the PD-AFCPM and all of the CD-AFCPM and the p-value between PD-AFCPM+GMM and CD-AFCPM+GMM are less than 0.0001.

B. Comparing PD-AFCPM and CD-AFCPM

Table IV also shows that phonetic-class AFCPM, regardless of the type of mapping functions, is superior to phoneme-dependent AFCPM. This confirms our earlier argument that when the amount of enrollment data is limited, we had better to enrich the amount of training data per model by grouping similar phonemes together. We advocate phonetic-class dependent AFCPMs (especially the one that uses mapping function $c = f^G_{\text{P+VQ}}(q)$) for two reasons. First, unlike phoneme-dependent AFCPM where training data are divided into 46 classes, data are divided into a maximum of 13 classes only in phonetic-class dependent AFCPM. As a result, a lot more data are available for training each phonetic-class dependent AFCPM, which leads to more reliable speaker models under limited enrollment data. Second, because of the small number of classes, phonetic-class dependent AFCPM is less sensitive to the accuracy of the phoneme recognizer. In phoneme-dependent AFCPM, acoustically confusable phonemes may cause the phoneme recognizer to make mistakes, leading to erroneous scores. However, some of the confusable phonemes may be mapped to the same phonetic class in the case of phonetic-class dependent AFCPM, which effectively alleviate the effect caused by phoneme recognition errors. There seems to be a tradeoff between the number of models per speaker and the representation ability of the models. In particular, a large number of models (e.g., 46 in PD-AFCPM) could lead to inferior performance, as evident in Table IV.

C. Choice of Relevance Factors

Both the phoneme-dependent and phonetic-class dependent AFCPM use MAP adaptation. The discriminative power of the resulting speaker models depends on the amount of adaptation, which in turn depends on the relevance factors in the adaptation equations. To investigate the sensitivity of the phonetic-class dependent AFCPM to the relevance factors, we randomly selected half of the data from NIST00 and varied the relevance factor r_β in Eq. 11. The EER performance is shown in Table V. 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 should not 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 380 in an attempt to avoid these extreme scenarios.

r_β	180	280	380	480	580
EER (%)	23.78	23.68	23.46	23.55	23.75

TABLE V
THE EFFECT OF VARYING THE RELEVANCE FACTOR r_β IN EQ. 11 ON THE SYSTEM PERFORMANCE.

D. Fusion of Low- and High-Level Features

Table IV shows that the UBM-GMM system that uses acoustic features as inputs achieves a significantly lower error rate as compared to the system that uses high-level features. The inferiority of high-level features is primarily due to the short verification utterances (15–45 seconds). However, fusing the scores obtained from these systems can lower the error rates further. The table also shows that fusion of phonetic-class AFCPM and GMM outperforms the fusion of phoneme-dependent AFCPM and GMM. The lowest error rate is achieved by fusing CD-AFCPM and GMM where the low-level features have been transformed by short-time Gaussianization (STG) and blind stochastic feature transformation (BSFT). Note that the p-values between the PD-AFCPM and all of the CD-AFCPMs are less than 0.0001, so as the p-value between PD-AFCPM and CD-AFCPM in the fusion cases. This suggests that fusion of low- and high-level features can bring significant performance gain, although the gain diminishes progressively when the low-level features become more robust. Making the low-level features robust, however, does not come without a price. It has been shown recently that using STG and fast BSFT as feature preprocessors requires 52 seconds to process a 53-second utterance on a Pentium IV 3.2GHz CPU, whereas processing the same utterance by the less powerful cepstral mean subtraction takes only 0.02 seconds [3].

The detection error tradeoff (DET) curves [35] corresponding to Table IV are shown in Figure 7. Evidently, the fusion of phonetic-class AFCPM and GMM achieves the best performance across a wide range of decision threshold. It is obvious that the high-level information captured by the phonetic-class dependent AFCPMs complements the short-term spectral information very well.

VII. CONCLUDING REMARKS

Phoneme-based AFCPM represents the pronunciation characteristics of speakers by building one discrete density function for each phoneme, which requires a large amount of training data to achieve high verification accuracy. Based on the observation that the AFCPM of some phonemes are very similar, this paper proposes a speaker verification system that uses phonetic class-based articulatory pronunciation models. Specifically, speaker models are represented by conditional probabilities of articulations given phonetic classes instead of phonemes. Three mapping functions that specify the relationship between phonemes and phonetic classes are proposed. Results show that among the three mapping functions, the one that combines the classical phoneme tree and Euclidean distance between AFCPMs achieves the best performance. Experimental results also show that phonetic-classes AFCPM achieves a significantly lower error rate as compared to conventional AFCPM.

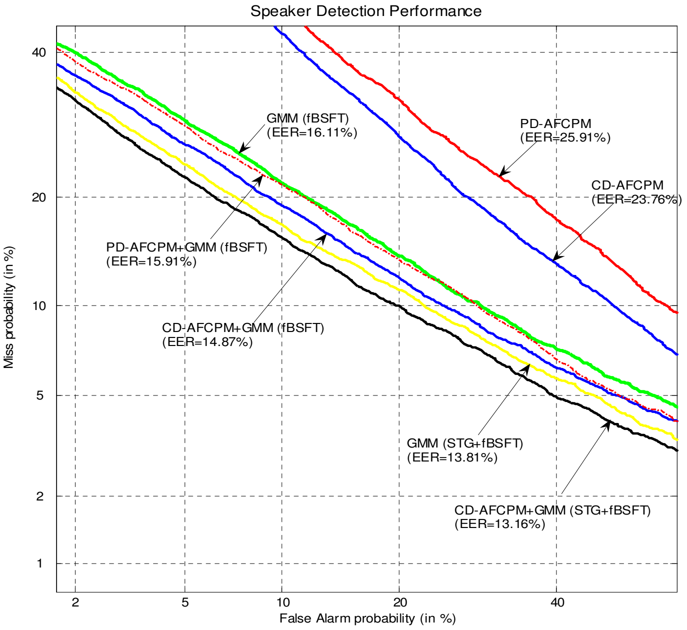

Fig. 7. Detection error tradeoff (DET) performance of phonetic-class dependent AFCPM (CD-AFCPM), phoneme-dependent AFCPM (PD-AFCPM), GMM (with fBSFT and STG applied), and their fusions. All curves are based on mix-gender scores.

APPENDIX I
PHONEMES AND PHONETIC CLASSES

Table VI shows the relationship between phonemes and phonetic classes obtained from the classical phoneme tree for three different number of classes.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their constructive comments and suggestions. This work was in part supported by the Research Grant Council of the Hong Kong SAR (Project Nos. PolyU 5230/05E and CUHK1/02C).

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] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *IEEE ICASSP*, 2003, vol. 2, pp. 6–10.

Phonetic Class c	**Phone Type**	**Phoneme** q
1	Vowels	iy, ih, ey, eh, ae, er, ax, ah, uw, uh, ow ao, aa
2	Fricatives	v, dh, z, zh, f, th, s, sh
3	Whisper	hh
4	Affricates	jh, ch
5	Diphthongs	ay, aw, oy
6	Semivowels	r, l, el, w, y
7	Consonants	b, d, g, p, t, k
8	Nasals	m, en, n, ng

(a)

Phonetic Class c	**Phone Type**	**Phoneme** q
1	Front Vowels	iy, ih, ey, eh, ae
2	Mid Vowels	er, ax, ah
3	Back Vowels	uw, uh, ow, ao, aa
4	Voice Fricatives	v, dh, z, zh
5	Unvoiced Fricatives	f, th, s, sh
6	Whisper	hh
7	Affricates	jh, ch
8	Diphthongs	ay, aw, oy
9	Semivowels	r, l, el, w, y
10	Consonants	b, d, g, p, t, k
11	Nasals	m, en, n, ng

(b)

Phonetic Class c	**Phone Type**	**Phoneme** q
1	Front Vowels	iy, ih, ey, eh, ae
2	Mid Vowels	er, ax, ah
3	Back Vowels	uw, uh, ow, ao, aa
4	Voiced Fricatives	v, dh, z, zh
5	Unvoiced Fricatives	f, th, s, sh
6	Whisper	hh
7	Affricates	jh, ch
8	Diphthongs	ay, aw, oy
9	Liquids Semivowels	r, l, el
10	Glides Semivowels	w, y
11	Voiced Consonants	b, d, g
12	Unvoiced Consonants	p, t, k
13	Nasals	m, en, n, ng

(c)

TABLE VI

THE RELATIONSHIP BETWEEN PHONEMES AND PHONETIC CLASSES OBTAINED FROM THE CLASSICAL PHONEME TREE [24] WHEN (A) $G = 8$, (B) $G = 11$, AND (C) $G = 13$.

[3] 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*, 2007.

[4] 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.

[5] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[6] 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.

[7] 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.

[8] 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.

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

[10] S. Shaiman, S. C. Adams, and M. D. Z. Kimelman, "Timing relationships of the upper lip and jaw across changes in speaking rate," *J. of Phonetics*, vol. 23, pp. 119–128, 1995.

[11] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eurospeech'01*, Aalborg, Sept. 2001, pp. 2521–2524.

[12] 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.

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

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

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

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

[17] J. P. Campbell and D. A. Reynolds, "Conditional pronunciation modeling in speaker detection," in *Proc. ICASSP'99*, 1999, vol. 2, pp. 829–832.

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

[19] 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*, 2003, vol. 4, pp. 792–795.

[20] D. Klusacek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," in *Proc. ICASSP'03*, 2003, vol. 4, pp. 804–807.

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

[22] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, PhD thesis, University of Bielefeld, 1999.

[23] P. Frber, "Quicknet on multispert: fast parallel neural network training," Tech. Rep. TR-97-047, ICSI, 1998.

[24] J. R. Deller Jr, J. G. Proakis, and J. H. L. Hansen, *Discrete-time Processing of Speech Signals*, Macmillan Pub. Company, 1993.

[25] R. Auckenthaler, E. Parris, and M. Carey, "Improving a GMM speaker verification system by phonetic weighting," in *Proc. IEEE ICASSP*, 1999, pp. 1440–1444.

[26] K. K. Yiu, M. W. Mak, M. C. Cheung, and S. Y. Kung, "Blind stochastic feature transformation for channel robust speaker verification," *J. of VLSI Signal Processing*, vol. 42, no. 2, pp. 117–126, 2006.

[27] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. ICASSP'02*, 2002, vol. 1, pp. 681–684.

[28] "The NIST year 1999 speaker recognition evaluation plan," in *http://www.nist.gov/speech/tests/spk/1999/doc*.

[29] "The NIST year 2000 speaker recognition evaluation plan," in *http://www.nist.gov/speech/tests/spk/2000/doc*.

[30] 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.

[31] 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.

[32] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, no. 4, pp. 357–366, August 1980.

[33] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP'89*, 1989, pp. 532–535.

[34] S. Y. Kung and M. W. Mak, "Machine learning for multi-modality genomic signal processing," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 117–121, May 2006.

[35] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech'97*, 1997, pp. 1895–1898.

Shi-Xiong Zhang received the B.Eng. degree in electrical engineering from Jilin University in 2005. Currently he is a graduate student in the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University. His research interests include speaker verification, speech recognition and machine learning.

Man-Wai Mak received a BEng(Hons) degree in Electronic Engineering from Newcastle Upon Tyne Polytechnic and a PhD degree in Electronic Engineering from the University of Northumbria at Newcastle. He joined the Department of Electronic Engineering at the Hong Kong Polytechnic University as a Lecturer in 1993 and as an Assistant Professor in 1995. He has authored more than 90 technical papers in speaker recognition, machine learning, and bioinformatics. Dr. Mak is also a co-author of the postgraduate textbook Biometric Authentication: A Machine Learning Approach, Prentice Hall, 2005. Dr. Mak received the Faculty of Engineering Research Grant Achievement Award in 2003. Since 1995, Dr. Mak has been an executive committee member of the IEEE Hong Kong Section Computer Chapter. He was the Chairman of the IEEE Hong Kong Section Computer Chapter in 2003-2005 and is currently a member of the IEEE Machine Learning for Signal Processing Technical Committee. Dr. Mak's research interests include speaker recognition, machine learning, and bioinformatics.

Helen M. Meng (M'98) received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge. She has been a Research Scientist at the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She joined The Chinese University of Hong Kong in 1998, where she is currently Professor in the Department of Systems Engineering and Engineering Management and Associate Dean of Research of the Faculty of Engineering. In 1999, she established the Human-Computer Communications Laboratory at CUHK and serves as director. In 2005, she established the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies and serves as Co-Director. Helen's research interest is in the area of human-computer interaction via multimodal and multilingual spoken language systems, as well as translingual speech retrieval technologies. She serves as Associate Editor of the IEEE Transactions of Speech, Audio and Language Processing. She is also a member of Sigma Xi and the International Speech Communication Association.