SPEAKER VERIFICATION USING ADAPTED ARTICULATORY FEATURE-BASED CONDITIONAL PRONUNCIATION MODELING

Ka-Yee Leung[1], Man-Wai Mak[1], Manhung Siu[2], and Sun-Yuan Kung[3]

[1]Center for Multimedia Signal Processing, Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
[2]Dept. of Electrical and Electronic Engineering, Hong Kong University of Science and Technology
[3]Dept. of Electrical Engineering, Princeton University

ABSTRACT

This paper proposes an articulatory feature-based conditional pronunciation modeling (AFCPM) technique for speaker verification. The technique captures the pronunciation characteristics of speakers by modeling the linkage between the actual phones produced by the speakers and the state of articulations during speech production. The speaker models, which consist of conditional probabilities of two articulatory classes, are adapted from a set of universal background models (UBMs) via MAP adaptation. This creates a direct coupling between the speaker and background models, which prevents over-fitting the speaker models when the amount of speaker data is limited. Experimental results demonstrate that MAP adaptation not only enhances the discriminative power of the speaker models but also improves their robustness against handset mismatches. Results also show that fusing the scores derived from an AFCPM-based system and a conventional spectral-based system achieves an error rate that is significantly lower than that can be achieved by the individual systems. This suggests that AFCPM and spectral features are complementary to each other.

1. INTRODUCTION

State-of-the-art text-independent speaker recognition systems typically use Gaussian mixture models (GMMs) [1] to represent the short-term spectral characteristics of speakers. The advantage of spectral-based systems is that promising results are obtainable from a limited amount of training data. However, except for spectral characteristics, these systems ignore other information in speech signals, which is useful for human to recognize speakers.

In recent years, researchers have started to investigate the use of high-level features, such as the usage or duration of particular words, prosodic features, etc., for speaker recognition [2]. Their work has demonstrated that these features contain different amount of speaker-dependent information and the best performance was achieved by a system that uses conditional pronunciation modeling (CPM) techniques [3]. Based on the fact that different speakers have different ways of pronouncing the same phoneme, CPM characterizes the pronunciation behavior of speakers by computing the correlation between the intended phonemes and the actual phones produced. The pronunciation behavior is encoded as discrete probability densities that are used for verifying speakers similar to the conventional GMMs in spectral-based systems. However, CPM requires multilingual speech data for training the phone models of different languages and long utterances for speaker enrollment and verification.

To avoid the requirement of multilingual training data, Leung et al. [4] proposed using articulatory feature (AF) streams to construct conditional pronunciation models. AFs are abstract classes describing the movements or positions of different articulators during speech production [5]. Compared to phone-based CPM in [3], AF-based CPM provides a more direct coupling between the pronunciation variations and the speech production process. Because the speech production process is a source of speaker variations, AF-based CPM is better than phone-based CPM in terms of speaker modeling. In addition, articulatory properties are the same irrespective of languages, therefore monolingual speech data are sufficient for determining their values. In Leung et al. [4], significantly shorter utterances were used to enroll and verify speakers when compared to those required in Klusáček et al. [3]. This has important computation implication for large-scale deployment.

In Leung et al. [4], the discrete distribution of each speaker model was estimated exclusively from the enrollment data of the corresponding speaker. This may lead to over-trained speaker models unless abundant enrollment data are available. To solve this problem, this paper proposes an adaptation approach in which the discrete distributions of speaker models are adapted from those of universal background models.

2. AF-BASED CPM

2.1. 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. AFs have been applied to speaker identification [6] and speaker verification [7]. In [6], speaker identification was performed by fusing the scores derived from seven speaker-dependent language models, each of which modeled the sequences of classes belonged to the same articulatory property by a discrete conditional distribution. For each utterance, seven articulatory class sequences were obtained from seven HMM-based recognizers, each responsible for one articulatory property. The usefulness of AFs in speaker verification was demonstrated in Leung et al. [7], where for each utterance, the probabilities of 26 articulatory classes determined from five mul-

This work was supported by The Hong Kong Polytechnic University, Grant No. GT860 and Research Grant Council of the Hong Kong SAR (Project No. CUHK 1/02C).

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

tilayer perceptrons (MLPs) were concatenated to form a sequence of articulatory feature vectors. The AF sequence was then fed to a GMM speaker model and a background model to compute a likelihood ratio for decision making.

2.2. Articulatory Feature Extraction

The AF extraction approach outlined in [4] was adopted. According to [4], only two articulatory properties, (i.e., the manner and place of articulations listed in Table 1) were used for pronunciation modeling.

The AF-MLPs take n consecutive frames of Mel-frequency cepstral coefficients (MFCCs) X_t (with consecutive frame indexes ranging from $t - \frac{n}{2}$ to $t + \frac{n}{2}$) as inputs at frame t. For a given X_t, the outputs of the two AF-MLPs, $P(Manner = m|X_t)$ and $P(Place = p|X_t)$, represent the posterior probabilities of different classes in the manner and place of articulation. The manner class label $l_t^M \in \mathcal{M}$ and the place class label $l_t^P \in \mathcal{P}$ (the sets of $\mathcal{M}$ and $\mathcal{P}$ are listed in Table 1) at frame t are determined by

$$l_t^M = \arg\max_{m \in \mathcal{M}} P(Manner = m|X_t) \text{ and} \quad (1)$$

$$l_t^P = \arg\max_{p \in \mathcal{P}} P(Place = p|X_t). \quad (2)$$

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 from $t = 1, \ldots, T$, where T is the total number of frames in the utterance.

2.3. Speaker Modeling

AF-based CPM (hereafter, referred to as AFCPM) aims to establish a relationship between the articulatory classes and the actual phonemes obtained from a phoneme-based recognizer. Because different speakers have different ways of pronunciation, their articulatory properties of the same phoneme can be varied.

2.3.1. Universal background models

For each phoneme, a set of universal background models (UBMs) is trained from the speech of a large number of speakers to represent the speaker-independent pronunciation characteristics corresponding to that phoneme. 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. For a particular phoneme q, the joint probabilities of the corresponding UBM are determined by

$$P_{bg}(Manner = m, Place = p|Phoneme = q) = \frac{\#((m, p, q) \text{ in the data of all background speakers})}{\#((*, *, q) \text{ in the data of all background speakers})} \quad (3)$$

where $m \in \mathcal{M}$, $p \in \mathcal{P}$, (m, p, q) denotes the condition for which $Manner = m$, $Place = p$ and $Phoneme = q$, $*$ represents all possible members in that class, and $\#(\ )$ represents the total number of frames with phoneme labels and AF labels fulfill the description inside the parentheses. The probabilities of unseen AF combinations are set to zero. For each phoneme, a total of 60 probabilities can be obtained. These probabilities are the products of 6 manner classes and 10 place classes. Therefore, a system with N phonemes has $60N$ probabilities in the UBMs.

2.3.2. Speaker models by MAP adaptation

In Leung et al. [4], speaker model was the joint probabilities of manner and place classes given the phoneme q estimated from the data of speaker s, which was expressed as

$$P_s(Manner = m, Place = p|Phoneme = q) = \frac{\#((m, p, q) \text{ in the data of speaker } s)}{\#((*, *, q) \text{ in the data of speaker } s)}. \quad (4)$$

However, the number of occurrences of some phonemes (e.g., /th/, /sh/ and /v/) are too low for an accurate estimation of the joint probabilities. As a result, the pronunciation models of these phonemes are less discriminative.

To overcome the data sparseness problem, speaker models can be adapted from the UBMs. This approach can also establish a tighter coupling between the speaker models and background models, which can result in a better verification performance [1].

Given the background model corresponding to phoneme q, the joint probabilities for speaker s are given by:

$$\hat{P}_s(Manner = m, Place = p|Phoneme = q) = \beta_s^q P_s(Manner = m, Place = p|Phoneme = q) + (1 - \beta_s^q) P_{bg}(Manner = m, Place = p|Phoneme = q), \quad (5)$$

where $\beta_s^q \in [0, 1]$ is a phoneme-dependent adaptation coefficient controlling the contribution of the speaker model (Eq. 4) and the background model (Eq. 3) on the adapted model. Similar to MAP adaptation of GMM-based systems [1], β_s^q is obtained by

$$\beta_s^q = \frac{\#((*, *, q) \text{ in the data of speaker } s)}{\#((*, *, q) \text{ in the data of speaker } s) + r}, \quad (6)$$

where r is a fixed relevance factor common to all phonemes and speakers. The purpose of r is to control the dependence of the adapted model's parameters on speaker's data. The estimation of r depends on the number of prior occurrences of $(*, *, q)$ of all q in the training data. If the number of occurrences of $(*, *, q)$ is much less than r, then β_s^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, then β_s^q will be very close to 1 and the the adapted model will become more dependent on speaker's data.

2.3.3. Verification

The verification score S_{AFCPM} of a test utterance is defined as:

$$S_{AFCPM} = \sum_{\substack{t=1, \\ p_s(X_t)\neq 0 \\ p_b(X_t)\neq 0 \\ q_t \neq \ silence}}^{T} (\log p_s(X_t) - \log p_b(X_t)), \quad (7)$$

where for each t, $p_s(X_t)$ and $p_b(X_t)$ are probabilities obtained from a speaker model of the claimed identity s and a background model, as follows:

$$p_s(X_t) = \begin{cases} P_s(Manner = l_t^M, Place = l_t^P | Phoneme = q_t) \\ \hat{P}_s(Manner = l_t^M, Place = l_t^P | Phoneme = q_t) \end{cases} \text{or} \quad (8)$$

and

$$p_b(X_t) = P_{bg}(Manner = l_t^M, Place = l_t^P | Phoneme = q_t). \quad (9)$$

In Eqs. 8 and 9, q_t is the phoneme at frame t. Because no speaker information is carried in the silence frames, they can be removed to improve the accuracy of the verification score. Moreover, only the "seen" AF combinations (i.e., $p_s(X_t) \neq 0$ and $p_b(X_t) \neq 0$) appeared in both speaker and background models are considered during verification.

3. FUSION OF FRAME-WEIGHTED SCORES

The AFCPM and the conventional spectral features (MFCCs) characterize speakers at two different levels; the former represent the pronunciation behaviors of individual speakers, whereas the latter look at their vocal tract's characteristics. Therefore, fusing the scores of AFCPM- and MFCC-based systems is expected to enhance speaker verification performance.

Scores from the AFCPM and MFCC systems were fused according to the frame-weighted fusion proposed in [4]. A frame-weighted fused score S_F^w is defined as

$$S_F^w = \frac{1}{W} \sum_{t=1}^{T} w(t) \overbrace{[(1-\alpha_u) s_{MFCC}(t) + \alpha_u s_{AFCPM}(t)]}^{S_F(t)} \quad (10)$$

where $\alpha_u \in [0,1]$ is a fusion weight, $W = \sum_{t'=1}^{T} w(t')$, and $w(t)$ represents the importance of the frame-based scores ($s_{MFCC}(t)$ and $s_{AFCPM}(t)$) with respect to the frame-weighted fused score S_F^w. According to Eq. 10, the introduction of $w(t)$ allows us to adjust the contribution of the frame-based fused scores $S_F(t)$ to the fused score S_F^w. It was suggested in [4] that the probabilities estimated from the manner MLP are more reliable than those from the place MLP. Therefore, probabilities of the manner MLP ($P(Manner = l_t^M | X_t)$) were adopted as $w(t)$.

4. EXPERIMENTS

The proposed approach was evaluated on the SPIDRE corpus [8]. Genuine verification trials involved one handset-match conversation and two handset-mismatch conversations from each of the 44 target speakers (speaker sp1007 was discarded due to corrupted data); impostor attempts involved 200 conversations from 160 nontarget speakers. The same set of nontarget speakers' conversations was applied to all target speaker models in the impostor attempts. Each of the testing utterances, which contains 5 minutes of speech (including silence), was split into short segments, with each segment ranging from 1 to 15 seconds according to the speaker turns labeled in the transcriptions [9]. All silence frames were removed by a voice activity detector.

The training conversation of all target speakers were used to train the phoneme models. The phoneme set consisted of 46 context-independent phonemes [9], including one silence and four noise, each of which was modeled by a three-state left-to-right HMM with 16 diagonal-covariance Gaussian mixtures per state. The HTK [10] was used to train the HMMs. Acoustic vectors of 39 dimensions—each comprising of 12 MFCCs, the normalized energy, and their first- and second-order derivatives—were used for training the phoneme models and for recognition.

The software Quicknet [11] was used to train two AF-MLPs, each of which was composed of 234 input nodes (nine frames of 26-dimensional MFCCs: 12 MFCCs, log energy, and the corresponding delta coefficients), 50 hidden nodes, and either 6 or 10 output nodes. To improve the robustness of AFs against handset variations, a total of 3,794 utterances randomly selected from all of the 10 handsets in the HTIMIT [12] corpus were used to train the AF-MLPs.

For the AFCPM systems, phoneme sequences of all training and testing utterances were obtained from a null-grammar recognizer. The phoneme recognition accuracy of the recognizer on all testing utterances was 37.69%. The aligned AF streams and phoneme sequences of all target speakers were used to train a set of UBMs (Λ_b^{AFCPM}) representing the probabilities of 60 manner and place class combinations conditioned on 41 phonemes (excluding the silence and noise) in the phone set. Two approaches were used to obtain an AFCPM-based speaker model Λ_s^{AFCPM}. For the first approach, the probabilities in Λ_s^{AFCPM} were computed based on the AF streams and phoneme sequences of a given speaker s according to Eq. 4. This approach was referred to as AFCPM. In the second approach, the speaker probabilities were adapted from those of Λ_b^{AFCPM} using the training data from speaker s according to Eqs. 5 and 6 with r set to 18. Hereafter, this adaptation approach is referred to as A-AFCPM.

For the MFCC system, 24-dimensional MFCC vectors were used as features. Each feature vector comprises 12 MFCCs and the corresponding delta coefficients computed every 14ms using a Hamming window of 28ms. A universal background GMM Λ_b^{MFCC} with 512 mixtures was trained using all training conversations of all target speakers. For a speaker s in the target speaker set, a speaker GMM Λ_s^{MFCC} was adapted from Λ_b^{MFCC} using MAP adaptation [1].

The fusion weights α_u were determined by K-fold cross validations. More specifically, the test data of the target and nontarget speakers were divided into K disjoint subsets, and the fusion weight was selected such that the average error obtained from the K-fold evaluations was minimized.

5. RESULTS AND DISCUSSIONS

Table 2 shows the experimental results of an MFCC system (the baseline for comparison), the AFCPM systems, and the fusion of these systems. When adaptation was adopted to obtain the AFCPM speaker models, the EER dropped from 25.83% to

Features	EER (%)		
	Matched	*Mismatch*	*All*
MFCC	7.59	18.08	15.29
AFCPM	19.52	27.69	25.83
A-AFCPM	18.07	26.69	24.04
MFCC+AFCPM	6.85	16.23	14.09
(error red. %)	(9.74)	(10.23)	(7.85)
MFCC+A-AFCPM	7.03	16.00	13.78
(error red. %)	(7.37)	(11.50)	(9.87)

Table 2. EERs and relative error reduction (in %) obtained from the MFCC system, the AFCPM systems, and the fusion of the two systems. *A-AFCPM* denotes the adaptive AFCPM system whose speaker models are adapted from the UBMs. *MFCC+AFCPM* (*MFCC+A-AFCPM*) denotes the fusion of frame-weighted MFCC scores and AFCPM (adaptive AFCPM) scores according to Eq. 10. *Matched* (*Mismatched*) refers to the cases where the handset used by a target speaker in a verification session is identical to (different from) the one used by himself or herself during the enrollment session. The test data from nontarget speakers under *Matched* and *Mismatched* are identical. *All* represents the overall EERs obtained from gathering all test data from the target speakers using both matched and mismatched handsets.

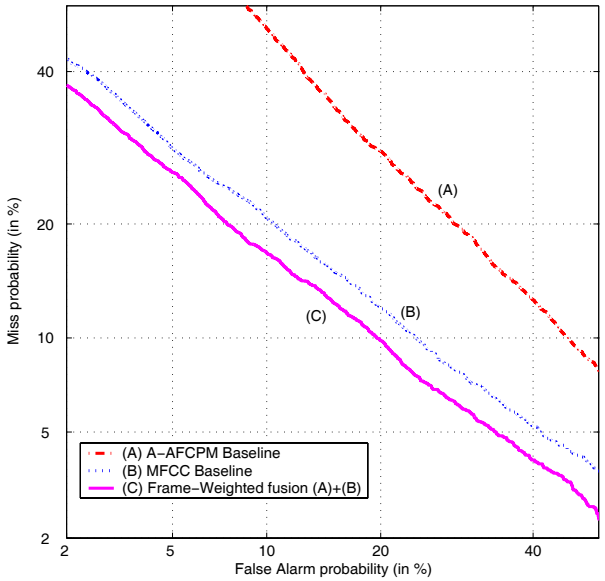

Fig. 1. Speaker detection performance of the A-AFCPM system, the MFCC system, and the fusion of the two systems.

24.04% (an 7.0% EER reduction). This reduction in EERs occurs in both matched and mismatched handsets, which suggests that better speaker models (in terms of capturing speaker characteristics and robustness against handset variations) can be obtained by adapting the UBMs. Through the adaptation, speaker models can become tightly coupled to the UBMs. This helps prevent overfitting the speaker models and improve their discriminative power.

Table 2 also shows that the frame-weighted fusion of MFCC and AFCPM scores is an effective means of combining the spectral- and AFCPM-based systems. Again, a more significant error reduction was obtained from *MFCC+A-AFCPM*, which demonstrates that a better representation of pronunciation characteristics can be achieved by estimating the speaker models via MAP adaptation. Therefore, it can be conclude that A-AFCPM provides more speaker-dependent information than AFCPM.

Figure 1 plots the detection error tradeoff (DET) curves [13] of the MFCC system, the A-AFCPM system, and the frame-weighted fusion of these two systems. Although there is a significant difference between the performance of the MFCC and A-AFCPM systems, the frame-weighted fusion results in lower miss probabilities for a wide range of false alarm probabilities.

6. CONCLUSIONS

This paper has presented an AFCPM speaker verification system in which speakers are distinguished by their pronunciation characteristics. This is achieved by the conditional pronunciation modeling of two articulatory property streams. Instead of directly estimating the conditional pronunciation probabilities of speakers, speaker models are adapted from universal background models via MAP adaptation. A better verification performance was achieved because speaker discrimination is enhanced by a tighter coupling between the speaker models and background models. A lower error rate was achieved by the frame-weighted fusion of conventional MFCC and the adapted AFCPM scores, which suggests that within an utterance, some frames may contain more speaker-dependent pronunciation characteristics than the others.

7. 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. Reynolds, et. al., "The superSID project: exploiting high-level information for high-accuracy speaker recognition," in *Proc. of ICASSP'03*, Hong Kong, April 2003, vol. 4, pp. 784–787.

[3] D. Klusáček, J. Navrátil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," in *Proc. of ICASSP'03*, Hong Kong, April 2003, vol. 4, pp. 804–807.

[4] K.Y. Leung, M.W. Mak, and S.Y. Kung, "Articulatory feature-based conditional pronunciation modeling for speaker verification," in *Proc. of ICSLP'04*, 2004.

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

[6] *http://www.clsp.jhu.edu/ws2002/groups/supersid/.*

[7] K.Y. Leung, M.W. Mak, and S.Y. Kung, "Applying articulatory features to telephone-based speaker verification," in *Proc. ICASSP'04*, Montreal, May 2004, vol. 1, pp. 85–88.

[8] J. P. Campbell and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proc. ICASSP'99*, 1999, vol. 2, pp. 829–832.

[9] *http://www.isip.msstate.edu/projects/switchboard/.*

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book for HTK 3.0," Tech. Rep., Microsoft Corporation, 2000.

[11] P. Färber, "Quicknet on multispert: fast parallel neural network training," Tech. Rep. TR-97-047, ICSI, 1997.

[12] 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.

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