2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Hybrid SVM/HMM Architectures for Statistical Model-based Voice Activity Detection

Ying-Wei Tan, Wen-Ju Liu, Wei Jiang and Hao Zheng

***Abstract*— The decision function of support vector machine (SVM) using the likelihood ratios (LRs) is successfully used for statistical model-based voice activity detection (VAD). It is known to incorporate an optimised nonlinear decision over two different classes, instead of comparing the geometric mean of the LRs for the individual frequency bands with a given threshold for speech detection. However, the inter-frame correlation of the voice activity is not taken into consideration. In this paper, we explore a hybrid SVM/hidden Markov model (HMM) approach for the VAD, which retains discriminative and nonlinear properties of SVM, while modeling the inter-frame correlation powerfully through a first-order HMM. Experimental results show the significant improvement of the performance of the proposed VAD in comparison with the SVM-based VAD.**

I. Introduction

Being an important module in many speech processing applications [1], [2], VAD has attracted a lot of attention in the research community over the last few decades. Different statistical model-based strategies are adopted for detecting speech in noise. The statistical model-based VAD approach originates from the speech enhancement algorithm [3]. A Gaussian statistical model [4] is applied to the VAD using the decision-directed (DD) method-based parameter estimation. On the other hand, VAD is essentially a binary classification problem. Therefore, machine learning algorithms are effective for solving it. In [5], SVM-based methods enable us to obtain the optimised hyperplane to minimise decision error, and speech is detected through the decision function using the LRs. In [6], the inter-frame correlation information of the voice activity is incorporated into the decision rule based on a first-order HMM. In [9], the statistical approaches and SVM with different features are combined for VAD. In [7], an HMM-based segmentation procedure with two model is used. Speech and non-speech are each modeled by five-state, left-to-right HMMs with no skip states. In [8], an improved voice activity detection (VAD) algorithm using wavelet and support vector machine (SVM) is proposed. In [10], A new voice activity detection (VAD) algorithm with soft decision output in Mel-frequency domain is developed based on hidden Markov model (HMM) and is incorporated in an HMM-based speech enhancement system. In [11], a robust voice activity detector (VAD) based on hidden Markov models (HMM) is presented.

YingWei Tan, WenJu Liu, Wei Jiang and Hao Zheng are with the Department of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (email: {ywtan,lwj,wjiang,hao.zheng}@nlpr.ia.ac.cn).

This research was supported in part by the China National Nature Science Foundation (No.91120303, No.61273267, No.90820011 and No.90820303).

In this paper, we present a new VAD algorithm based on hybrid SVM/HMM architectures, which use a first-order HMM for modeling the inter-frame correlation powerfully and augment the HMM with the SVM, that is trained discriminatively. The proposed VAD shows better performances in various noise environments.

In Section 2 the SVM-based VAD is described briefly. In Section 3 we show HMM-based MAP VAD concisely. In Section 4 we elaborate the VAD based on hybrid SVM/HMM architectures in detail. In Section 5 we describe experimental conditions and experimental results on evaluating our algorithm. In Section 6 the conclusions are drawn. Finally relation to the prior work is introduced.

II. SVM-based VAD

According to [5], let the noise signal $n(t)$ is added to the speech signal $x(t)$, with their sum being denoted by $y(t)$ in the time domain. By taking the discrete Fourier transform (DFT), we obtain

$$Y(t) = X(t) + N(t) \tag{1}$$

where $Y(t) = [Y_1(t), Y_2(t), \ldots, Y_L(t)]$, $X(t) = [X_1(t), X_2(t), \ldots, X_L(t)]$, and $N(t) = [N_1(t), N_2(t), \ldots, N_L(t)]$ denote the DFT coefficients of the noisy speech signal, clean speech signal, and the additive noise signal. Also, L is the total number of frequency bins. Given two hypotheses, H_0 and H_1, which, respectively, indicate speech absence and presence, it is assumed that

$$H_0 : \ speech\ absent : Y_l(t) = N_l(t) \tag{2}$$

$$H_1 : \ speech\ present : Y_l(t) = X_l(t) + N_l(t) \tag{3}$$

Assuming that each spectral component of speech and noise signals has complex Gaussian distribution, in which the noise is uncorrelated with the speech signal, the distributions of the noisy spectral components conditioned on both hypotheses are obtained as follows:

$$p(Y_l|H_0) = \frac{1}{\pi\lambda_{n,l}} \exp\left\{-\frac{|Y_l|^2}{\lambda_{n,l}}\right\} \tag{4}$$

$$p(Y_l|H_1) = \frac{1}{\pi\left[\lambda_{n,l} + \lambda_{x,l}\right]} \exp\left\{-\frac{|Y_l|^2}{\left[\lambda_{n,l} + \lambda_{x,l}\right]}\right\} \tag{5}$$

where $\lambda_{x,l}$ and $\lambda_{n,l}$ denote the variances of noise and speech for the individual frequency band, respectively. The likelihood ratio for the lth frequency band is

$$\Lambda_l \triangleq \frac{p(Y_l|H_1)}{p(Y_l|H_0)} = \frac{1}{1+\xi_l} \exp\left\{\frac{\gamma_l \xi_l}{1+\xi_l}\right\} \tag{6}$$

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

where $\xi_l = \frac{\lambda_{x,l}}{\lambda_{n,l}}$ and $\gamma = \frac{Y_l}{\lambda_{n,l}}$ denote the a priori SNR and a posteriori SNR, respectively. The a posteriori SNR γ_l is estimated using $\lambda_{n,l}$, and the a priori SNR ξ_l is estimated by the well-known DD method as follows:

$$\hat{\xi_l}(t) = \alpha \frac{\left|\hat{X}_l(t-1)\right|^2}{\lambda_{n,l}(t-1)} + (1+\alpha) P\left[\gamma_l(t) - 1\right] \tag{7}$$

where $\left|\hat{X}_l(t-1)\right|$ is the speech spectral amplitude estimate of the previous frame obtained using the minimum mean-square error (MMSE) estimator. Also, α is a weight that is usually determined in the range $(0.95, 0.99)$. The function $P[x] = x$ if $x \geq 0$ and $P[x] = 0$ otherwise. For the decision rule of the VAD, the LRs are incorporated as elements of feature vector characterised by SVM. Let $\Lambda(t) = [\Lambda_1(t), \Lambda_2(t), \ldots, \Lambda_L(t)]^T$ be the LRs obtained by (6) and Λ_m^* be the mth support vector of LRs obtained by training. Then

$$\begin{aligned} f(\Lambda(t)) &= (w^* \cdot \Lambda(t)) + b^* \\ &= \sum_{i=1}^{M} \alpha_i^* z_i (\Lambda_i^* \cdot \Lambda(t)) + b^* \underset{H_0}{\overset{H_1}{\gtrless}} \eta \end{aligned} \tag{8}$$

where w^* is the optimal weight vector, b^* is the bias, α_i^* is Lagrange multiplier, z_i is the corresponding class label, M is the number of support vector, and η is the threshold value. Comparing (8) and a given threshold value reveals the SVM-based decision statistic. It can be seen that decision statistic is derived by the use of the dot product between the given LR vector and the support vectors. In order to consider nonlinear input space, the various kernel function K has been addressed [12] rather than the linear kernel such that

$$K(\Lambda_i^*, \Lambda) = \Phi(\Lambda_i^*) \cdot \Phi(\Lambda) \tag{9}$$

Once the kernel function is specified as in (9), the decision statistic finally results in the following form

$$f(\Lambda(t)) = \sum_{i=1}^{M} \alpha_i^* z_i K(\Lambda_i^* \cdot \Lambda(t)) + b^* \tag{10}$$

The radius basis function (RBF) kernel is incorporated for the VAD due to the superior performance [13]

$$K_{RBF}(\Lambda_i^*, \Lambda(t)) = \exp\left(-\frac{1}{2\sigma^2} ||\Lambda_i^* - \Lambda(t)||^2\right) \tag{11}$$

where σ is the kernel width.

III. HMM-BASED MAP VAD

According to [6], the inter-frame correlation is strong. The sequence of voice activity states is modeled by a first-order HMM. The transition probability is defined as

$$a_{ij} = P(H(t) = H_j | H(t-1) = H_i) \tag{12}$$

for $i, j = 0, 1$, and the initialize probabilities are $P(H(1) = H_0) = P(H(1) = H_1) = 1/2$. The likelihood ratio of the observation $Y_g(t)$ at the tth frame is given by

$$\begin{aligned} \Lambda_g(t) &= \frac{P(Y_g(t)|H(t) = H_1)}{P(Y_g(t)|H(t) = H_0)} \\ &= \left(\prod_{l=1}^{L} \frac{P(Y_l(t)|H(t) = H_1)}{P(Y_l(t)|H(t) = H_0)}\right)^{1/L} \end{aligned} \tag{13}$$

The posterior probabilities of H_1 and H_0 given $Y_g(t)$ are derived as follows:

$$\begin{aligned} P(H(t) = H_1|Y_g(t)) &= \frac{\Lambda_g(t) P(H(t) = H_1)}{P(H(t) = H_0) + \Lambda_g(t) P(H(t) = H_1)} \\ P(H(t) = H_0|Y_g(t)) &= \frac{P(H(t) = H_0)}{P(H(t) = H_0) + \Lambda_g(t) P(H(t) = H_1)} \end{aligned} \tag{14}$$

where $P(H(t) = H_i)$ is the a priori probability. Based on the first-order HMM, the a priori probability is given by

$$P(H(t) = H_i) = \sum_j a_{ji} P(H(t-1) = H_j | Y_g(t-1)) \tag{15}$$

for $i, j = 0, 1$, and $P(H(1) = H_0) = \pi_0$, $P(H(1) = H_1) = \pi_1$ at the initial state. Finally, the decision rule is derived as

$$\Lambda_g(t) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\eta P(H(t) = H_0)}{(1-\eta) P(H(t) = H_1)} \tag{16}$$

where $\eta \in [1/2, 1)$ is the compensation factor.

IV. THE PROPOSED VAD BASED ON HYBRID SVM/HMM ARCHITECTURES

An important issue that had to be addressed in this hybrid system is the fact that SVM outputs a distance measure, while the decision rule uses likelihood ratio at (16). We therefore maps SVM distances to posterior probabilities based on a warping function. A simple approach to estimating the posterior is to assume that posterior takes the form of a sigmoid, and directly estimate the sigmoid [14] as

$$p(H(t)|\, f) = \frac{1}{1 + \exp(A\, f + B)} \tag{17}$$

In order to avoid severe bias in the distances for the training data, the free parameters, A and B are estimated on a cross-validation set. Once we have the posteriors, we obtain

$$\Lambda_{svm}(t) = \frac{P(H(t) = H_1|\, f)}{P(H(t) = H_0|\, f)} \tag{18}$$

By replacing the likelihood ratio Λ_g in (19) with Λ_{svm}, the new decision rule is

$$\Lambda_{svm}(t) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\eta P(H(t) = H_0)}{(1-\eta) P(H(t) = H_1)} \tag{19}$$

where $P(H(t))$ is derived by (15).

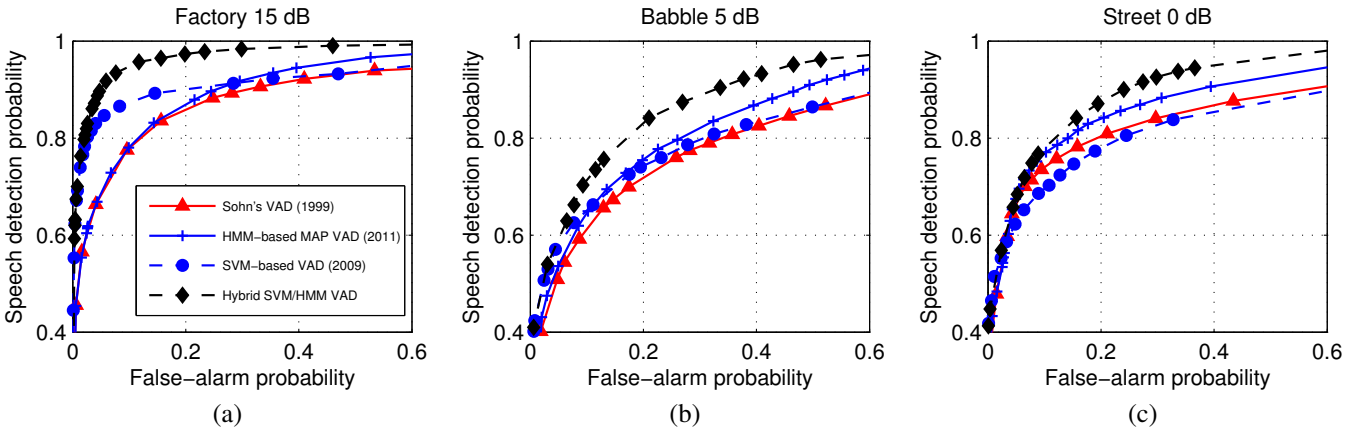

Fig. 1. ROC curves of the VAD approaches in (a) factory (SNR = 15 dB), (b) babble (SNR = 5 dB), and (c) street (SNR = 0 dB) noises.

Noise	SNR	Sohn's VAD	HMM-based MAP VAD	Jo's SVM VAD	SVM/HMM VAD
Factory	0 dB	58.07	57.50	46.36	**40.17**
	5 dB	48.06	47.50	38.19	**32.84**
	15 dB	32.05	30.16	25.28	**18.66**
Babble	0 dB	60.41	59.95	58.50	**54.49**
	5 dB	48.50	47.02	44.98	**38.64**
	15 dB	32.84	30.91	29.67	**24.94**
Street	0 dB	45.59	42.37	43.91	**36.88**
	5 dB	38.91	35.58	37.98	**30.24**
	15 dB	30.31	26.55	27.06	**22.85**

TABLE I

COMPARISON OF SPEECH DETECTION ERROR PROBABILITY ($P_e = (1 - P_d) + P_f$, %) IN DIFFERENT NOISE CONDITIONS.

V. EXPERIMENTS AND RESULTS

A. Experimental condition

For evaluating the proposed algorithm, experiments were conducted in various noisy environments including the factory, babble, and street noises at different signal-to-noise ratios (SNRs). The factory noise and babble noise are from the NOISEX-92 corpus [15], while the street noise was recorded by us on a busy street. The test set consists of 20 individual speakers' utterances in TIMIT test corpus [16]. These utterances are split into randomly into three groups for training, developing, and testing. These sentences in each group are concatenated, silence is inserted between sentences. As a result, 220 s, 200 s, and 180 s long clean utterances are obtained as the final training set, developing set, and testing set, respectively. These referenced labels are determined at every 10 ms frame by combining manual labels and energy-based VAD. Manual labels help to remove breath regions, while energy-based VAD helps to remove very low energy regions. The percentage of the marked speech frames in the training set is 66.69%, which consists of 31.47% voiced sound and 35.22% unvoiced sound frames, the percentage of the marked speech frames in the developing set is 65.91%, which consists of 31.04% voiced sound and 34.88% unvoiced sound frames, and the percentage of the marked speech frames in the testing set is 64.80%, which consists of 31.30% voiced sound and 33.50% unvoiced sound frames. Then, noise of each category was added at three different SNR levels (0 dB, 5 dB, 15 dB) to the three materials. We define P_d as the ratio of correct speech decisions to the marked speech frames, while P_f as that of false speech decisions to the marked noise frames. We investigate the receiver operating characteristic (ROC) curves, which shows the trade-off characteristic between the speech detection and false-alarm probabilities (P_d and P_f).

For the parameters in the SVM model, the best performance on the development set is picked up from the search of the parameters. The parameters C is set to 2^{12} and the kernel width σ is set to the average Euclidean distance from all feature samples. In the first-order HMM model, the transition

probabilities are obtained from training speech and the initial probabilities $P(H(1) = H_0) = P(H(1) = H_1) = 1/2$.

B. Results

For fair comparison, we do not consider any hangover scheme, as this can be added after the design of the decision rule.

By and large, Figure 1 (a)-(c) shows the proposed VAD based on SVM/HMM architectures yielded higher performance over all referenced VADs. As for the street noise (SNR = 0 dB), the performance of the proposed technique is not promoted significantly when only ($P_f < 0.06$). This means that the posterior probabilities derived by SVM are used appropriately, are fused into HMM architectures successfully, and contribute to the VAD performance improvement greatly.

In addition, the performance of the proposed VAD algorithm is evaluated by fixing the threshold. The operating point of the VAD is fixed to make the false-alarm probability of the proposed VAD sightly less than or equal to that of the conventional methods. As showed in Table I for various SNRs, the results confirm that the proposed VAD with hybrid SVM/HMM architectures outperforms other approaches in terms of the speech detection error probability (P_e), where both the false alarms and missing errors are incorporated.

VI. CONCLUSIONS

In this paper, we propose a VAD technique based on hybrid SVM/HMM architectures. The advantages of this work not only lie in making full use of powerful classifiers, SVMs, that are trained discriminatively, but also rest with modeling temporal evolution of data more effectively by the HMM. The experimental results show that the hybrid SVM/HMM VAD achieves an enhanced ability to discriminate speech and silences over the SVM-based VAD system in various noisy environments.

VII. RELATION TO PRIOR WORK

On the one hand, SVM-based VADs [5] consider nonlinear properties of the input data. However, one significant drawback in the SVMs is that, they are inherently static classifiers. They do not implicitly model temporal evolution of data. On the other hand, in [6], though nonlinear properties of SVM are not considered, the inter-frame correlation is taken into account by the VAD based on HMM, which have the advantage of handling dynamic data with certain assumptions about stationarity and independence. The algorithm presented here has focused on constructing the hybrid VAD system is to win both worlds by combining the discriminative strength and nonlinear properties of SVM-based VAD [5] with the ability of modeling the inter-frame correlation of voice activity based on the HMM. The advantages of hybrid SVM/HMM VAD are not considered in early studies. By these steps, we can obtain superior performance for VAD over the SVM-based VAD system.

REFERENCES

[1] R. Le Bouquin-Jeannès and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech communication*, vol. 16, no. 3, pp. 245–254, 1995.

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

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

[6] S. Deng, J. Han, T. Zheng, and G. Zheng, "A modified map criterion based on hidden Markov model for voice activity detecion," in *Proceedings of ICASSP*, 2011, pp. 5220–5223.

[7] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM spine evaluation system," in *Proceedings of ICASSP*, 2002, pp. I–53.

[8] S.-H. Chen, R. C. Guido, T.-K. Truong, and Y. Chang, "Improved voice activity detection algorithm using wavelet and support vector machine," *Computer Speech & Language*, vol. 24, no. 3, pp. 531–543, 2010.

[9] 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.

[10] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement," *IET signal processing*, vol. 6, no. 1, pp. 54–63, 2012.

[11] O. Varela Serrano, R. San Segundo Hernández, and L. A. Hernández, "Robust speech detection for noisy environments," *IEEE Aerospace and Electronic Systems Magazine*, vol. 26, no. 11, pp. 16–23, 2011.

[12] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[13] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *2002 6th International Conference on Signal Processing*, vol. 2, 2002, pp. 1124–1127.

[14] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[15] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[16] J. S. Garofolo *et al.*, "Getting started with the darpa TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.