

Adaptive Articulatory Feature-Based Conditional Pronunciation Modeling for Speaker Verification *

K. Y. Leung and M. W. Mak
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University.

M. H. Siu
Department of Electrical and Electronic Engineering,
Hong Kong University of Science and Technology.

S. Y. Kung
Department of Electrical Engineering,
Princeton University.

April 12, 2005

*This work was supported by Research Grant Council of the Hong Kong SAR (Project No. PolyU5214/04E and CUHK 1/02C). Correspondence should be sent to M. W. Mak, Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. Email: enmwak@polyu.edu.hk.

Abstract

Because of the differences in education background, accents, and so on, different persons have different ways of pronunciation. Therefore, the pronunciation patterns of individuals can be used as features for discriminating speakers. This paper exploits the pronunciation characteristics of speakers and proposes a new conditional pronunciation modeling (CPM) technique for speaker verification. The proposed technique establishes 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, which is followed by formulating the probabilities of articulatory classes conditioned on the phonemes as speaker-dependent discrete probabilistic models. The scores obtained from the AF-based pronunciation models are then fused with those obtained from spectral-based acoustic models. A frame-weighted fusion approach is introduced to weight the frame-based fused scores based on the confidence of observing the articulatory classes. The effectiveness of AF-based CPM and the frame-weighted approach is demonstrated in a speaker verification task.

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. In addition to spectral characteristics of speech, humans make use of high-level information such as dialect and lexical characteristics [2] to recognize speakers. This information has not been widely adopted in speaker verification systems because obtaining such information reliably from speech signals remains a challenging research problem.

In recent years, researchers have started to investigate the use of high-level features [3], such as the usage or duration of particular words [4], prosodic features [5, 6], etc., for speaker recognition. Their work has demonstrated that these features contain different amount of speaker-dependent information and some of them can be applied to identify speakers provided

that sufficient speech data are available. For example, in Campbell et al. [7], a significant improvement in speaker recognition accuracy was obtained when high-level features were fused with traditional spectral features using a simple perceptron.

Among all the high-level features evaluated in Reynolds et al. [3], the best performance was achieved by a system that uses conditional pronunciation modeling (CPM) techniques [8]. 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. In Klusáček et al. [8], the intended phonemes were obtained by a recognizer with lexical constraints and the actual phones were obtained by five null-grammar phone recognizers corresponding to five different languages. The pronunciation behaviors were encoded as discrete probability densities that were 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; as a result, a phoneme realized by different speakers could be recognized as different phones by the phone recognizers. However, CPM requires multilingual speech data for training the phone models of different languages; it also needs long utterances for speaker enrollment and verification. For example, in Klusáček et al. [8], 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, we have recently proposed using articulatory feature (AF) streams to construct conditional pronunciation models [9], where abstract classes that describe the movements or positions of different articulators during speech production were used as AFs [10]. Compared to phone-based CPM in [8], 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 expected to be better than phone-based CPM in terms of speaker modeling. In addition, articulatory properties are the same irrespective of languages, monolingual speech data are sufficient for determining their values. This helps to reduce resource requirements in large-scale deployment.

In our previous work [9], 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 adaptive approach in which the discrete distributions of speaker models are adapted from those of universal background models.

The proposed AF-based CPM was applied to a speaker verification task involving 44 target speakers and 160 impostors of the SPIDRE corpus [11]. Experimental results show that speaker information exists in AFs and that 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.

The remainder of the paper is organized as follows. Section 2 details the AF extraction process and the AF-based CPM verification system. In Section 3, an utterance-based fusion of the AF-based CPM system and a conventional spectral-based system is outlined. Besides the standard utterance-based fusion, a frame-weighted fusion is introduced. Section 4 explains how AF-based CPM can be applied to speaker verification and compares its effectiveness against conventional spectral-based techniques. Finally, the paper is concluded in Section 6.

2 AF-Based CPM

This section details the notion of articulatory features and their extraction methods. It also explains how AFs can be applied to model the pronunciation characteristics of speakers.

2.1 Articulatory Features

AFs are the representations of some important phonological properties appeared during speech production. More precisely, classes that describe the movements or positions of different articulators during speech production are treated as AFs. They have been used as features in speech recognition [10, 12] and language identification [13], where the phonological properties

during speech production rather than the spectral characteristics of the produced speech were considered.

AFs have also been applied to speaker identification [14] and speaker verification [15]. In [14], the characteristics of each speaker were modeled by seven speaker-dependent language models, each of which modeled the classes of one 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 of the seven articulatory properties. The usefulness of AFs in speaker verification was demonstrated in our previous work [15], where for each utterance, the probabilities of different articulatory classes determined from five multilayer perceptron-based AF classifiers 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

Kirchhoff [10] trained a classifier for each articulatory property to learn the mapping between the spectral features and the corresponding articulatory states (defined by a one-of- K coding scheme for an articulatory property with K classes). In this work, an approach similar to [10] was adopted to extract the AFs corresponding to the manner and place of articulations. More specifically, two multilayer perceptrons (MLPs)¹ with outputs representing the posterior probabilities of the classes defined in Table 1 were trained. Same as [10], we have adopted a softmax function as the output activation function of the AF-MLPs. To reduce the input dimension of the pronunciation models, only two out of five articulatory properties suggested in [15] were adopted. They were chosen because their combinations are able to distinguish consonants and most of the vowels.

The inputs to the two AF-MLPs are identical while their numbers of outputs are equal to the numbers of AF classes listed in the last column of Table 1. To ensure a more accurate estimation of the AF values, multiple frames of Mel-frequency cepstral coefficients (MFCCs) X_t

¹Referred to as AF-MLPs hereafter.

<i>Articulatory properties</i>	<i>Classes</i>	<i>Number of Classes</i>
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.

(with consecutive frame indexes ranging from $t - \frac{n}{2}$ to $t + \frac{n}{2}$) are served as the inputs to the AF-MLPs at frame t . In other words, the AFs at frame t are obtained from n consecutive frames of MFCCs centered at frame t . Rather than feeding the MFCCs directly to the AF-MLPs, they are normalized to zero mean and unit variance using a global mean vector and a variance vector. The normalization aims to confine the MLP inputs to within a suitable range so that the determination of MLP weights will not be dominated by those large magnitude inputs.

For a given X_t , the outputs of the two AF-MLPs, $P(\text{Manner} = m|X_t)$ and $P(\text{Place} = p|X_t)$, represent the posterior probabilities of different classes in the manner and place of articulation. The manner class label $l^M(X_t) \in \mathcal{M} = \{\text{Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral}\}$ and the place class label $l^P(X_t) \in \mathcal{P} = \{\text{Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal}\}$ at frame t are determined by

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

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

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 $l^M(X_t)$'s and $l^P(X_t)$'s from $t = 1, \dots, T$, where T is the total number of frames in the utterance. Table 2 shows an example of a 20-frame segment extracted from an utterance.

The two AF-MLPs can be trained from speech data with time-aligned phonetic labels. The phonetic alignments can be obtained from transcriptions or Viterbi decoding using phoneme

models. With the phoneme labels, articulatory classes can then be derived from a mapping between phonemes and their states of articulations [10].

2.3 Speaker Modeling

AF-based CPM (hereafter, referred to as AF-CPM) aims to establish a relationship between the 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. This subsection explains the procedures of training and testing the speaker and background models in an AF-CPM-based speaker verification system.

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 phoneme recognizer. For a particular phoneme q , the joint probabilities of the corresponding UBM are determined by

$$\begin{aligned}
 &P(\textit{Manner} = m, \textit{Place} = p | \textit{Phoneme} = q, \textit{Background}) \\
 &= \frac{\#((m, p, q) \text{ in the utterances of all background speakers})}{\#((*, *, q) \text{ in the utterances of all background speakers})}
 \end{aligned} \tag{3}$$

where $m \in \mathcal{M}$, $p \in \mathcal{P}$, (m, p, q) denotes the condition for which $\textit{Manner} = m$, $\textit{Place} = p$, and $\textit{Phoneme} = q$, $*$ represents all possible members in that class, and $\#()$ represents the total number of frames with phoneme labels and AF labels that fulfill the description inside the parentheses. For example, using the data in Table 2 and assuming only one background speaker, $P(\textit{Manner} = \textit{'Vowel'}, \textit{Place} = \textit{'Low'} | \textit{Phoneme} = /aa/) = 5/5 = 1$ and $P(\textit{Manner} = \textit{'Vowel'}, \textit{Place} = \textit{'Low'} | \textit{Phoneme} = /t/) = 1/6 = 0.167$. The probabilities of unseen AF combinations are set to zero or a small value (flooring). For each phoneme, a total of 60 probabilities can

Frame t	Phoneme q_t	AF class label and its probability			
		$l^M(X_t)$	$P(Manner = l^M(X_t) X_t)$	$l^P(X_t)$	$P(Place = l^P(X_t) X_t)$
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
12	v	Nasal	0.33	Coronal	0.71
13	v	Nasal	0.68	Coronal	0.50
14	v	Nasal	0.56	Coronal	0.45
15	v	Nasal	0.41	Labial	0.45
16	v	Fricative	0.47	Coronal	0.34
17	eh	Stop	0.43	Labial	0.32
18	eh	Stop	0.45	Labial	0.43
19	eh	Vowel	0.66	Labial	0.27
20	eh	Vowel	0.81	Middle	0.40

Table 2: A 20-frame example of an aligned phoneme sequence and its corresponding AF streams $\{l^M(X_t)$ and $l^P(X_t); t = 1, \dots, 20\}$. The manner class labels, $l^M(X_t)$'s, and place class labels, $l^P(X_t)$'s, are determined by Eqs. 1 and 2. These labels indicate the classes for which their corresponding posterior probabilities— $P(Manner = m|X_t)$ and $P(Place = p|X_t)$ —are maximum.

	Silence	Labial	Dental	Coronal	Palatal	Velar	Glottal	High	Middle	Low
Silence	4/6	0	0	0	0	0	0	0	0	0
Vowel	0	0	0	0	0	0	0	0	0	1/6
Stop	0	0	0	1/6	0	0	0	0	0	0
Nasal	0	0	0	0	0	0	0	0	0	0
Fricative	0	0	0	0	0	0	0	0	0	0
Approximant- Lateral	0	0	0	0	0	0	0	0	0	0

Table 3: The background model corresponding to phoneme /t/ obtained by using the utterance shown in Table 2 as the only enrollment data.

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. Table 3 shows the entries of the background model corresponding to phoneme /t/ based on the enrollment data in Table 2.

2.3.2 Speaker models

Similar to the UBMs, each speaker model consists of the joint probabilities of the manner and place classes. For a particular speaker s , the joint probabilities corresponding to phoneme q are given by

$$\begin{aligned}
 &P(\text{Manner} = m, \text{Place} = p | \text{Phoneme} = q, \text{Speaker} = s) \\
 &= \frac{\#((m, p, q) \text{ in the utterances of speaker } s)}{\#((*, *, q) \text{ in the utterances of speaker } s)}, \tag{4}
 \end{aligned}$$

where only the utterances from speaker s are used in the computation. The accuracy of the speaker-dependent joint probabilities is limited by the amount of training data available. For some phoneme (e.g., /th/, /sh/ and /v/), the number of occurrences is too low for an accurate estimation of the joint probabilities. As a result, the pronunciation models of these phonemes are less discriminative.

2.3.3 Speaker models by MAP adaptation

To overcome the data-sparseness problem mentioned in Section 2.3.2, 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:

$$\begin{aligned} \hat{P}(Manner = m, Place = p | Phoneme = q, Speaker = s) \\ = \beta_q P(Manner = m, Place = p | Phoneme = q, Speaker = s) \\ + (1 - \beta_q) P(Manner = m, Place = p | Phoneme = q, Background), \end{aligned} \quad (5)$$

where $\beta_q \in [0, 1]$ 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_q is obtained by

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

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. If the number of occurrences of $(*, *, q)$ is much less than r , then β_q will be very close to 0 and the estimation of the new model is less dependent on speaker’s data. On the contrary, if the number of occurrences of $(*, *, q)$ is significantly greater than r , then β_q will be very close to 1 and the adapted model will become more dependent on speaker’s data.

Figure 1 illustrates an example of MAP adaptation with the relevance factor r set to 18. The discrete distribution corresponding to phoneme /aa/ of speaker sp1013 in SPIDRE was selected for illustration. For ease of comparison, the 60 manner and place class probabilities were displayed as a 256-level grayscale image. Figures 1(a) and 1(b) depict the discrete distributions estimated from the training data of all background speakers (44 target speakers in SPIDRE) and speaker sp1013, respectively. The adapted distribution of sp1013, which is the weighted average of the background distribution and the speaker distribution, is shown in Figure 1(c). Because

some combinations of manner and place classes for a particular phoneme may occur rarely (or even none at all) in a single training utterance, the majority of entries in Figure 1(b) have very low or zero probabilities (boxes in white). According to Eq. 5, these entries have insignificant contributions to the adapted distribution (because $P(\text{Manner} = m, \text{Place} = p | \text{Phoneme} = q, \text{Speaker} = s) \approx 0$), which is reflected in Figure 1(c) where the gray levels of those white boxes in Figure 1(b) are almost identical to those in Figure 1(a). This is a reasonable result because if the amount of speaker’s data is insufficient for an accurate estimation of the distribution, the adapted distribution should depend more on the background model than on the speaker’s data.

2.3.4 Verification

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 :

$$S_{AFCPM} = S_s - S_b \quad (7)$$

$$= \sum_{\substack{t=1, \\ p_s(X_t) \neq 0, p_b(X_t) \neq 0 \\ q_t \neq \text{silence}}}^T (\log p_s(X_t) - \log p_b(X_t)), \quad (8)$$

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

$$p_b(X_t) = P(\text{Manner} = l^M(X_t), \text{Place} = l^P(X_t) | \text{Phoneme} = q_t, \text{Background}) \quad (9)$$

and $p_s(X_t)$ are either directly computed from the data of s or adapted from the background model. Therefore, $p_s(X_t)$ are computed from Eq. 4 in Section 2.3.2

$$p_s(X_t) = P(\text{Manner} = l^M(X_t), \text{Place} = l^P(X_t) | \text{Phoneme} = q_t, \text{Speaker} = s) \quad (10)$$

or alternatively based on Eq. 5 in Sections 2.3.3

$$p_s(X_t) = \hat{P}(\text{Manner} = l^M(X_t), \text{Place} = l^P(X_t) | \text{Phoneme} = q_t, \text{Speaker} = s). \quad (11)$$

In Eqs. 9-11, 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,

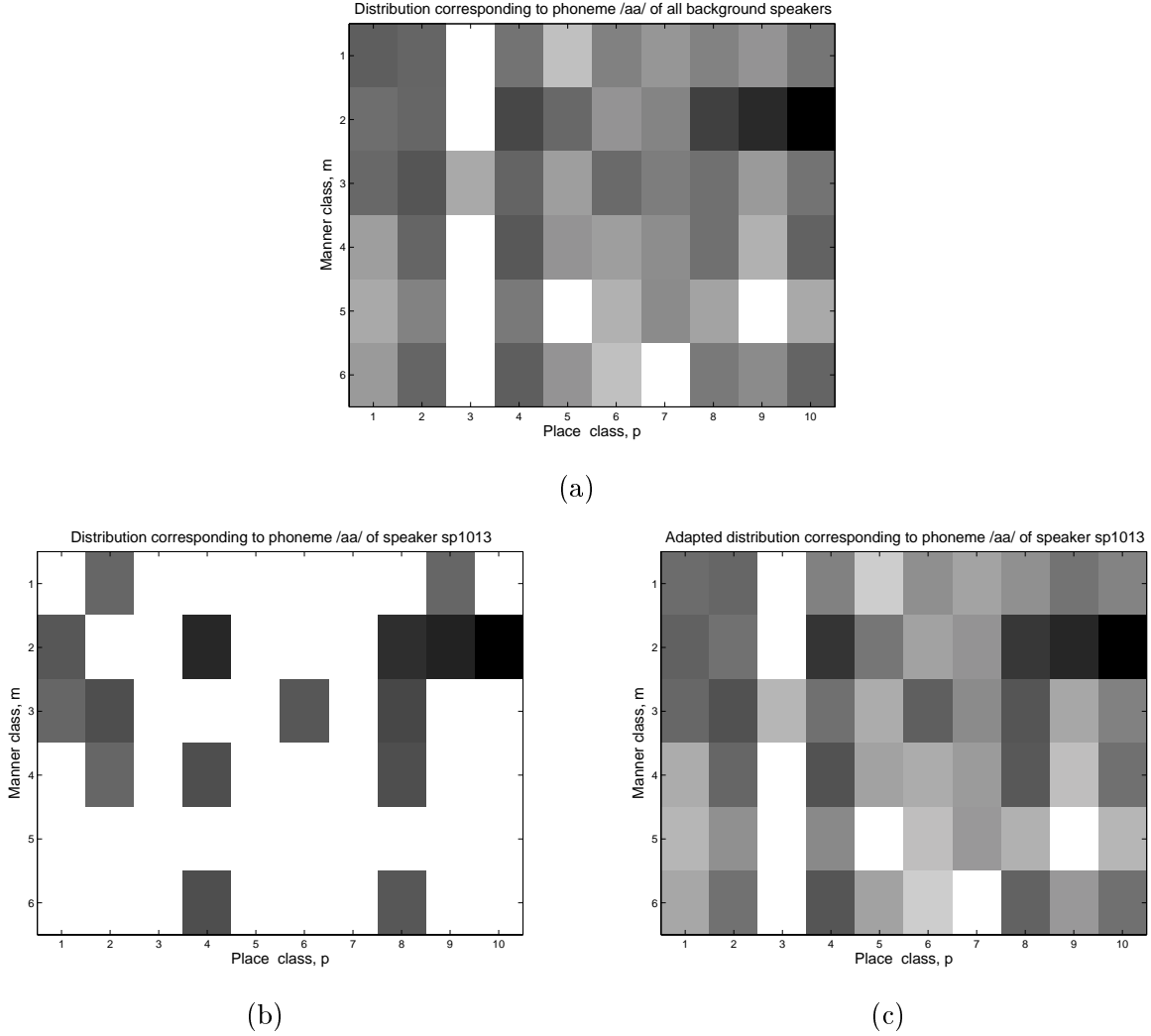


Figure 1: AFCPM of phoneme /aa/ based on (a) the training utterances of all background speakers in SPIDRE, (b) the training utterance of speaker sp1013 in SPIDRE, and (c) the adaptation of the distributions in (a) and (b) using Eqs. 5 and 6. The 60 discrete probabilities corresponding to the combinations of 6 manner and 10 place classes are nonlinearly quantized to 256 gray levels using a log scale, where white represents 0 and black represents 1. The 6 manner classes and 10 place classes, in ascending order of the axis labels are: {Silence, Vowel, Stop, Nasal, Fricative, Approximant-Lateral} and {Silence, Labial, Dental, Coronal, Palatal, Velar, Glottal, High, Middle, Low}.

only the “seen” AF combinations (i.e., $p_s(X_t) > 0$ and $p_b(X_t) > 0$) appeared in both speaker and background models are considered during verification. As can be observed from Eq. 8, the contributions of individual MFCC segments, X_t 's, to the verification score S_{AFCPM} are equally weighted.

S_{AFCPM} is then compared with a threshold for decision making. If S_{AFCPM} of an utterance is greater than or equal to a threshold, it is classified as speaker's speech, otherwise, it is classified as impostor's speech. For ease of comparison among different systems, the threshold is typically varied to obtain a speaker-independent equal error rate (EER), i.e., the point at which the false rejection rate (FRR) is equal to the false acceptance rate (FAR).

3 Fusion of MFCC and AFCPM scores

The traditional GMM-based MFCC system and the proposed AFCPM system characterize speakers at two different levels; the former models the vocal tract's characteristics of individual speakers, whereas the latter looks at the pronunciation behaviors of speakers. Therefore, fusing the scores of MFCC- and AFCPM-based systems is expected to enhance speaker verification performance. This section describes two approaches to fusing the scores: utterance-based fusion and frame-weighted fusion.

3.1 Fusion of Utterance-Based Scores

In the utterance-based fusion approach, the utterance scores (S_{MFCC} and S_{AFCPM}) obtained from the MFCC system and AFCPM system are linearly combined to obtain a fused score

$$S_F = (1 - \alpha_u)S_{MFCC} + \alpha_u S_{AFCPM}, \quad (12)$$

where $\alpha_u \in [0, 1]$ is a fusion weight. Note that the fused score can also be expressed as

$$S_F = (1 - \alpha_u) \frac{\sum_{t=1}^T s_{MFCC}(t)}{T} + \alpha_u \frac{\sum_{t=1}^T s_{AFCPM}(t)}{T} \quad (13)$$

$$= \frac{1}{T} \sum_{t=1}^T [(1 - \alpha_u) s_{MFCC}(t) + \alpha_u s_{AFCPM}(t)] \quad (14)$$

$$= \frac{1}{T} \sum_{t=1}^T s_F(t), \quad (15)$$

where $s_{MFCC}(t)$ and $s_{AFCPM}(t)$ are the verification scores at frame t obtained from the MFCC and AFCPM systems, respectively. More precisely, we have

$$s_{MFCC}(t) = \log p(\mathbf{x}_t | \Lambda_s^{MFCC}) - \log p(\mathbf{x}_t | \Lambda_b^{MFCC}) \quad (16)$$

and

$$s_{AFCPM}(t) = \log p_s(X_t) - \log p_b(X_t), \quad (17)$$

where \mathbf{x}_t is the MFCC vector at frame t , Λ_s^{MFCC} and Λ_b^{MFCC} are the MFCC-based speaker and background models, respectively. Therefore, the fusion of utterance-based scores (Eq. 12) can be considered as the fusion of frame-based scores (Eq. 14) with each frame being equally weighted.

3.2 Fusion of Frame-Weighted Scores

In Eq. 15, all frames in an utterance are considered equally important. As a result, they have equal contribution to the fused score S_F . However, in most utterances, some frames may contain more speaker-dependent information than the others. Therefore, a better score can be obtained by introducing a frame-dependent parameter $w(t)$ to weight the frame-based fused score $s_F(t)$ at frame t . Specifically, a frame-weighted fused score S_F^w is defined as

$$S_F^w = \frac{1}{\sum_{t'=1}^T w(t')} \sum_{t=1}^T w(t) \overbrace{[(1 - \alpha'_u) s_{MFCC}(t) + \alpha'_u s_{AFCPM}(t)]}^{S_F(t)} \quad (18)$$

$$= \sum_{t=1}^T \left[\frac{w(t)}{\sum_{t'=1}^T w(t')} (1 - \alpha'_u) \right] s_{MFCC}(t) + \sum_{t=1}^T \left[\frac{w(t)}{\sum_{t'=1}^T w(t')} \alpha'_u \right] s_{AFCPM}(t) \quad (19)$$

$$= (1 - \alpha'_u) \overbrace{\sum_{t=1}^T \frac{w(t)}{\sum_{t'=1}^T w(t')} s_{MFCC}(t)}^{S_{MFCC}^w} + \alpha'_u \overbrace{\sum_{t=1}^T \frac{w(t)}{\sum_{t'=1}^T w(t')} s_{AFCPM}(t)}^{S_{AFCPM}^w}, \quad (20)$$

where $\alpha'_u \in [0, 1]$ is a fusion weight 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. 18, 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 . Another interpretation of S_F^w can be obtained by considering the factors inside the square brackets of Eq. 19 as frame-variate fusion weights (although they do not sum to 1). In such interpretation, the introduction of $w(t)$ can lead to a more flexible fusion between the AFCPM and MFCC systems.

The weight $w(t)$ represents the importance of $s_{MFCC}(t)$ and $s_{AFCPM}(t)$ with respect to the overall fusion score S_F^w . In other words, $w(t)$ can be interpreted as a measure of the reliability of $s_{MFCC}(t)$ and $s_{AFCPM}(t)$. Because both the MFCC and AFCPM systems take MFCCs as input, the same confidence $w(t)$ 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 ($s_{AFCPM}(t)$) because these scores depend greatly on the output classes determined from the two AF-MLPs. Deriving $w(t)$ based on the articulatory classes probabilities can be interpreted as weighting the frame-based fused score according to the confidence of articulatory classes. For the t -th frame in a verification utterance, the maximum of the manner MLP’s outputs was adopted as $w(t)$ ’s, i.e.,

$$w(t) = \max_{m \in \mathcal{M}} P(\text{Manner} = m | X_t) \quad (21)$$

$$= P(\text{Manner} = l^M(X_t) | X_t). \quad (22)$$

Only the manner class probabilities were used because they exhibit more variation than the place class probabilities, and the manner class probabilities are generally higher than the place class probabilities (see Table 2), which is a sign of greater reliability and discriminative power. We have also investigated other frame-weight parameters, including the probabilities derived from the place MLP and the products of probabilities derived from the manner and place MLPs. However, it was found that the probabilities derived from the manner MLP give the best result.

4 Experiments

4.1 Speech Corpus

The AFCPM system was evaluated on the SPIDRE corpus [11], a subset of the Switchboard corpus. In the experiments, 44 out of 45 target speakers (speaker sp1007 was discarded due to corrupted data) and 160 nontarget speakers were used. Each target speakers has four 5-minute conversations (including silence) recorded from three different handsets. Among these four conversations, one was used for training, another one was used for matched-handset testing, and the remaining two were used for mismatched-handset testing. For the nontarget speakers, a total of 200 conversations were available for testing. A development set was used for determining the parameters α_u (Eq. 14), α'_u (Eq. 20), and r (Eq. 6). The set is composed of the test data of 4 target speakers and 20 impostors randomly selected from the speaker and impostor sets.

Another database, the HTIMIT corpus [16], 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 handset-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.

4.2 Speaker Enrollment

4.2.1 The AFCPM System

To obtain a set of phoneme sequences for CPM, the training conversations of all target speakers were used to train a set of phoneme models. The training of models was based on the phoneme labels converted from the word-level transcriptions and the lexicon in [17], where a total of 46 phonemes—including one silence, one background noise, one vocal-noise and one laughter—were adopted. 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. Each of the 46 context-independent phonemes was modeled by a three-state left-to-right HMM with 16 diagonal-covariance Gaussian mixtures per state. The HTK [18] was used to train the HMMs.

The software Quicknet [19] was used to train two AF-MLPs, each of which is 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 from 382 (192 female and 190 male) speakers randomly selected from all of the 10 handsets in the HTIMIT corpus were used to train the AF-MLPs. Speaker-independent AF-MLPs were adopted to ensure consistent assignment of class labels for all speakers.

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. The way to obtain the phoneme alignments of the training utterances was consistent with that of the verification utterances, which will be discussed in detail in Section 5.

Three approaches were adopted to obtain an AF-CPM-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 AF-CPM. 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. The verification performance evaluated on the development set was found to be insensitive to the choice of r in the range 10 to 25. Hereafter, this adaptation approach is referred to as A-AF-CPM. In the third approach, the zero entries in the speaker and background models were replaced by a small value f . We refer to this approach as AF-CPM with flooring. The value f was selected among 0.1, 0.01, and 0.001. It was found that $f = 0.1$ gave the best verification performance on the development set.

4.2.2 The Phoneme-based CPM System

A phoneme-based CPM system [8] was also implemented. In the system, speaker and background models were composed of conditional probabilities of actual phonemes given the intended phonemes.² In [8], the actual phone streams were obtained by a null-grammar phone recognizer while the intended phoneme streams were obtained by a recognizer with lexical constraints. To obtain results comparable to those of the AFCCPM system, one actual phoneme stream and one intended phoneme stream (both are in English) were obtained. The actual English phoneme streams were obtained from a null-grammar phoneme recognizer using the same phoneme models adopted by the AFCCPM system; the intended phoneme streams were obtained from forced aligning the word transcriptions from the lexicon [17] using the same phoneme models. This set-up achieved the best result because obtaining the intended phoneme streams from forced alignment was equivalent to adopting a perfect phoneme recognizer.

4.2.3 The MFCC System

For the MFCC (spectral-based) system, 24-dimensional MFCC vectors were used as features. Each feature vector \mathbf{x}_t comprises 12 MFCCs and the corresponding delta coefficients computed every 14ms using a Hamming window of 28ms. A 512-component universal background GMM Λ_b^{MFCC} 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].

4.3 Verification and Score Fusion

Genuine verification trials involved one handset-match conversation and two handset-mismatch conversations from each of the 44 target speakers; impostor attempts involved 200 conversations

²An enhanced phoneme-based CPM system was also proposed in [8], where the CPM resolution is improved by partitioning a phoneme into multiple states according to the phoneme duration. The results of the baseline phoneme-based CPM system were reported here because better performance was obtained from the baseline system. The baseline system performed better because the amount of enrollment data in SPIDRE was not sufficient for estimating the speaker models with fine resolution.

from 160 nontarget speakers. The same set of nontarget speakers’ conversations was applied to all target speaker models in the impostor attempts. 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 transcriptions [17]. All silence frames were removed by an energy- and zero-crossing-based voice activity detector.

Verification scores of the AFCCPM system were computed according to Eq. 8. Fusion of the AFCCPM and MFCC systems was also evaluated. Two fusion approaches were adopted: utterance-based fusion (as given in Eq. 15) and frame-weighted fusion (as given in Eq. 20). The fusion weights α_u and α'_u were determined by four-fold cross validation using the development data (see Section 4.1). More specifically, the development data was divided into four disjoint subsets, and α_u (or α'_u) was selected such that the average error obtained from the four-fold evaluations was minimized.

5 Results and Discussions

Table 4 summarized the performance of the MFCC system, the AFCCPM systems, the phoneme-based CPM system, and the fusion of the MFCC and A-AFCCPM systems. There are two sets of experimental results from the AFCCPM systems: *Recognized Alignment* and *Forced Alignment*. In the former, the phoneme sequences were obtained from a null-grammar phoneme recognizer with an accuracy (on all testing utterances) of 37.69%; in the latter, the word sequences were given and they were converted into phoneme sequences by looking up from the lexicon [17]. The forced alignments aim to minimize the effect of incorrect phoneme alignments on verification performance by assuming that a nearly perfect phoneme recognizer is available, thereby providing an upper bound performance of the AFCCPM and phoneme-based CPM systems. Note that the MFCC system does not require any phoneme alignments. The results of the MFCC system is the baseline for comparison.

When recognized alignments were used, an overall EER of 24.04% was obtained from AFCCPM with adaptation (labeled as A-AFCCPM). This represents a relative improvement of 7.0% and 3.2% when compared to the AFCCPM system (labeled as AFCCPM) and the AFCCPM system

	<i>Features</i>	<i>EER (%)</i>		
		<i>Matched</i>	<i>Mismatched</i>	<i>All</i>
	MFCC	7.59	18.08	15.29
Recognized Alignment	AFCPM	19.52	27.69	25.83
	AFCPM with flooring	20.00	26.63	24.82
	A-AFCPM	18.07	26.69	24.04
	Utterance-based MFCC + A-AFCPM (relative error reduction %)	7.59 (0.00)	17.14 (5.20)	14.67 (4.05)
	Frame-weighted MFCC + A-AFCPM (relative error reduction %)	7.09 (6.59)	16.31 (9.78)	13.77 (9.94)
	Forced Alignment	Phoneme-based CPM	30.73	32.85
Alignment	AFCPM	17.92	24.98	22.69
	AFCPM with flooring	18.85	24.63	22.85
	A-AFCPM	16.60	23.98	21.72
	Utterance-based MFCC + A-AFCPM (relative error reduction %)	7.59 (0.00)	16.36 (9.51)	14.20 (7.13)
	Frame-weighted MFCC + A-AFCPM (relative error reduction %)	6.85 (9.75)	14.89 (17.64)	12.92 (15.50)

Table 4: EERs and relative error reduction (in %) obtained from the MFCC system, the AFCPM system, the phoneme-based CPM system, and the fusion of the MFCC and AFCPM systems. *A-AFCPM* denotes the adaptive AFCPM system whose speaker models are adapted from the UBMs. *Utterance-based MFCC + A-AFCPM* denotes the fusion of MFCC scores and adaptive AFCPM scores using Eq. 15. *Frame-weighted MFCC + A-AFCPM* denotes the fusion of frame-weighted MFCC and adaptive AFCPM scores using Eq. 20. *Matched* (*Mismatched*) refers to the cases where the handset used by a claimant in a verification session is identical to (different from) the one used by the target speaker 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. Note that the MFCC system does not depend on any phoneme alignments.

with flooring, respectively. Interestingly, AFCPM with flooring achieves an EER comparable to A-AFCPM only when recognized alignments and mismatched handsets were used. This suggests that the flooring approach can provide robust AFCPM models under mismatched environment and less accurate phoneme alignments. However, the use of MAP adaptation generally results in lower EERs among the three AFCPM systems. MAP adaptation not only provides a data-driven estimation of the unseen probabilities but also results in better speaker models. Through the adaptation, speaker models can tightly couple to the UBMs, which prevents over-fitting the speaker models and enhances their discriminative power.

Referring to the overall EERs of the A-AFCPM system given in Table 4, the reduction from 24.04% (using recognized alignments) to 21.72% (using forced alignments) suggests that the accuracy of phoneme alignments is critical to the verification performance of the AFCPM system. In other words, improving the labeling accuracy will lead to a more precise modeling of the pronunciation characteristics. In the experiments, data from target speakers were used to train the phoneme models from which phoneme alignments were obtained. The phoneme models may need to be re-trained when new target speakers are added to the system. Re-training of the phoneme models can be avoided if they are trained from data not deriving from target speakers.

Given the same amount of enrollment data and the same set of phoneme models for obtaining the phoneme alignments, the AFCPM system achieved an EER that is significantly lower than that of the phoneme-based CPM system. This result suggests that when only a limited amount of enrollment data is available, using articulatory features is better than using recognized phonemes to capture the pronunciation characteristics of speakers. This is because the articulatory features describe the speech production properties that are closer to pronunciation behaviors than the recognized phonemes.

There are phonemes (e.g., /d/ and /t/, /sh/ and /zh/) whose articulatory properties are almost identical or very similar. Therefore, it makes sense to merge the AFCPM models of these phonemes. However, our investigation suggests that this approach results in poorly performed speaker models. This may be attributed to the pronunciation variations among these phonemes, although their manner and place classes are identical. Therefore, merging the models of these

phonemes may remove some useful speaker variations required for the verification of speakers.

The system based on the fusion of utterance-based scores (Eq. 15), with each frame being equally weighted, is labeled as *Utterance-based MFCC + A-AFCPM* in Table 4. The overall EERs were reduced to 14.67% (an 4.05% error reduction) and 14.20% (an 7.13% error reduction) when recognized alignments and forced alignments were adopted. This performance gain is mainly attributed to the difference in the speaker information captured by the two systems: the AFCPM system accounts for the variations in phoneme pronunciations while the MFCC system considers the spectral characteristics of speaker’s speech. The reduction of EERs in the fusion systems suggested that the information considered by the individual systems are complementary, which results in a more complete modeling of speaker variations.

The results corresponding to the matched and mismatched handsets clearly show that the fusion of spectral-based scores and A-AFCPM scores plays a more important role in reducing EERs under handset-mismatched conditions than under the handset-match conditions. Because spectral features are less reliable under handset mismatched conditions, pronunciation behaviors become more important, and they provide an alternative source of information for verification.

A significant error reduction was obtained by using Eqs. 20 and 22 to incorporate the frame-based manner class probabilities into the fused scores. With frame-weighted fusion and recognized alignments, an EER of 13.77% was achieved, which represents an error reduction of 9.94%. Figures 2(a) and (b) plot the detection error tradeoff (DET) curves [20] of the baseline, the A-AFCPM system, and the fusion systems corresponding to recognition alignments and forced alignments, respectively. Evidently, the frame-weighted fusion of MFCC scores and A-AFCPM scores performs better than the MFCC baseline as well as the equally-weighted, utterance-based fusion for a wide range false alarm probabilities. As shown in Figure 2(b), when forced alignments were used, frame-weighted fusion achieves the lowest EER, which demonstrates the effectiveness of adjusting the contribution of individual frames according to their score confidence.

The error rate reduction in the proposed system is less significant than that of [8]. This is mainly attributed to the differences between the two experiments. First, less data is

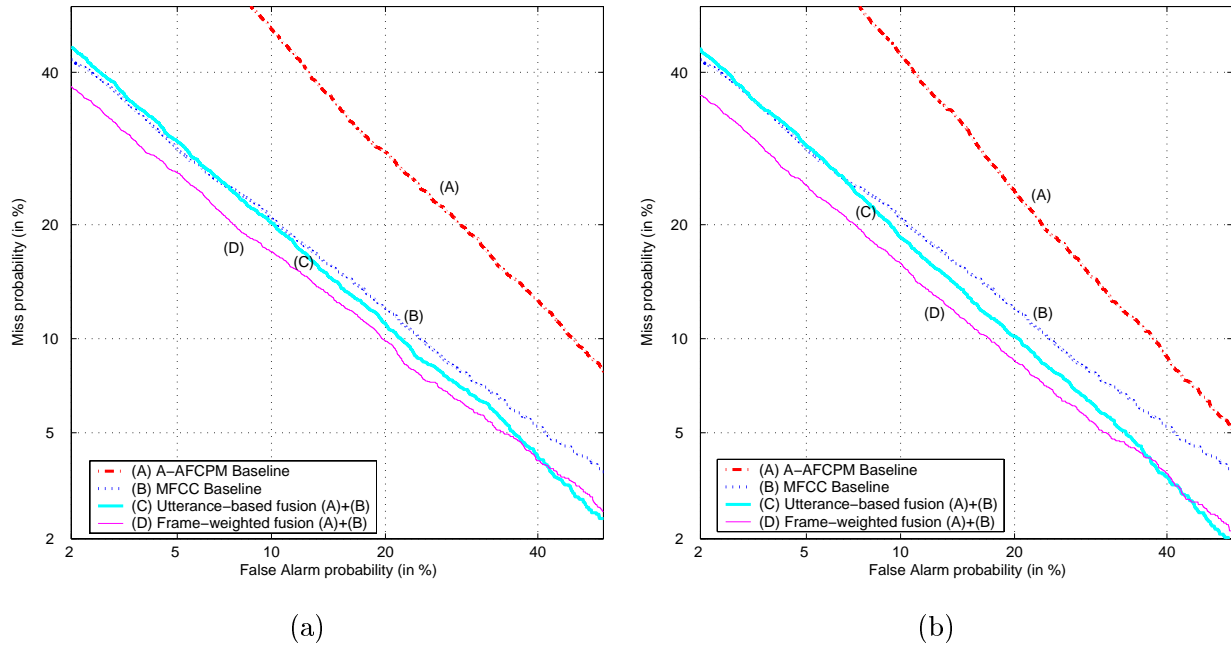


Figure 2: Speaker verification performance based on (a) recognition alignments and (b) forced alignments. For ease of comparison, methods in the legend are arranged in descending order of EERs.

available from the SPIDRE corpus for speaker enrollment in the proposed system. In [8], up to 40 minutes per speaker (1, 2, 4, 8, and 16 conversations, each with 2.5 minutes) were used for enrollment, whereas only 5 minutes enrollment data (an average of 2.5 minutes was left after silence removal) were available in the experiments reported here. The second difference is that short speech segments (1 to 15 seconds) were used for verification in our experiments, in contrast to an entire conversation (2.5 minutes) in [8]. In terms of data requirements, the proposed approach has merit because it requires a small amount of speech data for enrollment and short utterances for verification.

6 Conclusions

This paper has presented an AFCPM speaker verification system in which the conditional pronunciation probabilities of the speaker models are adapted from those of the universal background models. The system distinguishes speakers by capturing their pronunciation characteristics via the conditional pronunciation modeling of two articulatory property streams. 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 is proposed to improve the fusion of MFCC and AFCPM scores. The weights are determined from the output probabilities of the manner MLP. A lower error rate was achieved because the weighting scheme provides a flexible means of controlling the contribution of individual frame-based scores to the overall score.

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] G. R. Doddington., "Speaker recognition—identifying people by their voices," in *Proc. IEEE*, 1995, pp. 1651–1664.
- [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] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, "Phonetic speaker recognition using maximum likelihood binary decision tree models," in *Proc. ICASSP 2003*, 2003, vol. 4, pp. 796–799.
- [5] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusáček, D. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from jhu ws'02," in *Proc. ICASSP 2003*, 2003, vol. 4, pp. 792–795.
- [6] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP 2003*, 2003, vol. 4, pp. 788–791.
- [7] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proc. Eurospeech 2003*, 2003, pp. 2665–2668.

- [8] 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.
- [9] 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.
- [10] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, PhD thesis, University of Bielefeld, 1999.
- [11] 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.
- [12] K. Erler and L. Deng, “Hidden Markov model representation of quantized articulatory features for speech recognition,” *Computer Speech and Language*, vol. 7, no. 3, pp. 265–282, 1993.
- [13] S. Parandekar and K. Kirchhoff, “Multi-stream language identification using data-driven dependency selection,” in *Proc. ICASSP 2003*, 2003, vol. 1, pp. 28–31.
- [14] <http://www.clsp.jhu.edu/ws2002/groups/supersid/>.
- [15] 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.
- [16] 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.
- [17] <http://www.isip.msstate.edu/projects/switchboard/>.
- [18] 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.
- [19] P. Farber, “Quicknet on multispart: fast parallel neural network training,” Tech. Rep. TR-97-047, ICSI, 1997.
- [20] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” in *Proc. Eurospeech 1997*, 1997, pp. 1895–1898.