INTERSPEECH 2009 BRIGHTON

Optimization of Discriminative Kernels in SVM 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

10.21437/Interspeech.2009-380

ABSTRACT

An important aspect of SVM-based speaker verification systems is the design of sequence kernels. These kernels should be able to map variable-length observation sequences to fixed-size supervectors that capture the dynamic characteristics of speech utterances and allow speakers to be easily distinguished. Most existing kernels in SVM speaker verification are obtained by assuming a specific form for the similarity function of supervectors. This paper relaxes this assumption to derive a new general kernel. The kernel function is general in that it is a linear combination of any kernels belonging to the reproducing kernel Hilbert space. The combination weights are obtained by optimizing the ability of a discriminant function to separate a target speaker from impostors using either regression analysis or SVM training. The idea was applied to both low- and high-level speaker verification. In both cases, results show that the proposed kernels outperform the state-of-the-art sequence kernels. Further performance enhancement was also observed when the high-level scores were combined with acoustic scores.

Index Terms— speaker verification; optimal kernels; sequence kernels; SVM; high-level features.

1. INTRODUCTION

Many speaker verification systems (e.g., GMM-UBM [1]) compute the utterance-based score of a claimant by accumulating the frame-based log-likelihood (LR) scores. This frame-based scoring scheme has three drawbacks. First, because the goal of speaker verification is to minimize classification errors on the test utterances instead of on individual speech frames, treating speech frames independently may miss some important speaker information contained in the claimants' utterances. Second, consider every frame 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, e.g., AFCPM [2]), frame-based scoring is computationally inefficient because the same probability values will be repeatedly retrieved many times during the score accumulation process.

To mitigate these drawbacks, a number of sequence kernels—such as the generalized linear discriminant sequence (GLDS) kernel [3], n-gram kernel [4], linearized LR kernel, [5], GMM-supervector (GSV) kernel [6], and Fisher kernel [7]—have been proposed for speaker verification. All of these kernels can convert variable-length sequences into fixed-length vectors for classification (or scoring) by support vector machines (SVM). They are derived from similarity metrics between two sequences by assuming a specific form for the similarity (or discriminant) functions. For example, in GLDS, the discriminant function is assumed to be linear in the kernel-induced feature space.

In this paper, instead of assuming a fixed form for the discriminant functions, 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 regression analysis, leading to a kernel that is a general form of the GLDS, GSV, linearized LR or n-gram kernels. We further demonstrate that the discriminant function can also be optimized by the SVM training algorithm. Then, using the idea of empirical kernel map [8], the optimized discriminant function can satisfy the Mercer condition [9] for SVM scoring. Experimental results on the NIST2002 SRE are presented.

2. SEQUENCE KERNELS AND SIMILARITY METRICS

In speaker verification, speech utterances are typically represented by variable-length observations $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$. To apply SVM for classification, several sequence kernels have been proposed to convert variable-length sequences into fixed dimensional vectors:

$$\begin{aligned} K\left(utt_c, utt_s\right) &= \left\langle \boldsymbol{Q}^{-\frac{1}{2}}\phi\left(\boldsymbol{O}_c\right), \boldsymbol{Q}^{-\frac{1}{2}}\phi\left(\boldsymbol{O}_s\right)\right\rangle \\ &= \left\langle \boldsymbol{Q}^{-\frac{1}{2}}\overrightarrow{A}_c, \boldsymbol{Q}^{-\frac{1}{2}}\overrightarrow{A}_s\right\rangle, \end{aligned} \tag{1}$$

where $\boldsymbol{O}_c$ and $\boldsymbol{O}_s$ are the observations of claimant c and target speaker s, and $\phi\left(\boldsymbol{O}\right)$ is a function that maps $\boldsymbol{O}$ to a fixed dimensional supervector $\overrightarrow{A}$. The definition of $\overrightarrow{A}$ and $\boldsymbol{Q}$ for different kernels are summarized in Table 1.

Table 1. Definition of $\overrightarrow{A}$ and $\boldsymbol{Q}$ for different kernels. $\boldsymbol{p}(\cdot)$ is polynomial expansion; $\overrightarrow{A}_b$ and $\overrightarrow{A}_{b_i}$ are the supervectors representing the UBM and the i-th background speaker, respectively; M is the number of background speakers; $\Pr(i)$ is the probability of occurrences of the i-th combinations in n-grams; R is the number of combinations; $\boldsymbol{\mu}_i$ is the mean of the i-th Gaussian; $\boldsymbol{\Sigma}_b = \text{diag}\left[\lambda_{b,1}^{-1}\text{diag}\left(\boldsymbol{\Sigma}_{b,1}\right), \ldots, \lambda_{b,G}^{-1}\text{diag}\left(\boldsymbol{\Sigma}_{b,G}\right)\right]$, where $\lambda_{b,i}$ and $\boldsymbol{\Sigma}_{b,i}$ are the mixture weight and covariance matrix of the i-th Gaussian in the UBM, and G is the number of Gaussians.

Kernel Type	Supervector $\overrightarrow{A}$	Normalization Matrix $\boldsymbol{Q}$
GLDS [3]	$\overrightarrow{A} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{p}(\boldsymbol{o}_t)$	$\boldsymbol{Q} = \frac{1}{M}\sum_{i=1}^{M}\overrightarrow{A}_{b_i}\overrightarrow{A}_{b_i}^{\mathrm{T}}$
n-gram [4]	$\overrightarrow{A} = [\Pr(1), \ldots, \Pr(R)]^{\mathrm{T}}$	$\boldsymbol{Q} = \text{diag}\left\{\overrightarrow{A}_b\right\}$
GSV [6]	$\overrightarrow{A} = [\boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_G^{\mathrm{T}}]^{\mathrm{T}}$	$\boldsymbol{Q} = \boldsymbol{\Sigma}_b$

This work was in part supported by Center for Multimedia Signal Processing, The Hong Polytechnic University (1-BB9W) and Research Grant Council of the Hong Kong SAR (PolyU 5251/08E).

These sequence kernels can be derived from a similarity metric that computes a similarity score between two utterances through a specific similarity (or discriminant) function $f_s(\overrightarrow{A})$. For example, the (GLDS) kernel [3] is derived from a linear discriminant (scoring) function $f_s(\overrightarrow{A_c}) = \boldsymbol{w}_s^{\mathrm{T}}\overrightarrow{A_c}$, the n-gram kernel [4] and linearized LR kernel [5] can be derived from the log-likelihood ratio function $f_s(\overrightarrow{A_c}) = \left\langle \overrightarrow{A_c}, \log \overrightarrow{A_s}./\overrightarrow{A_b} \right\rangle$,[1] and the GSV kernel [6] can be derived from Mahalanobis distance function $d_{\mathrm{M}}^2(\overrightarrow{A_c}, \overrightarrow{A_s}) = \left(\overrightarrow{A_c} - \overrightarrow{A_s}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_b^{-1} \left(\overrightarrow{A_c} - \overrightarrow{A_s}\right)$.

3. OPTIMIZATION OF KERNELS

A common characteristic of the kernels in Section 2 is that they are all derived under the assumption that the discriminant function has a specific form. This constraint can be relaxed by using a general discriminant function $f_s(\overrightarrow{A})$. This section derives two new kernels, namely regression optimized kernel and maximum-margin empirical kernel, based on two different approaches to optimizing the general discriminant function.

3.1. Regression Optimized Kernel

For a target speaker s, our goal is to find the best discriminant function $\hat{f}_s(\overrightarrow{A})$:

$$\hat{f}_s = \arg\min_{f_s \in \mathcal{H}} \left\{ \sum_{i \in \{s, b_k\}_{k=1}^M} \gamma_i L\left(f_s(\overrightarrow{A_i}), y_i\right) + \lambda \parallel f_s \parallel^2 \right\} \tag{2}$$

where M is the number of background speakers, λ is a regularizing parameter, $L(\cdot,\cdot)$ is a loss function, and γ_i is to alleviate the unbalance between the two classes of data. According to [10], the optimal solution of Eq. 2 can be written as:

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

where $w_{s,i}$ are speaker-dependent weights and $k(\cdot, \overrightarrow{A_i}) : \mathbb{R}^N \times \mathbb{R}^N \mapsto \mathbb{R}$ are kernels in the reproducing kernel Hilbert space $\mathcal{H}$ such that

$$\left\langle f_s, k(\cdot, \overrightarrow{A_i}) \right\rangle_{\mathcal{H}} = f_s(\overrightarrow{A_i}) \quad \forall\, f_s \in \mathcal{H}. \tag{4}$$

When $L(\cdot,\cdot)$ is a squared loss function, the optimization problem amounts to finding the combination weights $w_{s,i}$ for which regression analysis using the least squares method is a natural solution. Eq. 3 suggests that supervector $\overrightarrow{A}$ is first mapped to an $(M+1)$-dim space defined by $k(\cdot, \overrightarrow{A_i})$. Eq. 3 and Eq. 4 suggest that

$$\begin{aligned} \parallel \hat{f}_s \parallel^2 &= \langle \hat{f}_s, \hat{f}_s \rangle = \left\langle \hat{f}_s, \sum_{i \in \{s, b_k\}_{k=1}^M} w_{s,i} k(\overrightarrow{A_i}, \cdot) \right\rangle \\ &= \sum_{i \in \{s, b_k\}_{k=1}^M} w_{s,i} \left(\sum_{j \in \{s, b_k\}_{k=1}^M} w_{s,j} k(\overrightarrow{A_i}, \overrightarrow{A_j}) \right). \end{aligned} \tag{5}$$

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

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

where

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

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 $k_{i,j} = k_{j,i} = k(\overrightarrow{A_i}, \overrightarrow{A_j})$. Taking the derivative with respect to $\mathbf{w}_s$ in Eq. 6 and setting it to zero, the solution of Eq. 6 is

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

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

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

Because γ^+ is a constant, it can be discarded without affecting the discriminative ability of $\hat{f}_s(\overrightarrow{A})$. Note that the matrix $\mathbf{K}_s$ and the vector $k(\overrightarrow{A}, \cdot)|_{(s,b_1,\ldots,b_M)}$ are target speaker-dependent.[2] Consider that these matrices and vectors are dominated by nontarget speaker data, to make $f_s(\overrightarrow{A}_c)$ symmetric and to reduce computation time and storage space, we perform the following approximations:

$$\mathbf{K}_s \approx \mathbf{K} = \begin{bmatrix} k_{b,b} & k_{b_1,b} & \cdots & k_{b_M,b} \\ k_{b,b_1} & k_{b_1,b_1} & \cdots & k_{b_M,b_1} \\ \vdots & \vdots & \ddots & \vdots \\ k_{b,b_M} & k_{b_1,b_M} & \cdots & k_{b_M,b_M} \end{bmatrix}, \tag{10}$$

and

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

where the universal background supervector $\overrightarrow{A_b}$ is used to approximate $\overrightarrow{A_s}$. With these approximations, the regression optimized kernel is written as:

$$\begin{aligned} K_{\mathrm{Reg}}(\overrightarrow{A_c}, \overrightarrow{A_s}) = \Big\langle &\left(\mathbf{K}\boldsymbol{\Gamma}\mathbf{K}^{\mathrm{T}} + \lambda \mathbf{K}^{\mathrm{T}}\right)^{-\frac{1}{2}} k(\overrightarrow{A_c}, \cdot)|_{(b,b_1,\ldots,b_M)}, \\ &\left(\mathbf{K}\boldsymbol{\Gamma}\mathbf{K}^{\mathrm{T}} + \lambda \mathbf{K}^{\mathrm{T}}\right)^{-\frac{1}{2}} k(\overrightarrow{A_s}, \cdot)|_{(b,b_1,\ldots,b_M)} \Big\rangle, \end{aligned} \tag{11}$$

where $\mathbf{K}$ and $\boldsymbol{\Gamma}$ are defined in Eqs. 10 and 7. $\left(\mathbf{K}\boldsymbol{\Gamma}\mathbf{K}^{\mathrm{T}} + \lambda \mathbf{K}^{\mathrm{T}}\right)^{-\frac{1}{2}}$ can be considered as a normalization matrix computed from the

[1] $\log \overrightarrow{A_s}./\overrightarrow{A_b}$ stands for element-wise division and logarithm.

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

background speakers. Note that $k_{i,j} = k(\overrightarrow{A_i}, \overrightarrow{A_j})$ should belong to $\mathcal{H}$. For Low-level system, one possibility is to use the GSV kernel. Fig. 1 illustrates the structure of regression optimized kernels.

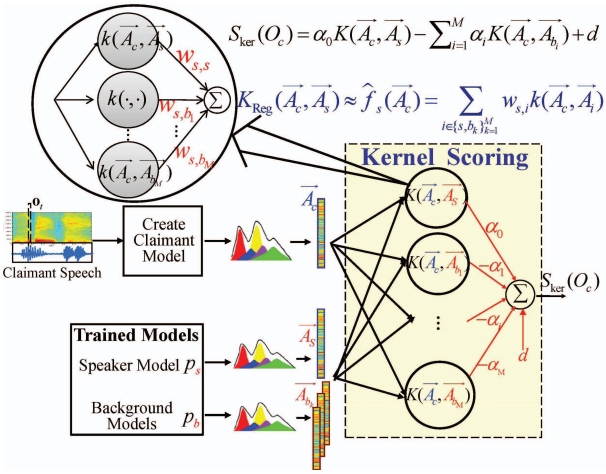

Fig. 1. The structure of regression optimized kernels.

The regression optimized kernel can be considered as a general form of the GLDS, n-gram and GSV kernels. Starting from Eq. 11, if $\mathbf{\Gamma} = \mathbf{0}$ and $\lambda = 1$, then the (i,j)-th element of the regression optimized kernel matrix $\mathbf{K}_{\text{Reg}}$ becomes:

$$\begin{aligned}\{\mathbf{K}_{\text{Reg}}\}_{i,j} &= K_{\text{Reg}}(\overrightarrow{A_i}, \overrightarrow{A_j}) \\ &= \left\langle \mathbf{K}^{-\frac{1}{2}} k(\overrightarrow{A_i}, \cdot)|_{(b,b_1,\ldots,b_M)},\ \mathbf{K}^{-\frac{1}{2}} k(\overrightarrow{A_j}, \cdot)|_{(b,b_1,\ldots,b_M)} \right\rangle \\ &= \left\langle \varphi(\overrightarrow{A_i}), \varphi(\overrightarrow{A_j}) \right\rangle .\end{aligned} \quad (12)$$

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

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

where $\mathbf{K}_s$ is defined in Eq. 8. Therefore, using Eq. 12, the regression optimized kernel matrix for target speaker s is:

$$\begin{aligned}\mathbf{K}^s_{\text{Reg}} &= \mathbf{\Omega}_s^{\text{T}} \mathbf{\Omega}_s = (\mathbf{K}^{-\frac{1}{2}} \mathbf{K}_s)^{\text{T}} (\mathbf{K}^{-\frac{1}{2}} \mathbf{K}_s) \\ &= \mathbf{K}_s^{\text{T}} \mathbf{K}^{-\frac{1}{2}} \mathbf{K}^{-\frac{1}{2}} \mathbf{K}_s \approx \mathbf{K}_s. \quad \text{(because Eq. 10: } \mathbf{K} \approx \mathbf{K}_s\text{)}\end{aligned} \quad (13)$$

Consider the elements of $\mathbf{K}_s$. If $k_{i,j} = k(\overrightarrow{A_i}, \overrightarrow{A_j}) = k_{\text{GSV}}(\overrightarrow{A_i}, \overrightarrow{A_j}) = \sum_{g=1}^{G} \left(\sqrt{\lambda_{b,g}} \mathbf{\Sigma}_{b,g}^{-\frac{1}{2}} \boldsymbol{\mu}_{i,g}\right)^{\text{T}} \left(\sqrt{\lambda_{b,g}} \mathbf{\Sigma}_{b,g}^{-\frac{1}{2}} \boldsymbol{\mu}_{j,g}\right)$, then the regression optimized kernel matrix $\mathbf{K}^s_{\text{Reg}}$ becomes the GSV kernel matrix $\mathbf{K}^s_{\text{GSV}}$. Therefore, for this special value of $\mathbf{\Gamma}$, λ, and $k(\overrightarrow{A_i}, \overrightarrow{A_j})$, the regression optimized kernel is equivalent to the GSV kernel. The above derivation can be generalized to other kernels.

3.2. Maximum-Margin Empirical Kernel

In the regression optimized kernel, supervectors that are mapped to points far away from the decision plane defined by $\{w_{s,i}\}$ in the $(M+1)$-dim space may have significant influence on the position and orientation of the plane, which may have undesirable effect on the kernel function. To avoid the influence of these extremes, we may use Vapnik's ϵ-insensitive loss function [11] as the loss function $L(x,y)$ in Eq. 2:

$$L(x,y) = \begin{cases} 0 & \text{if } |x-y| < \epsilon \\ |x-y| - \epsilon & \text{otherwise.} \end{cases}$$

It can be shown [10] that with $L(x,y)$ being the ϵ-insensitive loss function, the minimization in Eq. 2 is equivalent to the SVM training algorithm. Therefore, we can generalize Eq. 3 to

$$f_s(\overrightarrow{A}) = w_{s,0} k(\overrightarrow{A}, \overrightarrow{A_s}) - \sum_{i \in \mathcal{S}_b} w_{s,i} k(\overrightarrow{A}, \overrightarrow{A_i}) + d_s, \quad (14)$$

where $\mathcal{S}_b \subseteq \{b_k\}_{k=1}^M$ is a set of support vector indexes from the negative class, $w_{s,0}$ is the Lagrange multiplier corresponding to the (solely) positive support vector, and $w_{s,i}$, $i \in \mathcal{S}_b$, are the Lagrange multipliers corresponding to the negative support vectors.[3] Therefore, the optimal weights (Lagrange multipliers and bias) in Eq. 14 can be found by maximizing the margin of an SVM that separates the target speaker s from and background speakers $\{b_k\}_{k=1}^M$. We cannot, however, use Eq. 14 as a kernel, because it may not satisfy the Mercer's condition. One possible solution is to use empirical kernel map as follows.

Assume that we have M background speakers. We first train a UBM using these M speakers, which results in a supervector denoted $\overrightarrow{A_b}$. For the i-th background speaker, an SVM is trained to distinguish his/her voice from that of the other $M-1$ background speakers and the UBM. Similarly, an SVM is trained to distinguish the UBM from all of the M background speakers. Denote the output of the i-th background SVM as $f_{b_i}(\overrightarrow{A})$ and that corresponding to the UBM as $f_b(\overrightarrow{A})$, where we have replaced s in Eq. 14 by b_i and b. During enrollment, given an utterance from a target speaker s, we determine the corresponding supervector $\overrightarrow{A_s}$ and present it to the UBM's SVM and M background SVMs. We also present the UBM $\overrightarrow{A_b}$ and each of the background supervectors $\overrightarrow{A_{b_i}}$ to the speaker's SVM. The two sets of outputs are averaged to produce an $(M+1)$-dim vector:

$$\mathbf{f}_s = \frac{1}{2} \begin{bmatrix} f_b(\overrightarrow{A_s}) + f_s(\overrightarrow{A_b}) \\ f_{b_1}(\overrightarrow{A_s}) + f_s(\overrightarrow{A_{b_1}}) \\ \cdots \\ f_{b_M}(\overrightarrow{A_s}) + f_s(\overrightarrow{A_{b_M}}) \end{bmatrix}.$$

This vector represents the speaker class for training a linear scoring SVM. Vectors representing the impostor class are obtained by presenting each of the background speakers to the UBM's SVM and the M background SVMs, which results in M training vectors:

$$\mathbf{f}_{b_i} = \frac{1}{2} \begin{bmatrix} f_b(\overrightarrow{A_{b_i}}) + f_{b_i}(\overrightarrow{A_b}) \\ f_{b_1}(\overrightarrow{A_{b_i}}) + f_{b_i}(\overrightarrow{A_{b_1}}) \\ \cdots \\ f_{b_M}(\overrightarrow{A_{b_i}}) + f_{b_i}(\overrightarrow{A_{b_M}}) \end{bmatrix}, \ i = 1, \ldots, M.$$

The kernel of the scoring SVM is given by

$$K_{\text{MM-Emp}}(\overrightarrow{A_c}, \overrightarrow{A_s}) = < \mathbf{F}_b^{-\frac{1}{2}} \mathbf{f}_c, \mathbf{F}_b^{-\frac{1}{2}} \mathbf{f}_s >, \quad (15)$$

where $\mathbf{F}_b^{\text{T}} = [\mathbf{f}_b\ \mathbf{f}_{b_1}\ \ldots\ \mathbf{f}_{b_M}]$. We refer to $K_{\text{MM-Emp}}$ as the maximum-margin empirical kernel.

In Eq. 11, when $\mathbf{\Gamma} = \mathbf{0}$ and $\lambda = 1$, the regression optimized kernel becomes Eq. 12. A comparison between Eq. 12 and Eq. 15 suggests that the maximum-margin empirical kernel is a general form of the regression optimized kernel and other kernels.

[3] Note that $w_{s,i}$ are different from the weights in Eq. 3.

Table 2. Performance (EER) achieved by different scoring methods in low-level and high-level speaker verification. In low-level systems, GLDS supervectors are the second order polynomial expansions [3]. For other kernels in low-level systems, supervectors are the stacking of the Gaussians mean vectors. In high-level systems, supervectors are formed by stacking the values of probability mass functions (AFCPM [5]). For the scoring complexity, N is the supervectors' dimension, M is the number of background speakers, S is the number of support vectors and T is the number of frames.

Scoring Method	Kernel Type	Low-level	High-level	Scoring Complexity
Kernel Scoring	GLDS	14.56%	25.67%	$\mathcal{O}(N^3ST)$
	Linearized LR	18.14%	22.69%	$\mathcal{O}(N^2ST)$
	GSV	9.47%	23.41%	$\mathcal{O}(N^3ST)$
	Regression	**8.86%**	22.19%	$\mathcal{O}(N^3(M+1)^3ST)$
	Max-Margin	9.14%	**21.68%**	$\mathcal{O}(N^3(M+1)^3S^2T)$
LR Scoring	—	*9.42%*	*23.79%*	$\mathcal{O}(NT)$
Kernel + LR	Best{MM,Reg}	***7.90%***	***21.32%***	$\mathcal{O}(N^3(M+1)^3S^2T)$
High + Low	MM+Reg	***7.51%***		$\mathcal{O}(N^3(M+1)^3S^2T)$

4. EXPERIMENTS AND RESULTS

Datasets. NIST SRE 2001, NIST SRE 2002, SPIDRE, and HTIMIT were used in the experiments. NIST 2001 was used for creating background models, and NIST 2002 was used for creating speaker models and for performance evaluation in both high- and low-level speaker verification. HTIMIT and SPIDRE were used to train the MLPs and the phone recognizer for high-level speaker verification (AFCPM system [2]). For the low-level systems, cepstral mean normalization was applied to the MFCCs, followed by feature warping. Z-norm and T-norm were then applied to the scores to further reduce the effect of channel mismatch.

Parameters for Training Kernels. In Eq. 7, $\gamma^+ = \frac{M}{M+1}$ and $\gamma^- = \frac{1}{M+1}$, where M is the number of background speakers. Moreover, $\lambda = 0.8$ for high-level systems and $\lambda = 0.2$ for low-level systems. A small λ was chosen for low-level systems because their speaker models are more reliable; therefore less regularization is required. For the high-level systems, we used the linearized LR kernel [5] as the reproducing kernel in Eqs. 11 and 15; for the low-level systems, we used the GSV kernel [6] as the reproducing kernel.

EER and DET Performance. Table 2 and Fig. 2 show that the proposed optimized regression kernel and maximum-margin empirical kernel outperform the GSV kernel and LR scoring. 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., GSV kernel) and (b) assigning a specific form for the discriminant function as in the linearized LR kernel and the GLDS kernel. Results also show that the fusion of LR scoring and kernel scoring can further reduce the EER in both high- and low-level cases. Table 2 and Fig. 2 show that the performance can be further improved by linearly fusing the best high-level system and the best low-level system, resulting in an EER of 7.51%. To the best of our knowledge, this performance on NIST 2002 is better than the best result [12] reported in the literature. Although the kernel is evaluated on a speaker verification task, it is general enough for other classification problems.

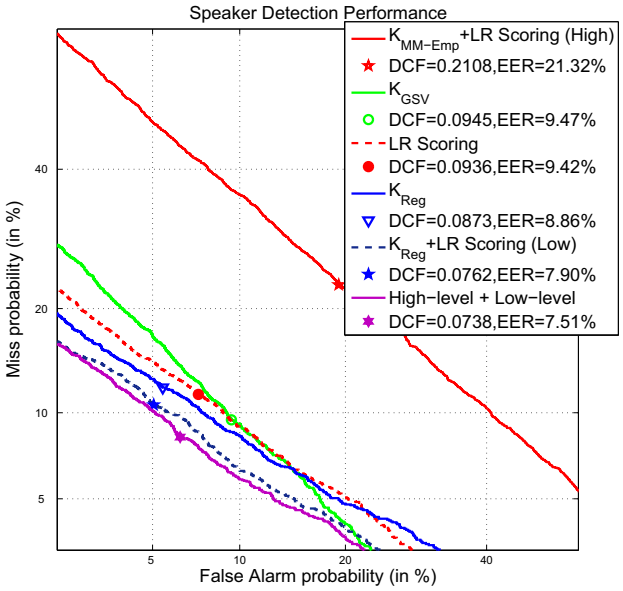

Fig. 2. DET performance of high- and low-level systems using different kernel scoring approaches and the fusion of the best high-level system ($K_{\text{MM-Emp}}$ + LR Scoring (High)) and the best low-level system (K_{Reg} + LR Scoring (Low)). The legends are arranged in decreasing EER.

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] S. X. Zhang and M. W. Mak, "High-level speaker verification via articulatory-feature based sequence kernels and SVM," in *Proc. Interspeech*, Brisbane, 2008, pp. 1393–1396.

[6] 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.

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

[8] B. Scholkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Muller, G. Ratsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, pp. 1000–1017, 1999, September.

[9] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Trans. of the London Philosophical Society (A)*, vol. 209, pp. 415–446, 1909.

[10] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural Computation*, vol. 10, pp. 1455–1480, 1998.

[11] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[12] C. Longworth and M. J. F. Gales, "Multiple kernel learning for speaker verification," in *Proc. ICASSP'2008*, 2008, pp. 1581–1584.