Toward Detecting Voice Activity Employing Soft Decision in Second-order Conditional MAP

Sang-Kyun Kim, Jae-Hun Choi, Sang-Ick Kang, Ji-Hyun Song, Joon-Hyuk Chang

School of Electronic Engineering
Inha University, Incheon, Korea
{skkim,jhchoi,sikang,jhsong}@dsp.inha.ac.kr, changjh@inha.ac.kr

Abstract

In this paper, we propose a novel approach to statistical model-based voice activity detection (VAD) that incorporates a second-order conditional maximum a posteriori (MAP) criterion. As a technical improvement for the first-order conditional MAP criterion in [1], we consider both the current observation and the voice activity decision in the previous two frames to take full consideration of the inter-frame correlation of voice activity. The soft decision scheme is incorporated to result in time-varying thresholds for further performance improvement. Experimental results show that the proposed algorithm outperforms the conventional CMAP-based VAD technique under various experimental conditions.

Index Terms: Voice Activity Detection, Second-order Conditional MAP, Soft Decision, Likelihood Ratio Test

1. Introduction

A voice activity detector (VAD) referring to a mean of a finite state machine with two states, “speech absent” and “speech present” is the most important part of discontinuous transmission (DTX), especially in mobile voice over internet protocol (VoIP) systems. Most traditional algorithms are based on the linear prediction coding (LPC) parameter, cepstral feature, and the periodicity measure [2]. More recently, as a novel strategy VADs based on likelihood ratio test (LRT) employing a statistical model have been proposed and shown to have superior performance in spite of the need for optimization of a few relevant parameters [3]-[7]. Note that the traditional LRT is eventually based on the maximum a posteriori (MAP) criterion, which selects the hypothesis having the maximum probability given a current observation. The previous approach by Shin et al. [1] considered the inter-frame correlation of the speech signal since the conventional approach based on the MAP characterizes each frame separately. Actually, this has been done by incorporating a simple conditional MAP (CMAP) criterion that chooses the hypothesis with the higher probability conditioned on the current data and the voice activity decision in the previous frame. This can be considered to be relevant because the decision threshold for the LRT has two different values adaptively according to the status of voice activity in the previous frame. On the other hand, Ramirez et al. [6] proposed an algorithm to incorporate long-term speech information to the LRT method. However, this approach has an inherent delay that is not adequate for the real-time speech communication scenario.

In this paper, we propose a novel technique for the LRT-based VAD based on the second-order CMAP, in which the LRT decision is performed in a frame using both the current observation and voice activity of the previous two frames. This is because taking higher-order into the CMAP into consideration enables more exact detection in the VAD. As a result, the idea behind this technique results in four different thresholds rather than two thresholds in [1], which can provide an additional freedom in decisions. From a number of experiments on VAD, we can see that the proposed approach shows better performance than the algorithm proposed by Shin et al. [1].

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. They are transformed by the discrete Fourier transform (DFT) as follows:

\[ Y(n) = X(n) + D(n) \] (1)

where

\[ Y(n) = [Y(0, n), Y(1, n), ..., Y(M - 1, n)] \]
\[ X(n) = [X(0, n), X(1, n), ..., X(M - 1, n)] \]
\[ D(n) = [D(0, n), D(1, n), ..., D(M - 1, n)] \]

At the nth frame, respectively. Assuming that speech is degraded by uncorrelated additive noise, two hypotheses, \( H_0(n) \) and \( H_1(n) \), indicate speech absence and presence in the noisy spectral component

\[ H_0(n): \: Y(k, n) = D(k, n) \] (2)
\[ H_1(n): \: Y(k, n) = X(k, n) + D(k, n) \] (3)

With the Gaussian probability density functions (pdf’s) assumption [7], the distributions of the noisy spectral components conditioned on both hypotheses are given by

\[ p(Y(k, n)|H_0(n)) = \frac{1}{\pi \lambda_d(k, n)} \exp \left\{ - \frac{|Y(k, n)|^2}{\lambda_d(k, n)} \right\} \] (4)
\[ p(Y(k, n)|H_1(n)) = \frac{1}{\pi (\lambda_d(k, n) + \lambda_s(k, n))} \exp \left\{ - \frac{|Y(k, n)|^2}{\lambda_d(k, n) + \lambda_s(k, n)} \right\} \] (5)

where \( \lambda_d(k, n) \) and \( \lambda_s(k, n) \) denote the variances of noise and speech for the individual frequency band, respectively. The LR of the kth frequency band is given by

\[ \Lambda(k, n) \equiv \frac{p(Y(k, n)|H_1(n))}{p(Y(k, n)|H_0(n))} \]
\[ = \frac{1}{1 + \xi(k, n)} \exp \left\{ \gamma(k, n) \xi(k, n) \right\} \] (6)

where \( \gamma(k, n) = \lambda_s(k, n)/\lambda_d(k, n) \) and \( \xi(k, n) \) denote 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 $\lambda(k, n)$ and the a priori SNR $\xi(k, n)$ is estimated by the well-known decision directed (DD) method as follows [5]:

$$\tilde{X}(k, n) = \alpha |\tilde{X}(k, n-1)|^2 \lambda(k, n-1) + (1 - \alpha)P[\gamma(k, n) - 1]$$

(7)

where $|\tilde{X}(k, n-1)|^2$ 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 determined 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 the conventional statistical model-based VADs has been achieved by the geometric mean of the LRs computed for the individual frequency bins [3]-[7] and is obtained by

$$\Lambda(n) = \frac{1}{M} \sum_{k=0}^{M-1} \log \Lambda(k, n) \geq \eta$$

(8)

where $\Lambda(n)$ denotes the a posteriori probability in the $n$th frame. This rule is changed to the following criterion in the LRT such that

$$P(Y(n)|H(n) = H_1(n)) P(H(n) = H_0(n)) \geq \frac{1}{\alpha} P(H(n) = H_1(n))$$

(9)

where $\alpha \geq 1$ [7].

This time, Shin et al. proposed a way to incorporate the inter-frame correlation of the voice activity into the MAP criterion. Specifically, the a posteriori probability $P(H(n)|Y(n))$ conditioned on not only the current observation $Y(n)$ but also the decision in the previous frame, $P(H(n)|Y(n), H(n-1))$. Then, it implies that

$$P(H(n) = H_1(n)|Y(n), H(n-1) = H_1) \geq \alpha, \quad i = 0, 1$$

(10)

$$P(H(n) = H_1(n)|Y(n), H(n-1) = H_2) \geq \alpha$$

(11)

where $\alpha$ is the threshold. Upper criterion could be expressed such that [1]

$$P(Y(n)|H(n) = H_1(n), H(n-1) = H_1) \geq \alpha$$

$$P(Y(n)|H(n) = H_1(n), H(n-1) = H_2) \geq \alpha$$

(12)

$$P(Y(n)|H(n) = H_2(n), H(n-1) = H_1) \geq \alpha$$

$$P(Y(n)|H(n) = H_0(n), H(n-1) = H_1) \geq \alpha$$

(13)

3. Proposed Algorithm

As for the derived statistic in the method of Shin et al., what we should note is that two separate thresholds are introduced according to the decision of the speech activity in the previous frame. Clearly, the multiple thresholds can give us an additional freedom that improves the performance of the VAD. Indeed, it is not sufficient for the single frame in the previous CMAPE to express the strong correlation in the consecutive occurrences of speech frames. Accordingly, one obvious motivation based on the upper derivation is that the second-order CMAP incorporating the VAD decisions in the previous two frames could perform better than the first-order CMAP scheme proposed by Shin et al. [1] and improve speech detection robustness. But, we do not consider the generalization to more than second-order since the performance improvement considering additional computation load was limited. Based on this, we derive the second-order CMAP as below:

$$P(Y(n)|H(n) = H_1(n), H(n-1) = H_1, H(n-2) = H_1) \geq \alpha$$

$$P(Y(n)|H(n) = H_0(n), H(n-1) = H_1, H(n-2) = H_1) \geq \alpha$$

(14)

This decision test is also changed to the following form:

$$P(Y(n)|H(n) = H_1(n), H(n-1) = H_1, H(n-2) = H_1) \geq \alpha$$

$$P(Y(n)|H(n) = H_0(n), H(n-1) = H_1, H(n-2) = H_1) \geq \alpha$$

(15)

In a similar reason with the conventional CMAP criterion, (15) can be approximated as follows:

$$P(Y(n)|H(n) = H_1(n)) \geq \alpha$$

$$P(Y(n)|H(n) = H_0(n)) \geq \alpha$$

(16)

Therefore, (16) can be finally expressed by

$$P(Y(n)|H(n) = H_1(n)) \geq \alpha$$

$$P(Y(n)|H(n) = H_0(n)) \geq \alpha$$

(17)

where $\gamma_{ij}$ is the threshold. Upper criterion could be expressed as follows:

$$P(Y(n)|H(n) = H_1(n)) \geq \alpha$$

$$P(Y(n)|H(n) = H_0(n)) \geq \alpha$$

(18)
Table 1: Comparison of voice activity detection probability of error ($P_{E}$), probability of miss ($P_{M}$) and false alarm probability ($P_{F}$) among the method of the statistical model-based, the CMAP-based and the proposed technique.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{E}$</td>
<td>$P_{M}$</td>
<td>$P_{F}$</td>
<td>$P_{E}$</td>
</tr>
<tr>
<td>Car</td>
<td>0</td>
<td>21.65</td>
<td>0.00</td>
<td>7.045</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>27.52</td>
<td>0.00</td>
<td>7.21</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>23.48</td>
<td>0.00</td>
<td>5.68</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>19.79</td>
<td>0.00</td>
<td>5.43</td>
</tr>
<tr>
<td>Street</td>
<td>0</td>
<td>23.53</td>
<td>0.00</td>
<td>10.90</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>23.03</td>
<td>0.00</td>
<td>10.33</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>19.52</td>
<td>0.00</td>
<td>9.34</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15.73</td>
<td>0.00</td>
<td>7.21</td>
</tr>
<tr>
<td>Office</td>
<td>0</td>
<td>28.51</td>
<td>0.00</td>
<td>17.87</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>26.45</td>
<td>0.00</td>
<td>17.85</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>22.74</td>
<td>0.00</td>
<td>13.46</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>19.28</td>
<td>0.00</td>
<td>11.25</td>
</tr>
<tr>
<td>Babble</td>
<td>0</td>
<td>34.48</td>
<td>0.00</td>
<td>22.51</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>27.81</td>
<td>0.00</td>
<td>13.64</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>22.65</td>
<td>0.00</td>
<td>8.25</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>18.62</td>
<td>0.00</td>
<td>6.27</td>
</tr>
</tbody>
</table>

In the figure, it can be seen that the threshold in the proposed method performs well by taking advantage of the soft decision scheme. Specifically, the new threshold $\gamma$ being adjusted by the SAPs show soft values along time, which yields better performance taking into account the manual VAD in Fig. 1 (c).

4. Experimental Results

Conventional methods and the proposed method were evaluated in a quantitative comparison under various noise environments. For the test material, we formed 456 s speech data 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 it manually at every 10 ms frame. The proportion of hand-marked speech frames was 57.1 % and that consisted of 44.0% voiced sounds and 13.1% unvoiced sounds. To consider various noise environments, car, street, and office noises were added to the clean speech data by varying SNR such as 0, 5, 10, and 15dB. Also, to simulate non-stationary noise in real environments, the babble noise is included at the same SNRs. In all cases, voice activity detection was conducted with $\gamma_{00} = 12.5$, $\gamma_{01} = 10$, $\gamma_{10} = 9.5$ and $\gamma_{11} = 8.5$ which are experimentally selected based on the correct speech/silence transcription on the 230 s length different speech material. Also, the initial value $\eta$ was set to 12.5 because we assumed that only noise exists on the initial several frames.

Table 1 including $P_{E}$ (probability of error), $P_{M}$ (probability of miss), and $P_{F}$ (false alarm probability) shows comparative results for the Sohn’s VAD, the first-order CMAP-based method and the proposed approach with either soft decision or not. In particular, to help someone to repeat the results, the standardized VAD such as ITU-T G.729 Annex B [10] is included. From the results, it is evident that the proposed VAD algorithm shows better performance than the previously reported VAD methods including the first-order CMAP [1] in most of the environmental conditions.
5. Conclusions

In this paper, we have proposed a novel VAD technique based on the second-order CMAP algorithm in which are incorporated the speech presence or absence probability of the previous two frames for a robust VAD decision. The proposed approach yields better performance than the conventional method in various noise environments.

6. Acknowledgements

This work was supported by National Research Foundation of Korea(NRF) grant funded by the Korean Government(MEST) (NRF-2009-0085162) 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]

7. References