Abstract: In this paper, we propose a novel approach to improve a statistical model-based voice activity detection (VAD) method based on a modified conditional maximum a posteriori (MAP) criterion incorporating the spectral gradient scheme. The proposed conditional MAP incorporates not only the voice activity decision in the previous frame as in Ref. [1] but also the spectral gradient of the observed spectra between the current frame and the past frames to efficiently exploit the inter-frame correlation of voice activity. As a result, the proposed VAD leads to six separate thresholds to be adaptively determined in the likelihood ratio test (LRT) depending on both the previous VAD result and the estimated spectral gradient parameter. Experimental results demonstrate that the proposed approach yields better results compared to those of the previous conditional MAP-based method.

Keywords: Voice activity detection; Spectral gradient; Conditional MAP; Likelihood ratio test

1 Introduction

Robust speech processing in adverse environments has been an important issue in recent years. Indeed, voice activity detection (VAD), which determines whether the input data is speech or noise, is a crucial component of speech processing systems [1-4]. Among the various VAD approaches, we focus on the likelihood ratio test (LRT), which uses the distributions of both speech and noise. Considering the statistical model of speech, the distributions of both noisy speech and noise are assumed to follow a statistical distribution such as Gaussian, Laplacian or generalized gamma [2-4]. Based on the assumed statistical model, the LRT is inherently established based on the maximum a posteriori (MAP) criterion, which chooses the hypothesis (speech or noise) with the higher probability. Recently, Shin et al. [1] used inter-frame correlation of speech signals. In contrast, conventional methods based on MAP depend on the independence of each frame. Specifically, this is achieved by incorporating a simple but rigorous rule such as a conditional MAP (CMAP) criterion, which is conditioned not only on the data of the current frame, but also the voice activity decision of the previous frame. This algorithm results in an adaptive decision threshold for the LRT based on the result of the voice activity in the previous frame. This method proved efficient in exploiting the inter-frame correlation. However, this approach does not fully consider spectral variation.

In this paper, we propose a novel technique for LRT by incorporating the spectral gradient in the decision criterion, in addition to the aforementioned result of voice activity in the previous frame. Consideration of the spectral gradient of the CMAP enables more exact detection since we can consider the spectral variation in terms of increases, decreases and sustenance of the spectra associated with the voice activity. As a result, the decision thresholds of the LRT have six different values depending on the status of voice activity in the previous frame and the spectral gradient between the current power spectrum and the averaged long-term power spectrum. Based on comparison of a number of experiments on VAD, the proposed approach shows better performance than the previous algorithm proposed by Shin et al. [1] does.

The paper is organized as follows. In Section 2, we briefly review the CMAP-based VAD. In Section 3, we present the spectral gradient scheme and apply it to the CMAP for VAD performance improvement. Finally, an objective evaluation of the previous method and our approach is performed in Section 4.

2 Review of CMAP-Based VAD

In the time domain, it is assumed that the noise signal $d(t)$ is added to the clean speech signal $x(t)$, with their sum being denoted by $y(t)$, which is called the noisy speech signal. These variables are transformed by the discrete Fourier transform (DFT) as follows:

$$Y(n) = X(n) + D(n),$$

where

$$Y(n) = [Y(0, n), Y(1, n), ..., Y(M-1, n)]$$

and

$$X(n) = [X(0, n), X(1, n), ..., X(M-1, n)],$$

and

$$D(n) = [D(0, n), D(1, n), ..., D(M-1, n)]$$

at the $n$th frame. Assuming that speech is degraded by uncorrelated additive noise, two hypotheses, $H_s(n)$ and $H_n(n)$, indicate speech absence and presence in the noisy spectral component, respectively.

$$H_s(n) : Y(k, n) = D(k, n)$$

and

$$H_n(n) : Y(k, n) = X(k, n) + D(k, n).$$
With the assumption of the Gaussian probability density function [7], the distributions of the noisy spectral components conditioned on both hypotheses are given by

\[
p(Y(k,n) | H_s(n)) = \frac{1}{\pi \lambda_s(k,n)} \exp \left\{ - \frac{\| Y(k,n) \|^2}{\lambda_s(k,n)} \right\}
\]

(4)

\[
p(Y(k,n) | H_i(n)) = \frac{1}{\pi \lambda_i(k,n) + \lambda_s(k,n)} \exp \left\{ - \frac{\| Y(k,n) \|^2}{\lambda_s(k,n)} \right\},
\]

(5)

where \( \lambda_s(k,n) \) and \( \lambda_i(k,n) \) denote the variances in noise and speech for each frequency bin, respectively. The LR of the 8th frequency band is achieved by

\[
\Lambda(k,n) = \frac{p(Y(k,n) | H_i(n))}{p(Y(k,n) | H_s(n))} = \frac{1 + \xi(k,n)}{1 + \xi(k,n)} \exp \left\{ \gamma(k,n) \xi(k,n) \right\},
\]

(6)

where \( \xi(k,n) \) and \( \gamma(k,n) \) represent the a priori signal-to-noise ratio (SNR) and the a posteriori SNR, respectively [7]. The a posteriori SNR \( \gamma(k,n) \) is estimated using the estimator of \( \hat{\lambda}(k,n) \) and the a priori SNR \( \hat{\xi}(k,n) \) is estimated by the well-known decision-directed (DD) method as follows [5]:

\[
\hat{\xi}(k,n) = \alpha \hat{X}(k,n-1) + (1-\alpha)P[y(k,n)-1]
\]

(7)

where \( \hat{X}(k,n-1) \) is the speech spectral amplitude estimate of the previous frame obtained using the minimum mean-square error (MMSE) estimator [7]. Also, \( \alpha \) is a weight that is usually chosen in the range (0.95, 0.99) [5], and the function \( P[x]=x \) if \( x \geq 0 \) and \( P[x]=0 \) otherwise. The final decision in conventional statistical model-based VADs is reached using the geometric mean of the LRs computed for the individual frequency bins [5-10] and is obtained by

\[
\Lambda(n) = \frac{1}{M} \sum_{k=1}^{M} \log \Lambda(k,n) > \eta, \quad \eta < \eta_s
\]

(8)

where an input frame is classified as speech if the geometric mean of the LRs is greater than a certain threshold value \( \eta \) and as non-speech otherwise.

On the other hand, the previous CMAP-based VAD originates from the conventional MAP according to the following decision rule:

\[
P(H_i(n) | Y(n)) > 1 - \eta_s,
\]

(9)

where \( H(n) \) denotes the correct hypothesis in the \( n \)th frame. This rule is changed to the following criterion in the LRT such that

\[
\frac{P(H_i(n) | Y(n))}{P(H_s(n) | Y(n))} > \frac{\alpha}{P(H_i(n) = H_i(n))} \quad \frac{P(H_s(n) = H_s(n))}{P(H_i(n) = H_i(n))} < \frac{\alpha}{P(H_s(n) = H_s(n))},
\]

(10)

where \( \alpha \geq 1 \).

Actually, Shin et al. proposed a way to incorporate the inter-frame correlation of the voice activity into the MAP criterion. More specifically, the a posteriori probability \( P(H(n) | Y(n)) \) is conditioned on both the current observation \( Y(n) \) and the decision in the previous frame. From that, \( P(H(n) | Y(n),H(n-1)) \) is derived. This implies that

\[
\frac{P(H_i(n) | Y(n),H(n-1))}{P(H_s(n) | Y(n),H(n-1))} > \frac{\alpha}{P(H_i(n) = H_i(n))} \quad \frac{P(H_s(n) = H_s(n))}{P(H_i(n) = H_i(n))} < \frac{\alpha}{P(H_s(n) = H_s(n))},
\]

(11)

where \( \alpha \) is the threshold. The upper criterion could be reexpressed such that [1]

\[
\frac{P(Y(n) | H(n) = H_i(n),H(n-1))}{P(Y(n) | H(n) = H_s(n),H(n-1))} > \frac{\alpha}{P(H_i(n) = H_i(n))} \quad \frac{P(H_s(n) = H_s(n))}{P(H_i(n) = H_i(n))} < \frac{\alpha}{P(H_s(n) = H_s(n))}, \quad i = 0,1,
\]

(12)

It is noted that the likelihoods \( P(Y(n) | H(n) = H_i(n),H(n-1)) \) and \( P(Y(n) | H(n) = H_s(n),H(n-1)) \) could be simplified for the dominant contribution of the distribution of \( Y(n) \) in the current frame as follows:

\[
\frac{P(Y(n) | H(n) = H_i(n))}{P(Y(n) | H(n) = H_s(n))} > \frac{\alpha}{P(H_i(n) = H_i(n))} \quad \frac{P(H_s(n) = H_s(n))}{P(H_i(n) = H_i(n))} < \frac{\alpha}{P(H_s(n) = H_s(n))}, \quad i = 0,1,
\]

(13)

where different thresholds are used depending on the speech activity in the previous frame \( (n-1) \).

3 Proposed method based on spectral gradient

The previous section shows that the method of Shin et al. derives two separate thresholds for the decision of speech activity in the previous frame. Here, we propose a way to incorporate the spectral gradient into the
conditional term in the CMAP, which considers the time-varying spectral change. For example, in the case of onset regions, the spectral power eventually increases, which can be relevant for detecting voice activity. For the rigorous algorithm, we first define the spectral gradient of each frame based on the difference between the current power spectrum and the average long-term power spectrum given by

$$\Delta(n) = \sum_{k=1}^{n} |Y(k,n)| - E(k,n)$$

(14)

where denotes the average long-term spectral estimate during previous frames, given by [11]

$$E(k,n) = \beta E(k,n-1) + (1-\beta) |Y(k,n)|$$

(15)

where $\beta = 0.8$ is a weight. Note that $E(k,1)$ is determined by $|Y(k,1)|$. Using $\Delta(n)$, we can categorize three cases by comparing the given threshold $\eta$

$$G(n) = \begin{cases} G_i, & \Delta(n) > \gamma, \\ G_{i1}, & -\gamma < \Delta(n) \leq \gamma, \\ G_{i2}, & \Delta(n) \leq -\gamma. \end{cases}$$

(16)

In this equation, $G_i$ implies an ascending spectral gradient since the current power sufficiently exceeds the average long-term power. Also, in the case of $G_{i1}$, the spectral gradient is descending.

In a similar way as in Ref. [1], the upper criterion is the basis for the following likelihood ratio test (LRT) using Bayes’ rule such that

$$P(H(n) = H_i | H(n-1) = H_j, G(n) = G_i) > \alpha, \quad i = 0,1 \text{ and } j = -1,0,1,$$

$$H_o$$

(17)

$P(H(n) = H_i | H(n-1) = H_j, G(n) = G_i)$

(18)

where $P(H(n))$ denotes the correct hypothesis at $n$th frame [7]. Based on this motivation, we propose the novel voice activity decision rule, which is analogous to Eq. (11), through the incorporation of $G(n)$:

$$P(H(n) = H_i | Y(n), H(n-1) = H_j, G(n) = G_i) > P(H(n) = H_i | Y(n), H(n-1) = H_j, G(n) = G_j),$$

$$H_i, \quad \alpha, \quad i = 0,1 \text{ and } j = -1,0,1.$$}

$$H_o$$

(19)

In Figure 2, we show the waveform and the CMAP. The waveform in (a) represents the input file (car noise, SNR=10 dB), and the CMAP in (d) shows the threshold of the proposed method. The plots in (b) and (c) show the manual VAD (silence=0, unvoiced=1, voiced=2)

![Figure 1](a) Waveform of the test file (street noise, SNR=10 dB); (b) Plot of the $|Y(k,n)|$ and $E(k,n)$

![Figure 2](a) Waveform of the test file (street noise, SNR=10 dB); (b) Manual VAD (silence=0, unvoiced=1, voiced=2); (c) Threshold of the proposed method; (d) Threshold of the CMAP
4 Experiments and results

Conventional methods and the proposed method were evaluated quantitatively in various noise environments. For the test material, 456 s of speech was recorded by four males and four females, and then it was sampled at 8 kHz. To evaluate the performance, we first made reference decisions on the clean speech material by labeling every 10 ms. The proportion of hand-marked speech frames was 58.2 % and consisted of 44.5 % voiced sounds and 13.4 % unvoiced sounds. To simulate various noise environments, car, office, and white noises were added to the clean speech data, resulting in SNRs of 5, 10, and 15 dB. The thresholds were experimentally set such that

\[ \psi_{00} = 26.2, \psi_{01} = 25, \psi_{10} = 26.8, \psi_{11} = 22.7, \psi_{12} = 24.7, \text{ and } \psi_{13} = 18. \]

Also, the conditions of the initial value \( \psi_{0} \) were set to \( \psi_{00} \) because we assumed that only noise existed in the initial several frames. Also, we further added the noise types such as street, babble, and subway of the widely labeled database, i.e., Aurora to show that the results are not dependent on the aforementioned training data set. Note that the Aurora database is a collection of short sentences containing digits. By concatenating the sentences, we made the additional test material of 140 s and the made reference decisions by hand labeling again.

Table I including \( P_e \) (probability of error), \( P_m \) (probability of miss), an \( P_f \) (false alarm probability) shows comparative results of the first-order CMAP-based method and the proposed approach. Indeed, to aid in the repeatability of the results, the standardized VAD, ITU-T G.729 Annex B [12], was included. Also, the results of the well-known standard VAD algorithm such as ETSI AMR VAD option 2 [13] to show the performance difference is practically acceptable. At first, from the results, it is evident that the proposed VAD algorithm shows better performance in most environmental conditions than the previously reported VAD methods do including the first-order CMAP [1], the G.729 Annex B and the AMR VAD. The proposed approach has also been found to exhibit superior performance with non-stationary noise. The test results confirm that the proposed method effectively enhances the performance of the statistical model-based VAD with rapidly changing noise types.

<table>
<thead>
<tr>
<th>Environments</th>
<th>G. 729B</th>
<th>AMR</th>
<th>CMAP [1]</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>SNR (dB)</td>
<td>( P_e )</td>
<td>( P_m )</td>
<td>( P_f )</td>
</tr>
<tr>
<td>Car</td>
<td>5</td>
<td>27.52</td>
<td>13.00</td>
<td>47.72</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>23.48</td>
<td>10.03</td>
<td>42.20</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>19.79</td>
<td>7.14</td>
<td>37.40</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>22.74</td>
<td>22.29</td>
<td>23.37</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>19.28</td>
<td>17.30</td>
<td>22.04</td>
</tr>
<tr>
<td>White</td>
<td>5</td>
<td>25.19</td>
<td>42.80</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>17.35</td>
<td>29.14</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>10.94</td>
<td>18.04</td>
<td>1.06</td>
</tr>
<tr>
<td>Street</td>
<td>5</td>
<td>27.20</td>
<td>15.18</td>
<td>39.37</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>24.50</td>
<td>9.27</td>
<td>39.92</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>21.22</td>
<td>5.96</td>
<td>36.67</td>
</tr>
<tr>
<td>Babble</td>
<td>5</td>
<td>25.28</td>
<td>14.82</td>
<td>35.86</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>24.17</td>
<td>8.15</td>
<td>40.39</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>23.94</td>
<td>5.59</td>
<td>42.53</td>
</tr>
<tr>
<td>Subway</td>
<td>5</td>
<td>25.29</td>
<td>17.33</td>
<td>33.89</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>21.08</td>
<td>11.92</td>
<td>30.99</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>19.51</td>
<td>7.30</td>
<td>32.72</td>
</tr>
</tbody>
</table>
5 Conclusions

In this paper, we have proposed a novel VAD technique based on the CMAP algorithm in which the spectral gradient is incorporated for a robust VAD decision. The proposed CMAP criterion determines a hypothesis with maximal conditional probability given the current observation, the voice activity in the previous frame and the condition derived from the spectral gradient. Tracking the spectral gradient involves determination of the difference between the current power and the averaged long-term power spectra. The proposed approach yields better performance than the conventional method in various noise environments.

Acknowledgements

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (MEST) (NRF 2011-0009182). And, this work was supported by the IT R&D program of MKE/KEIT. [2009-S-036-01, Development of New Virtual Machine Specification and Technology].

References