A MODIFIED MAP CRITERION BASED ON HIDDEN MARKOV MODEL FOR VOICE ACTIVITY DETECTION

Shiwen Deng¹², Jiqing Han¹, Tieran Zheng¹, Guibin Zheng¹

¹School of Computer Science and Technology, Harbin Institute of Technology
²School of Mathematical Sciences, Harbin Normal University
Email: dengswen@gmail.com, {jqhan, zhengtieran, zhengguibin}@hit.edu.cn

ABSTRACT

The maximum a posteriori (MAP) criterion is broadly used in the statistical model-based voice activity detection (VAD) approaches. In the conventional MAP criterion, however, the inter-frame correlation of the voice activity is not taken into consideration. In this paper, we propose a novel modified MAP criterion based on a two-state hidden Markov model (HMM) to improve the performance of the VAD, and the the inter-frame correlation of the voice activity is modeled. With the proposed MAP criterion, the decision rule is derived by explicitly incorporating the a priori, a posteriori, and inter-frame correlation information into the likelihood ratio test (LRT). In the LRT, a compensation factor for the hypothesis of speech presence is used to regulate the trade-off between the probability of detection and the false alarm probability. Experimental results show the superiority of the VAD algorithm based on the proposed MAP criterion in comparison with that based on the recent conditional MAP criterion (CMAP) under various noise conditions.

Index Terms— Voice activity detection, MAP criterion, likelihood ratio test, hidden Markov model.

1. INTRODUCTION

The voice activity detection, which refers to the problem of distinguishing active speech from nonspeech, has become an indispensable component for many applications of speech processing systems [1, 2], and various VAD algorithms have been developed. Most of the traditional VAD algorithms are based on energy, zero-crossing rate, and spectral difference in earlier literature [3–5]. However, the performance of these algorithms rapidly degrades in noisy environments.

Recently, the VAD approaches based on statistical models [6–10] have demonstrated impressive performances and have been carried out by incorporating a statistical model and a likelihood ratio test (LRT). In these algorithms, the LRT is derived from the maximum a posteriori (MAP) criterion, which chooses the hypothesis with the maximum probability given a single observation [6–8] or multiple observations [9, 10]. However, one of the drawbacks of these VADs based on the conventional MAP criterion is that the signal distribution characterizes each frame separately without considering the inter-frame correlation of the voice activity. More recently, Shin et al. [11] propose a conditional MAP criterion (CMAP) for improving the performance of the VAD, in which two separate thresholds for the LRT are used based on the voice active decision in the previous frame. Unfortunately, since the correlation is characterized by only two thresholds in the CMAP criterion, the capability of modeling the correlation is limited. A better performance will be obtained if more accurate information about the inter-frame correlation is taken into consideration.

In this paper, the inter-frame correlation of the voice activity in consecutive frames is modeled by a two-state hidden Markov model (HMM), which assumes that the current state depends only on the current observation and the previous state. With this model, the a priori probability and the a posteriori probability of each state, and the transition probability between the adjacent states are explicitly exploited to build the MAP criterion. From the point of view of the proposed MAP criterion, the CMAP criterion can be considered as a special case of the proposed MAP criterion by replacing the a posteriori probability of the hypothesis with the result of voice activity decision in the previous state. Moreover, a compensation factor for the hypothesis of speech presence is used to regulate the trade-off between the probability of detection and the false alarm probability. With the assumption that the distribution of the observation follows the complex Laplacian, the experimental results show the performance improvement of the VAD based on the proposed MAP criterion in comparison with that based on the CMAP criterion.

2. MAP CRITERION FOR DECISION RULE OF A SINGLE OBSERVATION

Let \(y(t)\) denote a noisy speech signal that is the sum of a clean speech signal \(x(t)\) and an uncorrelated additive noise signal...
n(t). By taking the discrete Fourier transform (DFT), we have
\[ Y(m) = X(m) + N(m) \]  
where \( Y(m) = [Y_1(m), \ldots, Y_K(m)]^T \), \( X(m) = [X_1(m), \ldots, X_K(m)]^T \), and \( N(m) = [N_1(m), \ldots, N_K(m)]^T \), are the DFT coefficients of the noisy speech, clean speech, and noise, respectively. \( K \) is the number of frequency bins, and \( m \) is the frame index. Given two hypotheses, \( H_0 \) and \( H_1 \), which respectively indicate speech absence and presence, it is assumed that

\[
H_0: \quad Y(m) = N(m) \\
H_1: \quad Y(m) = N(m) + X(m)
\]

Since the MAP criterion is discussed for the decision rule of a single observation in this session, for simplicity, the frame index is removed. By using the assumption that the DFT coefficients of speech and noise are independent random variables, the likelihood ratio is calculated as

\[
\Lambda = \frac{P(Y|H_1)}{P(Y|H_0)} = \frac{\prod_{k=1}^{K} P(Y_k|H_1)}{\prod_{k=1}^{K} P(Y_k|H_0)}
\]

Usually, the geometric mean of the above likelihood ratio is used:

\[
\Lambda_g = \left( \frac{P(Y|H_1)}{P(Y|H_0)} \right)^{1/K} = \left( \frac{\prod_{k=1}^{K} P(Y_k|H_1)}{\prod_{k=1}^{K} P(Y_k|H_0)} \right)^{1/K}
\]

We consider the geometric mean of the likelihood ratio, \( \Lambda_g \), as a new likelihood ratio of the observation \( Y_g \), which is associated with the original observation \( Y \). That is

\[
\Lambda_g = \frac{P(Y_g|H_1)}{P(Y_g|H_0)}
\]

Based on the total probability theorem and Bayes rule, we can derive the posterior probabilities of \( H_1 \) and \( H_0 \) given \( Y_g \) as follows

\[
P(H_1|Y_g) = \frac{P(Y_g|H_1)P(H_1)}{P(Y_g|H_0)P(H_0) + P(Y_g|H_1)P(H_1)} = \frac{P(Y_g|H_0)P(H_0) + P(Y_g|H_1)P(H_1)}{P(H_0)}
\]

\[
P(H_0|Y_g) = \frac{P(H_0)}{P(H_0) + \Lambda_g P(H_1)}
\]

where \( P(H_0) \) and \( P(H_1) \) are the \( a \ priori \) probabilities of the two hypothesis \( H_0 \) and \( H_1 \), respectively.

According to the MAP criterion, the decision rule is given by

\[
\frac{P(H_1|Y_g)}{P(H_0|Y_g)} \geq \frac{H_1}{H_0} \alpha
\]

where \( \alpha \geq 1 \) is used to compensate for the bias toward \( H_1 \). Note that \( P(H_0|Y_g) = 1 - P(H_1|Y_g) \) and the decision rule in Eq. (9) can be written as

\[
P(H_1|Y_g) \geq \frac{H_1}{H_0} \alpha \eta = \eta
\]

where \( \eta \) is the compensation factor for the above bias toward \( H_1 \). Moreover, note that \( \eta = 1/2 \) when \( \alpha = 1 \) and \( \eta = 1 \) when \( \alpha = +\infty \). With Eq. (6) and (10), the equivalent likelihood ratio decision rule based on the posterior probability is

\[
\Lambda_g \approx \frac{\eta P(H_0)}{\eta (1 - \eta) P(H_1)}
\]

where \( \eta \in [1/2, 1] \).

The decision rule in Eq. (11) depends on the selection of compensation factor and the estimation of the \( a \ priori \) probabilities at the current frame. The compensation factor is associated with the noise type and can be selected from \([1/2, 1]\) by regulating the trade-off between the probability of detection versus the false alarm probability. The \( a \ priori \) probabilities, in fact, are varying from frame to frame because of the inter-frame correlation and are calculated in the following session.

### 3. MAP CRITERION BASED ON HMM FOR VAD

Because the inter-frame correlation of voice activity in consecutive frames is strong, the sequence of voice activity states can be modeled as a first-order Markov process, which simply assumes that the current state only depends on the current data and the previous state. Moreover, as these states can not be directly observed from the observations, the sequence of voice activity states is modeled by a two-state discrete hidden Markov model (HMM). In this model, we assume that the transition matrix

\[
A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}
\]

are already known, where the transition probability is defined as

\[
a_{ij} = P(H(m) = H_j|H(m-1) = H_i)
\]

for \( i, j = 0, 1 \), and the initialize probabilities \( P(H(1) = H_0) = P(H(1) = H_1) = 1/2 \).

In order to derive the HMM-based MAP criterion, we discuss the decision in the \( m \)th frame as follows. Let \( H(m) \) denote the random variable of the voice activity in the \( m \)th frame. And let \( P(H(m) = H_i), P(Y_g(m)|H(m) = H_i), \) and \( P(H(m) = H_i|Y_g(m)) \) denote the \( a \ priori \) probability, observation probability, and the \( a \ posteriori \) probability in the \( m \)th frame for \( i = 0, 1 \), respectively. Thus, according
to Eq. (4) and (6) the likelihood ratio and the a posteriori probability at the \( m \)th frame are respectively given by

\[
\Lambda_y(m) = \left( \frac{P(Y(m)|H(m) = H_1)}{P(Y(m)|H(m) = H_0)} \right)^{1/K}
\]

\[
P(H(m) = H_1|Y_y(m)) = \frac{\Lambda_y(m)P(H(m) = H_1)}{\Lambda_y(m)P(H(m) = H_1) + \Lambda_y(m)P(H(m) = H_0)}
\]

The a priori probability is obtained based on the above HMM

\[
P(H(m) = H_1) = a_{01}P(H(m - 1) = H_0|Y_y(m - 1))
\]

\[
+ a_{11}P(H(m - 1) = H_1|Y_y(m - 1))
\]

and \( P(H(1) = H_0) = \pi_0, P(H(1) = H_1) = \pi_1 \) at the initial state. Hence, the decision rule at the \( m \)th frame is

\[
P(H(m) = H_1|Y_y(m)) \overset{H_1}{\geq} \eta \overset{H_0}{\geq} \eta
\]

With Eq. (15) and (17), the equivalent LRT decision rule is

\[
\Lambda_y(m) \overset{H_1}{\geq} \frac{\eta P(H(m) = H_0)}{1 - \eta P(H(m) = H_1)}
\]

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

Additionally, by replacing the a posteriori probabilities \( P(H(m - 1) = H_1|Y_y(m - 1)) \) in Eq. (16) with the decision results 0 or 1 at the \( (m-1) \)th frame, we will obtain the similar result based on the conditional MAP criterion whose two separate thresholds are respectively dependent on \( a_{00}/a_{01} \) and \( a_{10}/a_{11} \). Hence, the CMAP criterion is a special case of the proposed MAP criterion.

4. EXPERIMENTS AND RESULTS

4.1. Statistical model for observation

To perform the proposed MAP criterion for the VAD, the probability distributions of the observation under two hypotheses are modeled as the complex Laplacian distribution:

\[
Y_m \sim H_0: P(Y_k|H_0) = \frac{1}{\sqrt{2\pi\lambda_{N,k}}} e^{-\frac{2|Y_{kR}+|Y_{kI}^{2}|}{\lambda_{N,k}}}
\]

\[
Y_m \sim H_1: P(Y_k|H_1) = \frac{1}{\sqrt{2\pi\lambda_{N,k} + \lambda_{X,k}}} e^{-\frac{2|Y_{kR}+|Y_{kI}^{2}|}{\sqrt{2\pi\lambda_{N,k} + \lambda_{X,k}}}}
\]

where \( Y_{kR} \) and \( Y_{kI} \) denote the real and imaginary parts of the \( k \)th DFT coefficient \( Y_k \), and \( \lambda_{N,k} \) and \( \lambda_{X,k} \) denote the variances of the \( k \)th DFT coefficients for the noise and speech. Thus, the geometric mean of the likelihood ratio is given by

\[
\Lambda_y(m) = \prod_{k=1}^{K} \frac{1}{1 + \xi_k} e^{-\frac{2|Y_{kR}+|Y_{kI}^{2}|}{\sqrt{2\pi\lambda_{N,k} + \lambda_{X,k}}}} (1 - \frac{1}{\sqrt{1 + \xi_k}})^{1/K}
\]

where \( \xi_k \) is the a priori signal-to-noise ratio (SNR) which is estimated using the minimum mean-square error (MMSE) estimator, and \( \lambda_{N,k} \) is estimated as in [12].

Fig. 1. Receiver operating characteristic (ROC) curves at 0 dB SNR. (a) Babble noise. (b) Vehicle noise.

4.2. Experimental results

In order to evaluate the performance of the VAD algorithm based on the proposed MAP criterion, we compare the probability of detection \( (P_d) \) and false alarm probabilities \( (P_f) \) of the decision rule based on the proposed MAP and CMAP criterion without any hang-over scheme. The test material consists of 130 s long speech data by labeling manually at every 10 ms frame. The percentage of the hand-marked speech frames is 63.7\%, which consists of 43.3\% voiced sound and 20.4\% unvoiced sound frames. \( P_d \) and \( P_f \) are defined as the ratio of correct speech decisions to the hand-marked speech frames and the ratio of false speech decisions to the hand-marked noise frames, respectively. The above clean test speech is corrupted by the white, babble, and vehicular noise to simulate the adverse environments by varying SNR. Moreover, for the parameters in the HMM model, the transition probabilities are \( a_{00} = 0.8 \) and \( a_{11} = 0.9 \), which are obtained from a standard speech.

The performance of the VAD detectors can be analyzed by the receiver operating characteristic (ROC) curves which show the trade-off between the increase of \( P_d \) and the reduction of \( P_f \). Fig. 1 shows the ROC curves of the VAD based on the proposed MAP criterion and CMAP criterion for the babble noise at 0dB SNR in (a) and for the vehicular noise at 0dB SNR in (b), respectively. The HMM-based MAP criterion for the VAD exhibits an improvement in the ROC curves when the false alarm probability is same as the level of CMAP-based VAD. Moreover, the more detail test results for the white, babble, and vehicular noise are sum-
Evaluating the performance of different noise conditions, Table 1 shows the improvement in detection accuracy when using the CMAP-based MAP criterion. The table compares the HMM-based MAP method with the CMAP-based method across various noise conditions. The error reduction rate (ERR) is calculated for each noise type, indicating the relative improvement of the CMAP-based method over the HMM-based method.


table 1. Performance evaluation in different noise conditions

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR (dB)</th>
<th>Pe (%) HMM</th>
<th>Pm (%) HMM</th>
<th>Pf (%) HMM</th>
<th>Pe (%) CMAP</th>
<th>Pm (%) CMAP</th>
<th>Pf (%) CMAP</th>
<th>ERR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0</td>
<td>10.40</td>
<td>11.41</td>
<td>9.38</td>
<td>14.01</td>
<td>18.60</td>
<td>9.41</td>
<td>38.66</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7.61</td>
<td>6.28</td>
<td>8.94</td>
<td>10.59</td>
<td>11.90</td>
<td>9.27</td>
<td>47.23</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>6.86</td>
<td>5.06</td>
<td>8.65</td>
<td>9.09</td>
<td>9.22</td>
<td>8.95</td>
<td>45.12</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>6.47</td>
<td>4.62</td>
<td>8.32</td>
<td>8.98</td>
<td>9.04</td>
<td>8.92</td>
<td>48.89</td>
</tr>
<tr>
<td>Babble</td>
<td>0</td>
<td>15.28</td>
<td>14.96</td>
<td>15.59</td>
<td>18.52</td>
<td>20.00</td>
<td>17.03</td>
<td>25.02</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>12.75</td>
<td>11.00</td>
<td>14.49</td>
<td>15.70</td>
<td>15.65</td>
<td>15.74</td>
<td>29.71</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>11.45</td>
<td>8.74</td>
<td>14.16</td>
<td>14.52</td>
<td>13.53</td>
<td>15.48</td>
<td>35.50</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>10.50</td>
<td>7.30</td>
<td>13.69</td>
<td>13.77</td>
<td>13.31</td>
<td>14.22</td>
<td>45.15</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0</td>
<td>13.01</td>
<td>14.38</td>
<td>11.63</td>
<td>18.92</td>
<td>25.88</td>
<td>11.96</td>
<td>44.44</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>12.09</td>
<td>13.16</td>
<td>11.02</td>
<td>16.10</td>
<td>20.30</td>
<td>11.90</td>
<td>35.17</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10.87</td>
<td>11.80</td>
<td>9.94</td>
<td>15.17</td>
<td>20.27</td>
<td>10.06</td>
<td>41.79</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>9.20</td>
<td>9.88</td>
<td>8.52</td>
<td>13.77</td>
<td>17.59</td>
<td>9.95</td>
<td>43.83</td>
</tr>
</tbody>
</table>

4. CONCLUSION

In this paper, we have proposed a novel modified MAP criterion for VAD based on the two-state hidden Markov model which models the inter-frame correlation of the observation. In this criterion, all the a priori probability, a posteriori probability, and the inter-frame correlation of the voice activity have been explicitly incorporated into the derivation of the LRT. In this LRT, a compensation factor for the hypothesis is used to regulate the trade-off between the probability of detection and the false alarm probability. The experimental results show that the VAD algorithm based on the proposed MAP criterion outperforms the VAD based on the CMAP criterion under various noise conditions.

6. REFERENCES