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

Shi-Xiong Zhang[1], *Man-Wai Mak*[1], *and Helen Meng*[2]

[1]Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University
[2]Dept. of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

Abstract

Although articulatory feature-based conditional pronunciation models (AFCPMs) can capture the pronunciation characteristics of speakers, they requires one discrete density function for each phoneme, which may lead to inaccurate models when the amount of training data is limited. This paper proposes a phonetic-class based AFCPM in which the density functions in speaker models are conditioned on phonetic classes instead of phonemes. Phonemes are mapped to phonetic classes by (1) vector quantizing the phoneme-dependent universal background models, (2) grouping phonemes according to the classical phoneme tree, and (3) combination of (1) and (2). A new scoring method that uses an SVM to combine the scores of phonetic-class models is also proposed. Evaluations based on 2000 NIST SRE show that the proposed approach can effectively solve the data sparseness problem encountered in conventional AFCPM.

1. Introduction

Previous studies have shown that combining low-level acoustic information with high-level speaker information—such as the usage or duration of particular words, prosodic features and articulatory features—can improve speaker verification performance [1, 2]. In particular, it was found that the conditional pronunciation modeling (CPM) technique [3] is able to capture the variation in speakers' pronunciation, leading to the best performance in a benchmark evaluation [1]. CPM aims to model speaker-specific pronunciation by learning the relationship between what phonemes have been said and how these phonemes are pronounced. The idea is further improved in [4] where articulatory-feature (AF) streams are used to construct conditional pronunciation models, leading to the so-called AFCPM. Because AFs are closely related to the speech production process, they are suitable for capturing the pronunciation characteristics of speakers.

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 inaccurate speaker models and poor performance. Moreover, because 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 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 Fig. 1. Evidently, there is substantial similarity between the two background models (Figures 1(a) and 1(b)). Comparisons between Figures 1(c) and 1(d) and between Figures 1(e) and 1(f) also reveal that the models of speaker 1018 are very similar to those of speaker 3823.

This work was supported by the Research Grant Council of the Hong Kong SAR No. CUHK1/02C and HKPolyU No. A-PA6F.

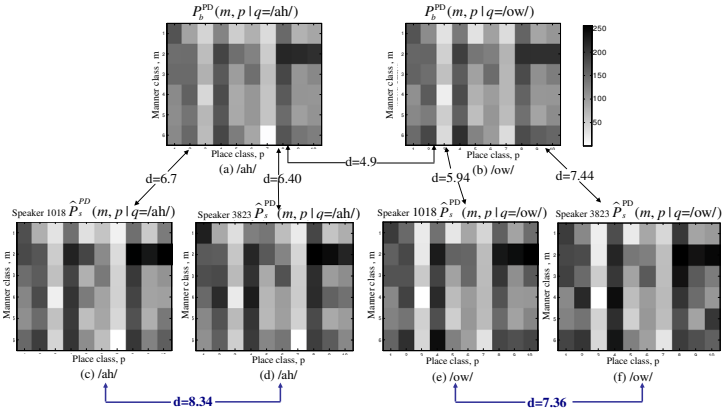

Figure 1: Phoneme-dependent AFCPM background models correspond to (a) phoneme /ah/ and (b) phoneme /ow/. (c) to (f): Phoneme-dependent speaker models adapted from (a) and (b). d represents the Euclidean distance between the models connected by the arrows. The 60 discrete probabilities corresponding to the combinations of the 6 manner and 10 place classes are nonlinearly quantized to 256 gray levels using log scale. See Section 2.1 for the details of manner and place classes.

To improve the accuracy of articulatory feature-based models, this paper proposes to group similar phonemes into phonetic classes by using a mapping function and to represent the background and speaker models as phonetic-class dependent density functions. Three approaches to determining the mapping function are evaluated. A new scoring method that uses an SVM to combine the scores obtained from different phonetic-class dependent models is also proposed. It was found that this phonetic-class AFCPM approach can effectively solve the data sparseness problem encountered in conventional AFCPM, resulting in a significantly lower error rate.

2. Phonetic-Class Dependent AFCPM

2.1. Articulatory Features

Articulatory features (AFs) are representations describing the movements or positions of different articulators during speech production. In Leung et al. [4], 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) shown in Fig. 2.

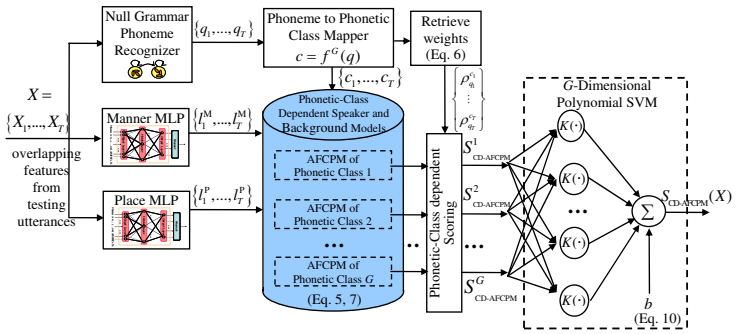

Figure 2: The verification phase of phonetic-class dependent AFCPM.

2.2. Mapping Functions

Because the AF of some phonemes are very similar, it makes sense to group these phonemes into a phonetic class such that the joint probabilities in the AFCPMs are conditioned on phonetic classes instead of individual phonemes. There are several ways of grouping phonemes: (1) according to the similarity (Euclidean distance) between the AFCPMs, (2) according to the phoneme properties as depicted in the classical phoneme tree [5], and (3) combination of (1) and (2).

Method 1: Grouping based on Euclidean distance. The phoneme-dependent UBMs, $P_b^{\text{PD}}(m,p|q)$ in [4], are vectorized to N 60-dimensional vectors called AFCPM vectors:

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

where $q \in \{\text{Phoneme } 1, \ldots, \text{Phoneme } N\}$, and L^{M} and L^{P} represent the manner and place labels, respectively. Then, K-means clustering or VQ can be applied to cluster the N AFCPM vectors into G ($< N$) classes. The mapping from a specific phoneme to its corresponding phonetic class index c is defined as a mapping function:

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

Method 2: Grouping based on phoneme properties. Because the phoneme grouping in the classical phoneme tree [5] 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

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

Table 1 shows the mapping between the phonemes and phonetic classes for three different values of G.

Method 3: Grouping based on Euclidean distance and phoneme properties. Note that Methods 1 and 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. Because these two ways of phoneme classification may complement each other, we propose to combine these two methods. Specifically, phonemes are firstly grouped by using phoneme properties to form a number of phoneme groups. The phonemes in the same group are then further divided into subgroups by VQ. For example, all phonemes belonging to 'Vowels' in Table 1 are grouped together and then divided into 3 subgroups by using VQ. This hybrid approach results in the third mapping function:

$$c = f_{\text{P+VQ}}^G(q), \quad c \in \{1, 2, \ldots, G\}. \tag{3}$$

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 1: The mapping between the phonemes and phonetic classes based on the classical phoneme tree.

2.3. Phonetic-Class Dependent UBMs

Given the mapping functions, the phonetic-class dependent UBM of phonetic class c can be determined by:

$$\begin{aligned} & P_b^{\text{CD}}(m,p|c) \\ &= P_b^{\text{CD}}(L^{\text{M}} = m, L^{\text{P}} = p|\text{PhoneClass} = c, \text{Background}) \\ &= \frac{\#((m,p,c)\text{in the utterances of all background speakers})}{\#((*,*,c)\text{in the untterances of all background speakers})} \\ &= \frac{\sum_{t\in\mathcal{T}_{\text{b}}} 1}{\sum_{t\in\mathcal{T}_{\text{b}}'} 1}, \qquad m \in \mathcal{M}, p \in \mathcal{P}, c \in \{1, \ldots, G\} \end{aligned} \tag{4}$$

where $\mathcal{T}_{\text{b}} = \{t : l_t^{\text{M}} = m, l_t^{\text{P}} = p, f^G(q_t) = c, X_t \in \text{all background speakers}\}$, $\mathcal{T}_{\text{b}}' = \{t : f^G(q_t) = c, X_t \in \text{all background speakers}\}$, L^{M} and L^{P} represent the manner and place labels, respectively, and l_t^{M} and l_t^{M} are the manner and place labels determined by the Manner and place MLPs, respectively. Eq. 4 suggests that all frames are weighted equally. However, frames that have a higher probability of belonging to phonetic class c should be given a higher weight. Therefore, it is intuitive to weight the contribution of frame t as follows:

$$P_b^{\text{CD}}(m,p|c) = \frac{\sum_{t\in\mathcal{T}_{\text{b}}} \rho_{q_t}^c}{\sum_{t\in\mathcal{T}_{\text{b}}'} \rho_{q_t}^c}, \tag{5}$$

where $\rho_{q_t}^c \equiv P(c|q_t)$ is the probability of phoneme q_t belonging to phonetic class c, which can be approximated by $P(c|\mathbf{a}_{q_t})$. Note that $P(c|\mathbf{a}_{q_t})$ is inversely proportional to $\|\mathbf{a}_{q_t} - \mathbf{m}_c\|$, where $\mathbf{m}_c$ is the centroid of phonetic class c. Therefore, we approximate the weights $\rho_{q_t}^c$ by:

$$\rho_{q_t}^c \approx P(c|\mathbf{a}_{q_t}) \approx \frac{\exp\left(-\frac{1}{2}\|\mathbf{a}_{q_t} - \mathbf{m}_c\|^2\right)}{\sum_{c'\in\mathcal{C}_i} \exp\left(-\frac{1}{2}\|\mathbf{a}_{q_t} - \mathbf{m}_{c'}\|^2\right)}, \; c \in \mathcal{C}_i \tag{6}$$

where $\mathcal{C}_i$ represents the phonetic classes in the i-th group.

2.4. Phonetic-Class Dependent Speaker Models

Target speaker models are obtained in two steps. In the first step, we compute:

$$\begin{aligned} & P_s^{\text{CD}}(m,p|c) \\ &= P_s^{\text{CD}}(L^{\text{M}}=m, L^{\text{P}}=p|\text{PhoneClass}=c, \text{Speaker}=s) \\ &= \frac{\sum_{t\in\mathcal{T}_s}\rho_{q_t}^c}{\sum_{t\in\mathcal{T}_s'}\rho_{q_t}^c}, \qquad m\in\mathcal{M}, p\in\mathcal{P}, c\in\{1,\ldots,G\} \end{aligned}$$

where $\mathcal{T}_s = \{t : l_t^{\text{M}} = m, l_t^{\text{P}} = p, f^G(q_t) = c,\ X_t \in \text{speaker } s\}$ and $\mathcal{T}_s' = \{t : f^G(q_t) = c,\ X_t \in \text{speaker } s\}$. Then in the second step, MAP adaptation is applied to obtain the model of target speaker s:

$$\widehat{P}_s^{\text{CD}}(m,p|c) = \beta_c P_s^{\text{CD}}(m,p|c) + (1-\beta_c)P_b^{\text{CD}}(m,p|c) \quad (7)$$

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. 5) on the MAP-adapted model. It is obtained by

$$\beta_c = \frac{\#((*,*,c) \text{ in the utterances of speaker } s)}{\#((*,*,c) \text{ in the utterances of speaker } s) + r_\beta} \quad (8)$$

where r_β is a fixed relevance factor common to all phonetic classes and speakers.

Fig. 3 shows the background model for phonetic class $c = 3$ of which phonemes /ah/ and /ow/ in Fig. 1 are members. Also shown are the phonetic-class speaker models of speakers 1018 and 3823 in NIST00. Figures 3(b) and 3(c) show that the two phonetic-class speaker models are more distinctive (therefore more discriminative) than the phoneme-dependent speaker models shown in Fig. 1. The Euclidean distance d between the phonetic-class speaker models (Figures 3(b) and 3(c)) is also larger than that of the phoneme-dependent models (Figures 1 (c)–(f)): 11.08 vs. 8.34 and 7.36. Moreover, the distances between the speaker 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.

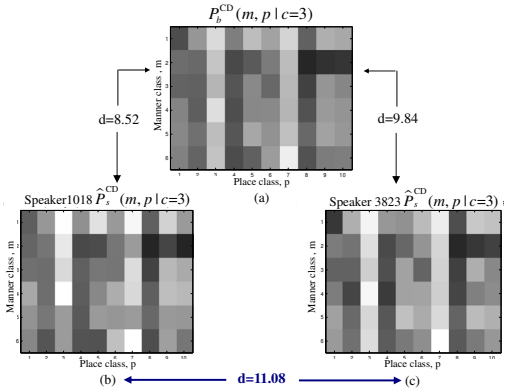

Figure 3: Phonetic-class dependent models in which the phonemes /ah/ and /ow/ are members of the phonetic class. The speaker models were obtained from the training utterances of speakers 1018 and 3823 in NIST00, using the mapping function $f_{\text{P+VQ}}^G(q)$. d represents the Euclidean distance between the models connected by arrows.

2.5. SVM Scoring

The conventional scoring methods sum the likelihood ratios of a test utterance $X = \{X_1, \ldots, X_T\}$ in a frame-by-frame basis:

$$\begin{aligned} S_{\text{CD-AFCPM}}(X) &= \sum_{t=1}^{T} \rho_{q_t}^c \left[\log \widehat{p}_s^{\text{CD}}(X_t) - \log p_b^{\text{CD}}(X_t)\right] \\ &= \sum_{c=1}^{G} \left(\sum_{t:f^G(q_t)=c} \rho_{q_t}^c \left[\log \widehat{p}_s^{\text{CD}}(X_t) - \log p_b^{\text{CD}}(X_t)\right] \right) \\ &= \sum_{c=1}^{G} S_{\text{CD-AFCPM}}^c \end{aligned} \quad (9)$$

where frame t is weighted by $\rho_{q_t}^c$, the probability of phoneme q_t belonging to phonetic class c. The speaker models (Eq. 7) and background models (Eq. 5) are used to compute the scores

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

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

where $c_t = f^G(q_t)$ is the phonetic class of frame t, and l_t^{M} and l_t^{P} are the AF labels determined by the AF-MLPs.

Eq. 9 suggests that the traditional scoring method treats all phonetic classes equally. In general, a summation of scores, as in Eq. 9, is likely to give suboptimal solutions. Better results may be obtained by applying an SVM to merge the phonetic-class dependent scores. Specifically, for each training utterance, the CD-AFCPM scores ($S_{\text{CD-AFCPM}}^c$) derived from the G phonetic classes form a G-dimensional score vector $\overrightarrow{S} = [S_{\text{CD-AFCPM}}^1, \ldots, S_{\text{CD-AFCPM}}^G]^T$. Vectors from target speakers and background speakers are then used to train an SVM (see Fig. 2):

$$S_{\text{CD-AFCPM}}(X) = \sum_{i=1}^{N} y_i \alpha_i K(\overrightarrow{S}, \overrightarrow{S_i}) + b. \quad (10)$$

3. Experiments and Results

3.1. Procedures

NIST99, NIST00 [6], SPIDRE [7], and HTIMIT [8] 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.

The phone recognizer uses standard 39-D vectors comprising MFCCs, energy, and their derivatives. The acoustic-based and AFCPM-based speaker models use 38-D vectors comprising 19-D MFCCs and their first derivative computed every 10ms. Cepstral mean subtraction (CMS), fast blind stochastic features transformation (fBSFT) [9] and short-time Gaussianization (STG) [10] were applied to the MFCCs to remove channel effects. Acoustic scores S_{GMM} were computed based on GMM-UBM framework [11]. The scores from AFCPMs and the acoustic GMMs were linearly combined to obtain the fused scores.

The training part of NIST99 was used for creating 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 then used for determining the mapping functions. Then, for each target speaker in NIST00, his/her speaker models were created using Eq. 7 and the 2-minute

enrollment speech based on the mapping functions and the phonetic-class dependent UBMs. 10-fold cross validation was used to train and evaluate a polynomial SVM for combining the scores of the phonetic-class models.

3.2. Results and Discussion

Table 2 shows that the mapping function $f^G_{\text{P+VQ}}(q)$ achieves the lowest error rates in CD-AFCPM, suggesting that phone properties and Euclidean distance between AF models 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.

Table 2 also shows that phonetic-class AFCPM 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. The insensitivity to the accuracy of the phoneme recognizer may attribute to the superiority of phonetic-class dependent AFCPMs. 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 2.

Table 2 shows that low-level features achieve a lower EER than the high-level features and that fusing the scores obtained from low- and high-level features can reduce the error rates further. The inferiority of high-level features is primarily due to the short verification utterances (15–45 seconds). The DET plots corresponding to Table 2 are shown in Fig. 4. Evidently, the fusion of phonetic-class AFCPM and GMM achieves the best performance across a wide range of decision threshold.

It is of interest to see how the speech recognizer affects SV performance, e.g., replace the null-grammar recognizer by a full-brown one. We are currently working in this direction.

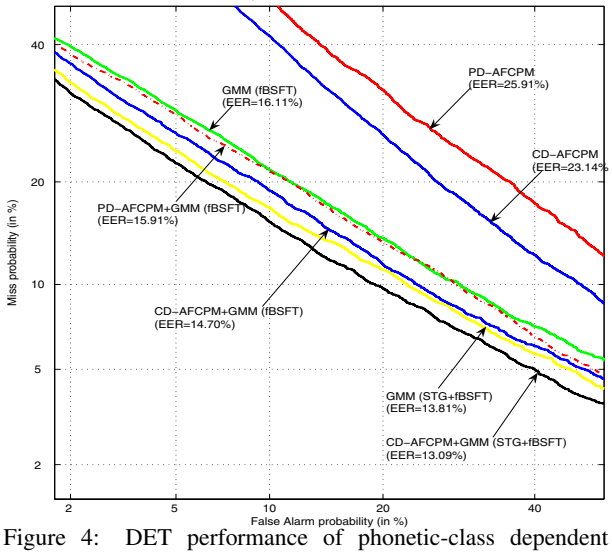

Figure 4: DET performance of phonetic-class dependent AFCPM (CD-AFCPM), phoneme-dependent AFCPM (PD-AFCPM), GMM (with fBSFT and STG applied), and their fusions. All curves are based on mix-gender scores.

	Phoneme-to-Phonetic Class Mapping Method	No. of Classes G	EER(%)			
			Female		Male	
			Old Scoring	SVM Scoring	Old Scoring	SVM Scoring
CD-AFCPM	VQ $c = f^G_{\text{VQ}}(q)$	8	26.72	26.33	23.85	23.68
		10	25.22	24.84	23.70	23.61
		12	25.64	25.30	23.73	23.66
	Phone Properties $c = f^G_{\text{P}}(q)$	8	25.04	24.85	24.32	24.07
		11	24.13	23.89	23.31	23.23
		13	24.48	24.25	23.09	23.20
	Phone Properties+VQ $c = f^G_{\text{P+VQ}}(q)$	12	23.42	23.31	22.88	22.81
			Mix gender: **23.14**			
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: **14.70**			
GMM (STG+fBSFT)			Mix gender: **13.81**			
PD-AFCPM + GMM (STG+fBSFT)			Mix gender: **13.71**			
CD-AFCPM + GMM (STG+fBSFT)			Mix gender: **13.09**			

Table 2: EERs obtained by acoustic GMM, phoneme-dependent AFCPM (PD-AFCPM) and phonetic-class dependent AFCPM (CD-AFCPM) using three different phoneme-to-phonetic class mappings. The fusion of PC-AFCPM and GMM is based on the PC-AFCPM that uses the mapping function $f^G_{\text{P+VQ}}$. 'Old Scoring' and 'SVM Scoring' mean that Eqs. 9 and 10 were used for scoring, respectively. 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.

4. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.