REGRESSION OPTIMIZED KERNEL FOR HIGH-LEVEL SPEAKER VERIFICATION

Shi-Xiong Zhang and Man-Wai Mak

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
SX.Zhang@inet.polyu.edu.hk, enmwmak@polyu.edu.hk

ABSTRACT

Computing the likelihood-ratio (LR) score of a test utterance is an important step in speaker verification. It has recently been shown that for discrete speaker models, the LR scores can be expressed as dot products between supervectors formed by the test utterance, target-speaker model, and background model. This paper leverages this dot-product formulation and the representer theorem to derive a general kernel, namely the regression optimized kernel, for computing utterance-based verification scores using support vector machines. The kernel is general in that it can be a linear combination of any kernels belonging to the reproduction kernel Hilbert space. The combination weights are obtained by maximizing the ability of a discriminant function in separating a target speaker from impostors. The regression optimized kernel was applied to high-level speaker verification using articulatory-feature based pronunciation models. Results show that the scores produced by the regression optimized kernel are not only superior but also complementary to the LR scores, resulting in better performance when the two types of scores are combined. The proposed regression optimized kernel can be easily applied to other SVM-based classification problems.

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

1. INTRODUCTION

Most speaker verification systems (e.g., acoustic-based GMM-UBM [1] and phonetic-based AFCPM [2]) compute the likelihood ratio (LR) scores of claimant's utterances by accumulating the frame-based LR scores. This frame-based scoring scheme has three drawbacks. First, treating speech frames independently may miss some important speaker information contained in the claimant's utterance. Bear in mind that the goal of speaker verification is to minimize classification errors on test utterances, not on speech frames. Second, consider every frame as equally important means that highly speaker-discriminative sounds will not receive more attention than less speaker-discriminative sounds. Third, for discrete generative models (commonly used in high-level systems), frame-based scoring is computationally inefficient because the same probability values will be repeatedly used many times during the score accumulation process.

This paper attempts to overcome these drawbacks by proposing a sequence-based scoring approach in which an utterance is considered as comprising a sequence of symbols and the utterance-based score is obtained from a support vector machine (SVM) through a specially designed kernel function called regression optimized kernel. The method extracts the articulatory feature (AF) supervectors from each target speaker to train a speaker-dependent SVM to discriminate the target speaker from background speakers in the AF-supervector space.

The proposed kernel is different from many other kernels (such as the generalized linear discriminant sequence (GLDS) kernel [3], n-gram kernel [4], GMM-supervector kernel [5], and Fisher kernel [6]) in that no specific form of the discriminant or scoring function for the similarity measure is assumed. In fact, any functions in the reproducing kernel Hilbert space are potential candidates. We show that the optimal discriminant function can be obtained by solving a functional optimization problem using the representer theorem [7], leading to a kernel that is a general form of several existing kernels.

The remainder of the paper derives the regression optimized kernel and provides the theoretical and experiment evidences to demonstrate that kernel-based scoring is superior to frame-based scoring. Experimental results on the NIST2000 SRE are presented.

2. DOT-PRODUCT FORMULATION OF LIKELIHOOD-RATIO

In high-level speaker verification, it is common to use sequences of labels (e.g., phone labels in [4] and AF labels in [2]) extracted from the utterances of a target speaker to train a discrete model for that target speaker. Without loss of generality, assume that the discrete model has two random variables $\{L_x, L_y\}$ and that the sample space is defined by $\{\mathcal{L}_x, \mathcal{L}_y\}$. Then, the model can be expressed as a probability mass function $f_{L_x,L_y}(l_x, l_y) = \Pr(L_x = l_x, L_y = l_y)$ where $l_x \in \mathcal{L}_x$ and $l_y \in \mathcal{L}_y$. For the AFCPMs in [2], $\mathcal{L}_x$ and $\mathcal{L}_y$ are the manner and place classes, respectively. Given a model, a supervector $\overrightarrow{A}$ can be obtained by stacking all of the probability entries in the model.

Assume that given a claimant's utterance, a sequence of 2-tuple labels $\ell_1^T = \{l_{x,t}, l_{y,t}\}_{t=1}^T$ is obtained, where $l_{x,t} \in \mathcal{L}_x$ and $l_{y,t} \in \mathcal{L}_y$. The likelihood ratio (LR) score of the utterance can be obtained by accumulating the frame-based likelihood ratio:

$$S_{\text{LR}}(\ell_1^T) = \frac{1}{T}\sum_{t=1}^{T}\left(\log\frac{\Pr(L_x = l_{x,t}, L_y = l_{y,t}|\text{Speaker } s)}{\Pr(L_x = l_{x,t}, L_y = l_{y,t}|\text{Background})}\right).$$

In [8], we have shown that the likelihood ratio can be expressed in terms of dot products as follows:

$$S_{\text{LR}}(\ell_1^T) = \left\langle \log\frac{\overrightarrow{A_s}}{\overrightarrow{A_b}}, \overrightarrow{A_c}\right\rangle = \left\langle \overrightarrow{A_c}, \log\overrightarrow{A_s}\right\rangle - \left\langle \overrightarrow{A_c}, \log\overrightarrow{A_b}\right\rangle \quad (1)$$

where $\overrightarrow{A_s}$, $\overrightarrow{A_b}$ and $\overrightarrow{A_c}$ are the supervectors corresponding to the

This work was in part supported by Center for Multimedia Signal Processing, The Hong Polytechnic University (1-BB9W and A-PA6F).

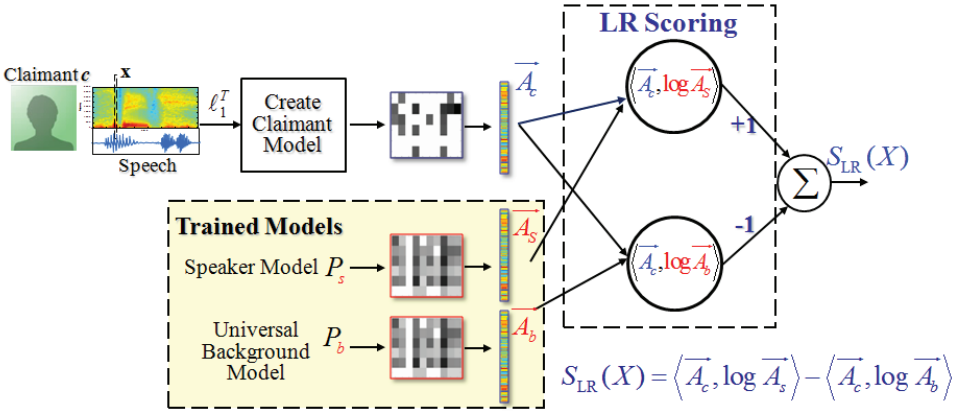

Fig. 1. The dot-product implementation of LR scoring in AFCPM speaker verification.

speaker, background, and claimant, respectively.[1] Fig. 1 illustrates the dot-product implementation of LR scoring.

3. REGRESSION OPTIMIZED KERNEL

Fig. 1 suggests a possible improvement of LR scoring: Replacing the fixed multiplication factors '+1' and '−1' by weights that are optimally determined by SVM training. The training procedure is shown in Fig. 2. In order to make sure that the SVM training algorithm converges to a stable solution, the function inside the circle in Fig. 1 should satisfy the Mercer condition [9]. Unfortunately, $f(\vec{X}, \vec{Y}) = \left\langle \vec{X}, \log \vec{Y} \right\rangle$ does not satisfy the Mercer condition because $\left\langle \vec{X}, \log \vec{Y} \right\rangle$ cannot be written as $\left\langle \Phi(\vec{X}), \Phi(\vec{Y}) \right\rangle$. Here, we propose a new general kernel to remedy this problem.

3.1. Optimal Discriminant Function

To derive a kernel, a similarity metric needs to be defined. If Mahalanobis distance is used, GMM-supervector kernel [5] will be obtained. Using the likelihood ratio scores will lead to the n-gram kernel [4] and the LR AF-kernel [8]. Kernels can also be derived from optimizing a discriminant function that separates target speaker's speech from impostors' speech in a high-dimensional kernel-induced feature space. The GLDS kernel is a typical example. A common characteristic of these kernels is that they are derived under the assumption that the discriminant functions or scoring functions have a specific form to measure the similarity. For kernels derived from discriminant functions, this constraint can be relaxed by using a general discriminant function $f_s(\vec{A})$. Therefore, given two set of training data $\left\{ \vec{A}_s, y_s = +1 \right\}$ and $\left\{ \vec{A}_{b_k}, y_{b_k} = 0 \right\}_{k=1}^{M}$, our goal is to find the best discriminant function $f_s(\vec{A})$ that solve the optimization problem:

$$\hat{f}_s = \arg \min_{f_s \in \mathcal{R}_K} \left\{ \sum_{i \in \{s, b_k\}_{k=1}^{M}} \gamma_i \left(f_s(\vec{A}_i) - y_i \right)^2 + \lambda \parallel f_s \parallel^2 \right\}, \tag{2}$$

[1] In Eq. 1, $\log \frac{\vec{Z}}{\vec{Y}} \equiv \left[\log \frac{z_1}{y_1}, \ldots, \log \frac{z_N}{y_N} \right]^{\mathrm{T}}$, where z_i and y_i are elements of $\vec{Z}$ and $\vec{Y}$, respectively.

where s and b_k denote the target speaker s and background speaker k, respectively, M is the number of background speakers, $\lambda > 0$ is a regularizing parameter, γ_i is used to alleviate the unbalance between the two classes of data, and $\mathcal{R}_K$ represents the reproducing kernel Hilbert space (RKHS) [9].

According to the representer theorem [7], if $f_s \in \mathcal{R}_K$, the solution to the functional optimization problem in Eq. 2 has the form:

$$\hat{f}_s(\vec{A}) = \sum_{i \in \{s, b_k\}_{k=1}^{M}} w_i^s k(\vec{A}, \vec{A}_i), \tag{3}$$

where w_i^s are speaker-dependent weights to be optimized and $k(\cdot, \cdot) : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$ is a kernel in $\mathcal{R}_K$ such that

$$\left\langle f_s, k(\vec{A}, \cdot) \right\rangle_{\mathcal{R}_K} = f_s(\vec{A}). \tag{4}$$

Eq. 3 and Eq. 4 suggest that

$$\begin{aligned} \parallel \hat{f}_s \parallel^2 &= \langle \hat{f}_s, \hat{f}_s \rangle = \left\langle \hat{f}_s, \sum_{i \in \{s, b_k\}_{k=1}^{M}} w_i^s k(\vec{A}_i, \cdot) \right\rangle \\ &= \sum_{i \in \{s, b_k\}_{k=1}^{M}} w_i^s \left\langle \hat{f}_s, k(\vec{A}_i, \cdot) \right\rangle = \sum_{i \in \{s, b_k\}_{k=1}^{M}} w_i^s \hat{f}_s(\vec{A}_i) \\ &= \sum_{i \in \{s, b_k\}_{k=1}^{M}} w_i^s \left(\sum_{j \in \{s, b_k\}_{k=1}^{M}} w_j^s k(\vec{A}_i, \vec{A}_j) \right). \end{aligned} \tag{5}$$

Therefore, the optimization problem in Eq. 2 can be formulated as:

$$\min_{\mathbf{w}_s \in \mathbb{R}^{M+1}} \left\{ (\mathbf{y} - \mathbf{K}_s \mathbf{w}_s)^{\mathrm{T}} \mathbf{\Gamma} (\mathbf{y} - \mathbf{K}_s \mathbf{w}_s) + \lambda \mathbf{w}_s{}^{\mathrm{T}} \mathbf{K}_s \mathbf{w}_s \right\} \tag{6}$$

where

$$\begin{aligned} &\mathbf{w}_s = [w_s^s, w_{b_1}^s, \ldots, w_{b_M}^s]^{\mathrm{T}}, \mathbf{y} = [1, 0, \ldots, 0]^{\mathrm{T}}_{(M+1)\times 1}, \\ &\mathbf{\Gamma} = diag\{\gamma_s, \gamma_{b_1}, \ldots, \gamma_{b_M}\} = diag\{\gamma^+, \gamma^-, \ldots, \gamma^-\}, \end{aligned} \tag{7}$$

and

$$\mathbf{K}_s = \begin{bmatrix} k_{s,s} & k_{b_1,s} & \cdots & k_{b_M,s} \\ k_{s,b_1} & k_{b_1,b_1} & \cdots & k_{b_M,b_1} \\ \vdots & \vdots & \ddots & \vdots \\ k_{s,b_M} & k_{b_1,b_M} & \cdots & k_{b_M,b_M} \end{bmatrix}, \tag{8}$$

where $k_{i,j} = k_{j,i} = k(\vec{A}_i, \vec{A}_j)$. Taking the derivative with respect to w_s in Eq. 6, the optimization solution of Eq. 6 is

$$\mathbf{w}_s = (\mathbf{K}_s \mathbf{\Gamma} \mathbf{K}_s^{\mathrm{T}} + \lambda \mathbf{K}_s^{\mathrm{T}})^{-1} (\mathbf{K}_s^{\mathrm{T}} \mathbf{\Gamma} \mathbf{y}). \tag{9}$$

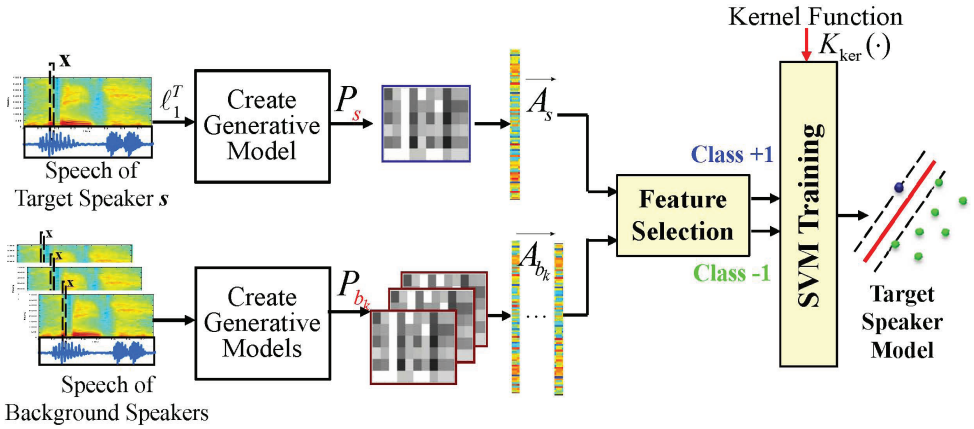

Fig. 2. The training procedure of an SVM-based speaker verification system. The procedure results in an SVM for each target speaker. See Section 3.1 for the derivation of the kernels. See Section 4.3 for the feature selection procedure.

3.2. Regression Optimized AF-Kernel

Using Eqs. 7–9, we can express the optimal discriminant function (Eq. 3) as:

$$\begin{aligned}\hat{f}_s(\vec{A}) &= \sum_{i\in\{s,b_k\}_{k=1}^{M}} w_i^s k(\vec{A},\vec{A}_i) \\ &= \left[(\mathbf{K}_s\boldsymbol{\Gamma}\mathbf{K}_s^{\mathrm{T}}+\boldsymbol{\lambda}\mathbf{K}_s^{\mathrm{T}})^{-1}(\mathbf{K}_s^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{y})\right]^{\mathrm{T}}_{(M+1)\times 1}\begin{bmatrix} k(\vec{A},\vec{A}_s)\\ k(\vec{A},\vec{A}_{b_1})\\ \vdots \\ k(\vec{A},\vec{A}_{b_M})\end{bmatrix} \\ &= \gamma^{+}\begin{bmatrix} k(\vec{A}_s,\vec{A}_s)\\ k(\vec{A}_s,\vec{A}_{b_1})\\ \vdots \\ k(\vec{A}_s,\vec{A}_{b_M})\end{bmatrix}^{\mathrm{T}}\left(\mathbf{K}_s\boldsymbol{\Gamma}\mathbf{K}_s^{\mathrm{T}}+\lambda\mathbf{K}_s^{\mathrm{T}}\right)^{-1}\begin{bmatrix} k(\vec{A},\vec{A}_s)\\ k(\vec{A},\vec{A}_{b_1})\\ \vdots \\ k(\vec{A},\vec{A}_{b_M})\end{bmatrix}.\end{aligned} \tag{10}$$

Note that because γ^{+} is a constant, it can be discarded without affecting the discriminative ability of $\hat{f}_s(\vec{A})$. Therefore, the the similarity between two supervectors $\vec{A}_c$ and $\vec{A}_s$ can be defined as:

$$\hat{f}_s(\vec{A}_c) = \begin{bmatrix} k(\vec{A}_c,\vec{A}_s)\\ k(\vec{A}_c,\vec{A}_{b_1})\\ \vdots \\ k(\vec{A}_c,\vec{A}_{b_M})\end{bmatrix}^{\mathrm{T}}\left(\mathbf{K}_s\boldsymbol{\Gamma}\mathbf{K}_s^{\mathrm{T}}+\lambda\mathbf{K}_s^{\mathrm{T}}\right)^{-1}\begin{bmatrix} k(\vec{A}_s,\vec{A}_s)\\ k(\vec{A}_s,\vec{A}_{b_1})\\ \vdots \\ k(\vec{A}_s,\vec{A}_{b_M})\end{bmatrix}. \tag{11}$$

Note that the matrix $\mathbf{K}_s$ and the vector $k(\vec{A},\cdot)|_{(s,b_1,\ldots,b_M)}$ are target speaker-dependent.[2] Because these matrices and vectors are dominated by nontarget speaker data, it is possible to make them speaker-independent by using the following approximations:

$$\mathbf{K}_s \approx \mathbf{K} = \begin{bmatrix} k_{b,b} & k_{b_1,b} & \cdots & k_{b_M,b}\\ k_{b,b_1} & k_{b_1,b_1} & \cdots & k_{b_M,b_1}\\ \vdots & \vdots & \ddots & \vdots \\ k_{b,b_M} & k_{b_1,b_M} & \cdots & k_{b_M,b_M}\end{bmatrix}, \tag{12}$$

and

$$k(\vec{A},\cdot)|_{(s,b_1,\ldots,b_M)} \approx k(\vec{A},\cdot)|_{(b,b_1,\ldots,b_M)},$$

where the universal background AF supervector $\vec{A}_b$ is used to approximate $\vec{A}_s$. Making these matrices speaker-independent have three advantages. First, substantial storage space can be saved because a speaker-independent matrix can be shared by all target speakers. Second, the matrix $\mathbf{K}$ can be pre-computed, saving computation significantly during recognition time. Third, this makes Eq. 11 symmetric between s and c.

Finally, the regression optimized kernel is written as:[3]

$$\begin{aligned}K_{\text{Reg}}(\vec{A}_c,\vec{A}_s) &= \langle\left(\mathbf{K}\boldsymbol{\Gamma}\mathbf{K}^{\mathrm{T}}+\lambda\mathbf{K}^{\mathrm{T}}\right)^{-\frac{1}{2}}k(\vec{A}_c,\cdot)|_{(b,b_1,\ldots,b_M)}, \\ &\qquad \left(\mathbf{K}\boldsymbol{\Gamma}\mathbf{K}^{\mathrm{T}}+\lambda\mathbf{K}^{\mathrm{T}}\right)^{-\frac{1}{2}}k(\vec{A}_s,\cdot)|_{(b,b_1,\ldots,b_M)}\rangle \\ &= \left\langle\varphi(\vec{A}_c),\varphi(\vec{A}_s)\right\rangle\end{aligned} \tag{13}$$

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

$$\varphi(\vec{A}) = \left(\mathbf{K}\boldsymbol{\Gamma}\mathbf{K}^{\mathrm{T}}+\lambda\mathbf{K}^{\mathrm{T}}\right)^{-\frac{1}{2}}k(\vec{A},\cdot)|_{(b,b_1,\ldots,b_M)}. \tag{14}$$

Note that $\left(\mathbf{K}\boldsymbol{\Gamma}\mathbf{K}^{\mathrm{T}}+\lambda\mathbf{K}^{\mathrm{T}}\right)^{-\frac{1}{2}}$ can be considered as a normalization matrix pre-computed from the background population.

According to the representer theorem, $k(\vec{A}_i,\vec{A}_j)$ should belong to $\mathcal{R}_K$. One possibility is to use the linearized LR AF-kernel [8], i.e.,

$$k(\vec{A}_i,\vec{A}_j) = K_{\text{LR}}(\vec{A}_i,\vec{A}_j) = \left\langle\frac{\sqrt{\vec{w}_b}.*\vec{A}_i}{\sqrt{\vec{A}_b}},\frac{\sqrt{\vec{w}_b}.*\vec{A}_j}{\sqrt{\vec{A}_b}}\right\rangle. \tag{15}$$

where $\vec{w}_b \in \mathbb{R}^D$ contains the phonetic-class weights obtained from the background speakers [8] and $\frac{\sqrt{\vec{X}}.*\vec{Y}}{\sqrt{\vec{Z}}}$ means element-wise multiplication and division.

3.3. Regression Optimized Kernel Vs. Other Kernels

The regression AF-kernel can be considered as a general form of the Euclidean, Mahalanobis (GMM-Supervector), GLDS, and LR AF-kernels [8]. This can be explained from the form of kernel functions. Starting from Eq. 13, if $\boldsymbol{\Gamma} = \mathbf{0}$ and $\lambda = 1$, then the (i,j)-th element

[2] $k(\vec{A},\cdot)|_{(s,b_1,\ldots,b_M)} \equiv \left[k(\vec{A},\vec{A}_s),k(\vec{A},\vec{A}_{b_1}),\cdots,k(\vec{A},\vec{A}_{b_M})\right]^{\mathrm{T}}$.

[3] Note that the regression optimized kernel (Eq. 13) is not limited to the articulatory features. It is also applicable to any SVM systems.

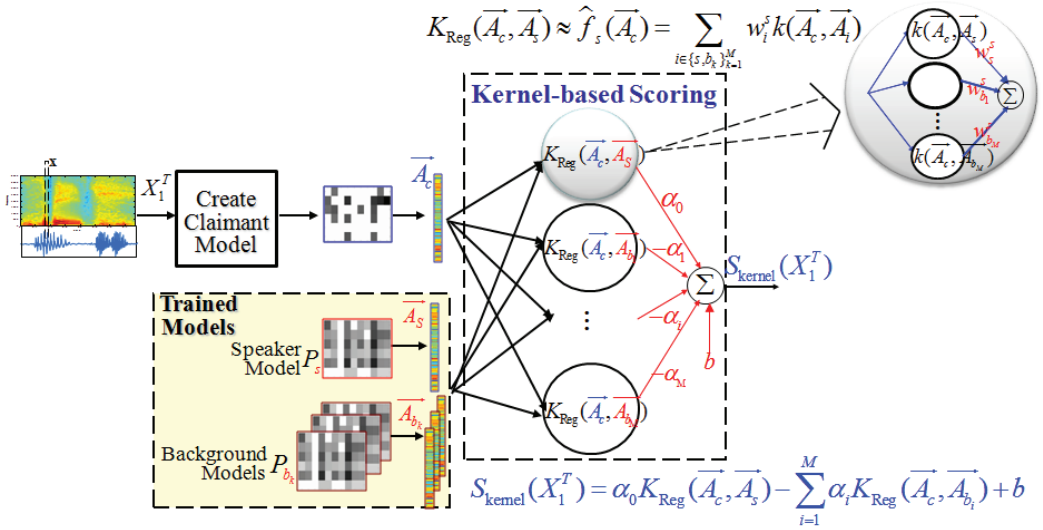

Fig. 3. The verification phase of an AF-kernel based speaker verification system. See Fig. 1 for a comparison.

of the regression optimized kernel matrix $\mathbf{K}_{\text{Reg}}$ becomes:

$$\begin{aligned} & K_{\text{Reg}}(\overrightarrow{A_i}, \overrightarrow{A_j}) \\ &= \left\langle \mathbf{K}^{-\frac{1}{2}} \begin{bmatrix} k(\overrightarrow{A_i}, \overrightarrow{A_b}) \\ k(\overrightarrow{A_i}, \overrightarrow{A_{b_1}}) \\ \vdots \\ k(\overrightarrow{A_i}, \overrightarrow{A_{b_M}}) \end{bmatrix}, \mathbf{K}^{-\frac{1}{2}} \begin{bmatrix} k(\overrightarrow{A_j}, \overrightarrow{A_b}) \\ k(\overrightarrow{A_j}, \overrightarrow{A_{b_1}}) \\ \vdots \\ k(\overrightarrow{A_j}, \overrightarrow{A_{b_M}}) \end{bmatrix} \right\rangle \\ &= \left\langle \hat{\varphi}(\overrightarrow{A_i}), \hat{\varphi}(\overrightarrow{A_j}) \right\rangle \end{aligned} \tag{16}$$

Define $\mathbf{\Omega} = \left[\hat{\varphi}(\overrightarrow{A_s}), \hat{\varphi}(\overrightarrow{A_{b_1}}), \ldots, \hat{\varphi}(\overrightarrow{A_{b_M}})\right]$. Then we have

$$\mathbf{\Omega} = \mathbf{K}^{-\frac{1}{2}} \begin{bmatrix} k_{s,b} & k_{b_1,b} & \cdots & k_{b_M,b} \\ k_{s,b_1} & k_{b_1,b_1} & \cdots & k_{b_M,b_1} \\ \vdots & \vdots & \ddots & \vdots \\ k_{s,b_M} & k_{b_1,b_M} & \cdots & k_{b_M,b_M} \end{bmatrix} \doteq \mathbf{K}^{-\frac{1}{2}} \mathbf{K}_s,$$

where $\mathbf{K}_s$ is defined in Eq. 8. Therefore, the regression optimized kernel matrix for target speaker s is:

$$\begin{aligned} \mathbf{K}^s_{\text{Reg}} &= \mathbf{\Omega}^{\text{T}}\mathbf{\Omega} = (\mathbf{K}^{-\frac{1}{2}}\mathbf{K}_s)^{\text{T}}(\mathbf{K}^{-\frac{1}{2}}\mathbf{K}_s) \\ &= \mathbf{K}_s^{\text{T}}\mathbf{K}^{-\frac{1}{2}}\mathbf{K}^{-\frac{1}{2}}\mathbf{K}_s \\ &\approx \mathbf{K}_s \quad \text{(because Eq. 12: } \mathbf{K} \approx \mathbf{K}_s\text{)}. \end{aligned} \tag{17}$$

Consider the elements of $\mathbf{K}_s$. If $k_{i,j} = k(\overrightarrow{A_i}, \overrightarrow{A_j}) = K_{\text{LR}}(\overrightarrow{A_i}, \overrightarrow{A_j})$, then the regression AF-kernel matrix $\mathbf{K}^s_{\text{Reg}}$ becomes the LR AF-kernel matrix $\mathbf{K}^s_{\text{LR}}$. For this special value of $\mathbf{\Gamma}$, λ, and $k(\overrightarrow{A_i}, \overrightarrow{A_j})$, the regression optimized kernel is equivalent to the LR kernel.

The above derivation can be generalized to other kernels. This generalization property can also be observed from the scoring procedure shown in Figures 1 and 3. For example, if the number of inner nodes in Fig. 3 reduces to one per outer node, then regression AF-kernel scoring reduces to Euclidean, Mahalanobis, GLDS, or LR AF-kernel scoring. Further, if the number of nodes in Fig. 3 reduces to two with $\alpha_0 = \alpha_1 = 1$, then AF-kernel scoring reduces to LR scoring.

3.4. Kernel-Scoring Vs. LR-Scoring

Fig. 3 shows the AF-kernel scoring procedure. The SVM output in the figure can be considered as a scoring function:

$$S_{\text{kernel}}(X_1^T) = \alpha_0 K_{\text{Reg}}\left(\overrightarrow{A_c}, \overrightarrow{A_s}\right) - \sum_{i=1}^{M} \alpha_i K_{\text{Reg}}\left(\overrightarrow{A_c}, \overrightarrow{A}_{b_i}\right) + b, \tag{18}$$

where α_0 is the Lagrange multiplier corresponding to the target speaker, and α_i $(i = 1, \ldots, M)$ are Lagrange multipliers (some of them may be zero) corresponding to the background speakers. Comparing Eqs. 1 and 18; Figs. 1 and 3 suggest that AF-kernel scoring is more general and is potentially better than LR scoring (Eq. 1). Note that Eq. 18 and Fig. 3 are not limited to AF-kernel. They are also applicable to acoustic-based GMM-SVM.

4. EXPERIMENTS AND RESULTS

4.1. 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.

4.2. Parameters for Training Kernels

In Eqs. 13 and 7, $\lambda = 0.8$, $\gamma^+ = \frac{M}{M+1}$ and $\gamma^- = \frac{1}{M+1}$, where M is the number of background speakers. Moreover, we used linearized LR kernel [8] (Eq. 15) as the reproducing kernel in Eq. 13.

4.3. 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.

4.4. EER and DET Performance

Fig. 4 shows that in the low false-alarm region, the performance of LR AF-kernel scoring (K_{LR}, solid black) is significantly better than that of LR scoring (solid red), although their performance is almost the same in the low miss-probability 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 LR 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 LR scoring, all target speakers share the same background model and the weight is always identical ($= -1$) across all target speakers. This explains the superiority of the AF-kernel scoring approach.

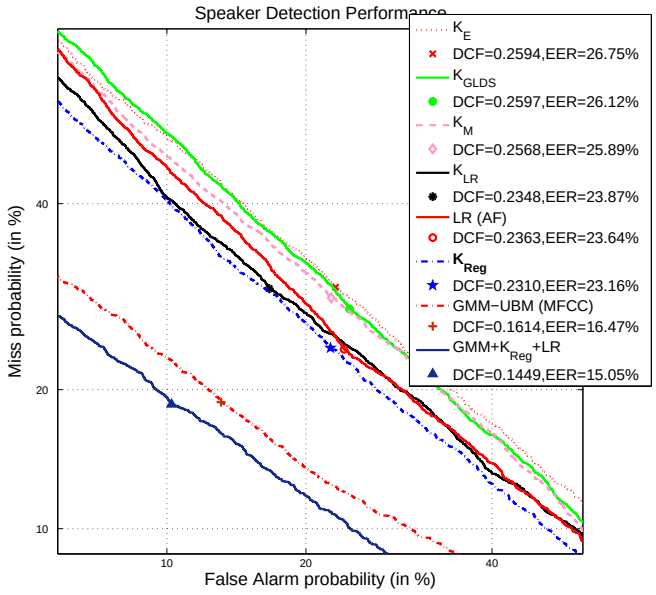

Fig. 4. DET produced by LR scoring, AF-kernel scoring, acoustic GMM-UBM, and their fusion. K_{E}, K_{M}, K_{GLDS}, K_{LR} and K_{Reg} denote the Euclidean, Mahalanobis, GLDS, LR and Regression Optimized AF-kernel, respectively. The Mahalanobis kernel K_{M} is the GMM supervector kernel in articulatory feature case.

Fig. 4 also shows that the regression AF-kernel K_{Reg} achieves the best performance among all the kernels. This suggests that optimizing a general discriminant function (Eq. 3) to derive a kernel is better than (a) using a specific distance metric (e.g., Euclidean AF-kernel K_{E} and Mahalanobis AF-kernel K_{M}) and (b) assigning a specific form for the discriminant function as in the LR AF-kernel K_{LR} and GLDS AF-kernel K_{GLDS}.[4]

[4]See the deviations and formulas of these kernels in [8].

4.5. Fusion of Low- and High-level Features

Because LR scoring and kernel scoring are based on different principles, their scores may compliment each other. To confirm this, we linearly fused the scores of the best performing AF-kernel with those obtained from LR scoring. Our result suggests that fusing the scores can improve performance. In particular, fusion can reduce the EER from 23.16% to 22.52% with a p-value smaller than 0.00001, suggesting that the differences in EERs are statistically significant.

The above fusion scores were further fused with the scores derived from a GMM-UBM acoustic system.[5] As evident in Fig. 4, fusion of low- and high-level system can further improve performance. Although the proposed kernel is evaluated on a speaker verification task, it is general enough for other classification problems.

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

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

[4] W. M. Campbell, J. R. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "High-level speaker verification with support vector machines," in *ICASSP*, May 2004, vol. 1, pp. 73–76.

[5] 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.

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

[7] B. Scholkopf, R. Herbrich, Alex J. Smola, and R. Williamson, "A generalized representer theorem," in *14th Annual Conference on Computational Learning Theory*, 2001, vol. 2111, pp. 416–426.

[8] S. X. Zhang and M. W. Mak, "High-level speaker verification via articulatory-feature based sequence kernels and SVM," in *Proc. Interspeech*, 2008.

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

[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.

[11] K. K. Yiu, M. W. Mak, M. C. Cheung, and S. Y. Kung, "Blind stochastic feature transformation for channel robust speaker verification," *J. of VLSI Signal Processing*, vol. 42, no. 2, pp. 117–126, 2006.

[12] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, 2001, pp. 213–218.

[5]Because we did not apply any feature transformation [11] or warping [12] on acoustic features, the EER of the GMM-UBM systems (red dashed-dot in Fig. 4) reported in this paper is slightly higher than the state-of-the-art.