Voice Activity Detection based on Support Vector Machine using Effective Feature Vectors

Q-Haing Jo, Yun-Sik Park, Kye-Hwan Lee, Ji-Hyun Song, Joon-Hyuk Chang

School of Electronic and Electrical Engineering
Inha University, Incheon, Korea
{qhjo, yspark, khlee, jhsong}@dsp.inha.ac.kr, changjh@inha.ac.kr

Abstract

In this paper, we propose effective feature vectors to improve the performance of voice activity detection (VAD) employing a support vector machine (SVM), which is known to incorporate an optimized nonlinear decision over two different classes. To extract the effective feature vectors, we present a novel scheme combining the \textit{a posteriori} SNR, \textit{a priori} SNR, and predicted SNR, widely adopted in conventional statistical model-based VAD. Based on the results of experiments, the performance of the SVM-based VAD using novel feature vectors is found to be better than that of ITU-T G.729B and other recently reported methods.

\textbf{Index Terms}: voice activity detection, support vector machine.

1. Introduction

Voice activity detection (VAD) has become a crucial component of variable-rate speech coding for efficient bandwidth reduction. Specifically, the realization of excellent variable-rate speech coding requires a VAD method offering superior performance with respect to the classification of speech/non-speech [1]. For this reason, various VAD algorithms have been proposed. Earlier algorithms for VAD were mostly based on linear prediction coding (LPC) parameters [2], energy levels, formant shape [3], zero crossing rate (ZCR) [4], cepstral feature [5], adaptive modeling of voice signals [6], and periodicity measure [7]. More recently, VAD approaches based on pattern recognition [8] and higher order cumulants of the LPC residual [9] have been presented as new strategies. Actually, the energy difference, ZCR, and spectral difference have been applied to the VAD of the ITU-T G.729 Annex B [10]. Similar approaches were adopted to the selectable mode vocoder in 3rd Generation Partnership Project2 (3GPP2) [11], and European Telecommunications Standards Institute Adaptive Multi-Rate (AMR) VAD option 2 [12].

Meanwhile, Enqing applied a support vector machine (SVM) to VAD using the ITU-T G.729B VAD parameters and showed better performance than the G.729B VAD [13]. As a statistical learning theory, SVM-based VAD (SVM-VAD) has been shown to give good generalization performance on classification problems. Recently, Ramirez improved the SVM-VAD using two alternative feature extraction processes based on subband SNR (Signal-to-Noise ratio) estimation and long-term SNR estimation in vehicular noisy conditions [14]. It was reported that the long-term SNR feature shows better performance than the subband SNR in [14]. However, the former cannot be directly applied to telecommunication systems due to its limitation regarding real-time processing.

In this paper, we propose novel features, \textit{a posteriori} SNR, \textit{a priori} SNR [15], and predicted SNR [16], as parameters of a VAD approach based on a conventional speech statistical model. First, the \textit{a posteriori} SNR is extracted using the noise power estimated in a speech absence segment and the current input signal power. Second, a decision-directed (DD) approach that extracts smoothing SNR estimation using the previous frame SNR and the current frame SNR [15] is applied to estimate the \textit{a priori} SNR. Also, the predicted SNR, which estimates the SNR on the current frame from the previous frame using the speech absence probability, is finally adopted as an effective SNR parameter [16]. Experimental results show improvements in speech/non-speech discrimination relative to the ITU-T G.729B VAD and other recently reported VADs.

2. Support Vector Machine

The SVM proposed by Vapnik is a pattern classification scheme based on statistical learning theory [17]. A conventional learning algorithm is based on empirical risk minimization (ERM) which optimizes the performance of data in a learning set. However, the SVM is based on structural risk minimization (SRM), which minimizes the classification error probability. Generally, binary classification of the linearly separable data has many hyperplanes while separating two binary classes. The SVM makes it possible to build an optimal hyperplane that is separated without the error where the distance between the closet vectors and the hyperplane is maximal. When given training data of $N$-dimensional patterns $x_i$ and class labels $y_i$ as $(x_1, y_1), \ldots, (x_L, y_L), x \in \mathbb{R}^N, y \in \{+1, -1\}$, the equation for the hyperplane is given by

$$w \cdot x + b = 0 \quad (1)$$

where $w$ is the weighted vector to the hyperplane and $b$ is the bias. The maximizing margin can be expressed via the following optimization problem:

Minimize $\Phi(w) = \frac{1}{2}(w \cdot w)$ \quad (2)

Subject to: $\{ (w \cdot x_i) + b \} y_i \geq 1, \forall i.$ \quad (3)

The solution to this optimization problem is given by the saddle point of the Lagrange functional

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{L} \alpha_i [(w \cdot x_i) + b] y_i - 1,$$

where $\alpha_i \geq 0, \forall i.$ \quad (4)

where $\alpha_i$ are Lagrange multipliers. Taking into account the Karush-Kuhn-Tucker (KKT) condition in (4), the optimization
Taking the discrete Fourier transform (DFT) gives us

\[ Y_k(n) = X_k(n) + D_k(n) \]

where \( k \) is the frequency-bin index \( (k = 0, 1, \ldots, k - 1) \) and \( n \) is the frame index \( (n = 0, 1, \ldots) \). Assuming that speech is degraded by uncorrelated additive noise, two hypotheses that the VAD should consider for each frame are

\[ H_0 : \text{speech absent} : Y(n) = D(n) \]
\[ H_1 : \text{speech present} : Y(n) = X(n) + D(n) \]

in which \( Y(n), D(n), \) and \( X(n) \) denote the DFT coefficients at the \( n \)th frame of the noisy speech, noise, and clean speech, respectively. With the complex Gaussian probability density functions (pdf’s) assumption [19], the distributions of the noisy spectral components conditioned on both hypotheses are given by

\[ p(Y_k|H_0) = \frac{1}{\pi \lambda_{d,k}} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{d,k}} \right\} \]  

(16)

\[ p(Y_k|H_1) = \frac{1}{\pi (\lambda_{d,k} + \lambda_{x,k})} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{d,k} + \lambda_{x,k}} \right\} \]  

(17)

where \( \lambda_{d,k} \) and \( \lambda_{x,k} \) denote the variances of \( X_k(n) \) and \( D_k(n) \), respectively. The \textit{a posteriori} SNR and \textit{a priori} SNR, to be used as parameters, are then defined by

\[ \gamma_k(n) \equiv \frac{|Y_k(n)|^2}{\lambda_{d,k}(n)} \]  

(18)

\[ \eta_k(n) \equiv \frac{\lambda_{x,k}(n)}{\lambda_{d,k}(n)} \]  

(19)

First, we consider the \textit{a posteriori} SNR \( \gamma_k(n) \) as the first feature vector, which is derived by the ratio of the input signal \( Y_k(n) \) and the variance \( \lambda_{d,k}(n) \) of the noise signal \( D_k(n) \) updated in the periods of speech absence. The second feature vector is the \textit{a priori} SNR \( \eta_k(n) \) based on the well-known DD approach as follows [15]:

\[ \tilde{\eta}_k(n) \equiv \alpha \frac{|X_k(n - 1)|^2}{\lambda_{d,k}(n - 1)} + (1 - \alpha)P[\gamma_k(n) - 1], \]  

(20)

where \(|X_k(n - 1)|\) is the amplitude estimator of the \( k \)th signal spectral component in the \((n - 1)\)th analysis frame, and \( P[\cdot] \) is an operator that is defined by

\[ P[x] = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \]  

(21)

The third feature vector is the predicted SNR, which is estimated by the long-term smoothed power spectra of the background noise and speech [16]. Assuming that \( \lambda_{d,k}(n) \) and \( \lambda_{x,k}(n) \) denote the power spectra at the \( n \)th analysis frame of noise and clean speech, respectively. The estimated noise and speech power for the predicted SNR estimation method are then given by

\[ \hat{\lambda}_{d,k}(n + 1) = \zeta_d \hat{\lambda}_{d,k}(n) + (1 - \zeta_d)E[|D_k(n)|^2|Y_k(n)] \]  

\[ \hat{\lambda}_{x,k}(n + 1) = \zeta_x \hat{\lambda}_{x,k}(n) + (1 - \zeta_x)E[|X_k(n)|^2|Y_k(n)] \]  

(22)

where \( \hat{\lambda}_{d,k}(n) \) and \( \hat{\lambda}_{x,k}(n) \) are the estimates for \( \lambda_{d,k}(n) \) and \( \lambda_{x,k}(n) \). \( \zeta_d \) and \( \zeta_x \) are the parameters for smoothing under a general stationarity assumption on \( D_k(n) \) and \( X_k(n) \). Based on (22) and statistical assumptions made on \( D_k(n) \) and \( X_k(n) \), we obtain

\[ E[|D_k(n)|^2|Y_k(n)] = E[|D_k(n)|^2|Y_k(n), H_0]p(H_0|Y_k(n)) \]  

\[ + E[|D_k(n)|^2|Y_k(n), H_1]p(H_1|Y_k(n)) \]  

(23)

\[ E[|X_k(n)|^2|Y_k(n)] = E[|X_k(n)|^2|Y_k(n), H_0]p(H_0|Y_k(n)) \]  

\[ + E[|X_k(n)|^2|Y_k(n), H_1]p(H_1|Y_k(n)) \]  

(24)
where

\[ E\left[ |D_k(n)|^2 | Y_k(n), H_0 \right] = |Y_k(n)|^2 \]  

(25)

and

\[ E\left[ |X_k(n)|^2 | Y_k(n), H_0 \right] = 0 \]  

(26)

\[ E\left[ |D_k(n)|^2 | Y_k(n), H_1 \right] = \left( \frac{\hat{\xi}_k(n)}{1 + \hat{\xi}_k(n)} \right) \hat{\lambda}_{d,k}(n) + \left( \frac{1}{1 + \hat{\xi}_k(n)} \right) |\hat{D}_k(n)|^2 \]

\[ = \left( \frac{\hat{\xi}_k(n)}{1 + \hat{\xi}_k(n)} \right) \hat{\lambda}_{d,k}(n) + \left( \frac{1}{1 + \hat{\xi}_k(n)} \right)^2 |\hat{Y}_k(n)|^2 \]  

(27)

\[ E\left[ |X_k(n)|^2 | Y_k(n), H_1 \right] = \left( \frac{1}{1 + \hat{\xi}_k(n)} \right) \hat{\lambda}_{x,k}(n) + \left( \frac{\hat{\xi}_k(n)}{1 + \hat{\xi}_k(n)} \right) |\hat{X}_k(n)|^2 \]

\[ = \left( \frac{1}{1 + \hat{\xi}_k(n)} \right) \hat{\lambda}_{x,k}(n) + \left( \frac{\hat{\xi}_k(n)}{1 + \hat{\xi}_k(n)} \right)^2 |\hat{Y}_k(n)|^2 \]  

(28)

with

\[ \hat{\xi}_k(n) = \frac{\hat{\lambda}_{x,k}(n)}{\hat{\lambda}_{d,k}(n)} \]  

(29)

where \( p(H_0|Y_k(n)) \) is the speech absence probability for each frame [16]. Finally, the estimate of the predicted SNR at the \( n \)th computed frame is based on the equations given in (22) and (29).

3.2. Training

In the feature vector extraction step, speech data spoken by four male and four female speakers were sampled at 8 kHz. We added vehicular, office, and street noises to 226 sec of clean speech data by varying SNR values between 5 dB and 25 dB. Feature vectors were extracted using (18), (19), and (29) for the training procedure. We organized the 12th feature vector (= 4 a priori SNRs + 4 a posteriori SNRs + 4 predicted SNRs) from 4 frequency bands considering the frequency subbands correlation and computational efficiency. A radius basis function (RBF) kernel is chosen for training in our experiments [14].

4. Experimental results

To evaluate the performance of the proposed VAD, the receiver operating characteristics (ROC) were applied to compare its performance with other algorithms presented in reference [14]. For the evaluation, we made reference decisions on clean speech material 230 sec long by manually labeling every 10 ms frame.
For a investigation of non-speech hit rate (HR0) and false-alarm rate (FAR0 = 1-HR0), we define HR0 as the ratio of correct non-speech decisions to the hand-marked non-speech frames and FAR0 as that of false non-speech decisions to the hand-marked speech frames. The percentage of hand-marked actual speech frames was 57.1% which consisted of 44.0% voiced sound frames and 13.1% unvoiced sound frames.

To simulate noisy environments, we added vehicular and street noises to the clean speech data by 5 dB SNR and 15 dB SNR. Figs.1-4 show the ROC curves of the proposed VAD and other recently reported VADs in noisy environments. The working points of the ITU-T G.729B VAD are also included. From Fig.1, it can be clearly seen that the presented scheme has performance better than or at least comparable to that of conventional methods in most of the street noisy condition. As the SNR is increased as depicted in Fig.2, we can see that our additional methods in most of the street noisy condition. The work-

5. Conclusion

This paper has presented effective features for SVM-based VAD based on a statistical model of the speech signal. An efficient combination of relevant SNRs is adopted to the SVM to improve the performance of VAD. The performance obtained using the proposed features was considerably higher than that of conventional approaches while imposing only a small additional computational load.

6. Acknowledgements

This work was partly supported by the IT R&D program of MIC/IITA [2005-S-096-02, Development of wearable PC interface technology for the disabled] and partly supported by ETRI SoC Industry Promotion Center.

7. References

dence on a DTW and HMM recognize,” in Proc. Eu-
tection using cepstral feature,” in Proc. IEEE TELCON,
tion,” in Proc. Int. Conf. Signal Process., vol. 2, pp. 1124-
ing advanced feature extraction and SVM learning,” in Proc. Int. Conf. Spoken Language Process., pp. 1662-