ARTICULATORY-FEATURE BASED SEQUENCE KERNEL FOR HIGH-LEVEL SPEAKER VERIFICATION

Shi-Xiong Zhang and Man-Wai Mak

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
{zhang.sx,enmwmak}@polyu.edu.hk

ABSTRACT

Research has shown that articulatory feature-based phonetic-class pronunciation models (AFCPMs) can capture the pronunciation characteristics of speakers. However, the scoring method used in AFCPMs does not explicitly use the discriminative information available in the training data. To harness this information, this paper proposes converting speaker models to supervectors by stacking the discrete densities in AFCPMs. An AF-kernel is constructed from the supervectors of target speakers, background speakers, and claimants. An AF-kernel based SVM is then trained to classify the supervectors. Results show that AF-kernel scoring is complementary to likelihood-ratio scoring, leading to better performance when the two scoring methods are combined.

Index Terms— Speaker verification, kernels, articulatory features, pronunciation models, SVM

1. INTRODUCTION

Speaker verification is a binary classification problem in which a speaker is authenticated based on his/her own voice. Text-independent speaker verification systems typically extract speaker features from short-term spectra of speech signals to build speaker-dependent Gaussian mixture models (GMMs) [1]. Studies have shown that combining low-level acoustic information with high-level speaker information—such as the usage or duration of particular words, prosodic features and articulatory features (AF)—can improve speaker verification performance [2–6].

However, in most systems (e.g., GMM-UBM [1], PD-AFCPM [5] and CD-AFCPM [6]), scoring is done at the frame-level, i.e., each frame of speech is scored separately and then frame-based scores are accumulated to produce an utterance-based score for classification. This frame-based scoring scheme has two drawbacks. First, treating the frames individually may not be able to fully capture the sequence information contained in the utterance. Second, the goal of speaker verification is to minimize classification errors on test utterances rather than on individual speech frames. These drawbacks motivate us to derive a sequence-based approach in which an utterance is considered comprising a sequence of symbols and the utterance-based score can be obtained from an SVM [7] through a kernel function of the sequence of symbols.

Support vector machines (SVMs) can produce complex decision functions without a large amount of training data. However, ordinary SVMs can only classify data of fixed dimensionality whereas speech utterances are typically parameterized as variable-length sequences. This leads to the idea of GMM supervectors in which variable-length observation sequences are mapped to fixed-dimensional vectors via stacking the parameters or scores of a GMM [8, 9].

This paper derives an articulatory-feature based sequence kernel and apply it to high-level speaker verification. The method begins with extracting the observation sequences (AF labels) from the utterances of target speakers. The AF sequences are then used to train the articulatory feature-based models (called AFCPM and CD-AFCPM) as proposed in [5, 6]. These models are then converted to fixed-dimensional AF supervectors. Redundant or irrelevant features are then removed from the AF supervectors. The reduced-dimension supervectors derived from both the target speaker and background speakers are used to train an SVM with a specially designed sequence kernel. The speaker-dependent SVM will be used to compute the verification scores. Since the kernel depends on the AF-based target speaker models, we refer it to as articulatory feature (AF) kernel. The remainder of the paper will derive the AF kernel and discuss the relationship between traditional frame-based log-likelihood (LR) scoring and AF-kernel based SVM scoring. Experimental results on the NIST2000 database are presented.

2. ARTICULATORY-FEATURE BASED SUPERVECTOR EXTRACTION

2.1. Articulatory Features

Articulatory features (AFs) are representations describing the movements or positions of different articulators during speech production. In [5, 6], the manner and place properties shown in Table 1 were used for pronunciation modeling. AFs can

This work was supported by the Research Grant Council of the Hong Kong SAR Project No. PolyU5230/05E and HKPolyU Project No. A-PA6F.

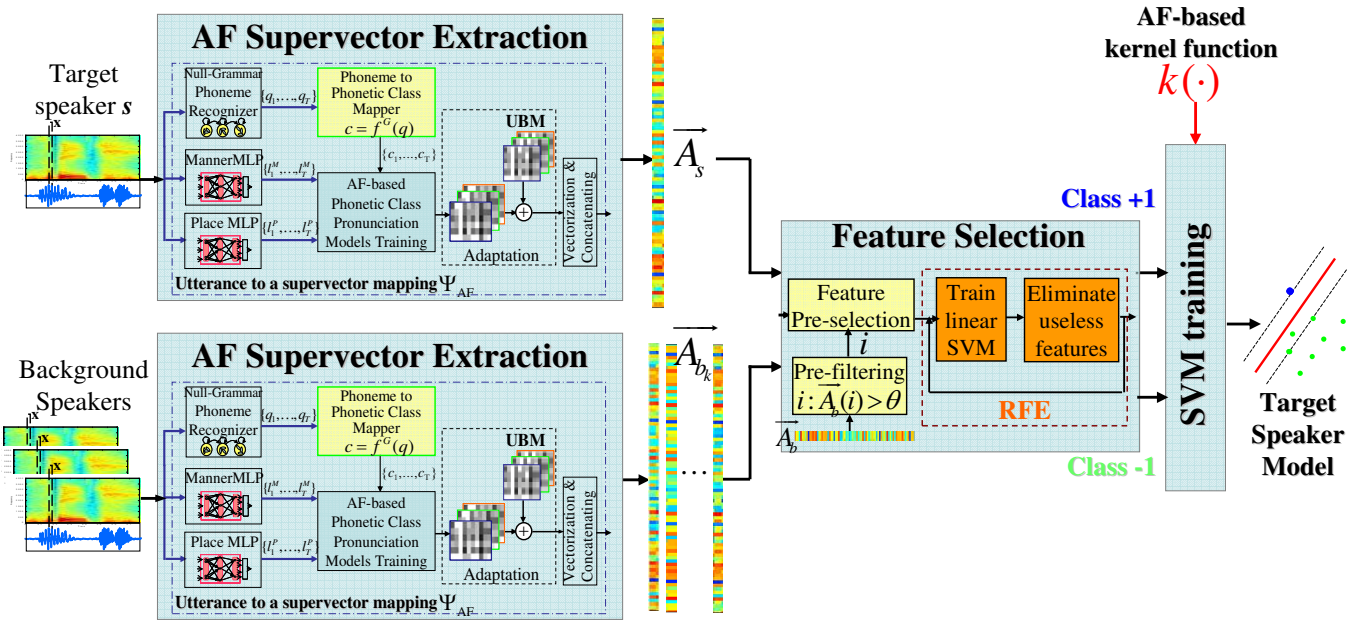

Fig. 1. The training procedure of the AF kernel-based high-level speaker verification system.

Articulatory Properties	Classes	Number of Classes
Manner($\mathcal{M}$)	Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral	6
Place($\mathcal{P}$)	Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal	10

Table 1. The manner and place properties.

be automatically determined from speech signals using AF-based multilayer perceptrons (MLPs) shown in Fig. 1.

2.2. AF-Based Supervectors

Fig. 1 shows the training phase of a high-level speaker verification system that uses AF-kernels. The first step is to create CD-AFCPM supervectors. In [6], G phonetic-class dependent AFCPMs were trained from the AF and phoneme streams of all background speakers to represent the speaker-independent pronunciation characteristics. Each CD-AFCPM comprises the joint probabilities of the manner and place classes conditioned on a phonetic class. The training procedure begins with aligning two AF streams (l_t^{M} and l_t^{P}) obtained from the AF-MLPs and a phonetic class sequence c_t obtained from a null-grammar recognizer and a mapping function. The joint probabilities of background models $P_b^{\mathrm{CD}}(m,p|c)$ corresponding to a particular phonetic class c is given by

$$\begin{aligned} &P_b^{\mathrm{CD}}(m,p|c) \\ &= P_b^{\mathrm{CD}}(L^{\mathrm{M}}=m, L^{\mathrm{P}}=p|\text{Phonetic Class}=c,\text{background}) \\ &= \frac{\#((m,p,c) \text{ in the data of all background speakers})}{\#((*,*,c) \text{ in the data of all background speakers})}, \end{aligned} \quad (1)$$

where $m \in \mathcal{M}, p \in \mathcal{P}, (m,p,q)$ denotes the condition for which $L^{\mathrm{M}}=m$, $L^{\mathrm{P}}=p$, and Phonetic Class $=c$, $*$ represents all possible members in that class, and $\#()$ represents the total number of frames with phoneme labels and AF labels fulfill the description inside the parentheses. We can see for each phonetic class, a total of 60 probabilities can be obtained, and a collection of G CD-AFCPM forms a AF-based universal background model (UBM).

The unadapted speaker model $P_s^{\mathrm{CD}}(m,p|c)$ are created in the same way. And the MAP adaptation described in [10] is applied to create the adapted speaker models

$$\widehat{P}_s^{\mathrm{CD}}(m,p|c) = \beta_c P_s^{\mathrm{CD}}(m,p|c) + (1-\beta_c) P_b^{\mathrm{CD}}(m,p|c), \quad (2)$$

where, $\beta_c \in [0,1]$ is a phonetic class-dependent adaptation coefficient controlling the contribution of the speaker data and the background models (Eq. 1) on the MAP-adapted model.

For each target speaker, G speaker models (each model has 60 parameters) are stacked to form a single supervector, called CD-AFCPM Supervector.[1] The process maps a test utterance to a point in $60G$-dimensional vector space. The procedure of CD-AFCPM supervector extraction is illustrated in Fig. 1.

2.3. Feature Selection

The dimensionality of CD-AFPCM and PD-AFCPM is 720 (when $G=12$) and 2760 (46 phonemes in English), respectively. Many of these features, however, may be redundant or having low discriminative power. Therefore, extracting the relevant features from the high-dimensional supervectors is expected to improve verification performance.

[1]The procedure is also applicable to PD-AFCPM with $G=46$. For clarity, we focus on CD-AFCPM in the sequel.

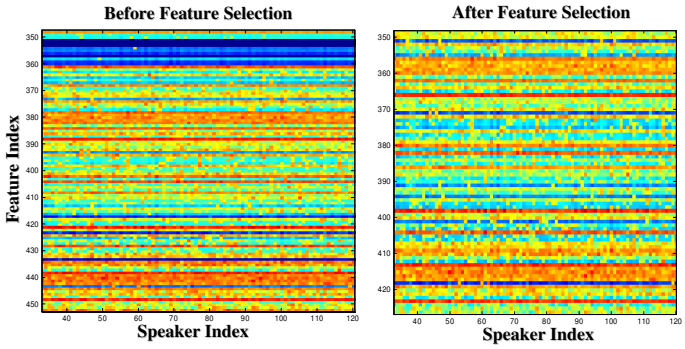

Fig. 2. Effect of feature selection on CD-AFCPM supervectors. A row with small variation (almost identical color intensity) suggests that the corresponding feature is not speaker dependent and therefore can be removed without scarifying classification accuracy.

The feature selection process is divided into two steps. In Step 1, irrelevant features are weeded out using a pre-filtering approach. More precisely, the background CD-AFCPMs are vectorized and feature elements with value smaller than a threshold are removed. This step avoids the numerical difficult that may encounter when the CD-AFCPMs supervectors are normalized during the evaluation of the kernel function (see Eqs. 12 and 13). Then, in Step 2, based on the remaining features in Step 1, a CD-AFCPM supervector is constructed for each target speaker, and 618 CD-AFCPM supervectors are constructed from 618 background speakers. Then, for each speaker, feature elimination is done by applying SVM-RFE [11] to the dataset formed by the speaker's CD-AFCPM supervector (positive class) and background speakers' CD-AFCPM supervectors (negative class). Note that Step 2 is applied to each of the target speakers, meaning that each speaker has their own feature set.

The effect of feature selection is shown in Fig. 2, where each column corresponds to a portion of a speaker's CD-AFCPM supervector. A row with small variation (almost identical color intensity) suggests that the corresponding feature is not speaker dependent and therefore can be removed without sacrifying classification performance. We can see from Fig. 2 (right panel) that features of low discriminative power have been eliminated.

3. ARTICULATORY FEATURE-BASED KERNELS

3.1. Articulatory Feature-Based LR Scoring

Given a test utterance $X_1^T = \{X_1, \ldots, X_t, \ldots, X_T\}$, we can express the frame-based likelihood-ratio (LR) score as follows:

$$S_{\text{CD-AFCPM}}(X_1^T) = \frac{1}{T}\sum_{t=1}^{T}\left(\log\frac{\widehat{P}_s^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|c_t)}{P_b^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|c_t)}\right) \quad (3)$$

where $f^G(q_t)$ is one of the mapping function proposed in [6], c_t is the phonetic class determined by a null-grammar recognizer, and $l_t^{\text{M}} \in \mathcal{M}$ and $l_t^{\text{P}} \in \mathcal{P}$ are the labels determined by the manner and place MLPs, respectively. Grouping all frames that are classified to the same phonetic class, we can further express the LR score as:

$$\begin{aligned}
&S_{\text{CD-AFCPM}}(X_1^T)\\
&= \sum_{c=1}^{G}\frac{1}{T}\left(\sum_{\substack{m\in\mathcal{M}\\ p\in\mathcal{P}}}\left(\sum_{t:\begin{cases} f^G(q_t)=c\\ l_t^{\text{M}}=m,\\ l_t^{\text{P}}=p\end{cases}}\left(\log\frac{\widehat{P}_s^{\text{CD}}(l_t^{\text{M}}=m, l_t^{\text{P}}=p|c)}{P_b^{\text{CD}}(l_t^{\text{M}}=m, l_t^{\text{P}}=p|c)}\right)\right)\right)\\
&= \sum_{c=1}^{G}\frac{T_c}{T}\left\{\frac{1}{T_c}\sum_{i=1}^{60}\left(\left(\log\frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_i|c)}{P_b^{\text{CD}}(\mathcal{L}_i|c)}\right)\sum_{t:\begin{cases} f^G(q_t)=c\\ \mathcal{L}_i\end{cases}} 1\right)\right\}\\
&= \sum_{c=1}^{G}\frac{T_c}{T}\left\{\frac{1}{T_c}\sum_{i=1}^{60}\left(\left(\log\frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_i|c)}{P_b^{\text{CD}}(\mathcal{L}_i|c)}\right)N_{i,c}\right)\right\}\\
&= \sum_{c=1}^{G}\frac{T_c}{T}\left\{\sum_{i=1}^{60}\left(\left(\log\frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_i|c)}{P_b^{\text{CD}}(\mathcal{L}_i|c)}\right)\frac{N_{i,c}}{T_c}\right)\right\}
\end{aligned} \quad (4)$$

where $\mathcal{L}_1 = \{l_t^{\text{M}} = \text{'Vowel'}, l_t^{\text{P}} = \text{'High' for any } t\}, \ldots,$ $\mathcal{L}_{60} = \{l_t^{\text{M}} = \text{'Lateral'}, l_t^{\text{P}} = \text{'Glottal' for any } t\}$, $N_{i,c}$ is the number of frames belonging to phonetic class c and $\mathcal{L}_i$, and T_c is the number of frames belonging to phonetic class c.

Assume that a test utterance is produced by a claimant cl claiming a speaker identity s. Then, we can obtain the CD-AFCPM of the claimant as follows:

$$\begin{aligned}
&P_{cl}^{\text{CD}}(m,p|c)\\
&= P_{cl}^{\text{CD}}(L^{\text{M}}=m, L^{\text{P}}=p|\text{PhoneClass}=c, \text{claimant}=cl)\\
&= P_{cl}^{\text{CD}}(\mathcal{L}_i|c)\\
&= \frac{\#((m,p,c)\text{ in the utterances of the claimant})}{\#((*,*,c)\text{ in the utterances of the claimant})}\\
&= \frac{N_{i,c}}{T_c},
\end{aligned} \quad (5)$$

where index i corresponds to the i-th combination of the manner and place class (m,p). Setting $G = 12$ and substituting Eq. 5 into Eq. 4, we obtain:

$$\begin{aligned}
&S_{\text{CD-AFCPM}}(X_1^T)\\
&= \sum_{c=1}^{G}\frac{T_c}{T}\left(\sum_{i=1}^{60}\left(\left(\log\frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_i|c)}{P_b^{\text{CD}}(\mathcal{L}_i|c)}\right)P_{cl}^{\text{CD}}(\mathcal{L}_i|c)\right)\right)\\
&= \sum_{c=1}^{G}\left\langle\begin{bmatrix}\log\frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_1|c)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_1|c)}\\ \log\frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_2|c)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_2|c)}\\ \cdots\\ \log\frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_{60}|c)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_{60}|c)}\end{bmatrix}_{60}, \begin{bmatrix}\frac{T_c}{T}P_{cl}^{\text{CD}}(\mathcal{L}_1|c)\\ \frac{T_c}{T}P_{cl}^{\text{CD}}(\mathcal{L}_2|c)\\ \cdots\\ \frac{T_c}{T}P_{cl}^{\text{CD}}(\mathcal{L}_{60}|c)\end{bmatrix}_{60}\right\rangle
\end{aligned}$$

$$= \left\langle \begin{bmatrix} \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_1|c=1)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_1|c=1)} \\ \cdots \\ \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_{60}|c=1)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_{60}|c=1)} \\ \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_1|c=2)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_1|c=2)} \\ \cdots \\ \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_{60}|c=2)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_{60}|c=2)} \\ \cdots \\ \cdots \\ \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_{60}|c=12)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_{60}|c=12)} \end{bmatrix}_{720}, \begin{bmatrix} w_1 P_{cl}^{\text{CD}}(\mathcal{L}_1|c=1) \\ \cdots \\ w_1 P_{cl}^{\text{CD}}(\mathcal{L}_{60}|c=1) \\ w_2 P_{cl}^{\text{CD}}(\mathcal{L}_1|c=2) \\ \cdots \\ w_2 P_{cl}^{\text{CD}}(\mathcal{L}_{60}|c=2) \\ \cdots \\ \cdots \\ w_{12} P_{cl}^{\text{CD}}(\mathcal{L}_{60}|c=12) \end{bmatrix}_{720} \right\rangle,$$

where $w_i = \dfrac{T_i}{T}$ $(i = 1, \ldots, 12)$, which gives

$$S_{\text{CD-AFCPM}}(X_1^T) = \left\langle \log \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b}, \overrightarrow{A}'_c \right\rangle = \left\langle \log \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b}, \overrightarrow{w} .* \overrightarrow{A}_c \right\rangle \tag{6}$$

where $\overrightarrow{A}_s$, $\overrightarrow{A}_b$ and $\overrightarrow{A}_c$ stand for the AF supervector of the

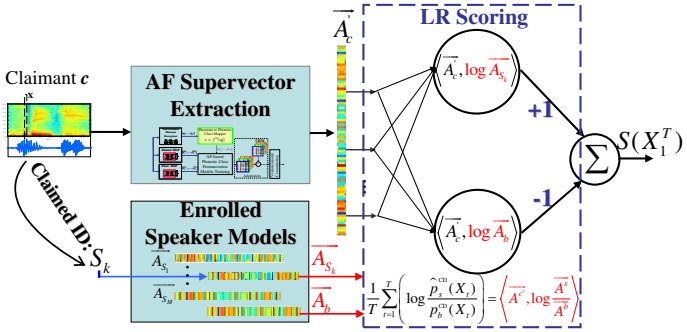

Fig. 3. An alternative implementation of the traditional log-likelihood scoring in CD-AFCPM speaker verification.

speaker, background and claimant, respectively, $\log \dfrac{\overrightarrow{X}}{\overrightarrow{Y}} \equiv \left[\log \dfrac{x_1}{y_1}, \ldots, \log \dfrac{x_N}{y_N} \right]^{\text{T}}$ and $\overrightarrow{X} .* \overrightarrow{Y} \equiv \left[x_1 y_1, \ldots, x_N y_N \right]^{\text{T}}$, where x_i and y_i are elements of $\overrightarrow{X}$ and $\overrightarrow{Y}$, respectively. We can see from Eq. 6 that the traditional frame-based LR scoring for discrete models can be computed in another way: compute the dot product between a speaker-dependent supervector derived from the models of the target speakers and a weighted supervector obtained from the claimant's model.

Further, denote $\overrightarrow{A}'_c$ as the weighted AF supervector of the claimant, we have

$$\begin{aligned} S_{\text{CD-AFCPM}}(X_1^T) &= \left\langle \log \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b}, \overrightarrow{A}'_c \right\rangle \\ &= \left\langle \overrightarrow{A}'_c, \log \overrightarrow{A}_s \right\rangle - \left\langle \overrightarrow{A}'_c, \log \overrightarrow{A}_b \right\rangle. \end{aligned} \tag{7}$$

Eq. 7 suggests an alternative approach to implementing the traditional LR scoring. This is shown in Fig. 3. It shows that if we can replace the fixed '+1' and '−1' multiplication factors in the LR scoring block by varying weights, the result may probably be improved. These weights can be optimally determined by SVM training. In order to apply SVM and to make sure that the training algorithm converges to a stable solution, the function inside the cycle in Fig. 3 should satisfy the Mercer condition [7]. Unfortunately, $f(\overrightarrow{X}, \overrightarrow{Y}) = \left\langle \overrightarrow{X}, \log \overrightarrow{Y} \right\rangle$ does not satisfy the Mercer condition because $\left\langle \overrightarrow{X}, \log \overrightarrow{Y} \right\rangle$ cannot be written as $\left\langle \Phi(\overrightarrow{X}), \Phi(\overrightarrow{Y}) \right\rangle$. Therefore, we derive an AF kernel function that satisfies the Mercer condition in the next section.

3.2. Deriving Kernels from Similarity Scores

Given AF-based supervectors $\overrightarrow{A}_s$ and $\overrightarrow{A}_b$ obtained from the utterances of speaker s and background speakers into a fixed-dimension input space, the similarity score between the model deriving from the test utterance X_1^T of claimant c and the model of speaker s can be computed by a similarity (discriminant) function:

$$\text{Similarity}(\overrightarrow{A}_c, \overrightarrow{A}_s) = \left\langle \overrightarrow{A}'_c, \log \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \right\rangle \tag{8}$$

where Eq. 7 has been used in the derivation.

Our goal is to make Eq. 8 symmetric and satisfy the Mercer condition. To this end, we expand $\log(x)$ at $x = 1$ as a Taylor series:

$$\begin{aligned} \log(x) &= \sum_{n=0}^{\infty} \frac{\log^{(n)}(1)}{n!}(x-1)^n \\ &= (x-1) - \frac{1}{2}(x-1)^2 + \mathcal{O}\left((x-1)^3\right). \end{aligned} \tag{9}$$

Because the speaker models are adapted from the UBMs, $\dfrac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \to \overrightarrow{1}$. Therefore, we can ignore the high orders of $\left(\overrightarrow{A}_s / \overrightarrow{A}_b - \overrightarrow{1}\right)$ and approximate Eq. 8 as:

$$\begin{aligned} &\text{Similarity}(\overrightarrow{A}_c, \overrightarrow{A}_s) \\ &= \left\langle \overrightarrow{A}'_c, \log \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \right\rangle \approx \left\langle \overrightarrow{A}'_c, \left(\frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} - \overrightarrow{1} \right) \right\rangle \\ &= \left\langle \overrightarrow{A}'_c, \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \right\rangle - \left\langle \overrightarrow{A}'_c, \overrightarrow{1} \right\rangle = \left\langle \overrightarrow{A}'_c, \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \right\rangle - 1, \end{aligned} \tag{10}$$

where the last term is due to

$$\sum_{i=1}^{60G} A'_{c,i} = \sum_{k=1}^{G} \sum_{i=60(k-1)+1}^{60k} w_i A_{c,i} = \frac{1}{T} \sum_{k=1}^{G} T_k = 1.$$

Because the last term can be dropped without affecting the classification, we have

$$\text{Similarity}(\overrightarrow{A}_c, \overrightarrow{A}_s) \approx \left\langle \frac{\overrightarrow{A}_c}{\overrightarrow{A}_b}, \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \right\rangle = \left\langle \frac{\overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle$$
$$= \left\langle \frac{\overrightarrow{w}. * \overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle \approx \left\langle \frac{\sqrt{\overrightarrow{w}_b}. * \overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\sqrt{\overrightarrow{w}_b}. * \overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle \tag{11}$$

where $\overrightarrow{w}_b = \left[\overbrace{\frac{T_1^b}{T}, \cdots, \frac{T_1^b}{T}}^{60}, \overbrace{\frac{T_2^b}{T}, \cdots, \frac{T_2^b}{T}}^{60}, \cdots, \overbrace{\frac{T_G^b}{T}, \cdots, \frac{T_G^b}{T}}^{60} \right]^{\mathrm{T}}_{60G}$ contains the phonetic-class weights obtained from the UBM, which is used to approximate $\overrightarrow{w}$ in Eq. 6. The approximation aims to make the similarity measure symmetric.

Finally, we write the similarity function (Eq. 11) as a kernel:

$$K_{\mathrm{AF}}(\overrightarrow{A}_c, \overrightarrow{A}_s) = \left\langle \frac{\sqrt{\overrightarrow{w}_b}. * \overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\sqrt{\overrightarrow{w}_b}. * \overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle = \left\langle \varphi(\overrightarrow{A}_c), \varphi(\overrightarrow{A}_s) \right\rangle \tag{12}$$

where the mapping $\varphi(\cdot)$ is defined as:

$$\varphi(\overrightarrow{X}) = \frac{\sqrt{\overrightarrow{w}_b}. * \overrightarrow{X}}{\sqrt{\overrightarrow{A}_b}}. \tag{13}$$

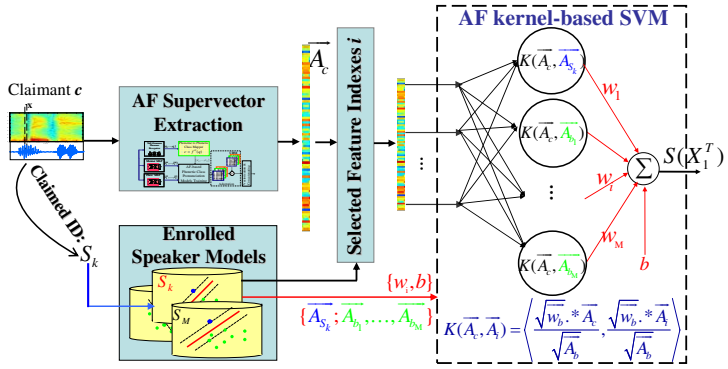

Fig. 4. The verification phase of an AF-kernel based speaker verification system.

Figures 5(a) and 5(b) show the un-normalized $\overrightarrow{A}_s$ and the normalized $\overrightarrow{A}_s$ (i.e., $\varphi(\overrightarrow{A}_s)$) for 150 speakers, respectively. For clarity, only feature elements with indexes between 531 and 660 are shown. Evidently, without normalization, some features have a large but almost constant value across all speakers (e.g., rows with dark-red color). These features will cause problems in SVM classification because they will dictate the decision boundary of the SVM, even though they contain little speaker-dependent information. This problem has been largely alleviated by the normalization, as demonstrated in Fig. 5(b). In particular, the normalization has the effect of keeping all features within a comparable range, which helps prevent the large but almost constant features from dominating the classification decision. Figures 4 shows the scoring procedure during the verification phase. Comparing Fig. 4 and Fig. 3 suggests that scoring based on the AF-kernel is more general. The scores produced by the AF-kernel may also complement the ones produced by the LR-based method, which will be demonstrated in the next section.

Note that Eq. 12 depends on the speaker model, which in turn depends on the target speaker's utterances. Denote $\overrightarrow{O}_s$ as the stacking of all acoustic vectors of speaker s. We can consider the training of speaker s's CD-AFCPM as finding a function $\Psi_{\mathrm{AF}}(\overrightarrow{O}_s)$ that maps the variable-dimension supervector $\overrightarrow{O}_s$ in $\Re^{TD}$ to an AF-supervector $\overrightarrow{A}_s$ of fixed-dimension in $\Re^{60G}$, where T is the number of acoustic frames in the training utterances, D is the dimension of the acoustic vectors, and G is the number of phonetic classes. Therefore, we can express the AF kernel as:

$$\widetilde{K}_{\mathrm{AF}}(\overrightarrow{O}_c, \overrightarrow{O}_s) = \left\langle \frac{\sqrt{\overrightarrow{w}_b}. * \Psi_{\mathrm{AF}}(\overrightarrow{O}_c)}{\sqrt{\Psi_{\mathrm{AF}}(\overrightarrow{O}_b)}}, \frac{\sqrt{\overrightarrow{w}_b}. * \Psi_{\mathrm{AF}}(\overrightarrow{O}_s)}{\sqrt{\Psi_{\mathrm{AF}}(\overrightarrow{O}_b)}} \right\rangle$$
$$= \left\langle \Phi_{\mathrm{AF}}(\overrightarrow{O}_c), \Phi_{\mathrm{AF}}(\overrightarrow{O}_s) \right\rangle \tag{14}$$

where the mapping $\Phi_{\mathrm{AF}}(\cdot)$ is defined as:

$$\Phi_{\mathrm{AF}}(\overrightarrow{O}) = \frac{\sqrt{\overrightarrow{w}_b}. * \Psi_{\mathrm{AF}}(\overrightarrow{O})}{\sqrt{\Psi_{\mathrm{AF}}(\overrightarrow{O}_b)}}. \tag{15}$$

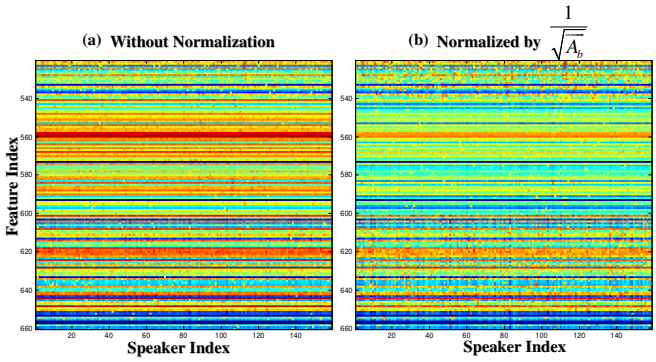

Fig. 5. The effect of the normalization term $\frac{1}{\sqrt{A_b}}$ in the mapping $\varphi(\cdot)$.

4. EXPERIMENTS AND RESULTS

4.1. Procedures

NIST99, NIST00, SPIDRE, and HTIMIT were used in the experiments. NIST99 was used for creating the background models and mapping functions, and the female part of NIST00 was used for creating speaker models and for performance evaluation. HTIMIT and SPIDRE were used for training the

AF-MLPs and the null-grammar phone recognizer, respectively. The phone recognizer uses standard 39-D vectors comprising MFCCs, energy, and their derivatives. The AF-MLPs use 38-D vectors comprising 19-D MFCCs and their first derivative computed every 10ms.

4.2. Results and Discussions

Kernel Function	CD-AFCPM Supervector	PD-AFCPM Supervector
$K(\overrightarrow{A}_c, \overrightarrow{A}_s) = \langle \overrightarrow{A}_c, \overrightarrow{A}_s \rangle$	26.12%	28.63%
$K(\overrightarrow{A}_c, \overrightarrow{A}_s) = \left\langle \frac{\sqrt{\overrightarrow{w}_b} .* \overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\sqrt{\overrightarrow{w}_b} .* \overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle$	24.14%	27.14%

Table 2. The EERs of AF kernel-based speaker verification systems using PD-AFCPM and CD-AFCPM supervectors without feature selection.

The results of using AF-kernels (without feature selection) for computing the verification scores are shown in Table 2. Evidently, normalization helps reduce the EER significantly. Similar to the results in LR-based scoring approach, CD-AFCPM is superior to PD-AFCPM under the AF-kernel framework. Comparing the result in [6] (EER $= 23.46\%$) and the 2nd row of Table 2 (EER $= 24.14\%$) suggests that without feature selection, the AF-kernel is inferior to the conventional LR-based scoring.

Kernel Function	No Feature Selection	Feature Selection
$K(\overrightarrow{A}_c, \overrightarrow{A}_s) = \left\langle \frac{\sqrt{\overrightarrow{w}_b} .* \overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\sqrt{\overrightarrow{w}_b} .* \overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle$	24.14%	23.90%

Table 3. EER achieved by the AF kernel-based speaker verification system using CD-AFCPM supervectors with and without feature selection.

We conjecture that the inferiority is caused by the irrelevant features in the supervectors. To verify this conjecture, we performed the same experiment but with the irrelevant features removed by SVM-RFE [11]. The results are shown in Table 3. Although feature selection can reduce the EER from 24.14% to 23.90%, it is still higher than that achieved by the LR-based method, which is 23.46%. From the experiment we can see that the EER can be reduced by reducing the dimensionality of the supervectors. To further reduce the EER, we fused the LR-based method and the AF-kernel scoring method. The EER is reduced to 23.02%, as shown in the DET plots in Fig. 6. The p-values [12] between the EERs of the fusion and non-fusion cases are all smaller than 0.00001, suggesting that the differences in EERs are statistically significant. This result illustrates that the AF-kernel based method and LR-based method are complementary to each other.

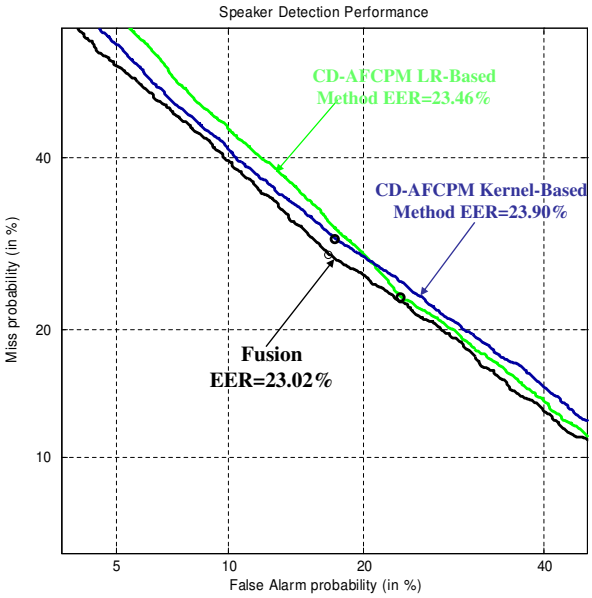

Fig. 6. DET produced by the AF-kernel scoring method with and without feature selection. CD-AFCPM supervectors were used in the AF-kernel.

5. 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] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proc. Eurospeech*, 2003, pp. 2665–2668.

[4] D. Klusacek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," in *Proc. ICASSP'03*, 2003, vol. 4, pp. 804–807.

[5] K. Y. Leung, M. W. Mak, and S. Y. Kung, "Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification," *Speech Communication*, vol. 48, no. 1, pp. 71–84, 2006.

[6] S. X. Zhang, M. W. Mak, and Helen H. Meng, "Speaker verification via high-level feature based phonetic-class pronunciation modeling," *IEEE Trans. on Computers*, vol. 56, no. 9, pp. 1189–1198, 2007.

[7] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006, May.

[9] V. Wan and S. Renals, "SVMSVM: Support vector machine speaker verification methodology," in *Proc. ICASSP'03*, 2003, vol. II, pp. 221–224.

[10] S.X. Zhang and M. W. Mak, "A new adaptation method for speaker-model creation in high-level speaker verification," in *Advances in Multimedia Information Processing (PCM'2007)*, Hong Kong, Springer LNCS 4810, 2007, pp. 325–335.

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[12] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP'89*, 1989, pp. 532–535.