Statistical model-based voice activity detection using support vector machine

Q.-H. Jo\textsuperscript{1} J.-H. Chang\textsuperscript{1} J.W. Shin\textsuperscript{2} N.S. Kim\textsuperscript{2}

\textsuperscript{1}School of Electronic and Electrical Engineering, Inha University, Incheon 402-751, South Korea
\textsuperscript{2}School of Electrical Engineering, Seoul National University, Seoul 151-742, South Korea
E-mail: changjh@inha.ac.kr

Abstract: From an investigation of a statistical model-based voice activity detection (VAD), it is discovered that a simple heuristic way like a geometric mean has been adopted for a decision rule based on the likelihood ratio (LR) test. For a successful VAD operation, the authors first review the behaviour mechanism of support vector machine (SVM) and then propose a novel technique, which employs the decision function of SVM using the LRs, while the conventional techniques perform VAD comparing the geometric mean of the LRs with a given threshold value. The proposed SVM-based VAD is compared to the conventional statistical model-based scheme, and shows better performances in various noise environments.

1 Introduction

A voice activity detector (VAD) that detects the presence of speech in a noisy signal has become an essential part of the variable-rate speech coding for bandwidth efficiency [1]. Various types of VAD algorithms have been proposed based on the energy level difference, zero-crossing rate (ZCR), and spectral difference [2]. Recently, much work to improve the performance of the VAD by incorporating a statistical model has been carried out with the simple decision rule derived from the likelihood ratio test (LRT). Specifically, the statistical model-based VAD approach originated from the Ephraim and Malah’s speech enhancement algorithm [3]. Sohn \textit{et al.} [4] applied a Gaussian statistical model to the VAD using the decision-directed (DD) method-based parameter estimation and have reported that it produced a high-detection performance despite the need for optimisation of a few relevant parameters. On the other hand, Enqing \textit{et al.} [5] applied support vector machine (SVM) to VAD based on the VAD parameters of the ITU-T G.729B and showed better performance than that of the original G.729B. More recently, Ramirez \textit{et al.} [6, 7] presented an improved SVM-based VAD using two different feature extraction methods employing sub-band signal-to-noise ratio (SNR) and the long-term SNR estimation.

In this correspondence, we propose a novel VAD technique based on SVM by treating the likelihood ratios (LRs) computed in each frequency bin as the elements of the feature vector. This is different from the conventional statistical model-based VAD in that we obtain the optimised hyperplane to minimise decision error using the SVM-based decision function rather than the simple decision rule being derived from the LRT [4, 8–13]. The performance of the proposed VAD approach is evaluated in terms of an objective measure.

2 Review of a statistical model-based VAD

We assume that a noise signal $d(t)$ is added to a speech signal $x(t)$, with their sum being denoted by $y(t)$, that is

$$y(t) = x(t) + d(t)$$

(1)

Taking the discrete Fourier transform (DFT) gives us

$$Y_k(n) = X_k(n) + D_k(n)$$

(2)

where $k$ is the frequency-bin index ($k = 0, 1, \ldots, L - 1$) and $n$ is the frame index ($n = 0, 1, \ldots$). Assuming that speech is degraded by uncorrelated additive noise, two hypotheses $H_0$ and $H_1$ that indicates speech absence and presence in the noisy spectral component $Y_k(n)$, respectively, for each frame
are given by [4, 8–11]

\[ H_0: \text{speech absent}; Y_j(n) = D_j(n) \]  
\[ H_1: \text{speech present}; Y_j(n) = X_j(n) + D_j(n) \]  

With the complex Gaussian probability density functions (pdf’s) assumption [4], the distributions of the noisy spectral components are given by

\[ p(Y_j|H_0) = \frac{1}{2\pi \lambda_{d,j}} \exp \left\{ -\frac{|Y_j|^2}{\lambda_{d,j}} \right\} \]  
\[ p(Y_j|H_1) = \frac{1}{2\pi (\lambda_{d,j} + \lambda_a,j)} \exp \left\{ -\frac{|Y_j|^2}{\lambda_{d,j} + \lambda_a,j} \right\} \]

where \( \lambda_{a,j} \) and \( \lambda_{d,j} \) denote the variances of \( X_j \) and \( D_j \), respectively. The LR of the \( k \)th frequency band is derived as

\[ \Lambda_k = \frac{p(Y_k|H_1)}{p(Y_k|H_0)} = \frac{1}{1 + \xi_k} \exp \left\{ \frac{\gamma_k \xi_k}{1 + \xi_k} \right\} \]

where \( \xi_k = \lambda_{a,k}/\lambda_{d,k} \) and \( \gamma_k = Y_k/\lambda_{d,k} \) are called the a priori and a posteriori SNR’s, respectively [4]. The a posteriori SNR \( \gamma_k(n) \) is obtained by minimum mean square error (MMSE) [4], and \( P[x] = x \) if \( x \geq 0 \) and \( P[x] = 0 \) otherwise.

In previous literature, the decision statistic of VAD is established from the geometric mean of the LRs over all frequency bands as follows [4, 8–13]

\[ \log \Lambda(n) = \frac{1}{L} \sum_{j=0}^{L-1} \log \Lambda_j(n) \geq \eta \]

where \( \eta \) is a threshold for speech detection. As depicted in Fig. 1, which shows the decision flowchart of a statistical model-based VAD, an input frame is classified as voice active if the geometric mean of the LRs is greater than the threshold and voice-inactive otherwise.

### 3 Proposed SVM-based VAD

We briefly review the notion of SVM that was proposed in [14, 15]. Generally, binary classification of the linearly separable data has many hyperplanes separating two binary classes. It is known that SVM makes it possible to build an optimal hyperplane such that the distance between the closest vectors and the hyperplane is maximal. Given training data consisting of

- Input Speech \( Y_k(n) \) (\( k = 0, \ldots, L - 1 \))
- Estimation of \( \lambda_{d,k}(\tau) \)
- Estimation of \( \gamma_k(n), \xi_k(n) \)
- Computation of \( \Lambda_k(n) \)
- Computation of the Geometric Mean

\[ \log \Lambda(n) = \frac{1}{L} \sum_{j=0}^{L-1} \log \Lambda_j(n) \geq \eta \]

\[ \phi(\omega) = \frac{1}{2} (\omega \cdot \omega) \]

Subject to: \( [(\omega \cdot x_i) + b]z_i \geq 1, \forall i \)  

The solution to this optimisation problem is given by the saddle point of the Lagrange functional as follows

\[ L(\omega, b, \alpha_i) = \frac{1}{2} (\omega \cdot \omega) - \sum_{i=1}^{L} \alpha_i [(\omega \cdot x_i) + b]z_i - 1, \alpha_i \geq 0, \forall i \]  

where \( \alpha_i \) are Lagrange multipliers. Taking into account the Karush–Kuhn–Tucker (KKT) condition in (13), the optimisation problem can be converted to a dual maximisation problem [14]

Maximise: \( Q(\alpha) = \sum_{i=1}^{L} \alpha_i - \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j z_i z_j (x_i \cdot x_j) \)

Subject to: \( \sum_{i=1}^{L} \alpha_i z_i = 0, \alpha_i \geq 0, \forall i \)
In practice, it is needed as an approach to make the training of SVMs on problems, to decompose the problem into a series of smaller optimisation tasks. We adopt an approach called sequential minimal optimisation (SMO) for the SVM training [14]. Substituting \( \alpha^*_i \) that maximises \( Q(\alpha) \) into (16) and (17) derived from the KKT conditions, we obtain the optimal weight vector \( \mathbf{w}^* \) and the bias \( b^* \) as follows:

\[
\mathbf{w}^* = \sum_{i=1}^{M} \alpha^*_i z_i x_i^* \quad (16)
\]

\[
b^* = 1 - (\mathbf{w}^* \cdot \mathbf{x}_i^*) \quad z_i = 1 \quad (17)
\]

where \( x_i^* \) represents the support vector and \( M \) denotes the number of the support vectors. In [8], it was reported that the LR is the novel feature in the VAD problem in terms of the consistency (It is known that the DD method of (8) provides smoother estimates of the a priori SNR and consequently reduces the unwanted fluctuation of the estimated LRs during noise-only periods [11].) and the tracking performance, especially for less stationary noise. For this reason, for the decision rule of the VAD, we present a novel technique to incorporate the LRs as elements of feature vector characterised by SVM. Let \( \Lambda(n) = [\Lambda_0(n), \Lambda_1(n), \ldots, \Lambda_{L-1}(n)]^T \) be the LRs obtained by (7) and \( \Lambda^*_m \) be the \( m \)th support vector of LRs obtained by training. Then

\[
f(\Lambda(n)) = (\mathbf{w}^* \cdot \Lambda(n)) + b^* = \sum_{i=1}^{M} \alpha^*_i z_i (\Lambda^*_i \cdot \Lambda(n)) + b^* \quad \text{if} \quad \frac{H_1}{H_0} \mathbf{f} \quad (18)
\]

Comparing (18) and a given threshold value reveals the presented SVM-based decision statistic, which is analogous to (9). It can be seen that we derive a novel decision statistic by the use of the dot product between the given LR vector and the support vectors. Fig. 2 shows an overall structure of the proposed VAD built on the basis of the SVM paradigm.

On the other hand, from the extensive scatter plot analysis illustrated in Fig. 3 as a representative case, it was found that the proposed LR cannot be separated by a linear function since there is considerable class overlap in the input space. In order to consider nonlinear input space, the various kernel function \( K \) has been addressed [14] rather than the linear kernel such that

\[
K(\Lambda^*_i, \Lambda) = \Phi(\Lambda^*_i) \cdot \Phi(\Lambda) \quad (19)
\]

Once the kernel function is specified as in (19), the decision statistic finally results in the following form

\[
f(\Lambda(n)) = \sum_{i=1}^{M} \alpha^*_i z_i K(\Lambda^*_i, \Lambda(n)) + b^* \quad (20)
\]

In our approach, the radius basis function (RBF) kernel is incorporated among the various kernels due to the superior performance [5]

\[
K(\Lambda^*_i, \Lambda) = \exp\left(-\frac{||\Lambda^*_i - \Lambda||^2}{\sigma^2}\right) \quad (21)
\]

where \( \sigma \) is the kernel width. In practice, this would be a major reason that the proposed LRs-based SVM significantly outperforms the conventional geometric mean of the LRs for the VAD.

4 Experiments and results

Performance of the proposed VAD approach was evaluated on the NTT database that consists of a number of speech...
samples [10, 16]. Let us explain the NTT database for the training part. All the training data used for building support vectors were recorded in a quiet environment. Utterances of the NTT database spoken by from four male speakers and four female speakers were used to construct 226 s long speech data. It can be pointed that there is consensus on the use of the NTT database since the NTT corpus is faithfully designed for evaluating the performance of the speech coder [17, 18].

Each utterance (8 s) consisted of two different meaningful sentences with silence periods was concatenated [16]. For training, we made a reference decisions on the clean speech materials by labelling manually at every 10 ms frame. The proportions of voiced, unvoiced and silence frames of the training materials are 45.6, 13.7 and 40.7%, respectively. After making reference decisions, we added vehicular, babble, street and white noises to whole signal and SNR values of each noise are 5, 10 and 20 dB. In our experiments, the input signal was sampled at 8 kHz and the analysis window size was 10 ms with 3 ms overlap. Each frame of the windowed signal was transformed to its corresponding spectrum through a 128-point DFT after zero-padding. For the RBF kernel, the kernel width $\sigma$ was set to 1.0.

For the test, we made reference decisions on a different clean speech material 349 s long by labelling manually at every 10 ms frame by concatenating different speech materials of the NTT database. The percentage of hand-marked active speech frames was 56.7%, which consisted of 44.0% voiced sound frames and 12.7% unvoiced sound frames. To simulate noisy conditions, vehicular, babble, street and white noises are added to the clean speech data by 5 dB SNR. To evaluate the performance of the proposed VAD compared with the previous statistical model-based VAD with the geometric mean, we investigated the receiver operating characteristics (ROC’s), which shows the trade-off characteristic between the speech detection and false-alarm probabilities ($P_d$ and $P_f$). We define $P_d$ as the ratio of correct speech decisions to the hand-marked speech frames and $P_f$ as that of false speech decisions to the hand-marked non-speech frames.

For fair comparison, we do not consider any hangover scheme, as this can be added after the design of the decision rule. Figs. 4–7 show the ROC curves of the proposed SVM-based VAD with the linear and RBF kernel and the conventional VAD methods by Sohn [4]. Also, performances of other SVM-based approaches by Enqing et al. [5] and Ramirez et al. [7] with the ITU-T G.729 Annex B (G.729B) are plotted for the purpose of gaining relative performance evaluation. The previous methods are achieved by adopting similar parameter settings such that the frame size is 10 ms and the kernel width $\sigma$ is 1.0.

In all testing conditions, the proposed SVM-based VAD using the linear kernel yielded higher performance over all than the Sohn’s method based on the geometric mean [4]. It is evident from the result that SVM is a very desirable way to establish the decision rule for the statistical model-based VAD. Especially, we observe that the RBF kernel in most of the tested conditions significantly improved the performance of the proposed SVM-based VAD while the polynomial-based VAD did not give us a consistent performance improvement. Specifically, in the case of the vehicular noise depicted in Fig. 4, the proposed technique outperformed other approaches except for $P_f < 0.04$ while the Ramirez’s algorithm is the next best one. Fig. 5 illustrates the proposed approach with the RBF kernel significantly outperformed other methods under the babble noise condition. This time, the detection accuracy of the method produced by Ramirez was much poorer than that of other methods. For the highly non-stationary street noise type in Fig. 6, again, the best performance was achieved with the proposed scheme with the RBF kernel. It is noted that the difference in recognition performance between the RBF kernel based-method and the linear kernel-based algorithm is much bigger compared to other noise types. As for the white noise (Fig. 7), the
The proposed approach with the RBF kernel consistently outperformed other methods except for the low false-alarm probabilities ($P_f < 0.08$). In addition, The Ramirez’s method was found to be ineffective in improving the detection performance for white type of noise. Finally, we show the superiority of the presented SVM-based VAD to establish the decision rule for the statistical model-based VAD. Also, the induced RBF kernel is an effective way to improve the performance of the proposed algorithm. This phenomenon is attributable to the fact that the nonlinear drawback of the input data (LRs) is successfully resolved by the RBF kernel, which makes the input data linearly separable. Summarising the overall results, we can see that the proposed technique performs well except for very low false-alarm probability regions ($P_f < 0.08$) in all cases. In this regard, the proposed technique can be considered practically acceptable, since the performance of detection (or miss) is more important (or critical) than that of the false-alarm in speech coding and the usual range for the false-alarm probabilities was 0.07 to 0.28 (Average $= 0.17$) when we investigated the performance of the practical methods such G.729B VAD, ETSI adaptive multi-rate (AMR) VAD and so on in the literature [4, 11, 18].

In addition, we evaluated the performance of the proposed VAD algorithm by fixing the threshold. The operating point of the VAD was fixed to make the false alarm probability $P_f$ of the proposed VAD slightly less than or equal to that of the conventional approaches [4, 11, 18]. As summarised in Table 1 for various SNRs, the results confirm that the proposed method with the RBF kernel outperforms other approaches in terms of the speech detection error probability ($P_e$), where both the false alarms and missing errors are incorporated.

### Table 1 Comparison of speech detection error probability ($P_e = (1 - P_d) + P_f$) under the various noise condition

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR, dB</th>
<th>Proposed (RBF), %</th>
<th>Proposed (linear), %</th>
<th>Ramirez (RBF), %</th>
<th>Enqing (RBF), %</th>
<th>Sohn, %</th>
<th>G.729B, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>vehicle</td>
<td>5</td>
<td>14.4</td>
<td>17.1</td>
<td>15.4</td>
<td>24.1</td>
<td>22.8</td>
<td>54.7</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>11.0</td>
<td>12.0</td>
<td>11.7</td>
<td>18.8</td>
<td>12.0</td>
<td>45.7</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>9.4</td>
<td>10.6</td>
<td>11.1</td>
<td>13.0</td>
<td>10.4</td>
<td>34.9</td>
</tr>
<tr>
<td>babble</td>
<td>5</td>
<td>36.3</td>
<td>44.7</td>
<td>55.2</td>
<td>50.9</td>
<td>46.2</td>
<td>49.6</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>35.8</td>
<td>39.4</td>
<td>33.7</td>
<td>38.4</td>
<td>41.7</td>
<td>44.0</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20.9</td>
<td>21.9</td>
<td>24.0</td>
<td>22.7</td>
<td>29.3</td>
<td>31.4</td>
</tr>
<tr>
<td>street</td>
<td>5</td>
<td>24.9</td>
<td>29.0</td>
<td>33.3</td>
<td>37.1</td>
<td>34.3</td>
<td>43.7</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>20.7</td>
<td>24.8</td>
<td>23.5</td>
<td>27.4</td>
<td>34.2</td>
<td>35.1</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>14.4</td>
<td>15.2</td>
<td>15.7</td>
<td>16.4</td>
<td>19.9</td>
<td>22.1</td>
</tr>
<tr>
<td>white</td>
<td>5</td>
<td>30.7</td>
<td>32.8</td>
<td>60.9</td>
<td>28.8</td>
<td>34.8</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>22.7</td>
<td>28.0</td>
<td>31.3</td>
<td>20.4</td>
<td>31.6</td>
<td>42.4</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>21.6</td>
<td>26.9</td>
<td>14.9</td>
<td>13.1</td>
<td>22.7</td>
<td>27.3</td>
</tr>
</tbody>
</table>
5 Conclusions

In this study, we have presented effective VAD using SVM, which 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. The RBF kernel as well as the linear kernel is also adopted to consider nonlinear properties of the input data. The proposed approach has been found to significantly improve the conventional statistical model-based VAD under various noise conditions.

6 Acknowledgments

This work was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Centre) support programme supervised by the IITA (Institute for Information Technology Advancement) (IT7A-2008-C1090-0902-0010) and this work was partly supported by the IT R&D program of MKE/IITA [2008-F-045-01].

7 References


