Speaker Verification via Articulatory Feature-based Conditional Pronunciation Modeling with Vowel and Consonant Mixture Models

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

10.21437/Interspeech.2005-663

Abstract

Articulatory feature-based conditional pronunciation modeling (AFCPM) aims to capture the pronunciation characteristics of speakers by modeling the linkage between the states of articulation during speech production and the actual phones produced by a speaker. Previous AFCPM systems use one discrete density function for each phoneme to model the pronunciation characteristics of speakers. This paper proposes using a mixture of discrete density functions for AFCPM. In particular, the pronunciation characteristics of each phoneme is modeled by two density functions: one responsible for describing the articulatory features that are more relevant to vowels and the other for consonants. Verification scores are the weighted sum of the outputs of the two models. To enhance the resolution of the pronunciation models, four articulatory properties (front-back, lip-rounding, place of articulation, and manner of articulation) are used for pronunciation modeling. The proposed AFCPM is applied to a speaker verification task. Results show that using four articulatory features achieves a lower error rate as compared to using two features (manner and place of articulation) only. It was also found that dividing the articulatory properties into two groups is an effective means of solving the data-sparseness problem encountered in the training phase of AFCPM systems.

1. Introduction

The rationale behind using conditional pronunciation modeling (CPM) [1] for speaker verification is that different speakers have different ways of pronouncing the same phoneme. In contrast to the conventional speaker recognition systems in which short-term spectral features are modeled by Gaussian mixture models (GMMs) [2], CPM-based systems identify speakers based on the speakers' pronunciation characteristics. The pronunciation characteristics are encoded as discrete probability densities from which verification scores are computed. Because of the differences in the features in spectral-based and CPM-based systems, fusing the scores obtained from these systems can achieve performance better than that of single-feature systems [3].

In [4, 5], we proposed an articulatory feature-based CPM speaker verification system. Two articulatory feature (AF) streams were used (instead of multilingual phoneme streams as in [1]) for modeling the pronunciation characteristics of speakers. AFs are some abstract classes describing the vocal tract properties or the articulator motion during speech production. Because AFs are directly related to the speech production process, applying them to CPM facilitates the modeling of speaker's pronunciation characteristics. Another advantage of using AFs for CPM is that multilingual training data are not required, because articulatory properties are the same irrespective of languages.

The AF-based CPM system in [4, 5] uses phoneme-dependent discrete densities of two articulatory properties (manner and place of articulation) for speaker modeling. More precisely, for each phoneme, a two-dimensional discrete density function of the classes in the manner and place of articulations is built to model the way and location that air-stream along the vocal tract is obstructed, shaped, and modified by the articulators. Although these two articulatory properties are sufficient for describing how consonant phonemes are produced, their ability to describe vowel phonemes is limited. Research has found that vowel phonemes contain useful speaker information that are crucial for speaker recognition [6]. To achieve better pronunciation modeling, this paper proposes adding another two articulatory properties (front-back and lip-rounding), which are more relevant to the vowel phonemes, to the CPM system in [4, 5]. Instead of building a discrete density function of four dimensions for each phoneme as in [4, 5], the four articulatory properties are separated into two groups, with one group responsible for capturing the pronunciation characteristics of vowel phonemes and the other group responsible for consonant phonemes. Therefore, each phoneme has a two-mixture density function (one mixture for each group) and the verification scores for each phoneme are the weighted sum of the two mixtures. Experimental results show that adding front-back and lip-rounding and dividing the four articulatory properties into two groups improve the resolution of CPM and thus achieve a lower error rate as compared to the original AF-based CPM system.

2. Articulatory Feature-Based CPM

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. They have been applied to speaker identification in [7], where seven speaker-dependent AF-based language models were used to model the distributions of articulatory classes. Articulatory class probabilities can also be used as features in GMM-based speaker verification systems [8]. Because AFs are closely related to the speech production process,

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 (AFs)*	*Number of Classes*
Front-back ($\mathcal{FB}$)	Silence, Front, Back, Nil	4
Rounding ($\mathcal{R}$)	Silence, Rounded, Not Rounded, Nil	4
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. Because front-back and rounding are relevant to vowels only, a label "Nil" is added in case the phoneme is a consonant.

they are suitable for capturing the pronunciation characteristics of speakers.

2.1. Articulatory Feature Extraction

In our previous work [4, 5], the manner and place of articulation, which describe the way and location that the air-stream along the vocal tract is constricted by the articulators, were used for CPM. In this work, besides the *manner* and *place* of articulation, two more articulatory properties—the tongue position in the horizontal axis (*front-back*) and *lip-rounding*—were introduced to provide a better representation of the pronunciation characteristics of vowels and consonants. These four articulatory properties and their classes are listed in Table 1.

The AFs are automatically determined from speech signals using AF-based multilayer perceptrons (MLPs) as follows. For each articulatory property, an AF-MLP takes n consecutive frames of MFCCs X_t (with consecutive frame indexes ranging from $t-\frac{n}{2}$ to $t+\frac{n}{2}$) as input to determine the posterior probabilities of the output classes at frame t. For example, given X_t at frame t, the manner MLP determines six posterior probabilities of the output classes, i.e., $P(L^M = m|X_t)$. With these probabilities, the manner class label $l_t^M \in \mathcal{M}$ at frame t is determined by

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

where the set $\mathcal{M}$ is defined in Table 1. The four AF streams for creating the conditional pronunciation models are formed by concatenating the front-back labels l_t^{FB}, rounding labels l_t^R, manner labels l_t^M, and place labels l_t^P from $t = 1, \ldots, T$, where T is the total number of frames in the utterance.

2.2. Speaker Modeling

The AF-based CPM system (hereafter, referred to as AFCPM system) proposed in [5] considers the combinations of AFs belonging to the manner and place properties (see $\mathcal{M}$ and $\mathcal{P}$ in Table 1). The system aims to establish a link between the two articulatory properties 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.

In this work, we extended the AFCPM system in [5] to a two-mixture AFCPM system. The pronunciation characteristics of vowel and consonant phonemes are modeled separately using different articulatory properties. In addition to the manner and place of articulations, two articulatory properties—the tongue position in front-back axis and the lip-rounding—that are more relevant to vowel speech are adopted. The four articulatory properties are divided into two sets. The front-back, rounding, and place are grouped together as a single model to capture the pronunciation characteristics of vowel phonemes. The reason of such grouping is that the front-back and rounding properties are only meaningful to vowels and the classes *high, middle*, and *low* in the place property are the AFs that describe vowels. The AFs of manner and place, which have been used in the previous AFCPM system, form another model to capture the pronunciation characteristics of consonant phonemes.

Grouping the AFs into two sets rather than considering all of them together reduces the number of probabilities in the models significantly. Comparing to the single-mixture AFCPM in [5], the proposed two-mixture AFCPM has a better resolution in representing the pronunciation characteristics of speakers.

2.2.1. Universal background models

For each phoneme, two sets of universal background models (UBMs) are trained from the speech of a large number of speakers to represent the speaker-independent pronunciation characteristics corresponding to that phoneme. The training procedure begins with aligning the AF streams obtained from the MLPs and a phoneme sequence obtained from a phoneme recognizer. For a particular phoneme q, the joint probabilities of the two UBMs are determined by

$$\begin{aligned} &P_{bg}(m, p|q) \\ &= P_{bg}(L^M = m, L^P = p|Phoneme = q) \\ &= \frac{\#((m, p, q) \text{ in the data of all background speakers})}{\#((q) \text{ in the data of all background speakers})} \end{aligned} \tag{2}$$

and

$$\begin{aligned} &P_{bg}(fb, r, p|q) \\ &= P_{bg}(L^{FB} = fb, L^R = r, L^P = p|Phoneme = q) \\ &= \frac{\#((fb, r, p, q) \text{ in the data of all background speakers})}{\#((q) \text{ in the data of all background speakers})} \end{aligned} \tag{3}$$

where $m \in \mathcal{M}$, $p \in \mathcal{P}$, $fb \in \mathcal{FB}$, and $r \in \mathcal{R}$. (m,p,q) denotes the condition for which $L^M = m$, $L^P = p$ and $Phoneme = q$, and (fb,r,p,q) denotes $L^{FB} = fb$, $L^R = r$, $L^P = p$, and $Phoneme = q$. $\#(\,)$ represents the total number of frames with phoneme labels and AF labels (or phoneme labels only in the denominator) fulfill the description inside the parentheses. The probabilities of unseen AF combinations are set to zero.

For each phoneme, the two-mixture AFCPM system involves 60 (total number of AF combinations in manner and place properties) + 160 (total number of AF combinations in front-back, rounding, and place properties) = 220 probabilities. Therefore, a system with N phonemes has $220N$ probabilities in the UBMs. Although the number of probabilities in the UBMs is larger than that in [5], the amount of data used for estimating each probability remains the same. This is because the four AF streams are independent (i.e., a frame can belong to four articulatory properties); as a result, training data can be shared to estimate the probabilities of the two models. For example, a frame that is recognized as phoneme /aa/ and assigned AFs L^{FB}=back, L^R=rounded, L^M=vowel, and L^P=low can be used to estimate both P_{bg}(vowel,low|/aa/) and P_{bg}(back,rounded,low|/aa/).

2.2.2. Speaker models by MAP adaptation

MAP adaptation is applied to obtain the speaker models [5]. Given the background model corresponding to phoneme q, the joint probabilities for speaker s are given by

$$\begin{aligned}&\hat{P}_s(m,p|q)\\&=\hat{P}_s(L^M=m,L^P=p|Phoneme=q)\\&=\beta_s^q\left[\frac{\#((m,p,q)\text{ in the data of speaker }s)}{\#((q)\text{ in the data of speaker }s)}\right]+\\&\quad(1-\beta_s^q)P_{bg}(m,p|q),\end{aligned}\tag{4}$$

and

$$\begin{aligned}&\hat{P}_s(fb,r,p|q)\\&=\hat{P}_s(L^{FB}=fb,L^R=r,L^P=p|Phoneme=q)\\&=\beta_s^q\left[\frac{\#((fb,r,p,q)\text{ in the data of speaker }s)}{\#((q)\text{ in the data of speaker }s)}\right]+\\&\quad(1-\beta_s^q)P_{bg}(fb,r,p|q).\end{aligned}\tag{5}$$

In Eqs. 4 and 5, $\beta_s^q \in [0,1]$ is a phoneme-dependent adaptation coefficient controlling the contribution of the speaker data and the background models (Eqs. 2 and 3) on the MAP-adapted model. It is obtained by

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

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 value of β_s^q depends on the number of occurrences of (q) in the training data. The probabilities given in Eqs. 4 and 5 are obtained from the same set of training data with the same number of occurrences of q. Therefore, a same value of β_s^q is applied to both Eqs. 4 and 5.

2.2.3. Verification

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

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

where for each t, q_t is the phoneme label, $\log p_s(X_t)$ and $\log p_{bg}(X_t)$ are log-probabilities obtained from the speaker models of the claimed identity s and the background models. Due to the coarticulation effect, the change of AFs is asynchronous at the phoneme boundaries [9]. Therefore, instead of making a hard vowel or consonant decision in which $p(X_t)$ depends exclusively on $P(l_t^{FB},l_t^R,l_t^P|q_t)$ for vowel frames or on $P(l_t^M,l_t^P|q_t)$ for consonant frames, a soft decision is adopted during score computation. More specifically, $\log p_s(X_t)$ and $\log p_{bg}(X_t)$ are the weighted sum of the log-probabilities obtained from the two CPM models as follows:

$$\begin{aligned}&\log p_s(X_t)\\&\quad=P(vowel|X_t)\log\hat{P}_s(l_t^{FB},l_t^R,l_t^P|q_t)+\\&\quad[1-P(vowel|X_t)]\log\hat{P}_s(l_t^M,l_t^P|q_t)\end{aligned}\tag{8}$$

and

$$\begin{aligned}&\log p_{bg}(X_t)\\&\quad=P(vowel|X_t)\log P_{bg}(l_t^{FB},l_t^R,l_t^P|q_t)+\\&\quad[1-P(vowel|X_t)]\log P_{bg}(l_t^M,l_t^P|q_t).\end{aligned}\tag{9}$$

In Eqs. 8 and 9, the weight $P(vowel|X_t)$ (i.e., the probability of vowel) is obtained from the output of the manner MLP given X_t. Therefore, the contribution of the two CPM models can be flexibly adjusted according to the vowel probabilities.

3. Fusion of Frame-Weighted Scores

The AFCPM and the conventional spectral features (MFCCs) characterize speakers at two different levels. The former represents the pronunciation behaviors of individual speakers, whereas the latter looks 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

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

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. 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 [10]. 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 [11]. A development set, which was used for finding appropriate values for r and α_u, was formed by the test data of 4 target speakers and 20 impostors randomly selected from the speaker and impostor sets. 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 consists of 46 context-independent phonemes [11] (including one silence and four types of noise), each of which was modeled by a three-state left-to-right HMM with 16 diagonal-covariance Gaussian mixtures per state. 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 [12] was used to train four AF-MLPs, each comprising 234 input nodes (nine frames of 26-

Features	EER (%)		
	Matched	*Mismatch*	*All*
MFCC	7.59	18.08	15.29
AFCPM	18.07	26.69	24.04
2M-AFCPM	16.69	25.91	23.23
MFCC + AFCPM	7.09	16.31	13.77
(error red. %)	(6.59)	(9.78)	(9.94)
MFCC + 2M-AFCPM	7.15	15.68	13.34
(error red. %)	(5.79)	(13.27)	(12.75)

Table 2: EERs and relative error reduction (in %) obtained by the MFCC system, the AFCPM systems, and the fusion of the two systems. *AFCPM* denotes the MAP-adapted AFCPM system [5]. *2M-AFCPM* denotes the MAP-adapted AFCPM system with two-mixture models proposed in this paper. *MFCC + AFCPM* (*MFCC + 2M-AFCPM*) denotes the fusion of frame-weighted MFCC scores and AFCPM (AFCPM with two-mixture models) 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.

dimensional MFCCs: 12 MFCCs, log-energy, and the corresponding delta coefficients), 50 hidden nodes, and different numbers (4, 6, or 10) of 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 [13] 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 on all testing utterances was 37.69%. The aligned AF streams and phoneme sequences of all target speakers were used to train the UBMs (Eqs. 2 and 3) corresponding to 41 phonemes (excluding the silence and noise) in the phone set. The MAP-adapted speaker models were obtained according to Eqs. 4–6 with r set to 18.

For the MFCC system, 24-dimensional MFCC vectors (12 MFCCs and their delta coefficients computed every 14ms using a Hamming window of 28ms) were used as features. 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 [2].

The fusion weights α_u were determined by four-fold cross validations based on the development set. More specifically, the data of the development set were divided into four disjoint subsets, and the fusion weight was selected such that the average error obtained from the four-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. Using two-mixture AFCPM reduces the overall EER from 24.04% to 23.23% (an 3.37% EER reduction). This performance gain is mainly attributed to the additional articulatory properties in the two-mixture AFCPM system, which lead to finer resolution in the pronunciation models. The results also suggest that considering the pronunciation characteristics of vowel and consonant phonemes separately can lead to better verification systems.

The fusion results given in 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 + 2M-AFCPM*, which demonstrates that a better representation of pronunciation characteristics can be achieved by replacing the single-mixture models with the two-mixture models in the AFCPM system.

6. Conclusions

This paper has presented an extended AFCPM speaker verification system in which speakers are distinguished by their pronunciation characteristics in terms of the articulatory properties. Pronunciation characteristics of vowels and consonants captured separately by phoneme-dependent discrete distributions of articulatory streams. Verification scores were computed from the mixture of these two models. The proposed scheme achieves better verification performance than the original AFCPM system.

7. References

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

[2] 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.

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

[4] K.Y. Leung, M.W. Mak, and S.Y. Kung, "Articulatory feature-based conditional pronunciation modeling for speaker verification," in *Proc. ICSLP 2004*, 2004, pp. 516–519.

[5] K. Y. Leung, M. W. Mak, M. Siu, and S. Y. Kung, "Speaker verification using adapted articulatory feature-based conditional pronunciation modeling," in *Proc. ICASSP 2005*, Philadelphia, PA, USA, March 2005, vol. 1, pp. 181–184.

[6] J. P. Eatock and J. S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Proc. ICASSP 1994*, 1994, vol. 1, pp. 133–136.

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

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

[9] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," *Phonetica*, vol. 49, pp. 153–180, 1992.

[10] 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.

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

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

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