# High-Level Speaker Verification via Articulatory-Feature based Sequence Kernels and SVM

Shi-Xiong Zhang and Man-Wai Mak

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

## Abstract

Articulatory-feature based pronunciation models (AFCPMs) are capable of capturing the pronunciation variations among different speakers and are good for high-level speaker recognition. However, the likelihood-ratio scoring method of AFPCMs is based on a decision boundary created by training the target speaker model and universal background model (UBM) separately. Therefore, the method does not fully utilize the discriminative information available in the training data. To fully harness the discriminative information, this paper proposes training a support vector machine (SVM) for computing the verification scores. More precisely, the models of target speakers, individual background speakers, and claimants are converted to AF-supervectors, which form the inputs to an AF-based kernel of the SVM for computing verification scores. Results show that the proposed AF-kernel scoring is complementary to likelihood-ratio scoring, leading to better performance when the two scoring methods are combined. Further performance enhancement was also observed when the AF scores were combined with acoustic scores derived from a GMM-UBM system.

## 1. Introduction

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 [1–5]. However, in most systems (e.g., GMM-UBM [6] and CD-AFCPM [5]), 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 a support vector machine (SVM) through a kernel function of the sequence of symbols.

This paper derives an articulatory-feature based sequence kernel and apply it to high-level speaker verification. For each target speaker, the observation sequences (AF labels) derived from his/her utterances are used to train a phonetic-class dependent articulatory feature-based pronunciation model (CD-AFCPM) [5]. These models are then converted to fixed-dimension AF supervectors for training a speaker-dependent SVM to discriminate the target speaker from background speakers in the AF-supervector space. To enhance the discrimination, a kernel that computes the similarity between the target speaker's supervector and the claimant's supervector is derived for the SVM. During verification, the AF labels derived from the speech of a claimant are used to build a CD-AFCPM of the claimant, which together with the target speaker model form the inputs to the speaker-dependent SVM to compute the verification scores. Because the kernel depends on the AF models of both the target speaker and the background speakers, we refer to it as 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. Phonetic-Class Dependent AFCPM

### 2.1. Articulatory-Feature Based Supervectors

Articulatory features (AFs) are representations describing the movements or positions of different articulators during speech production. Typically, the manner and place of articulation are used for pronunciation modeling. Manner has 6 classes: $\mathcal{M}$ ={Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral}, and place has 10 classes: $\mathcal{P}$ ={Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal}. AFs can be automatically determined from speech signals using AF-based multilayer perceptrons (MLPs) [4]. More specifically, given a sequence of acoustic vectors (MFCCs) $\mathbf{x}_t$ where $t = 1, \ldots, T$, the MLPs produce a sequence of manner labels $l_t^{\mathrm{M}} \in \mathcal{M}$ and a sequence of place labels $l_t^{\mathrm{P}} \in \mathcal{P}$ (see Fig. 1).

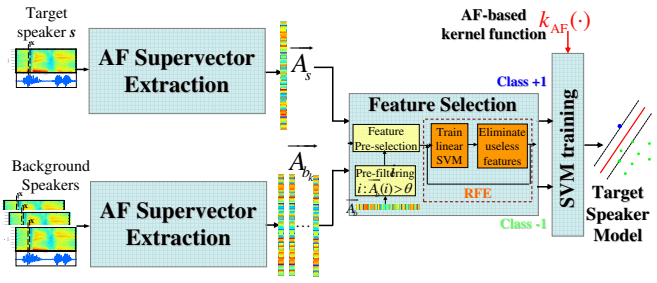


Figure 1: The training procedure of the AF kernel-based high-level speaker verification system.

The characteristics of background speakers are represented by $G$ (= 12 in this work) CD-AFCPMs. Each model comprises

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

the joint probabilities of manner $m \in \mathcal{M}$ and place $p \in \mathcal{P}$ conditioned on a phonetic class $k$:

$$P_b^{\text{CD}}(m,p|k) \qquad k = 1,\ldots,G$$
$$= \frac{\#(m,p|k) \text{ in background speakers}}{\sum_{m' \in \mathcal{M}, p' \in \mathcal{P}} \#(m',p'|k) \text{ in background speakers}} \quad (1)$$

where $\#(m,p|k)$ denotes the number of times the combination $(m,p)$ appears in phonetic class $k$.[1] A collection of $G$ background CD-AFCPMs is referred to as a universal background model (UBM).

Given the utterance of a target speaker $s$, $G$ speaker-dependent CD-AFCPMs can be obtained by

$$\widehat{P}_s^{\text{CD}}(m,p|k) = \beta_k P_s^{\text{CD}}(m,p|k) + (1-\beta_k) P_b^{\text{CD}}(m,p|k), \quad (2)$$

where $k = 1,\ldots,G$, $P_s^{\text{CD}}(m,p|k)$ is a model obtained from the target speaker utterance, and $\beta_k \in [0,1]$ controls the contribution of the speaker utterance and the background model on the target speaker model [5]. A collection of $G$ target-speaker dependent CD-AFCPMs is referred to as a target-speaker model. The elements of $G$ CD-AFCPMs $\{\widehat{P}_s^{\text{CD}}(m,p|k),\ k = 1,\ldots,G\}$ of a target speaker are concatenated to form a $60G$-dim supervector $\overrightarrow{A}_s$, namely CD-AFCPM supervector.

### 2.2. AF-Based Likelihood-Ratio Scoring

Denote a test utterance from a claimant as $X_1^T = \{X_1,\ldots,X_t,\ldots,X_T\}$, where $X_t$ contains 9 frames of MFCCs centered on frame $t$ of the utterance. Also denote $\widehat{P}_s^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|k)$ as the output of the $k$-th CD-AFCPM of the target speaker given that $X_t$ belongs to the $k$-th phonetic class, where $l_t^{\text{M}} \in \mathcal{M}$ and $l_t^{\text{P}} \in \mathcal{P}$ are the labels determined by the manner and place MLPs, respectively.[2]

The log likelihood-ratio (LR) score can be expressed as:

$$S_{\text{LR}}(X_1^T) = \sum_{k=1}^{G} \left( \frac{1}{T} \sum_{t: f^G(q_t)=k} \left( \log \frac{\widehat{P}_s^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|k)}{P_b^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|k)} \right) \right) \quad (3)$$

where $f^G(q_t)$ is a function that maps phoneme $q_t$ to phonetic class $k$ [5] and $q_t$ is determined by a null-grammar phoneme recognizer. Grouping frames according to $\mathcal{M}$ and $\mathcal{P}$, we have

$$S_{\text{LR}}(X_1^T) = \sum_{k=1}^{G} \frac{1}{T} \sum_{\substack{m \in \mathcal{M} \\ p \in \mathcal{P}}} \sum_{t: \begin{cases} f^G(q_t)=k, \\ l_t^{\text{M}}=m, l_t^{\text{P}}=p \end{cases}} \log \frac{\widehat{P}_s^{\text{CD}}(l_t^{\text{M}}=m, l_t^{\text{P}}=p|k)}{P_b^{\text{CD}}(l_t^{\text{M}}=m, l_t^{\text{P}}=p|k)}$$

$$= \sum_{k=1}^{G} \frac{T_k}{T} \left\{ \frac{1}{T_k} \sum_{i=1}^{60} \left( \left( \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_i|k)}{P_b^{\text{CD}}(\mathcal{L}_i|k)} \right) \sum_{t: \begin{cases} f^G(q_t)=k \\ \mathcal{L}_i \end{cases}} 1 \right) \right\}$$

$$= \sum_{k=1}^{G} \frac{T_k}{T} \left\{ \sum_{i=1}^{60} \left( \left( \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_i|k)}{P_b^{\text{CD}}(\mathcal{L}_i|k)} \right) \frac{N_{i,k}}{T_k} \right) \right\} \quad (4)$$

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

$$\frac{N_{i,k}}{T_k} = \frac{\#(\mathcal{L}_i|k) \text{ in the claimant}}{\sum_j \#(\mathcal{L}_j|k) \text{ in the claimant}} = P_c^{\text{CD}}(\mathcal{L}_i|k) \quad (5)$$

[1] We can see that for each phonetic class, there are $6 \times 10 = 60$ probabilities in the model.

[2] A similar notation is also applied to $\widehat{p}_b^{\text{CD}}(l_t^{\text{M}}, l_t^{\text{P}}|k)$.

where $P_c^{\text{CD}}(\mathcal{L}_i|k)$ is a claimant model and the index $i$ corresponds to the $i$-th combination of the manner and place class $(m,p)$. Substituting Eq. 5 into Eq. 4, we have

$$S_{\text{LR}}(X_1^T) = \sum_{k=1}^{G} \frac{T_k}{T} \left\{ \sum_{i=1}^{60} \left( \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_i|k)}{P_b^{\text{CD}}(\mathcal{L}_i|k)} \right) P_c^{\text{CD}}(\mathcal{L}_i|k) \right\}$$

$$= \sum_{k=1}^{G} \left\langle \begin{bmatrix} \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_1|k)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_1|k)} \\ \cdots \\ \log \frac{\widehat{P}_s^{\text{CD}}(\mathcal{L}_{60}|k)}{\widehat{P}_b^{\text{CD}}(\mathcal{L}_{60}|k)} \end{bmatrix}_{60}, \begin{bmatrix} \frac{T_k}{T} P_c^{\text{CD}}(\mathcal{L}_1|k) \\ \cdots \\ \frac{T_k}{T} P_c^{\text{CD}}(\mathcal{L}_{60}|k) \end{bmatrix}_{60} \right\rangle \quad (6)$$

$$= \left\langle \log \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b}, \overrightarrow{w} .* \overrightarrow{A}_c \right\rangle = \left\langle \overrightarrow{A}'_c, \log \overrightarrow{A}_s \right\rangle - \left\langle \overrightarrow{A}'_c, \log \overrightarrow{A}_b \right\rangle$$

where $\overrightarrow{w} = [T_1/T,\ldots,T_1/T,\ldots,T_G/T,\ldots,T_G/T]^{\text{T}}$; $\overrightarrow{A}_s$, $\overrightarrow{A}_b$ and $\overrightarrow{A}_c$ stand for the AF supervector of the speaker, background, and claimant, respectively; $\log \frac{\overrightarrow{X}}{\overrightarrow{Y}} \equiv \left[\log \frac{x_1}{y_1},\ldots,\log \frac{x_N}{y_N}\right]^{\text{T}}$; and $\overrightarrow{X} .* \overrightarrow{Y} \equiv [x_1 y_1,\ldots,x_N y_N]^{\text{T}}$, where $x_i$ and $y_i$ are elements of $\overrightarrow{X}$ and $\overrightarrow{Y}$, respectively. Eq. 6 suggests that the LR score can be obtained by computing a dot product. Fig. 2 illustrates the implementation of LR scoring.

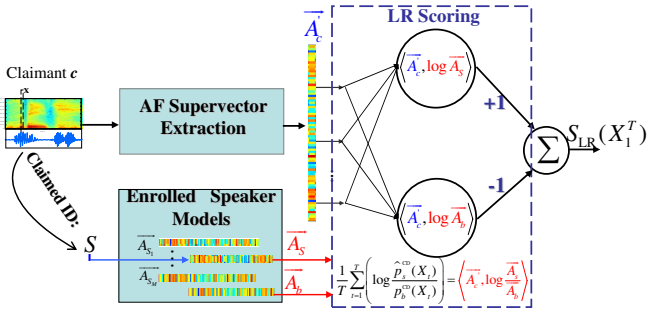


Figure 2: A dot-product implementation of the traditional log-likelihood scoring in CD-AFCPM speaker verification.

## 3. Articulatory Feature-Based Kernels

Fig. 2 suggests a possible improvement of LR scoring: Replacing the fixed multiplication factors '+1' and '−1' by weights that are optimally determined by SVM training. This strategy, however, requires the function inside the 'circle' in Fig. 2 to satisfy the Mercer's condition [7]. Unfortunately, the function $f(\overrightarrow{X}, \overrightarrow{Y}) = \left\langle \overrightarrow{X}, \log \overrightarrow{Y} \right\rangle$ does not satisfy the Mercer's condition because it cannot be written as $\left\langle \Phi(\overrightarrow{X}), \Phi(\overrightarrow{Y}) \right\rangle$. We propose 3 approaches to remedying this problem.

### 3.1. Euclidean AF-Kernel

The simplest type of Mercer AF-kernel is a linear kernel:

$$K_{\text{AF-E}}(\overrightarrow{A}_c, \overrightarrow{A}_s) = \left\langle \overrightarrow{A}_c, \overrightarrow{A}_s \right\rangle. \quad (7)$$

Essentially, this kernel can be derived from the Euclidean distance between projected vectors in the feature space [7]; therefore we refer to it as Euclidean AF-Kernel.

### 3.2. Mahalanobis AF-Kernel

Kernel can also be derived using Mahalanobis distance:

$$d_{\text{M}}(\overrightarrow{A}_c, \overrightarrow{A}_s) = \sqrt{(\overrightarrow{A}_c - \overrightarrow{A}_s)^{\text{T}} \boldsymbol{\Sigma}^{-1} (\overrightarrow{A}_c - \overrightarrow{A}_s)}$$
$$= \sqrt{K_{\text{AF-M}}(\overrightarrow{A}_c, \overrightarrow{A}_c) - 2K_{\text{AF-M}}(\overrightarrow{A}_c, \overrightarrow{A}_s) + K_{\text{AF-M}}(\overrightarrow{A}_s, \overrightarrow{A}_s)},$$

where

$$\boldsymbol{\Sigma} = \frac{1}{M}\sum_{i=1}^{M} \overrightarrow{A}_{b_i} \overrightarrow{A}_{b_i}^{\text{T}} - \left(\frac{1}{M}\sum_{i=1}^{M} \overrightarrow{A}_{b_i}\right)\left(\frac{1}{M}\sum_{i=1}^{M} \overrightarrow{A}_{b_i}\right)^{\text{T}} \tag{8}$$

is a covariance matrix computed from background models and

$$K_{\text{AF-M}}(\overrightarrow{A}_c, \overrightarrow{A}_s) = \left\langle \boldsymbol{\Sigma}^{-\frac{1}{2}} \overrightarrow{A}_c, \boldsymbol{\Sigma}^{-\frac{1}{2}} \overrightarrow{A}_s \right\rangle \tag{9}$$

is a kernel function. Comparing with Eq. 7, the dimensions of the supervectors are now normalized by the variances of the background models. Note also that this kernel is similar to the GMM-supervector kernel [8]. If we discard the substraction of the means in Eq. 8, we will obtain the GLDS kernel [9].

### 3.3. Likelihood-Ratio AF-Kernel

The above two kernels are derived from distance metric. Kernels can also be derived from similarity metric such as likelihood ratio. To this end, we ensure that Eq. 6 can satisfy the Mercer condition by the following approximation:

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

The approximation is valid because the speaker models are adapted from the UBM $\overrightarrow{A}_b$ and therefore $\frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \rightarrow \overrightarrow{1}$. Dropping the constant in Eq. 10 that does not affect verification decisions, we define a likelihood-ratio (LR) based AF-kernel (because this kernel is derived from LR scoring, we refer to it as LR AF-kernel):

$$K_{\text{AF-LR}}\left(\overrightarrow{A}_c, \overrightarrow{A}_s\right) \equiv \left\langle \overrightarrow{A}'_c, \frac{\overrightarrow{A}_s}{\overrightarrow{A}_b} \right\rangle = \left\langle \frac{\overrightarrow{A}'_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle$$
$$= \left\langle \frac{\overrightarrow{w}_b . * \overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle \approx \left\langle \frac{\sqrt{\overrightarrow{w}_b} . * \overrightarrow{A}_c}{\sqrt{\overrightarrow{A}_b}}, \frac{\sqrt{\overrightarrow{w}_b} . * \overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}} \right\rangle$$

where $\overrightarrow{w}_b = \left[\overbrace{\frac{T_1^b}{T}, \cdots, \frac{T_1^b}{T}}^{60}, \overbrace{\frac{T_2^b}{T}, \cdots, \frac{T_2^b}{T}}^{60}, \cdots, \overbrace{\frac{T_G^b}{T}, \cdots, \frac{T_G^b}{T}}^{60}\right]^{\text{T}}_{60G}$ contains the phonetic-class weights obtained from the background speakers, $T_k^b$ is the number of times phonetic class $k$ appears in the utterances of background speakers, and $\sqrt{\overrightarrow{X}} \equiv [\sqrt{x_1}, \ldots, \sqrt{x_N}]^{\text{T}}$. The approximation aims to make the similarity measure symmetric. Fig. 3 shows the scoring procedure during the verification phase. Figs. 4(a) and 4(b) show the un-normalized supervectors $\overrightarrow{A}_s$ and the normalized supervectors $\frac{\sqrt{\overrightarrow{w}_b} . * \overrightarrow{A}_s}{\sqrt{\overrightarrow{A}_b}}$ for 150 speakers, respectively. For

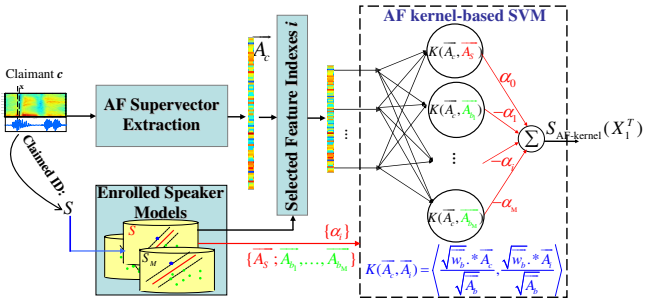


Figure 3: The verification phase of an AF-kernel based speaker verification system.

clarity, only 120 features 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 affect 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. 4(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.

### 3.4. Comparing AF-Kernel Scoring and LR-scoring

The SVM output can be considered as a scoring function:

$$S_{\text{AF-kernel}}(X_1^T) = \alpha_0 K_{\text{AF}}\left(\overrightarrow{A}_c, \overrightarrow{A}_s\right) - \sum_{i=1}^{M} \alpha_i K_{\text{AF}}\left(\overrightarrow{A}_c, \overrightarrow{A}_{b_i}\right), \tag{11}$$

where $K_{\text{AF}}$ is any of the three AF-kernels mentioned earlier, $\alpha_0$ is the Lagrange multiplier corresponding to the target speaker, and $\alpha_i$ $(i = 1, \ldots, M)$ are Lagrange multipliers (some of them may be zero) corresponding to the background speakers. Comparing Eqs. 6 and 11 and comparing Figs. 2 and 3 suggest that AF-kernel scoring is more general and is potentially better than LR scoring (Eq. 6) in two aspects. First, the SVM optimally selects the most appropriate background speakers through the non-zero $\alpha_i$. Second, instead of using a single background model that contains the average characteristics of all background speakers, a specific set of background speakers is used for each target speaker for scoring. This is to some extends analogous to cohort scoring. However, the cohort set is now discriminatively and optimally determined by SVM training, and the contribution of the selected background models is also optimally weighted through the Lagrange multipliers $\alpha_i$.

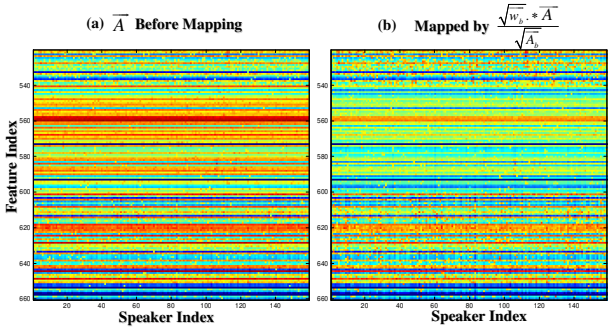


Figure 4: The effect of the normalization term $\sqrt{\overrightarrow{A}_b}$ and the weighting term $\vec{w}_b$ on the AF supervectors.

## 4. Experiments and Results

**Datasets**. 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.

**Feature Selection**. We applied SVM-RFE [10] to select 600 features from 720 features in the AF supervectors and found that the EER can be reduced from 24.14% to 23.87%. Because of this encouraging result, feature selection was applied to all experiments.

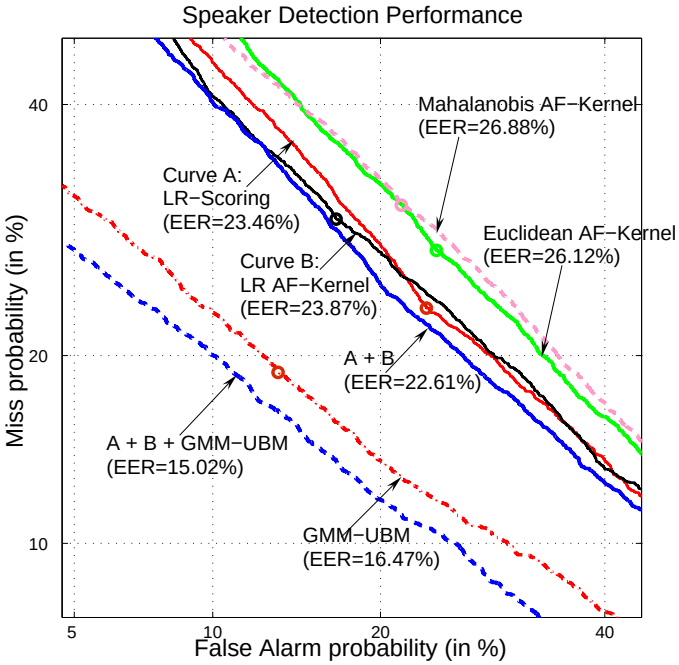


Figure 5: DET produced by LR scoring, AF-kernel scoring, acoustic GMM-UBM, and their fusion.

**EER and DET Performance**. Fig. 5 shows the performance of likelihood-ratio (LR) scoring, kernel-based scoring, and their fusion with an MFCC-based GMM-UBM system. Results show that the Euclidean kernel performs slightly better than the Mahalanobis kernel. This may be attributed to the inaccurate covariance matrix. Unlike the MFCCs in GMM-supervectors, there are significant correlation among the features in the AF supervectors; therefore, a full covariance matrix should be used. However, this will demand extensive amount of training data to estimate the matrix accurately. Insufficient training data could lead to singular matrix. We solved this problem by assuming diagonal covariance, but the assumption is too crude for articulatory features.

The results also show that scoring based on the LR AF-kernel $K_{\text{AF-LR}}$ (Curve B) outperforms LR scoring (Curve A) at the low false-alarm region, whereas the situation is reverse at the low miss-probability region. This suggests that the two scoring methods are complementary to each other, which is evident by the superior performance (Curve A+B) when the scores resulting from the two scoring methods are fused.

At the low-miss probability region, LR AF-kernel scoring is only slightly worse than LR scoring, but it is significantly better than LR scoring in the low false alarm region. This suggests that LR AF-kernel scoring is generally better than LR scoring, which is mainly attributed to the explicitly use of discriminative information in the kernel function of the SVM and to the optimal selection of background speakers by SVM training. Although LR scoring also considers the impostor information, it can only implicity use this information through the UBM. In AF-kernel scoring, on the other hand, the SVM of each target speaker is discriminatively trained to differentiate the target speaker from all of the background speakers. The SVM effectively provides an optimal set of weights for this differentiation. On the other hand, in log-likelihood scoring, all target speakers share the same background model and the weight is always equal ($= -1$) across all target speakers. This explains the superiority of the AF-kernel scoring approach.

Interestingly, LR AF-kernel scoring outperforms Euclidean AF-kernel and Mahalanobis AF-kernel scoring. This suggests that normalizing the features of AF-supervectors by the background models can prevent some features (with large numerical values) from dominating the SVM scoring.

Among the four scoring methods, LR scoring is the fastest and the Mahalanobis kernel is the slowest, 0.11sec vs. 0.65sec per utterance.

## 5. Conclusions

An AF-based kernel scoring method that explicitly uses the discriminative information available in the training data was proposed. Experimental results on NIST2000 suggests that the method is superior to the conventional likelihood ratio scoring method and that the method is readily fusible with low-level acoustic systems.

## 6. References


[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambrige, 2004.

[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] W.M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP'02*, 2002, vol. 1, pp. 161–164.

[10] 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.