GENERALIZED LINEAR KERNELS FOR ONE-VERSUS-ALL CLASSIFICATION: APPLICATION TO SPEAKER RECOGNITION

Andrew O. Hatch[1,2] *and Andreas Stolcke*[1,3]

[1]The International Computer Science Institute, Berkeley, CA, USA
[2]The University of California at Berkeley, USA
[3]SRI International, Menlo Park, CA, USA

ahatch@icsi.berkeley.edu, stolcke@speech.sri.com

ABSTRACT

In this paper, we examine the problem of kernel selection for one-versus-all (OVA) classification of multiclass data with support vector machines (SVMs). We focus specifically on the problem of training what we refer to as *generalized linear kernels*—that is, kernels of the form, $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{R} \mathbf{x}_2$, where $\mathbf{R}$ is a positive semidefinite matrix. Our approach for training $k(\mathbf{x}_1, \mathbf{x}_2)$ involves first constructing a set of upper bounds on the rates of false positives and false negatives at a given score threshold. Under various conditions, minimizing these bounds leads to the closed-form solution, $\mathbf{R} = \mathbf{W}^{-1}$, where $\mathbf{W}$ is the expected within-class covariance matrix of the data. We tested various parameterizations of $\mathbf{R}$, including a diagonal parameterization that simply performs per-feature variance normalization, on the 1-conversation training condition of the SRE-2003 and SRE-2004 speaker recognition tasks. In experiments on a state-of-the-art MLLR-SVM speaker recognition system [1], the parameterization, $\mathbf{R} = \hat{\mathbf{W}}_s^{-1}$, where $\hat{\mathbf{W}}_s$ is a smoothed estimate of $\mathbf{W}$, achieves relative reductions in the minimum decision cost function (DCF) [2] of up to 22% below the results obtained when $\mathbf{R}$ does per-feature variance normalization.

1. INTRODUCTION

One of the central problems in the study of support vector machines (SVMs) is kernel selection—that is, the problem of choosing or training an appropriate kernel function for a particular dataset. Recent efforts to address this issue for binary classification have largely focused on a general setting, where the training data are divided into two classes that can be arbitrarily selected as either "target" or "impostor." Selected techniques for kernel selection in this setting are described in [3, 4, 5, 6, 7].

In this paper, we examine kernel selection for tasks involving one-versus-all (OVA) classification of multiclass data. OVA tasks arise naturally in a large number of applications (e.g. speaker verification, relevance testing for documents, image authentication, etc.) In these tasks, the goal is to determine whether or not a given test example belongs to a given target class. The OVA setting differs from the general binary classification setting in that the impostor data is composed of multiple, predefined classes whose identities are known, a priori, in the training data.

In the following report, we develop a theoretical framework for training what we refer to as *generalized linear kernels*—that is, kernels of the form $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{R} \mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors in the input space, and $\mathbf{R}$ is a positive semidefinite matrix. Our approach involves first constructing a set of upper bounds on the rates of false positives and false negatives at a given score threshold. We show that under various conditions, minimizing these bounds leads to the closed-form solution, $\mathbf{R} = \mathbf{W}^{-1}$, where $\mathbf{W}$ is the expected within-class covariance matrix of the data. This solution for $\mathbf{R}$ is particularly applicable to OVA tasks where little is known about the target classes, but where the second-order statistics of any given target class can be naively estimated from the impostor classes. We tested various parameterizations of $\mathbf{R}$, including a diagonal parameterization that simply performs per-feature variance normalization, on data for two such tasks: the 1-conversation training condition of the SRE-2003 and SRE-2004 speaker recognition sets. In experiments on a state-of-the-art MLLR-SVM speaker recognition system [1], the parameterization, $\mathbf{R} = \hat{\mathbf{W}}_s^{-1}$, where $\hat{\mathbf{W}}_s$ is a smoothed estimate of $\mathbf{W}$, achieves relative reductions in the minimum decision cost function (DCF) [2] of up to 22% below the results obtained when $\mathbf{R}$ does per-feature variance normalization.

The paper is organized as follows: In section 2, we provide a brief overview of SVMs. This is followed by a description of our problem setting in section 3. Sections 4 and 5 describe the theory behind our approach. Finally, a set of experiments and conclusions are described in sections 6 and 7.

2. SUPPORT VECTOR MACHINES

SVMs provide a means of training decision boundaries for binary classification problems. In the standard SVM formulation, the decision boundary between two classes is formed by the set, $\{\mathbf{x} : f(\mathbf{x}) = 0\}$, where f is defined as

$$f(\mathbf{x}) \triangleq \mathbf{w}^T \Phi(\mathbf{x}) + b.$$

Here, $\mathbf{x}$ represents an input feature vector, Φ represents a particular feature mapping, and $\mathbf{w}$ and b represent trainable SVM parameters. In SVMs, and in all kernel machines, the feature mapping Φ can be expressed in terms of the corresponding kernel function, k,

$$k(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1)^T \Phi(\mathbf{x}_2).$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ represents input feature vectors. Thus, for the purpose of performing SVM classification, the problem of optimizing k for a particular dataset is equivalent to the problem of optimizing Φ. Additional information about SVMs can be found in [8, 9].

This material is based upon work supported by the National Science Foundation under grant No. 0329258

3. SETTING

Given a multiclass training set composed of M disjoint classes, we would like to use an SVM with a generalized linear kernel (i.e. a kernel of the form, $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{R} \mathbf{x}_2$, where $\mathbf{R}$ is a positive semidefinite matrix) to train an OVA classifier for some arbitrary target class, i. Our goal is to derive an $\mathbf{R}$ matrix that is optimized for this purpose. To do this, we first examine the general problem of training a linear, OVA classifier for a given target class. By constructing various bounds on classification error, we arrive at a formulation for training linear classifiers that is equivalent to a *hard margin* SVM [9] with a generalized linear kernel. Under various conditions, minimizing the bounds on classification error leads to the solution, $\mathbf{R} = \mathbf{W}^{-1}$, where $\mathbf{W}$ is the expected within-class covariance matrix.

We begin by defining the function f_i to be an affine *scoring function*, which we use to perform OVA classification for target class i:

$$f_i(\mathbf{x}) \triangleq \mathbf{v}_i^T \mathbf{x} + b_i.$$

Here, $\mathbf{x}$ represents an input feature vector, $\mathbf{v}_i$ represents a weight vector, and b_i represents a bias term. We assume that $\mathbf{v}_i$ and b_i are trainable parameters. Given f_i, all test examples where $f_i(\mathbf{x}) \geq 0$ are classified as belonging to target class i, and all test examples where $f_i(\mathbf{x}) < 0$ are classified as belonging to the set of impostor classes.

We can evaluate the binary classification performance of f_i by defining the risk metric, $\mathcal{R}(f_i, \mu)$, as

$$\mathcal{R}(f_i, \mu) \triangleq \mu \cdot p(f_i(\mathbf{x}) > 0 \mid \mathbf{x} \notin C_i) + (1 - \mu) \cdot p(f_i(\mathbf{x}) < 0 \mid \mathbf{x} \in C_i).$$

In the above equation, $p(f_i(\mathbf{x}) > 0 \mid \mathbf{x} \notin C_i)$ and $p(f_i(\mathbf{x}) < 0 \mid \mathbf{x} \in C_i)$ represent the expected rates of false positives and false negatives at a score threshold of zero (here, we use the notation, C_i, to represent "class i"). The parameter $\mu \in [0, 1]$ weights the relative importance of false positives versus false negatives in computing the overall risk. Our goal in this paper is to minimize some upper bound on $\mathcal{R}(f_i, \mu)$ with respect to f_i—that is, with respect to $\mathbf{v}_i$ and b_i—for some μ.

3.1. Notation and Additional Definitions

The equations in sections 4-7 use the following notation: Let $\mathbf{x}_i$ be a random draw from class i, and let $\bar{\mathbf{x}}_i \triangleq \mathbb{E}\, \mathbf{x}_i$, where the expectation, $\mathbb{E}\, \mathbf{x}_i$, is taken over all vectors in class i. We define Σ_i to be the within-class covariance matrix for class i, Σ_i^{imp} to be the expected within-class covariance matrix over all classes $j \neq i$, and $\mathbf{W}$ to be the expected within-class covariance matrix over all classes:

$$\Sigma_i \triangleq \mathbb{E}\, (\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_i - \bar{\mathbf{x}}_i)^T \quad \forall i,$$

$$\Sigma_i^{imp} \triangleq \frac{\sum_{j \neq i} p(j) \cdot \Sigma_j}{\sum_{j \neq i} p(j)} \quad \forall i,$$

$$\mathbf{W} \triangleq \sum_{i=1}^{M} p(i) \cdot \Sigma_i.$$

Here, $p(j)$ represents the prior probability of class j. We use the term, $\mathbf{T}$, to denote the overall covariance matrix over all of the data.

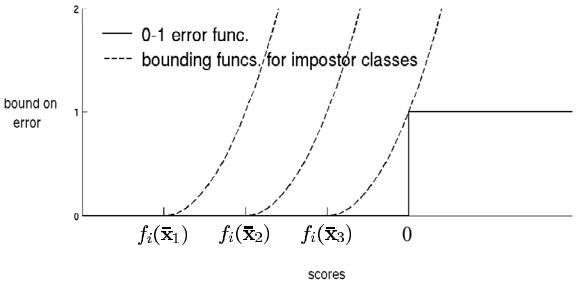

Fig. 1. Illustration of the $0-1$ error function and the corresponding bounding functions for Bound 1.

4. BOUNDING FUNCTIONS

In this section, we construct a set of upper bounds on $p(f_i(\mathbf{x}) > 0 \mid \mathbf{x} \notin C_i)$ and on $p(f_i(\mathbf{x}) < 0 \mid \mathbf{x} \in C_i)$ which we then use to bound the risk metric, $\mathcal{R}(f_i, \mu)$. To simplify the bounds, we assume throughout the following sections that each class is symmetrically distributed about its mean (*i.e.* $p(\mathbf{x}_j - \bar{\mathbf{x}}_j = \delta) = p(\mathbf{x}_j - \bar{\mathbf{x}}_j = -\delta)\ \forall\, i, \delta$). This implies that the within-class score distributions are symmetrical as well, since f_i is an affine function. Hence, we can treat a symmetrical loss function on the scores of a particular class, j, as a one-sided loss function by scaling it by $\frac{1}{2}$, assuming that the loss function is centered at $f_i(\bar{\mathbf{x}}_j)$.

Bound 1. *Given $\mathbf{v}_i$ and b_i, if $f_i(\bar{\mathbf{x}}_j) < 0 \ \ \forall \ \ j \neq i$ and if $\mathbf{x}_j$ is symmetrically distributed about its mean for all $j \neq i$, then the following bound holds.*

$$p(f_i(\mathbf{x}) > 0 \mid \mathbf{x} \notin C_i) \leq \max_{k \neq i} \frac{1}{2} \cdot \frac{\sum_{j \neq i} p(j) \cdot \mathbf{Var}(f_i(\mathbf{x}_j))}{(\sum_{j \neq i} p(j)) \cdot f_i(\bar{\mathbf{x}}_k)^2} \tag{1}$$

$$= \max_{k \neq i} \frac{1}{2} \cdot \frac{\sum_{j \neq i} p(j) \cdot \mathbb{E}\, (\mathbf{v}_i^T (\mathbf{x}_j - \bar{\mathbf{x}}_j))^2}{(\sum_{j \neq i} p(j)) \cdot (\mathbf{v}_i^T \bar{\mathbf{x}}_k + b_i)^2} \tag{2}$$

$$= \max_{k \neq i} \frac{1}{2} \cdot \frac{\mathbf{v}_i^T \Sigma_i^{imp} \mathbf{v}_i}{(\mathbf{v}_i^T \bar{\mathbf{x}}_k + b_i)^2}. \tag{3}$$

Proof. Given $\mathbf{v}_i$ and b_i, if $f_i(\bar{\mathbf{x}}_j) < 0 \ \ \forall \ \ j \neq i$ and if $\mathbf{x}_j$ is symmetrically distributed about its mean for all $j \neq i$, then we see by inspection that the following is true:

$$\mathbf{1}\Big(f_i(\mathbf{x}_j) > 0\Big) \leq \max_{k \neq i} \frac{1}{2} \cdot \left(\frac{f_i(\mathbf{x}_j) - f_i(\bar{\mathbf{x}}_j)}{f_i(\bar{\mathbf{x}}_k)}\right)^2 \ \ \forall \ j \neq i.$$

Here, the $\frac{1}{2}$ comes from the fact that the within-class score distribution for class j (i.e. the distribution of $f_i(\mathbf{x}_j)$) and the corresponding 2nd-order bounding function in the above inequality are both symmetrical and both centered at $f_i(\bar{\mathbf{x}}_j)$. Thus, we can treat the 2nd-order bounding function as a one-sided function where we only count the right-hand side (see Figure 1). Taking the expectation over all $j \neq i$ of the above inequality gives us the bound in (1). □

The 2nd-order bounding functions for the bound in (1)—which are shown in Figure 1, along with the corresponding $0-1$ error function—are simply shifted versions of one-another and are therefore quite loose for the impostor classes whose mean scores

are far from 0. Although the bounding functions for the individual impostor classes could be made tighter, we have chosen the functional form in (1) because it leads to a simple SVM-based formulation for optimizing $\mathbf{v}_i$ and b_i, as we will show in section 5.

We assume that the total number of bounding functions in (1)—which we denote by L—can be less than or equal to $M-1$, the total number of impostor classes. For example, we could treat all impostor classes as a single class (i.e. the $L=1$ scenario), in which case we only have one 2nd-order bounding function for all of the impostor scores. However, as implied by (1), the $L>1$ case leads to tighter bounds than $L=1$ when the mean of the within-class variances of the impostor scores is small compared with their overall variance.

Bound 2. *Given $\mathbf{v}_i$ and b_i, if $f_i(\bar{\mathbf{x}}_i) > 0$ and if $\mathbf{x}_i$ is symmetrically distributed about its mean, then the following bound holds.*

$$\begin{aligned} p(f_i(\mathbf{x}) < 0 \mid \mathbf{x} \in C_i) &\leq \frac{1}{2} \cdot \frac{\mathbf{Var}(f_i(\mathbf{x}_i))}{f_i(\bar{\mathbf{x}}_i)^2} \\ &= \frac{1}{2} \cdot \frac{\mathbb{E}\left(\mathbf{v}_i^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)\right)^2}{(\mathbf{v}_i^T \bar{\mathbf{x}}_i + b_i)^2} \\ &= \frac{1}{2} \cdot \frac{\mathbf{v}_i^T \Sigma_i \mathbf{v}_i}{(\mathbf{v}_i^T \bar{\mathbf{x}}_i + b_i)^2}. \end{aligned} \tag{4}$$

The derivation for Bound 2 is similar to that of Bound 1 and is therefore omitted. Putting together Bound 1 and Bound 2 gives us the following bound for $\mathcal{R}(f_i, \mu)$:

$$\begin{aligned} \mathcal{R}(f_i, \mu) \leq \max_{k \neq i} \frac{\mu}{2} \cdot \frac{\mathbf{v}_i^T \Sigma_i^{imp} \mathbf{v}_i}{(\mathbf{v}_i^T \bar{\mathbf{x}}_k + b_i)^2} + \\ \frac{(1-\mu)}{2} \cdot \frac{\mathbf{v}_i^T \Sigma_i \mathbf{v}_i}{(\mathbf{v}_i^T \bar{\mathbf{x}}_i + b_i)^2}. \end{aligned} \tag{5}$$

5. OPTIMIZATION

We can now train a linear, OVA classifier for target class i by minimizing the bound in (5) with respect to $\mathbf{v}_i$ and b_i. As we will soon show, minimizing (5) over $\mathbf{v}_i$ and b_i leads to a modified form of Vapnik's *hard margin SVM* [8, 9]. This modified SVM implicitly defines a kernel function of the form, $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{R} \mathbf{x}_2$, where $\mathbf{R}$ is a positive semidefinite matrix. We also show that under various conditions, the solution for the optimized $\mathbf{R}$ is independent of the given target class, which means that k can be implemented for any choice of target class by applying a single feature mapping to all input features.

Before we address the problem of how to minimize (5), we note that in most speaker recognition tasks (including the tasks that we consider in this paper), the amount of training data available for any given target speaker is very limited—typically no more than 8 conversation sides of around 2.5 minutes each. Given this limited amount of training data, the task of coming up with a robust estimate of the covariance matrix Σ_i for target speaker i can be very difficult, especially when the dimensionality of the feature space is very high. One way of getting around this, in the absense of any other information, is to simply assume that Σ_i is equal to the expected within-class covariance matrix over all impostor speakers (i.e. classes):

$$\Sigma_i = \Sigma_i^{imp}.$$

For simplicity, we will use this assumption in the following derivation, where we minimize (5) with respect to $\mathbf{v}_i$ and b_i. A similar derivation can be performed for the more general case where Σ_i and Σ_i^{imp} are not assumed to be equal.

Assuming that $\Sigma_i = \Sigma_i^{imp}$, we can minimize (5) with respect to $(\mathbf{v}_i, b)$ by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{v}_i, b_i} \quad & \frac{1}{2} \mathbf{v}_i^T \Sigma_i^{imp} \mathbf{v}_i \\ subject\ to \quad & 1 \leq y_j (v_i^T \bar{\mathbf{x}}_j + b_i) \quad \forall j. \end{aligned} \tag{6}$$

Here, y_j is defined as

$$y_j = \begin{cases} 1, & \text{if } j = i \\ -1, & \text{if } j \neq i \end{cases} \quad \forall j.$$

If Σ_i^{imp} is full-rank, then the problem in (6) can be converted to a more familiar form by defining the vector $\mathbf{w}_i$ and the matrix $\mathbf{U}$ as follows:

$$\begin{aligned} \mathbf{v}_i &\triangleq \mathbf{U} \mathbf{w}_i, \\ \mathbf{U}\mathbf{U}^T &\triangleq \Sigma_i^{imp^{-1}}. \end{aligned} \tag{7}$$

Substituting $\mathbf{U}\mathbf{w}_i$ in for $\mathbf{v}_i$ in (6) gives us

$$\begin{aligned} \min_{\mathbf{w}_i, b_i} \quad & \frac{1}{2} \mathbf{w}_i^T \mathbf{w}_i \\ subject\ to \quad & 1 \leq y_j (w_i^T \mathbf{U}^T \bar{\mathbf{x}}_j + b_i) \quad \forall j. \end{aligned} \tag{8}$$

Here, we see that the optimization problem in (8) has the same form as Vapnik's hard margin SVM (see for example, [8, 9]), except that the feature vector, $\bar{\mathbf{x}}_j$, has been replaced with $\mathbf{U}^T \bar{\mathbf{x}}_j$. Thus, the hard margin SVM in (8) defines the following feature mapping Φ and kernel function, k:

$$\begin{aligned} \Phi(\mathbf{x}) &= \mathbf{U}^T \mathbf{x}, \\ k(\mathbf{x}_1, \mathbf{x}_2) &= \mathbf{x}_1^T \mathbf{U}\mathbf{U}^T \mathbf{x}_2. \end{aligned} \tag{9}$$

Putting together (7) and (9), we end up with the following solution for $\mathbf{R}$ in the generalized linear kernel, $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{R} \mathbf{x}_2$:

$$\mathbf{R} = \Sigma_i^{imp^{-1}}.$$

Note that if $\Sigma_i = \sigma \cdot \Sigma_i^{imp}$ for some $\sigma > 0$, then $\mathbf{R} = \Sigma_i^{imp^{-1}}$ is optimal for any choice of $\mu \in [0, 1]$ in the bound on $\mathcal{R}(f_i, \mu)$ in (5). Otherwise, if Σ_i is *not* proportional to Σ_i^{imp}, then we can easily show that $\mathbf{R} = \Sigma_i^{imp^{-1}}$ is only optimal, in general, for the upper bound on false positives in (3). That is, if Σ_i is not proportional to Σ_i^{imp}, then $\mathbf{R}$ is only optimal for the case where $\mu = 1$. Hence, the solution, $\mathbf{R} = \Sigma_i^{imp^{-1}}$, is particularly applicable—at least in theory—to multiclass settings where the covariance matrix for any given target class is unknown (or simply hard to estimate), but where the classes are related in such a way that Σ_i^{imp} can be used as a naive estimate for Σ_i.

If the entries of Σ_i are bounded above and below for all i, then $\Sigma_i^{imp} \rightarrow \mathbf{W}$ as $p(i) \rightarrow 0$ when $L = M-1$ (i.e. the case where each impostor class gets its own loss function), and $\Sigma_i^{imp} \rightarrow \mathbf{T}$, where $\mathbf{T}$ is the overall covariance matrix of the data, as $p(i) \rightarrow 0$ when $L = 1$. Thus, when $p(i)$ is very small, we can obtain a single solution for $\mathbf{R}$ that is independent of the given target class by choosing either $\mathbf{R} = \mathbf{W}^{-1}$ or $\mathbf{R} = \mathbf{T}^{-1}$. The

tradeoffs between these two choices have to do with which value of L is better at bounding false positives given the particular dataset (this was discussed in section 4), and with which choice provides the better approximation to Σ_i^{-1}. In cases where we have little or no information about Σ_i, we can argue that it's better to choose $\mathbf{R} = \mathbf{W}^{-1}$ than $\mathbf{R} = \mathbf{T}^{-1}$, since $\mathbf{W}$ is the mean of Σ_i over all i, and there exists no such relationship, in general, between $\mathbf{T}$ and Σ_i. We investigate both choices for $\mathbf{R}$ in the experiments of the following section.

Note that the solution, $\mathbf{R} = {\Sigma_i^{imp}}^{-1}$, applies specifically to hard margin SVM training on the class *means* (see the optimization problem in (8)). We have not shown that the same solution applies to the more typical SVM training scenario, where real-world class observations are used to train *soft margin* SVMs (i.e. SVMs where the training examples are not required to be linearly separable in feature space [8, 9]). Nonetheless, the experiments in the following section deal specifically with the latter case.

6. EXPERIMENTS

Experiments were performed on two NIST-defined speaker recognition tasks where the goal is to correctly decide whether or not a given pair of conversation sides belong to the same speaker. In these tasks, one of the conversation sides in each pair is used as the "target class," while the other is used as a test example. We train an SVM-based scoring function for every target class using a fixed pool of held-out training data, which is taken from hundreds of impostor classes, as the negative examples. Note that the classes in these experiments represent speakers.

We used a version of the state-of-the-art MLLR-SVM system described in [1] to extract one 12480-dimensional feature vector from every conversation side. These features can be divided into eight disjoint groups of 1560 features each, where each group is associated with a particular set of speech phonemes. We used held-out data from the NIST SRE-2003 dataset to compute the empirical expected within-class covariance matrix $\hat{\mathbf{W}}$ and the empirical overall covariance matrix, $\hat{\mathbf{T}}$. Note that both $\hat{\mathbf{W}}$ and $\hat{\mathbf{T}}$ were estimated in a block-diagonal fashion, where the covariance between any two features i and j, where i and j belong to different phoneme groups, was set to zero. The resulting covariance matrices were then smoothed using the models,

$$\hat{\mathbf{W}}_s = \rho_W \cdot \hat{\mathbf{W}} + (1 - \rho_W) \cdot diag(\hat{\mathbf{W}}), \quad \rho_W \in [0, 1],$$
$$\hat{\mathbf{T}}_s = \rho_T \cdot \hat{\mathbf{T}} + (1 - \rho_T) \cdot diag(\hat{\mathbf{T}}), \quad \rho_T \in [0, 1],$$

where $diag(\mathbf{A})$ is the diagonal of the square matrix $\mathbf{A}$. The parameters, ρ_W and ρ_T, were independently tuned to a value of 0.30 by performing cross-validation on held-out data from the SRE-2003 dataset.

Testing was performed on a subset of the SRE-2003 task and dataset and on the entire SRE-2004 task and dataset for the 1-conversation training condition. Results are shown in Table 1 for two standard error metrics: equal-error rate (EER) and minimum *decision cost-function* (DCF)—a standard metric used by NIST to measure classification error when the relative rate of false positives is high [2]. Note that both error metrics are computed on the pooled set of SVM output scores obtained from the various target classes. As shown in Table 1, the $\mathbf{R} = \hat{\mathbf{W}}_s^{-1}$ case shows a substantial improvement over the $\mathbf{R} = \hat{\mathbf{T}}_s^{-1}$ case and over the baseline, where each feature is normalized to have unit variance (i.e. $\mathbf{R} = diag(\hat{\mathbf{T}})^{-1}$). The improvement on SRE-2004 is significantly smaller than that obtained on SRE-2003. However, this is to be expected, since both $\hat{\mathbf{W}}$ and $\hat{\mathbf{T}}$ were estimated only on SRE-2003 data, which represents a different set of channel and recording conditions than SRE-2004 (see [1] for more details about the system, datasets, and tasks).

	SRE-03 subset		SRE-04	
kernel	EER%	DCF	EER%	DCF
$\mathbf{R} = diag(\hat{\mathbf{T}}_s)^{-1}$ (baseline)	4.36	0.0166	9.84	0.0347
$\mathbf{R} = diag(\hat{\mathbf{W}}_s)^{-1}$	4.21	0.0151	9.56	0.0338
$\mathbf{R} = \hat{\mathbf{T}}_s^{-1}$	4.15	0.0141	9.56	0.0348
$\mathbf{R} = \hat{\mathbf{W}}_s^{-1}$	3.80	0.0128	9.28	0.0322
relative improvement	**12.8%**	**22.9%**	**5.7%**	**7.2%**

Table 1. EERs and minimum DCFs for various generalized linear kernels. Here, "relative improvement" compares the performance of $\mathbf{R} = \hat{\mathbf{W}}_s^{-1}$ with the baseline.

7. CONCLUSIONS

The preceding report describes an approach for training generalized linear kernels of the form, $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{R} \mathbf{x}_2$, for OVA classification tasks. We develop a set of error bounds which, under various conditions, are minimized by choosing $\mathbf{R} = \mathbf{W}^{-1}$. This particular choice for $\mathbf{R}$ achieves substantial reductions in EER and minimum DCF [2] when applied to a state-of-the-art MLLR-SVM system on various speaker recognition tasks.

8. REFERENCES

[1] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Interspeech*, 2005.

[2] *The NIST Year 2005 Speaker Recognition Evaluation Plan*, NIST, 2005, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf.

[3] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," in *Journal of Machine Learning Research*, 2004.

[4] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. of ICML*, 2004.

[5] C. S. Ong, A. Smola, and R. Williamson, "Hyperkernels," in *Neural Information Processing Systems*, 2003.

[6] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," in *Machine Learning 46*, 2002.

[7] Y. Grandvalet and S. Canu, "Adaptive scaling for feature selection in SVMs," in *Neural Information Processing Systems*, 2003.

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

[9] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, 2000.