Maximum Margin Clustering Based Statistical VAD With Multiple Observation Compound Feature

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

***Abstract*—In this letter, we propose a new robust feature and an unsupervised learning approach for statistical voice activity detection (VAD). Maximum margin clustering (MMC), as an unsupervised classifier, can improve the robustness of support vector machine (SVM) based VAD while requiring no data labeling for model training. In the MMC framework, the multiple observation compound feature (MO-CF) is proposed to improve accuracy. MO-CF is composed of two subfeatures—multiple observation signal-to-noise ratio (MO-SNR) and multiple observation maximum probability (MO-MP). The contributions of the two subfeatures are balanced by a factor which is chosen to yield the largest area under the ROC curve (AUC) of the performance. The proposed approach obtains improved performance over seven commonly used VAD techniques in the experiments covering various noisy scenarios with low SNRs.**

***Index Terms*—Maximum margin clustering, multiple observation compound feature, support vector machine, unsupervised learning, voice activity detection.**

I. Introduction

VOICE activity detection (VAD), which is used to identify the speech portion of utterances, finds its applications in a wide spectrum of modern speech communication systems. A statistical model based VAD approach with impressive performance was proposed by Sohn [1]. Later on, multiple observation likelihood ratio test (MO-LPT) was developed in [2], [3] to further improve its robustness. In addition, Jo [4] and Shin [5] combined the statistical approaches and SVM with different features. The SVM based VADs rely on the labels of training samples. The labeling process could be expensive, time-consuming and the labels might be unreliable. Besides, the features extracted from the statistical model usually take no consideration of the contexts of the observations. In this letter, we propose an unsupervised learning approach for VAD by combining maximum margin clustering (MMC) [6], [7] and multple observation compound feature (MO-CF) attempting to improve its overall robustness and accuracy.

II. Maximum Margin Clustering Based VAD

A. Review of MMC

Given l training samples in the N-dimensional space $\bar{\mathbf{x}} \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_l\} \in \mathcal{X}$, and a possibly nonlinear kernel function $\mathcal{Q}$, an embedding kernel space $\mathcal{F}$ is defined over $\mathcal{X}$ with a mapping function $\phi(\cdot) : \mathcal{X} \to \mathcal{F}$, where $\mathcal{Q}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j)$ is the kernel matrix with each entry defined as $\mathcal{Q}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$. MMC extends the idea of SVM, and aims at finding not only the maximum margin hyperplane $(\mathbf{w}, b)$ in the feature space but also the optimal label vector $\mathbf{y} = \{y_1, \ldots, y_l\}$ that maximizes the margin among all possible label vectors. It can be formulated as the following optimization problem [6]:

$$\min_{\mathbf{y}\in\{-1,+1\}^l} \min_{\mathbf{w},b,\xi_i\geq 0} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{l}\sum_{i=1}^{l}\xi_i \quad \text{s.t. } y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i)+b\right) \geq 1-\xi_i \qquad (1)$$

where C is a user defined constant. Equation (1) can be solved by *semidefinite programming*, which has a high computational complexity. Some recent work has been conducted towards more efficient solutions. For instance, Wang [7] presented an equivalent form of (1) which can be solved iteratively by efficient *quadratic programming*. In this letter, we follow [7] for the implementation of MMC.

B. MMC Based VAD

If the labeling is perfect in SVM based VAD, one maximum margin hyperplane could be found in the feature space $\mathcal{F}$ which will lead to the minimal classification error. From the MMC theory, the hyperplane found by MMC is very close to that found by SVM with perfect labeling. However, in practical applications, especially under noisy conditions, the perfect labeling requirement is hard to satisfy. There are always some wrong labels (noisy labeling) in the training samples which result in a suboptimal maximum margin hyperplane, and therefore can not obtain the minimal classification error. If the labeling errors increase beyond certain level, the performance of SVM based VAD could be be inferior to its MMC counterpart. The MMC based VAD is presented as follows.

The MMC in [7] is only applicable to linear kernel, or to the original sample $\mathbf{x}_i$, but nonlinear kernels, such as the radius basis function (RBF) kernel, have shown superior performance on VAD [4], [5]. Hence, in order to use the nonlinear kernel, the kernel principle component analysis (KPCA) basis [8] is used to calculate the coordinates of each training samples, or $\phi(\mathbf{x}_i)$, in the high-dimensional kernel space $\mathcal{F}$, and then the MMC [7] uses $\phi(\mathbf{x}_i)$ as the feature for training and

Manuscript received December 05, 2010; revised February 19, 2011; accepted February 20, 2011. Date of publication February 28, 2011; date of current version March 07, 2011. This work was supported by The National High-Tech R&D Program of China (863 Program) under Grant 2006AA010104. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Gales.

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).

Digital Object Identifier 10.1109/LSP.2011.2119482

1070-9908/$26.00 © 2011 IEEE

classification instead of the original samples $\mathbf{x}_i$. There are two steps to obtain. $\phi(\mathbf{x}_i)$.

1) Construct $\mathcal{F}$ from training data. The training kernel matrix $\mathcal{Q}_{\rm tr} \triangleq \mathcal{Q}(\bar{\mathbf{x}}_{\rm tr}, \bar{\mathbf{x}}_{\rm tr})$ is calculated from $\bar{\mathbf{x}}_{\rm tr} = \{\mathbf{x}_1^{\rm tr}, \ldots, \mathbf{x}_u^{\rm tr}\}$, and then $\mathcal{Q}_{\rm tr}$ is diagonalized [8] to $\tilde{\mathcal{Q}}_{\rm tr}$. For efficiency, the g largest eigenvalues $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_g\}$ of $\tilde{\mathcal{Q}}_{\rm tr}$ corresponding with their eigenvectors $\boldsymbol{\mathcal{A}} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g\}$ are chosen to construct $\mathcal{F}$ approximately.
2) Calculate $\phi(\mathbf{x}_i)$ in $\mathcal{F}$. The test kernel matrix $\mathcal{Q}_{\rm te} \triangleq \mathcal{Q}(\bar{\mathbf{x}}_{\rm te}, \bar{\mathbf{x}}_{\rm tr})$ is calculated from $\bar{\mathbf{x}}_{\rm te} = \{\mathbf{x}_1^{\rm te}, \ldots, \mathbf{x}_v^{\rm te}\}$ and $\bar{\mathbf{x}}_{\rm tr}$, and then $\mathcal{Q}_{\rm te}$ is diagonalized to $\tilde{\mathcal{Q}}_{\rm te}$ with the help of $(\bar{\mathbf{x}}_{\rm tr}, \mathcal{Q}_{\rm tr}, \boldsymbol{\mathcal{A}})$. Finally, the coordinates are computed as $\phi(\bar{\mathbf{x}}_{\rm te}) = \tilde{\mathcal{Q}}_{\rm te}\boldsymbol{\mathcal{A}}$, with $\bar{\mathbf{x}}_{\rm te} = \bar{\mathbf{x}}_{\rm tr}$ in model training procedure and $\bar{\mathbf{x}}_{\rm te}$ being the feature $\mathbf{x}_i$ of any observation $\mathbf{o}_i$ in VAD detection procedure.

In MMC training procedure, after getting $\phi(\bar{\mathbf{x}}_{\rm tr}) = \{\phi(\mathbf{x}_1^{\rm tr}), \ldots, \phi(\mathbf{x}_u^{\rm tr})\}$, the maximum margin hyperplane $(\mathbf{w}, b)$ is calculated by [7], and the MMC model is defined as $(\mathbf{w}, b, \bar{\mathbf{x}}_{\rm tr}, \mathcal{Q}_{\rm tr}, \boldsymbol{\mathcal{A}})$. In the VAD detection procedure, the coordinates $\phi(\mathbf{x}_i)$ of the VAD feature $\mathbf{x}_i$ is calculated from step 2 and the decision rule is defined as

$$f(\mathbf{x}_i) \triangleq \mathbf{w}^T\phi(\mathbf{x}_i) + b \underset{h_i=0}{\overset{h_i=1}{\gtrless}} \eta \tag{2}$$

where $h_i = 0$ (or 1) denotes the speech absence (or presence) in the ith observation, $f(\mathbf{x}_i)$ is regarded as the distance between $\phi(\mathbf{x}_i)$ and the hyperplane or regarded as the soft output of MMC, η is used to tune the operating point of VAD.

We note that the diagonalization [8] of $\mathcal{Q}$ is important for the realization of the VAD, and the approximation of $\mathcal{F}$ is significant for the VAD's efficiency.

III. Feature Extraction

Inspired by [2], [3], [5], a new feature called multiple-observation compound feature (MO-CF) is proposed. It takes the advantages of the statistical model and the multiple-observation techniques. Specifically, it consists of two subfeatures. The first subfeature of MO-CF is the multiple observation signal-to-noise ratio (MO-SNR) feature $\boldsymbol{\rho}^{\sf MO}$ which is derived from single-observation SNR (SO-SNR) [5]. It has a better control over the randomness of the SNR estimation and leads to better performance on speech detection rate (SD) than SO-SNR. However, MO-SNR increases the false alarm rate (FA) simultaneously. To overcome this drawback, multiple observation maximum probability (MO-MP) $\boldsymbol{\varphi}$ is included as the second subfeature. The $\boldsymbol{\varphi}$ vector is derived from revised MO-LRT (RMO-LRT) [3] and inherits the good ability of RMO-LRT on FA. The major difference between MO-MP and RMO-LRT is that MO-MP consists of LRT scores of all DFT-bins under the maximum probabilistic global hypotheses [3] while RMO-LRT is a sum of the LRT scores. Obviously, the former is more informative than the latter. Although MO-MP could yield higher SD than RMO-LRT, it is still inferior to MO-SNR on SD. In order to combine the merits of the two proposed subfeatures, the MO-CF is defined as

$$\mathbf{x} \triangleq \left\{\frac{1}{\beta}\boldsymbol{\rho}^{{\sf MO}^T}, \boldsymbol{\varphi}^T\right\}^T \tag{3}$$

where the factor β is to balance the contributions of the two subfeatures for the best overall performance. The construction of MO-CF is different from the scheme that the MO-SNR and MO-MP are applied individually. In that case, the MO-SNR and MO-MP are separately applied and added up with scaling by a soft decision scoring function $f(\cdot)$ of MMC. The latter is simply one compromising scheme. The optimal balance factor β^* is defined as the β that will yield the largest area under the ROC curve (AUC) [9]. In practice, β^* is obtained by grid search in a range of $(0, +\infty)$. The two subfeatures are presented as follows.

For MO-SNR, it is assumed that the speech is corrupted by uncorrelated additive noise. The discrete Fourier transform (DFT) analysis is applied on each observation. The DFT coefficients of the ith observation $\mathbf{o}_i$ are denoted as $\mathbf{z}_i = \{z_{i,1}, \ldots, z_{i,N}\}$. For the nth DFT-bin of $\mathbf{o}_i$, we take the same definition of *a posteriori* SNR $\gamma_{i,n}$ and *a priori* SNR $\zeta_{i,n}$ as [10] and estimate them in the same way as [1]. Finally, the SO-SNR feature $\boldsymbol{\rho}_i^{\sf SO}$ is obtained similarly as [5]

$$\boldsymbol{\rho}_i^{\sf SO} = \{\gamma_{i,1}, \ldots, \gamma_{i,N}, \zeta_{i,1}, \ldots, \zeta_{i,N}\}^T. \tag{4}$$

The proposed MO-SNR feature $\boldsymbol{\rho}_i^{\sf MO}$ is defined as the moving average of the observation vectors $\{\mathbf{o}_{i-m}, \ldots, \mathbf{o}_i, \ldots, \mathbf{o}_{i+m}\}$:

$$\rho_{i,n}^{\sf MO} \triangleq \frac{1}{2m+1}\sum_{j=i-m}^{i+m}\rho_{j,n}^{\sf SO}. \tag{5}$$

For MO-MP, the pth global hypotheses $\mathbf{k}_p$ of an arbitrary observation vector with $2m+1$ observations is defined as $\{k_{p,1}, \ldots, k_{p,2m+1}\}$ with the central local hypotheses $k_{p,m}$ representing the class of the present observation, where $k_{p,q}$ is defined as [3]:

$$k_{p,q} = 0 : \text{speech absent in the } q\text{th position}$$
$$k_{p,q} = 1 : \text{speech present in the } q\text{th position.}$$

It is trivial to show that there are $2^{(2m+1)}$ global hypotheses in total. Under the same assumption as [3] that there is up to one speech-to-noise or noise-to-speech transition in a small enough window, the number of global hypotheses is reduced to $2(2m+1)$. By selecting the global hypotheses whose central element is 0 (or 1) from $\{\mathbf{k}_1, \ldots, \mathbf{k}_{2(2m+1)}\}$, the global hypotheses set $\mathbf{K}^0$ (or $\mathbf{K}^1$), which represents the speech absent (or present), can be formed as $\mathbf{K}^0 = \{\mathbf{k}_1^0, \ldots, \mathbf{k}_{2m+1}^0\}$ (or $\mathbf{K}^1 = \{\mathbf{k}_1^1, \ldots, \mathbf{k}_{2m+1}^1\}$). Then for the nth DFT bin of $\mathbf{o}_i$, under $\mathbf{K}^0$ and $\mathbf{K}^1$ sets, two probability vectors $\mathbf{c}_{i,n}^0$ and $\mathbf{c}_{i,n}^1$ can be defined as

$$\mathbf{c}_{i,n}^0 = \left[\log p\left(\vec{\mathbf{z}}_{i,n}\left|\mathbf{k}_1^0\right.\right)\ldots\log p\left(\vec{\mathbf{z}}_{i,n}\left|\mathbf{k}_{2m+1}^0\right.\right)\right]^T \tag{6}$$

$$\mathbf{c}_{i,n}^1 = \left[\log p\left(\vec{\mathbf{z}}_{i,n}\left|\mathbf{k}_1^1\right.\right)\ldots\log p\left(\vec{\mathbf{z}}_{i,n}\left|\mathbf{k}_{2m+1}^1\right.\right)\right]^T \tag{7}$$

with $\vec{\mathbf{z}}_{i,n}$ for $\{z_{i-m,n}, \ldots, z_{i,n}, \ldots, z_{i+m,n}\}$ and $p(\vec{\mathbf{z}}_{i,n}|\mathbf{k}_p)$ computed as

$$p(\vec{\mathbf{z}}_{i,n}|\mathbf{k}_p) = \prod_{j=i-m}^{i+m} p\left(z_{j,n}|k_{p,j-(i-m)+1}\right). \tag{8}$$

If the Gaussian model is used in (8), the conditional probability is calculated in the same way as [1]. Finally, two matrices is defined as

$$\mathbf{C}_i^0 = \left[\mathbf{c}_{i,1}^0\ldots\mathbf{c}_{i,N}^0\right], \quad \mathbf{C}_i^1 = \left[\mathbf{c}_{i,1}^1\ldots\mathbf{c}_{i,N}^1\right]. \tag{9}$$

For the maximum probability of the observation $\mathbf{o}_i$ under the "0" or "1" global hypotheses set, two row indices are given:

$$p_0^* = \arg\max_p \sum_{n=1}^{N} \left(\mathbf{C}_i^0\right)_{p,n}, p_1^* = \arg\max_p \sum_{n=1}^{N} \left(\mathbf{C}_i^1\right)_{p,n} \tag{10}$$

where $(\cdot)_{p,n}$ means the element of the pth row and the nth column of matrix. The proposed MO-MP feature is defined as

$$\boldsymbol{\varphi}_i \triangleq \{\omega_{i,1}, \ldots, \omega_{i,N}\}^T \tag{11}$$

where $\omega_{i,n} = (\mathbf{C}_i^1)_{p_1^*,n} - (\mathbf{C}_i^0)_{p_0^*,n}$.

To summarize, given the MO-SNR in (5) and the MO-MP in (11), the proposed MO-CF of $\mathbf{o}_i$ is constructed as (3).

IV. Experiments and Performance Analysis

TIMIT [11] corpus is used for the experiments with its word transcription for the VAD evaluation. All recorded speech signals are downsampled from 16 kHz to 8 kHz. 120 clean utterances are randomly selected from the training set of TIMIT. Half of them are concatenated for VAD training with another half for parameter development. Another 60 randomly chosen utterances from the test set of TIMIT are concatenated for the VAD evaluation. After the above processing, three concatenated long utterances are obtained with each of them being about 260 s long with about 63% of speech signals. In order to simulate the practical noisy environments, the original long utterances and NOISEX-92 [12] corpus are filtered by intermediate reference system (IRS) [13] to simulate the phone handset. The SNR estimation algorithm based on active speech level [14] is used to add babble and vehicle noises at an SNR level of 5 dB. After DFT analysis of an observation, the spectrum is divided into 16 critical bands ($N = 16$) which is analogous to that of the IS-127 speech enhancement technique [15], so that the MO-CF feature has a dimension of 48. The Gaussian RBF kernel $\mathcal{Q}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2))$ is chosen as the kernel operator of MMC or SVM based VADs. The number of the selected eigenvectors g is set to 10. We only randomly extract part of the samples that have low spectral energy in the training set for MMC\SVM training [16]. Based on our experimental environments, setting the window length m to 16 helps the MO-based VADs reach the optimal performance.

Fig. 1 gives an example of the optimization process of β in vehicle noise at a 5 dB SNR. From the figure, the MO-SNR feature has the advantage of detecting trivial speeches but obviously has a higher FA. On the other hand, the MO-MP feature has a good ability of controlling FA but a lower SD than MO-SNR. The MO-CF with β^* is able to take the advantages from both of them. Also, the β^*s in the development set and the test set appear in the same position, which proves the robustness of the feature. Note that the development set and the test set might match well in our task. In other applications, mismatching might happen, but the worst case is no worse than using each subfeature separately.

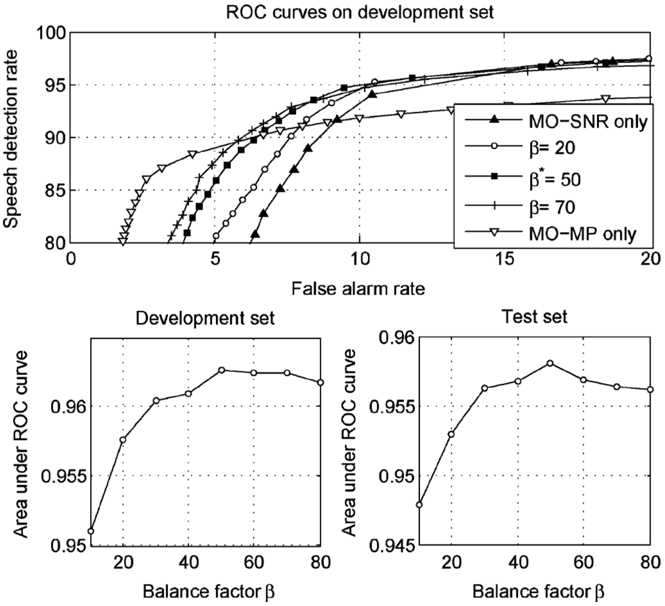

Fig. 1. Optimization process of β (vehicle noise, $\mathrm{SNR} = 5$ dB).

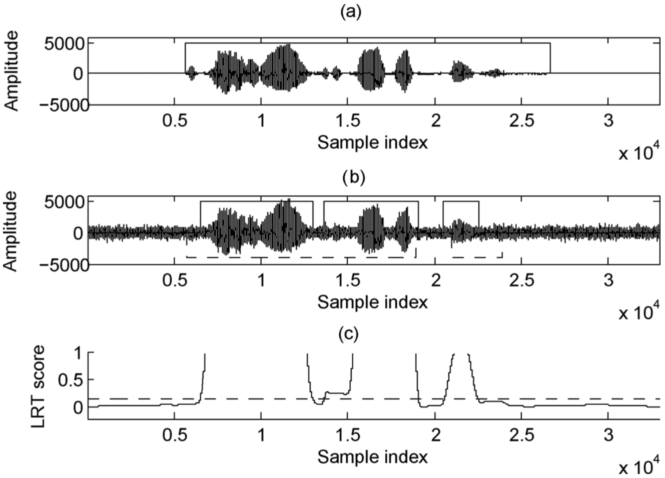

Fig. 2. Noisy labeling assumption. (a) Manual labeling in clean environment; (b) machine labeling by Ramirez VAD [2] (solid line) and the best manual labeling by human experience (dashed line); (c) error labeling assumption: the speech observations whose LRT scores are very small (below the dashed line) will be labeled as noise observations in probability. The utterance is randomly chosen from TIMIT with "/train/dr1/fvmh0/sx206.wav" as its directory.

As mentioned previously, when the SNR is low, the labels would be obviously inaccurate (noisy). Fig. 2 gives an example of comparison of manual labeling in clean environment and noisy labeling in vehicle noise. From Fig. 2(b), if we label the utterance in a noisy environment, there will be 33.08% of speech wrongly labeled as noise by Ramirez VAD [2] and 24.19% of that by human with significant efforts. Therefore, error labeling can not be totally avoided. Fig. 3 shows that the performances of the SVM-based and unsupervised SVM (USSVM) based VADs would degrade under noisy labeling assumption, while the proposed VAD shows robust performance. The USSVM means that the labels are generated by some available classifier (Here, Ramirez VAD [2] is the classifier). The same phenomenon is also observed in other scenarios.

For general comparison, besides the proposed VAD and the SVM based VAD by using MO-CF, the G.729B VAD, VAD for noise estimation from ETSI AFE (AFE WF VAD) [17], Sohn VAD [1], MO-LRT VAD [2], and RMO-LRT VAD [3] proposed by Ramirez, Tahmasbi VAD [18], and Jo VAD [4] are also tested and compared. Note that the Jo VAD is an SVM based method. Figs. 4 and 5 show the performance comparisons of the proposed VAD with other referenced VADs in vehicle and babble noise environments. From Fig. 4, although MO-CF based VADs

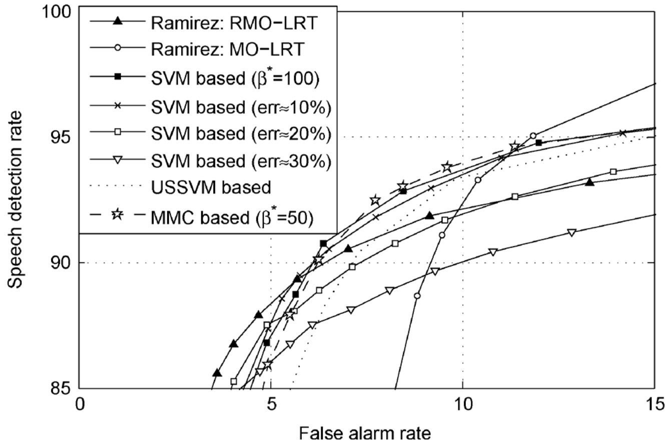

Fig. 3. Performance comparison of the MMC-based VAD and the SVM-based VAD under noisy labeling assumption in vehicle noise ($\mathrm{SNR} = 5$ dB).

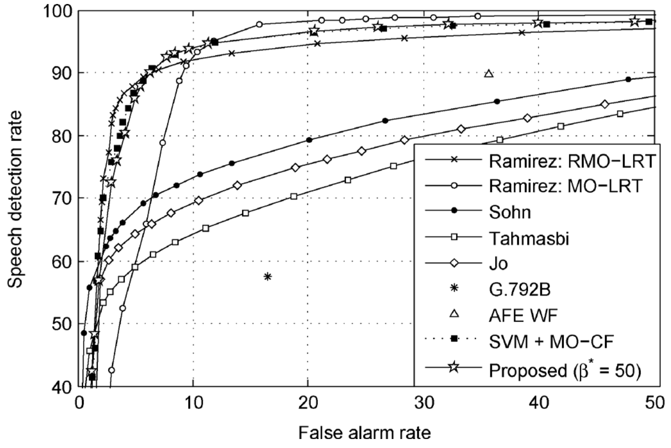

Fig. 4. Performance comparison of the VADs in vehicle noise ($\mathrm{SNR} = 5$ dB).

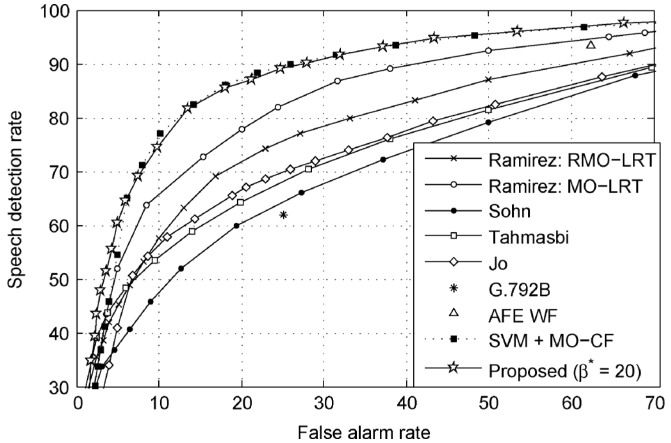

Fig. 5. Performance comparison of the VADs in babble noise ($\mathrm{SNR} = 5$ dB).

show some degradation on SD when compared with MO-LRT, they have better operating points and have a better control on FA. From Fig. 5, the MO-CF based VADs yield superior performances over all referenced VADs.

The main purpose of constructing the data set artificially from an open corpus is to simulate a continuous speech detection task with long detection time, which is the main working environment of VAD in practice. Although the proposed approach in this paper is aimed for a task-independent generic VAD solution and the optimization process of the feature parameter has shown to be quite consistent in both development and test sets, it's possible that the proposed approach is still biased towards certain tasks. To that end, besides the experiments on TIMIT, we also tested the algorithm on several real-world records of the telephone dialogs and directory enquiries. It achieved competitive performance compared to other VADs.

V. Conclusion

In this letter, we present a statistical VAD approach based on the unsupervised MMC algorithm. Compared to SVM based VAD, it does not rely on labeling of the training data and can improve the robustness of the VAD. Furthermore, a new robust feature with two complimentary subfeatures is proposed. Experimental results show that the proposed VAD could achieve good performances at a low level SNR.

References

[1] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[2] J. Ramírez, J. C. 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.

[3] J. Ramírez, J. C. Segura, J. M. Górriz, and L. García, “Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2177–2189, 2007.

[4] Q. Jo, J. Chang, J. Shin, and N. Kim, “Statistical model-based voice activity detection using support vector machine,” *IET Signal Process.*, vol. 3, no. 3, pp. 205–210, 2009.

[5] J. Shin, J. Chang, and N. Kim, “Voice activity detection based on statistical models and machine learning approaches,” *Comput. Speech Lang.*, 2009.

[6] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, “Maximum margin clustering,” in *Advances in Neural Inform. Process. Systems (NIPS)*, 2005, vol. 17, pp. 1537–1544.

[7] F. Wang, B. Zhao, and C. S. Zhang, “Linear time maximum margin clustering,” *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 319–332, 2010.

[8] B. Schölkopf, A. Smola, and K. Müller, “Kernel principal component analysis,” *Artificial Neural NetworksłICANN’97*, pp. 583–588, 1997.

[9] T. Yu and J. H. L. Hansen, “Discriminative training for multiple observation likelihood ratio based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 897–900, 2010.

[10] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, “DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM,” NTIS Order PB91-100354 1993.

[12] Rice Univ.. Houston, TX, Noisex-92 Database [Online]. Available: http://spib.rice.edu/spib

[13] *Specifications for an Intermediate Reference System*, ITU-T Rec. P.48, Mar. 1989.

[14] *Objective Measurement of Active Speech Level*, ITU-T Rec. P.56, 1993.

[15] Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spectrum Digital Systems Tech. Rep. 3GPP2 C.S0014-A, Apr. 2004, TIA/EIA/IS-127.

[16] J. H. Chang, Q. H. Jo, D. K. Kim, and N. S. Kim, “Global soft decision employing support vector machine for speech enhancement,” *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 57–60, 2008.

[17] ETSI, “Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms,” *ETSI ES*, vol. 202, no. 50.

[18] R. Tahmasbi and S. Rezaei, “A soft voice activity detection using GARCH filter and variance Gamma distribution,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1129–1134, 2007.