Articulatory Feature-Based Conditional Pronunciation Modeling for Speaker Verification

Ka-Yee Leung, Man-Wai Mak

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University, USA

Abstract

Because of the differences in education background, accents, etc., different persons have their unique way of pronunciation. This paper exploits the pronunciation characteristics of speakers and proposes a new conditional pronunciation modeling (CPM) technique for speaker verification. The proposed technique aims to establish a link between articulatory properties (e.g., manners and places of articulation) and phoneme sequences produced by a speaker. This is achieved by aligning two articulatory feature (AF) streams with a phoneme sequence determined by a phoneme recognizer, and formulating the probabilities of articulatory classes conditioned on the phonemes as speaker-dependent probabilistic models. The scores obtained from the AF-based pronunciation models are then fused with those obtained from a spectral-based speaker verification system, with the frame-by-frame fused scores weighted by the confidence of the pronunciation models. Evaluations based on the SPIDRE corpus demonstrate that AF-based CPM systems can recognize speakers even with short utterances and are readily combined with spectral-based systems to further enhance the reliability of speaker verification.

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 speaker-dependent information in speech signals during verification.

In recent years, researchers have started to investigate the use of high-level speaker information, such as the usage or duration of particular words, prosodic features, etc., for speaker recognition [2]. Their work has demonstrated that various degree of speaker information is imparted in these features and some of them can be applied to identify speakers accurately when enough data are available. Among those evaluated features, the best result was obtained from a system that uses the conditional pronunciation modeling (CPM) technique [3]. CPM aims to characterize the pronunciation behaviors of a speaker by computing the correlation between the intended phonemes and the actual phones produced by the speaker, where the phonemes were obtained by a recognizer with lexical constraints and the phones were obtained by five null-grammar phone recognizers corresponding to five different languages. The pronunciation behaviors are encoded as discrete probability densities, which are subsequently used for verifying speakers similar to the conventional GMMs in spectral-based systems. It was found that pronunciation modeling is applicable to speaker detection because different speakers have different ways of pronouncing the same phoneme. However, one problem of CPM is that it requires multilingual speech data for training the phone models of different languages. CPM systems also require long utterances for speaker enrollment and verification. For example, in [3], up to 40 minutes of speech was used to enroll a speaker and 2.5 minutes of speech was used in each verification trial.

To avoid the requirement of a large amount of multilingual training data, this paper proposes 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 [4]. Compared to the phone-based CPM in [3], AFs provide a more direct coupling between the pronunciation variations and the speech production process, which is a source of speaker variations. Because articulatory properties are the same irrespective of languages, monolingual speech data is sufficient for determining their values.

Experimental results based on the SPIDRE corpus show that speaker information exists in the AFs and the proposed AF-based CPM technique is an effective approach to modeling the pronunciation variations of speakers. It was also found that the use of AF streams for CPM allows shorter utterances to be used for enrollment and verification.

2. AF-Based CPM System

AFs are abstract classes describing the movements or positions of different articulators during speech production [4]. Preliminary work of applying AFs for speaker identification has been reported in [5], where seven parallel Tri-gram models obtained from seven AF streams were derived. The usefulness of AFs in speaker verification was also demonstrated in our previous work [6], where the features for verification were formed by concatenating the outputs of five AF classifiers.

The AF-based CPM system (hereafter referred to as AFCPM system) is based on the CPM system proposed in [3]. However, the aims of the two systems are different. The aim of our AFCPM system is to establish the relationship between the articulatory properties and the actual phonemes obtained from a phoneme recognizer. As different speakers have their unique way of pronunciation, their articulatory properties can be varied even when they pronounce the same phoneme.

The AFCPM system is operated as follows. For each utterance, a time-aligned phoneme sequence, $\mathbf{Q} = \{q_1, q_2, \ldots, q_T\}$, was obtained from a null-grammar phoneme recognizer. This phonetic sequence represents what the speaker has said at every frame. The AF streams that track the speaker's articulatory properties were ob-

This work was supported by The Hong Kong Polytechnic University, Grant No. A-PE44 and Research Grant Council of the Hong Kong SAR (Project No. CUHK 1/02C).

Articulatory properties	*Classes*	*No. of Classes*
Manner	Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral	6
Place	Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal	10

Table 1: *Articulatory properties and their associated classes.*

tained in parallel from two AF classifiers (see Section 2.1 and [6] for the details of AF classifiers). To reduce the dimensionality of conditional pronunciation models, only two out of the five articulatory properties suggested in [6] were adopted. These two properties are the manner and place of articulation, and they were chosen because their combinations can be used to classify consonants and most of the vowels.

2.1. Articulatory Feature Extraction

To extract AFs from speech, classifiers were trained to learn the mapping between the articulatory states and the spectral features derived from acoustic signals. In this work, AFs were extracted from spectral features based on an approach similar to Kirchhoff [4]. Specifically, to obtain the AFs of an utterance, a sequence of spectral vectors was fed to two multilayer perceptrons (MLPs) with outputs representing the posterior probabilities of different classes in the manner and place articulatory properties (see Table 1).

The inputs to the two AF-MLPs are identical and their numbers of outputs are equal to the numbers of AF classes listed in the last column of Table 1. To improve the accuracy of AFs, nine consecutive frames of normalized MFCCs centered at frame t are served as the inputs to the AF-MLPs. More specifically, the AFs at frame t are obtained by presenting a vector sequence $X_t = \{\mathbf{x}_{t-4}, \ldots, \mathbf{x}_{t+4}\}$ to the two MLPs. Rather than feeding the MFCCs directly to the AF-MLPs, they were normalized to zero mean and unit variance. The normalization aims to remove the variations of different MLP inputs so that the training of MLP weights will not be dominated by those large magnitude inputs.

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 $l_t^m \in \mathcal{M}$ = {'Silence', 'Vowel', 'Stop', 'Fricative', 'Nasal', 'Approximant-Lateral'} and the place class label $l_t^p \in \mathcal{P}$ = {'Silence', 'High', 'Middle', 'Low', 'Labial', 'Dental', 'Coronal', 'Palatal', 'Velar', 'Glottal'} at frame t are determined by:

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

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

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. Table 2 shows an example of an 11-frame segment extracted from an utterance.

2.2. Speaker Modeling

Each target speaker in the system is assigned a set of speaker models, and all target speakers share the same set of universal background models. Each of the models comprises of a set of joint probabilities of the manner and place classes conditioned on a phoneme q. 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 a speaker s were obtained from:

Frame t	Phoneme q_t	AF class label and its probability			
		l_t^m *and its' prob.*		l_t^p *and its' prob.*	
1	aa	Vowel	0.75	Low	0.25
2	aa	Vowel	0.79	Low	0.30
3	aa	Vowel	0.92	Low	0.38
4	aa	Vowel	0.88	Low	0.49
5	aa	Vowel	0.79	Low	0.54
6	t	Vowel	0.63	Low	0.38
7	t	Silence	0.52	Silence	0.54
8	t	Silence	0.89	Silence	0.44
9	t	Silence	0.47	Silence	0.28
10	t	Silence	0.48	Silence	0.26
11	t	Stop	0.65	Coronal	0.47

Table 2: *An 11-frame example of an aligned phoneme sequence and its corresponding AF streams.*

$$P(Manner = m, Place = p|q, s) = \frac{\#(q \text{ with Manner} = m \text{ and Place} = p \text{ from } s\text{'s utterances})}{\#(q \text{ from } s\text{'s utterances})} \quad (3)$$

where $\#(\ )$ represents the total number of frames with phoneme labels and AF labels fulfill the description inside the parentheses. Similarly, the joint probabilities of a background model for a particular phoneme are given by $P(Manner = m, Place = p|q, bkg)$ and they were determined using all utterances of all target speakers in the training set. For example, based on the data in Table 2, $P(Manner =$ 'Vowel', $Place =$ 'Low'|'aa') $= 1.0$ while $P(Manner =$ 'Vowel', $Place =$ 'Low'|'t') $= 1/6 = 0.167$. The probabilities of unseen AF combinations were set to zero. For each phoneme, a total of 60 probabilities were obtained. These probabilities are the products of 6 manner classes and 10 place classes. Therefore, for a system with N phonemes, there are $60N$ CPM probabilities per speaker.

The verification score S_{AFCPM} of a test utterance is defined as the difference between the speaker score S_s and background score S_b:

$$S_{AFCPM} = S_s - S_b \quad (4)$$

$$= \sum_{\substack{t=1, \\ p_s(t)>0,\ p_b(t)>0 \\ q_t \neq silence}}^{T} (\log p_s(t) - \log p_b(t)) \quad (5)$$

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

$$p_s(t) = P(Manner = l_t^m, Place = l_t^p|q_t,\ s) \quad \text{and} \quad (6)$$

$$p_b(t) = P(Manner = l_t^m, Place = l_t^p|q_t,\ bkg). \quad (7)$$

As no speaker information is carried in the silence frames, they were removed to improve the accuracy of the verification score. Moreover, only the seen AF combinations (i.e. $p_s(t) > 0$ and $p_b(t) > 0$) appeared in both speaker and background models were considered

during verification. As can be observed from (5), the contributions of individual MFCC segments, X_t's, to the verification score S_{AFCPM} are equally weighted.

3. Fusion of AFCPM and MFCC scores

AFCPM scores can be fused with the scores of a spectral-based system to enhance verification performance because the former are derived from high-level speaker information whereas the latter from low-level information. In most utterances, some frames may contain more speaker-dependent information than the others. Therefore, we propose introducing a frame-dependent parameter $\beta(t)$ to weight the fusion scores at frame t. Specifically, the frame-weighted fusion score S_F^w is obtained from:

$$S_F^w = \alpha \sum_{t=1}^{T} \beta(t) \overbrace{[(1-w_{af}) s_{MFCC}(t) + w_{af} s_{AFCPM}(t)]}^{S_F(t)} \quad (8)$$

$$= \sum_{t=1}^{T} [\alpha\beta(t)(1-w_{af})]\, s_{MFCC}(t) + \sum_{t=1}^{T} [\alpha\beta(t) w_{af}]\, s_{AFCPM}(t) \quad (9)$$

$$= (1-w_{af}) \overbrace{\sum_{t=1}^{T} \alpha\beta(t) s_{MFCC}(t)}^{S_{MFCC}^w} + w_{af} \overbrace{\sum_{t=1}^{T} \alpha\beta(t) s_{AFCPM}(t)}^{S_{AFCPM}^w} \quad (10)$$

where $s_{MFCC}(t)$ is the MFCC score obtained from the spectral-based system at frame t, $\alpha = 1/\sum_{t'=1}^{T} \beta(t')$ is a normalization constant and w_{af} is a fusion weight determined from K-fold cross validation. More specifically, the test data of the target and non-target speakers were divided into K disjoint subsets. w_{af} was selected such that the average error obtained from the K-fold evaluations was minimized. According to (8), the introduction of $\beta(t)$ enables us to adjust the contribution of the frame-based fusion scores $S_F(t)$ to the utterance-based fusion score S_F^w. Another interpretation of S_F^w can be obtained by considering the factors inside the square brackets of (9) as frame-dependent fusion weights (although they do not sum to 1.0). In such interpretation, the introduction of $\beta(t)$ enables us to achieve a more flexible fusion between the AFCPM and MFCC systems.

The weight $\beta(t)$ represents a measure of how important the two scores ($s_{MFCC}(t)$ and $s_{AFCPM}(t)$) are with respect to the overall fusion score S_F^w. In this paper, $\beta(t)$ is interpreted as the confidence of MFCC and AFCPM scores, i.e. the reliabilities of $s_{MFCC}(t)$ and $s_{AFCPM}(t)$. As both the MFCC and AFCPM systems take MFCCs as input, the same confidence $\beta(t)$ was assigned to $s_{MFCC}(t)$ and $s_{AFCPM}(t)$ at frame t. Among the two systems, it is easier to derive a confidence measure for AFCPM scores as these scores depend greatly on the outputs of the two AF-MLPs. Therefore, the manner MLP's outputs were adopted as $\beta(t)$'s, i.e. $\beta(t) = P(Manner = l_t^m | X_t)$. Only the manner class probabilities were used because they exhibit more variation than the place class probabilities, and for each frame the manner class probabilities are also higher than the place class probabilities (see Table 2), which is a sign of greater reliability and discriminative power.

	Features	*Matched*	*Mismatched*	*All*
	MFCC	8.55	18.18	15.84
Recognized Alignment	AFCPM	19.52	27.69	25.83
	MFCC+AFCPM	8.50	16.61	14.44
	(error red. %)	(0.58)	(8.63)	(8.83)
Forced Alignment	AFCPM	17.92	24.98	22.69
	MFCC+AFCPM	8.08	15.24	13.26
	(error. red. %)	(5.49)	(16.17)	(16.28)

Table 3: *EERs (in %) and error reduction (error red.) obtained from the MFCC system, AFCPM system and fusion of the two systems.* MFCC + AFCPM *denotes the fusion of frame-weighted MFCC and AFCPM scores given in (10).* Matched *(*Mismatched*) means the enrollment handset is identical to (different from) the verification handsets. The test data from non-target 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. Note that the MFCC system does not require phoneme alignments.*

4. Experiments

The AFCPM system was evaluated on the SPIDRE corpus [7], which is a subset of the Switchboard corpus. In the experiments, 44 out of 45 target speakers (speaker sp1007 was discarded due to corrupted data) and 100 non-target speakers were used. Each target speakers has four 5-minute conversations (including silence) recorded from three different handsets: one for training, one for matched-handset testing and the remaining two for mismatched-handset testing. For each non-target speaker, only one conversation is available for testing. In the experiments, the testing conversations were split into short segments, with each segment ranging from 1 to 15 seconds according to the speaker turns labeled in the transcription [8].

Another database, the HTIMIT corpus [9], was used to train the AF-MLPs. HTIMIT was constructed by playing a gender-balanced subset of the TIMIT corpus through nine telephone handsets and a Sennheizer head-mounted microphone. This set-up introduces real transducer distortion in a controlled manner but without losing the time-aligned phonetic transcriptions of the TIMIT corpus. This facilitates the training of AF-MLPs by mapping the time-aligned phoneme labels to their corresponding articulatory classes.

To obtain a set of phoneme sequences for CPM, HTK [10] was used to train a set of phoneme models using the training conversations of all target speakers from the SPIDRE corpus. The training of models was based on the phoneme labels converted from the word-level transcriptions and lexicon obtained from [8]. According to the lexicon, there are a total of 46 phonemes, including one silence and four types of noise. Acoustic vectors of 39 dimensions, each comprising 12 MFCCs, the normalized power as well as their first- and second-order derivatives, were used for phoneme model training and recognition. Each of the 46 context-independent phonemes was modeled by a three-state left-to-right hidden Markov model with 16 diagonal-covariance Gaussian mixtures per state.

The two AF-MLPs were trained using the Quicknet [11]. The MLPs are 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 HTIMIT handsets were used to train the AF-MLPs.

For a given speaker s, his/her AF streams and phoneme sequences were time-aligned to compute the joint probabilities (3) to create a set of speaker models $\mathcal{M}_s^{AFCPM}$, which includes the probabilities of 60 manner and place class combinations conditioned on the 41 phonemes (excluding the silence and noise phonemes) in the phone set. A set of universal background models $\mathcal{M}_b^{AFCPM}$ was obtained using the aligned AF streams and phoneme sequences from all target speakers. The way to obtain the phoneme alignments of the training utterances was consistent with that of the verification utterances, which will be discussed in Section 5.

The scores from the AFCPM system were fused with those from an MFCC system. The features for the MFCC system were 24-dimensional MFCC vectors, with each vector $\mathbf{x}_t$ comprising of 12 MFCCs and the corresponding delta coefficients computed every 14ms using a Hamming window of 28ms. A 128-center universal background GMM, $\mathcal{M}_b$, was trained using all training conversations from all target speakers. For a speaker s in the target speaker set, a speaker GMM, $\mathcal{M}_s$, was adapted from $\mathcal{M}_b$ using MAP adaptation [1].

5. Results

Table 3 lists two sets of experimental results. The two experiments are different in the way the phoneme alignments for CPM were obtained. One set of results is summarized under *Recognized Alignment* and the other under *Forced Alignment* in Table 3. The table is divided into three columns: "Matched", "Mismatched" and "All". "Matched" means the same handset was used for both training and testing, whereas "Mismatched" means the testing handsets are different from the training handsets; "All" represents the results were obtained by combining the test data from "Matched" and "Mismatched".

When recognized alignments were used (i.e., the phoneme sequences were obtained from a null-grammar recognizer during both enrollment and verification), an overall EER of 25.83% was obtained. To minimize the effect of incorrect phoneme alignments on verification performance, a nearly perfect phoneme recognizer was assumed to be available. This was achieved by forced aligning the phoneme sequences of all enrollment and verification utterances with the transcribed word sequences and lexicon obtained from [8]. The results of using forced phoneme alignments are summarized under the heading *Forced Alignment* in Table 3. The overall EER is reduced to 22.69%. The reduction from 25.83% to 22.69% suggests that the accuracy of phoneme alignments is critical to the verification performance of the AFCPM system.

The experimental results of the MFCC system and the fusion systems are also summarized in Table 3. The fusion weights of the systems were determined from a four-fold cross validation using the test data from target speakers and non-target speakers. Note that the MFCC system does not require any phoneme alignments. Results show that the MFCC system achieves an EER of 15.84% on all test data, which is the baseline for comparison.

Significant error reductions were obtained by incoperating the frame-based manner class probabilities into the fusion scores (10). With frame-weighted fusion scores and recognized alignments, an EER of 14.44% was achieved, which represents an error reduction of 8.83%. When forced alignments were used, the fusion system achieved an EER of 14.92%, which represents a 16.28% error reduction. The results demonstrate the effectiveness of adjusting the contribution of individual frames according to the score confidence in each frame. From the results corresponding to matched handsets and mismatched handsets in Table 3, it is clear that fusion of the spectral-based scores and AFCPM scores plays an important role in reducing the EER under handset mismatched conditions. As the spectral features are less reliable under handset mismatched conditions, the speaker information from AFCPM becomes more important.

The error rate reduction in our system is less significant as compared to [3]. This is mainly due to the differences between the two experiments. First, less data is available from the SPIDRE corpus for speaker enrollment in our system. In [3], up to 40 minutes per speaker (1, 2, 4, 8 and 16 conversations each with 2.5 minutes were available) were used for enrollment, whereas only 5 minutes enrollment data were available in our experiments. The second difference is that relatively short speech segments (1 to 15 seconds) were used for verification in our experiments, in contrast to an entire conversation (2.5 minutes) in [3]. In terms of data requirements, our proposed approach has merit because it requires a small amount of speech data for enrollment and short utterances for verification.

6. Conclusions

We have proposed an AFCPM speaker verification system that distinguishes speakers based on their pronunciation characteristics. Experimental results have demonstrated the effectiveness of the AFCPM system in telephone-based speaker verification. It was found that AFCPM provides speaker-dependent information complementary to spectral features. It was also found that AFCPM is especially effective under handset mismatched conditions. A frame-based confidence weighting scheme was proposed to fuse MFCC and AFCPM scores. The weights were determined from the output probabilities of the manner classifier. A lower error rates was achieved because the weighting scheme can optimally controls the contribution of individual frame-based fusion scores.

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. International Conference on Audio, Speech, and Signal Processing*, 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. IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2003, vol. 4, pp. 804–807.

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

[5] JHU WS'2002 SuperSID group website *http://www.clsp.jhu.edu/ws2002/groups/supersid/*

[6] K.Y. Leung, M.W. Mak, and S.Y. Kung, "Applying articulatory features to telephone-based speaker verification," in *Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing*, Montreal, May 2004, vol. 1, pp. 85–88.

[7] J. P. Campbell and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1999, vol. 2, pp. 829–832.

[8] SWITCHBOARD transcription *http://www.isip.msstate.edu/projects/switchboard/*

[9] 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.

[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.