Efficient Multiple Kernel Support Vector Machine Based Voice Activity Detection

Ji Wu, *Member, IEEE*, and Xiao-Lei Zhang, *Student Member, IEEE*

***Abstract*—In this letter, we propose a multiple kernel support vector machine (MK-SVM) method for multiple feature based VAD. To make the MK-SVM based VAD practical, we adapt the multiple kernel learning (MKL) thought to an efficient cutting-plane structural SVM solver. We further discuss the performances of the MK-SVM with two different optimization objectives, in terms of minimum classification errors (MCE) and improvement of receiver operating characteristic (ROC) curves. Our experimental results show that the proposed method not only leads to better global performances by taking the advantages of multiple features but also has a low computational complexity.**

***Index Terms*—Data fusion, multiple kernel learning, receiver operating characteristic, voice activity detection.**

I. Introduction

TO achieve robust voice activity detection (VAD) against its background noise is one of the key issues in practical speech system. There are many kinds of VAD features. If we stick them together, we might get better performance than using them separately. Recently, the feature level data fusion methods have been used in the support vector machine (SVM) based VADs [1], [2].

In this letter, we propose to use a kernel level data fusion method, called multiple kernel SVM (MK-SVM) [3]–[5], to further improve the performance of VAD. However, traditional MK-SVM algorithm is too computationally expensive for VAD in practice. To make the MK-SVM based VAD practical, we adapt the multiple kernel learning (MKL) thought to a structural SVM solver [6]–[9], where the computational burden can be overcome by the cutting-plane algorithm. Furthermore, after the adaptation, different optimization objectives are easily implemented, such as minimum classification error (MCE) and improvement of receiver operating characteristics (ROC) curves [10]. We will discuss the performances of the MK-SVM with above two objectives.

II. Review of MK-SVM and Problem Formulation

Given D dimensional observations $\bar{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathcal{X} \subseteq \mathcal{R}^D$, their labels $\bar{y} = \{y_i\}_{i=1}^n$ with $y_i \in \{-1, +1\} \triangleq \mathcal{Y}$ representing speech absence and presence at $\mathbf{x}_i$ respectively, and Q mapping function $\phi_q(\cdot) : \mathcal{X} \rightarrow \mathcal{F}_q$, $q = 1, \ldots, Q$, where $\mathcal{F}_q$ is the qth kernel space, the original definition of the classification-MK-SVM problem is formulated as [4], [5][1], [2]

$$\min_{\boldsymbol{\theta} \geq \mathbf{0}} \min_{\mathbf{w}_q, \xi_i \geq 0} \frac{1}{2} \sum_{q=1}^{Q} \frac{\|\mathbf{w}_q\|^2}{\theta_q} + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$
$$s.t.\ \forall i : y_i \left(\sum_{q=1}^{Q} \mathbf{w}_q^T \phi_q(\mathbf{x}_i) \right) \geq 1 - \xi_i; \sum_{q=1}^{Q} \theta_q = 1 \quad (1)$$

where C is a user defined constant, θ_q is the weight of the qth mapping function $\phi_q(\cdot)$, and $\{\xi_i\}_{i=1}^n$ are the n-slacks. A common method of solving (1) is to do the following two steps iteratively until convergence. For **Step 1**, we fix $\boldsymbol{\theta}$ and solve the Lagrange dual [11] of (1):

$$\max_{\mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{C}{n} \cdot \mathbf{1}_n} \mathbf{1}_n^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \left(\sum_{q=1}^{Q} \theta_q \mathbf{H}_q \right) \boldsymbol{\alpha} \quad (2)$$

with $\mathbf{w}_q$ formulated as $\mathbf{w}_q = \theta_q \sum_{i=1}^n \alpha_i y_i \phi_q(\mathbf{x}_i)$, where $\boldsymbol{\alpha}$ is a vector of Lagrangian variables, $\mathbf{1}_n$ is a vector with all entries equivalent to 1, and the Gram matrix $\mathbf{H}_q$ is defined as

$$H_{q,i,j} = y_i y_j K_q(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \ldots, n \quad (3)$$

where $K_q(\mathbf{x}_i, \mathbf{x}_j) \triangleq \langle \phi_q(\mathbf{x}_i) \phi_q(\mathbf{x}_j) \rangle$ is the kernel inner product of the qth kernel. For **Step 2**, we update $\boldsymbol{\theta}$ to further decrease the objective value of (1).

For a VAD problem, an observation can have different acoustic feature expressions, such as different kinds of energy, which can be regarded as different mapping values of the observation. We might improve VAD's performance by taking multiple feature expressions into the MK-SVM.

In most VAD applications, real-time detection is a strong demand. But from the formulation of $\mathbf{w}_q$, the time complexity for predicting a single observation is even as high as $\mathcal{O}(Qn)$.[3]

In VAD evaluation, ROC curves reflect the global performance of VAD directly. The maximum of the area under ROC curve (MaxAUC) is considered as the optimal VAD performance regardless of any tunable parameters, thresholds, and environmental conditions. Recently, Yu [10] proposed a discriminative training method for MaxAUC. However, this objective was not studied in SVM based VAD yet.

III. Structural MK-SVM Based VAD

A. Efficient Structural MK-SVM

In [6]–[9] Joachims took the relationship of observations into consideration, and proposed the structural single kernel SVM

Manuscript received May 03, 2011; revised June 02, 2011; accepted June 06, 2011. Date of publication June 13, 2011; date of current version June 21, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Constantine L. Kotropoulos.

The authors are with the Multimedia Signal and Intelligent Information Processing Laboratory, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China (e-mail: wuji_ee@tsinghua.edu.cn; huoshan6@126.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2011.2159374

[1]We omit the bias term b of the classification hyperplane according to [6].

[2]We didn't consider the L_2-norm term of the kernel weight $\boldsymbol{\theta}$'s constraint in [5] for simplicity.

[3]The computation of each kernel inner product $K(\mathbf{x}_i, \mathbf{x})$ is counted as 1.

(SK-SVM) which aimed at finding the maximum margin between different structures of the observations. In this subsection, we extend the structural SK-SVM to multiple kernel scenario for the above two problems. After defining the kernel feature tuple $\phi_q(\bar{\mathbf{x}}) \in \bar{\mathcal{F}}_q = \mathcal{F}_q \times \ldots \times \mathcal{F}_q$ and the label tuple $\bar{y} \in \bar{\mathcal{Y}} = \mathcal{Y} \times \ldots \times \mathcal{Y}$, the objective of 1-slack structural SK-SVM [6], [8] is easily extended to MK-SVM

$$\begin{aligned} &\min_{\boldsymbol{\theta}\geq\mathbf{0}} \min_{\mathbf{w}_q,\xi\geq 0} \frac{1}{2}\sum_{q=1}^{Q}\frac{\|\mathbf{w}_q\|^2}{\theta_q} + C\xi \\ &s.t. \sum_{q=1}^{Q}\theta_q = 1; \forall \bar{y}' \in \bar{\mathcal{Y}} \setminus \bar{y}: \\ &\quad \frac{1}{n}\sum_{q=1}^{Q}\mathbf{w}_q^T\left(\Psi_q(\bar{\mathbf{x}},\bar{y}) - \Psi_q(\bar{\mathbf{x}},\bar{y}')\right) \geq \frac{1}{n}\Delta(\bar{y},\bar{y}') - \xi \end{aligned} \quad (4)$$

where $\Psi_q(\bar{\mathbf{x}},\bar{y}')$ is a feature vector that can be seen as a similarity measure between $\phi_q(\bar{\mathbf{x}})$ and the pseudo label tuple $\bar{y}'$ [8], $\Delta(\bar{y},\bar{y}')$ is the loss between $\bar{y}$ and $\bar{y}'$. We define $\Psi_q(\bar{\mathbf{x}},\bar{y}') = \sum_{i=1}^{n} y_i'\phi_q(\mathbf{x}_i)$ which is the same as [6].

Here, we focus on **Step 1** of solving (4) only. Given fixed $\boldsymbol{\theta}$, (4) can be solved by cutting-plane algorithm [6], [8], [9]. Given the current cutting-plane working constraint set $\boldsymbol{\Omega} = \{\bar{y}_k'\}_{k=1}^{|\Omega|}$, the cutting-plane MK-SVM solves problem (5) and adds the most violated constraint $\bar{y}_{|\Omega|+1}'$ to $\boldsymbol{\Omega}$ iteratively until convergence occurs.

$$\begin{aligned} &\min_{\mathbf{w}_q,\xi\geq 0} \frac{1}{2}\sum_{q=1}^{Q}\frac{\|\mathbf{w}_q\|^2}{\theta_q} + C\xi \\ &s.t.\ \forall \bar{y}_k' \in \boldsymbol{\Omega}: \sum_{q=1}^{Q}\mathbf{w}_q^T\bar{\Psi}_{q,k} \geq \bar{\Delta}_k - \xi \end{aligned} \quad (5)$$

where $\bar{\Psi}_{q,k}$ is short for $(1/n)(\Psi_q(\bar{\mathbf{x}},\bar{y}) - \Psi_q(\bar{\mathbf{x}},\bar{y}_k'))$, $\bar{\Delta}_k$ is short for $(1/n)\Delta(\bar{y},\bar{y}_k')$. Because it's often difficult to get the explicit expression of $\phi_q(\cdot)$, we solve (5) in its dual as

$$\max_{\boldsymbol{\alpha}\geq\mathbf{0}} \bar{\bar{\Delta}}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\left(\sum_{q=1}^{Q}\theta_q\mathbf{H}_q\right)\boldsymbol{\alpha};\ s.t.\ \mathbf{1}_{|\Omega|}^T\boldsymbol{\alpha} \leq C \quad (6)$$

with $\mathbf{w}_q = \theta_q\sum_{k=1}^{|\Omega|}\alpha_k\bar{\Psi}_{q,k}$, where $\bar{\bar{\Delta}} = [\bar{\Delta}_1 \ldots \bar{\Delta}_{|\Omega|}]^T$, the entry of the Gram matrix $\mathbf{H}_q$ is defined as

$$H_{q,k,l} = \bar{\Psi}_{q,k}^T\bar{\Psi}_{q,l}, \quad k,l = 1,\ldots,|\Omega| \quad (7)$$

The derivation is in Appendix. The cutting-plane algorithm usually converges very fast with an upper bound convergence rate of $\mathcal{O}(1/\epsilon)$ [7], which means $|\Omega| \ll n$ and is irrelevant to n, where ϵ is a user defined cutting-plane solution precision. Therefore, compared to (2), the scale of the Gram matrix in (6) is very small and irrelevant to the training set size.

Another expensive $\mathcal{O}(n)$ scaling behavior at $\bar{\Psi}_{q,k}$, which causes a computational complexity of $\mathcal{O}(n^2)$ for $\mathbf{H}_q$ and $\mathcal{O}(n)$ for $\mathbf{w}_q$, can be eliminated by substituting $\bar{\Psi}_{q,k}$ with its sparse estimation $\hat{\Psi}_{q,k} = \beta_{q,k}\phi_q(\mathbf{b}_{q,k})$,[4] where $(\beta_{q,k},\mathbf{b}_{q,k})$ is the basis vector estimated by the cutting-plane subspace pursuit (CPSP) algorithm [9].

Algorithm 1: Efficient Structural MK-SVM

(**Step 1**)

1: **repeat**:

(**Step 2**)

2: $\Omega \leftarrow \emptyset, |\Omega| \leftarrow 0$

3: **repeat**:

4: **for** $q = 1,\ldots,Q$ **do**

5: $\mathbf{H}_q \leftarrow (H_{q,k,l})_{1\leq k,l\leq|\Omega|}$, where $H_{q,k,l} = \hat{\Psi}_{q,k}^T\hat{\Psi}_{q,l}$.

6: **end for**

7: Solve the quadratic programming (6) and get α.

8: Calculate $h(\bar{\mathbf{x}}) : h(\mathbf{x}) \leftarrow \sum_{q=1}^{Q}\hat{\mathbf{w}}_q^T\phi_q(\mathbf{x})$, where $\hat{\mathbf{w}}_q = \theta_q\sum_{k=1}^{|\Omega|}\alpha_k\hat{\Psi}_{q,k}$.

9: Calculate the most violated constraint $\bar{y}_{|\Omega|+1}'$ from $h(\bar{\mathbf{x}})$ as [6].

10: Renew $\boldsymbol{\Omega} : \boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} \cup \bar{y}_{|\Omega|+1}'$.

11: $|\Omega| \leftarrow |\Omega| + 1$.

12: Calculate $\bar{\Delta}_{|\Omega|}$ (Section III-C).

13: **for** $q = 1,\ldots,Q$ **do**

14: Estimate $\hat{\Psi}_{q,|\Omega|} = \beta_{q,|\Omega|}\phi_q(\mathbf{b}_{q,|\Omega|})$ from $\bar{\Psi}_{q,|\Omega|}$ (Section III-C) by CPSP [9].

15: **end for**

16: **until** convergence of the cutting-plane algorithm.

17: **update** $\boldsymbol{\theta}$ by the level-method [5].

18: **until** convergence of the level-method.

We summarize the efficient structural MK-SVM algorithm briefly in Algorithm 1, which is easily implemented by firstly modifying the SVM^{perf} TOOLBOX [12] (Algorithm 3 of [9][5] to multiple kernel case, and then adding the algorithm of updating $\boldsymbol{\theta}$ [4], [5] as the outer loop.

B. MK-SVM Based VAD

Given an observation $\mathbf{o}$, we might get P different acoustic feature expressions $\{\mathbf{x}^1,\ldots,\mathbf{x}^P\}$. Then, each feature $\mathbf{x}^p$ is further mapped to Q_p kernel spaces by (possibly conventional) mapping functions $\{\phi_i^p(\mathbf{x}^p)\}_{i=1}^{Q_p}$. The kernel feature group $\{\{\phi_i^1(\mathbf{x}^1)\}_{i=1}^{Q_1},\ldots,\{\phi_i^P(\mathbf{x}^P)\}_{i=1}^{Q_P}\}$ is the input of the MK-SVM. It can be regarded as a serial generalized kernel expressions $\{\phi_q(\mathbf{o})\}_{q=1}^{Q}$ with $Q = \sum_{p=1}^{P}Q_p$, where the acoustic feature extraction methods have been fused into the design of $\phi_q(\cdot)$.

[4]Only one basis vector is used for the estimation of a single $\bar{\Psi}$.

[5]The Algorithm 3 of [9] is only for classification-SVM.

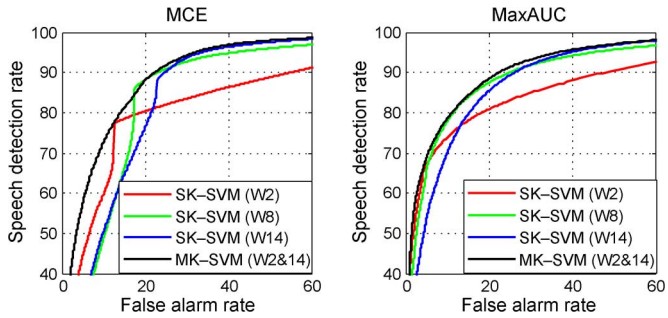

Fig. 1. ROC curve comparison of SK\MK-SVM based VAD with different optimization objectives in car noise ($\mathrm{SNR} = 5$ dB). W# is short for the MO-MP features with different window lengths.

The prediction rule of the MK-SVM based VAD is defined as

$$h(\mathbf{o}) \triangleq \sum_{q=1}^{Q} \hat{\mathbf{w}}_q^T \phi_q(\mathbf{o}) = \sum_{p=1}^{P} \sum_{q=1}^{Q_p} \theta_q^p \sum_{k=1}^{|\Omega|} \alpha_k \beta_{q,k}^p K_q^p \left(\mathbf{b}_{q,k}^p, \mathbf{x}^p\right) \underset{\hat{y}=-1}{\overset{\hat{y}=1}{\gtrless}} \eta \quad (8)$$

where $\{\theta_q^p, \{(\beta_{q,k}^p, \mathbf{b}_{q,k}^p)\}_{k=1}^{|\Omega|}\}$ and $K_q^p(\cdot,\cdot)$ are the estimated parameter group and the kernel function respectively relating to the qth kernel of the pth acoustic feature of $\mathbf{o}$, $\hat{y} = -1$ (or 1) denotes the speech absence (or presence) of $\mathbf{o}$, $h(\mathbf{o})$ is regarded as the soft output of MK-SVM, and η is used to tune the operating point of VAD. From the prediction rule, it's easy to know that the proposed method can classify a single observation only in time $\mathcal{O}(Q|\Omega|)$ and has a storage complexity of only $\mathcal{O}(Q|\Omega|D)$, which makes it practical.

C. Different Optimization Objectives

After combining MKL thought with the structural SK-SVM, using the structural MK-SVM to pursue MCE and MaxAUC is easily implemented as [6] did. For integrity of this letter, we present them briefly as follows.

For MCE, $\bar{\Delta}_k = (1/2n)\sum_{i=1}^{n} |y_i - y'_{k,i}|$, $\bar{\Psi}_{q,k} = (1/2n)\sum_{i=1}^{n}(y_i - y'_{k,i})\phi_q(\mathbf{x}_i)$. For MaxAUC, all positive (speech) labeled observations are denoted as $L_s = \{\mathbf{x}_i\}_{i=1}^{n_s}$, and all negative (noise) labeled ones are denoted as $L_d = \{\mathbf{x}_j\}_{j=1}^{n_d}$. After defining a new structure on the observations as $\phi_q(\mathbf{x}_{i,j}) = \phi_q(\mathbf{x}_i) - \phi_q(\mathbf{x}_j)$ with $y_{i,j} = 1$, the similarity is defined as $\bar{\Psi}_{q,k} = (1/2n_s n_d)\sum_{i=1}^{n_s}\sum_{j=1}^{n_d}(1 - y'_{k,i,j})\phi_q(\mathbf{x}_{i,j})$, and the loss is defined as $\bar{\Delta}_k = (1/2n_s n_d)\sum_{i=1}^{n_s}\sum_{j=1}^{n_d}(1 - y'_{k,i,j})$. An efficient calculation method was presented in Algorithm 3 of [6].

IV. Experimental Analysis

All experiments are conducted with MATLAB 7.8 on a 2.4 GHZ Itel(R) Core(TM)2 Duo PC running Windows XP with 2 GB main memory. Seven noisy test corpora of the AU-RORA2 [13] are used. The signal-to-noise ratio (SNR) level is about 5 dB. Each test corpus contains 1001 utterances, which are split randomly into three groups for training, developing and evaluation respectively. Each training set and development set consist of 300 utterances respectively. Each evaluation set consists of 401 utterances. We concatenate all short utterances in each data set into a long one so as to simulate the real-world application environment of VAD. Eventually, the length of each long utterance is in a range of (450 750)s with about 65% speeches. The observation is 25 ms with an overlap of 10 ms. In our previous work, we have implemented the structural SK-SVM in the same way as the $\mathrm{SVM}^{\mathrm{perf}}$ [12].

In every noise scenario, we run the SK\MK-SVMs ten times and report the average results. For each independent run, 6000 observations are randomly extracted from the training set for training. Then, the classifier that yields the best performance on the development set is picked up from the grid search of the parameters. For the SK-SVM based VAD, the parameter C is set to 2^{12}. The Gaussian RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2))$ is used. The kernel width σ is searched from $[0.5A, A, 2A]$, where A is the average Euclidean distance from all feature samples. For the MK-SVM based VAD, we take two kinds of features ($P = 2$). C is also set to 2^{12}. Each feature uses only one RBF kernel ($Q_1 = Q_2 = 1$). The best classifier is picked up from searching the kernel width pair $[\sigma_1, \sigma_2]$, where σ_p, $p = 1, 2$, is obtained in the same way as the SK-SVM based VAD. At last, we report the performance of the selected classifier on the evaluation set.

Two experiments with different features are conducted.

In the **first** experiment, the multiple-observation maximum probability (MO-MP) feature [2], which is extracted from the revised multiple-observation likelihood ratio test (RMO-LRT) [14], is used for performance analysis. Because the MO-MP features with different window lengths yield different ROC curves, we use three MO-MP features with window lengths of {2, 8, 14} respectively. Because the MO-MP with a window length of eight achieves the optimal performance in many noise scenarios of AURORA2 according to [14], [15], we use two inferior MO-MP features with window lengths of 2 and 14 respectively as the inputs of MK-SVM to show the power of the proposed method.

Fig. 1 gives an example of the ROC curve comparisons of the SK\MK-SVM based VADs in car scenario. From the figure, the MK-SVM based VAD yields better ROC curves than SK-SVM based VAD at both of the optimization objectives.

Table I lists the performance comparisons of the SK\MK-SVM based VADs with MCE as the objective. Table II lists the comparisons with MaxAUC as the objective. From the two tables respectively, we can conclude that the MK-SVM

TABLE I
PERFORMANCE COMPARISON (%) OF THE SK\MK-SVM BASED VADS WITH **MCE** AS THE OBJECTIVE. "**W#**" IS SHORT FOR THE MO-MPS WITH DIFFERENT WINDOW LENGTHS. THE VALUES IN BRACKETS ARE STANDARD DEVIATIONS

Accuracy				
Noise	**SK (W2)**	**SK (W8)**	**SK (W14)**	**MK (W2&14)**
Babble	56.86 (5.67)	74.24 (1.48)	73.70 (1.70)	**75.04** (0.85)
Car	81.67 (0.48)	**84.63** (0.48)	83.57 (0.25)	84.60 (0.16)
Restaurant	71.13 (1.07)	73.22 (1.52)	73.39 (1.57)	**73.94** (0.88)
Street	55.17 (1.08)	59.96 (5.37)	**62.06** (5.70)	61.40 (3.80)
Airport	73.79 (0.45)	73.96 (1.26)	73.70 (0.58)	**74.70** (0.49)
Train	72.89 (1.80)	74.70 (1.12)	73.90 (1.39)	**75.77** (0.61)
Subway	71.14 (1.17)	74.15 (1.52)	**76.28** (1.37)	75.77 (1.57)
Corresponding AUC				
Noise	**SK (W2)**	**SK (W8)**	**SK (W14)**	**MK (W2&14)**
Babble	72.42 (8.37)	80.74 (3.48)	79.36 (4.05)	**83.49** (2.40)
Car	85.75 (1.88)	89.35 (1.43)	88.65 (1.59)	**91.06** (1.32)
Restaurant	79.43 (0.52)	81.94 (1.27)	82.16 (1.08)	**83.18** (0.57)
Street	69.57 (8.72)	78.33 (3.64)	77.25 (6.92)	**79.58** (3.06)
Airport	79.86 (1.66)	80.39 (2.02)	79.89 (1.54)	**82.01** (1.26)
Train	78.98 (2.57)	80.18 (2.76)	79.41 (3.18)	**82.93** (1.15)
Subway	79.37 (0.91)	83.53 (1.02)	84.08 (1.48)	**84.66** (0.87)

TABLE II
PERFORMANCE COMPARISON (%) OF THE SK\MK-SVM BASED VADS WITH **MaxAUC** AS THE OBJECTIVE. "**W#**" IS SHORT FOR THE MO-MPS WITH DIFFERENT WINDOW LENGTHS. THE VALUES IN BRACKETS ARE STANDARD DEVIATIONS

Corresponding accuracy				
Noise	**SK (W2)**	**SK (W8)**	**SK (W14)**	**MK (W2&14)**
Babble	77.57 (1.39)	**78.79** (0.75)	77.53 (0.68)	78.68 (0.60)
Car	81.94 (0.63)	84.16 (0.94)	83.43 (0.46)	**84.46** (0.36)
Restaurant	70.19 (2.70)	73.19 (3.61)	74.21 (3.09)	**75.39** (1.33)
Street	71.07 (2.49)	72.36 (6.78)	**75.10** (1.60)	74.43 (1.28)
Airport	71.45 (4.62)	73.19 (3.48)	73.16 (1.04)	**74.70** (0.66)
Train	70.53 (5.70)	74.68 (1.33)	73.65 (1.69)	**75.95** (0.70)
Subway	71.62 (2.49)	76.06 (1.81)	75.77 (3.41)	**76.57** (1.89)
AUC				
Noise	**SK (W2)**	**SK (W8)**	**SK (W14)**	**MK (W2&14)**
Babble	83.90 (2.15)	85.42 (1.41)	83.77 (0.98)	**86.21** (0.74)
Car	87.78 (0.33)	90.16 (0.57)	90.53 (0.38)	**92.09** (0.55)
Restaurant	75.91 (2.97)	78.36 (4.87)	79.00 (3.85)	**81.30** (1.46)
Street	77.04 (2.75)	77.65 (9.81)	**81.91** (1.71)	81.10 (1.62)
Airport	76.40 (5.99)	79.20 (4.63)	79.76 (1.16)	**82.26** (0.82)
Train	76.11 (7.52)	81.75 (1.74)	79.93 (2.14)	**83.28** (0.69)
Subway	75.85 (5.76)	81.89 (2.18)	80.41 (5.65)	**82.49** (2.54)

TABLE III
CPU TIME COMPARISON (IN SECONDS) BETWEEN FEATURE EXTRACTION, CLASSIFIER TRAINING AND TEST OF THE SK\MK-SVM BASED VADS IN **Babble Noise**. NOTE THAT THE TIME FOR FEATURE EXTRACTION IS ON THE TEST SET WHICH IS 721.14 S LONG

		SK (W2)	**SK (W8)**	**SK (W14)**	**MK (W2&14)**
Feature extraction		101.37	102.18	103.22	103.68
MCE	Training	8.44	24.98	28.18	79.50
	Test	0.38	0.39	0.39	0.78
MaxAUC	Training	20.87	31.39	27.62	126.15
	Test	0.38	0.38	0.35	0.75

TABLE IV
AVERAGE PERFORMANCES (%) OF THE SK\MK-SVM BASED VADS OVER ALL NOISE SCENARIOS WITH **MaxAUC** AS THE OBJECTIVE

	SK (MO-MP)	**SK (MO-SNR)**	**MK**
Corresponding accuracy	76.06	76.25	**77.26**
AUC	82.06	80.38	**83.68**

based VAD can achieve higher accuracies and larger AUCs than that of the SK-SVM based one in most noise scenarios.

Comparing Table I with Table II, we can conclude that 1) for the SK-SVM based VAD, taking MaxAUC as the optimization objective is generally better than MCE. 2) For the MK-SVM based VAD, it's clear that MaxAUC is an overwhelmingly better objective than MCE on both metrics.

Note that the optimization objective MCE aims to minimize the classification error at a certain threshold, while the objective MaxAUC aims to optimize the whole ROC curve but not a single operating point on the ROC curve. Therefore, if MaxAUC is adopted as the objective, the optimal prediction threshold η^* should be decided on-the-fly. From this point of view, the accuracies in Table I are obtained at $\eta^* = 0$, while the accuracies in Table II and their corresponding η^* are obtained by searching η in a wide range.

Table III lists the average CPU time on feature extraction, classifier training and test of the SK\MK-SVM based VADs in babble noise. Thanks to the CPSP algorithm [9], although the test time of the MK-SVM classifier is about 2 times slower than the SK-SVM, it's also very efficient in practice. Note that the same phenomenon is observed in other noise scenarios.

In the **second** experiment, the MO-MP and the multiple-observation SNR (MO-SNR) feature [2] are taken as two input features of the MK-SVM. Both of the two features use a window length of 8. For simplicity, Table IV lists the average results over all noise scenarios with MaxAUC as the objective. The experimental result is consistent with the first experiment.

V. CONCLUSIONS

In this letter, we proposed a structural MK-SVM method for multiple feature based VAD. Specifically, we used the MKL method for the multiple feature fusion. Because the structural SK-SVM not only was very efficient but also could be optimized with different objectives, we adapted the MKL thought to it, which led to the structural MK-SVM. The experimental results showed that the MK-SVM based VAD achieved better global performances than the SK-SVM based VAD and met the real-time demand of the VAD.

APPENDIX

By using the Karush-Kuhn-Tucker (KKT) conditions [11], the Lagrangian of (5) is formulated as

$$\mathcal{L}(\mathbf{w}_q, \xi) = \frac{1}{2}\sum_{q=1}^{Q}\frac{\|\mathbf{w}_q\|^2}{\theta_q} + C\xi + \sum_{k=1}^{|\Omega|}\alpha_k\left(\bar{\Delta}_k - \xi - \sum_{q=1}^{Q}\mathbf{w}_q^T\bar{\Psi}_{q,k}\right) - \lambda\xi \quad (9)$$

where $\{\alpha_k\}_{k=1}^{|\Omega|}$, λ are non-negative Lagrangian variables. Calculating the partial derivatives with respect to the primal variables $\partial\mathcal{L}/\partial\mathbf{w}_q = 0$, $\partial\mathcal{L}/\partial\xi = 0$, we can get

$$\mathbf{w}_q = \theta_q\sum_{k=1}^{|\Omega|}\alpha_k\bar{\Psi}_{q,k}, \quad C - \sum_{k=1}^{|\Omega|}\alpha_k - \lambda = 0 \quad (10)$$

Substituting (10) to (9) can get the dual form of (5) as (6).

REFERENCES

[1] J. W. Shin, J. H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.

[2] J. Wu and X. L. Zhang, "Maximum margin clustering based statistical VAD with multiple observation compound feature," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 283–286, 2011.

[3] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.

[4] Z. Xu, R. Jin, I. King, and M. R. Lyu, "An extended level method for efficient multiple kernel learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, vol. 21, pp. 1825–1832.

[5] H. Yang, Z. Xu, J. Ye, I. King, and M. R. Lyu, "Efficient sparse generalized multiple kernel learning," *IEEE Trans. Neural Netw.*, no. 3, pp. 433–446, 2011.

[6] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 384–392.

[7] T. Joachims, "Training linear SVMs in linear time," in *Proc. 12th ACM Int. Conf. Knowl. Disc., Data Min.*, 2006, pp. 226–235.

[8] T. Joachims, T. Finley, and C. N. J. Yu, "Cutting-plane training of structural SVMs," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, 2009.

[9] T. Joachims and C. N. J. Yu, "Sparse kernel SVMs via cutting-plane training," *Mach. Learn.*, vol. 76, no. 2, pp. 179–193, 2009.

[10] T. Yu and J. H. L. Hansen, "Discriminative training for multiple observation likelihood ratio based voice activity detection," *IEEE Signal Process. Lett.*, vol. 17, pp. 897–900, Nov. 2010.

[11] B. Schölkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[12] T. Joachims, svm_perf toolbox [Online]. Available: http://svmlight.joachims.org/svm_perf.html

[13] D. Pearce and H. Hirsch *et al.*, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP00*, 2000, vol. 4, pp. 29–32.

[14] J. Ramírez, J. Segura, J. Góorriz, and L. García, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 8, pp. 2177–2189, 2007.

[15] J. Ramírez, J. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, 2005.