Robust Statistical Voice Activity Detection Using a Likelihood Ratio Sign Test*

Shiwen Deng 1, 2, Jiqing Han 1

1 School of Computer Science and Technology, Harbin Institute of Technology, China
2 School of Mathematical Sciences, Harbin Normal University, China
dengswen@gmail.com, jghan@hit.edu.cn

Abstract

Voice activity detection (VAD) plays an important role on the performance of speech processing systems. Recently, more and more works focused on the statistical model-based VAD algorithms have been presented in literature, which make a decision of speech and nonspeech based on the likelihood ratio (LR). However, all the statistical models used in those algorithms are unable to exactly describe the statistics of noisy speech and various type noises. In this paper, a novel VAD algorithm is proposed based on the nonparametric detection theory by incorporating the likelihood ratio into the sign test to provide a new decision rule. Meanwhile, an optimal threshold of the proposed method is derived and the selections of relevant parameters are discussed as well. Experimental results show that the proposed VAD algorithm outperforms the conventional statistical model-based VAD.

Index Terms: voice activity detection, likelihood ratio, sign test, nonparametric detection

1. Introduction

Voice activity detection (VAD), which refers to the problem of distinguishing active speech from nonspeech, plays an important role on the performance of speech processing systems such as robust speech recognition, speech enhancement, and coding systems [1, 2]. Various traditional VAD algorithms have been proposed 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, much work for improving the performance of the VADs in various high noise environments has been carried out by incorporating a statistical model and a likelihood ratio test (LRT). Those algorithms assume that the spectra distributions of the noise and the noisy speech are specified in terms of some certain parametric models such as complex Gaussian [6], complex Laplacian [7], generalized Gaussian [8], or generalized Gamma distribution [9]. Most of those above methods make a decision by comparing the likelihood ratio (LR) of two models involving a single observation vector with a fixed threshold. Comparing with the algorithms based on a single observation, some more general algorithms based on the multiple observation likelihood test (MO-LRT) are proposed and have better performance [10-12].

The LRT used in the statistical model-based VAD for the classification problem is the optimal method if the signal exactly follows the statistical model that the above algorithms have assumed. Unfortunately, the assumption of exact statistical models is impractical due to the non-stationary of the speech and noise in a real environment. As a result, the inexact statistical models assumed in the above algorithms decrease the robustness and accuracy of VAD.

In this paper, a novel approach for VAD based on the signal nonparametric detection theory [13] is presented. Instead of comparing the LR with a fixed threshold in the conventional statistical model-based method [6-12], the statistics of LR are used for the classification task in despite of the difficulty in modeling its distribution. The nonparametric detection, or named as sign detection in some case, only requires the rough statistical information of a signal instead of the exact statistical model. This motivates the use of the sign test for the LR, which can be considered as a random variable. Specifically, multiple observation vectors are adopted and the LR calculated from each vector is mapped into the sign (1 or 0) according to the decision results by comparing with the fixed threshold. Subsequently, the sign test makes a decision based on the sign sequence from multiple observation vectors. Furthermore, the optimal threshold is derived for the sign test based on the Bayes minimum error. For the evaluation of our work, MO-LRT is used as the reference in our experiments not only because of its better performance but because it is a more general method than the single observation likelihood ratio test methods. For a fair comparison, the hang-over scheme for VAD is not used in our experiments, considering that this paper only focuses on the issues of the statistical model and decision rule. The experimental results indicate that the proposed method achieves a better performance than the conventional statistical model-based algorithms.

2. Likelihood ratio sign test

Being an optimal decision rule that minimizes the error probability in a two-hypothesis test, Bayesian criterion is widely used in VAD [6-12]. Given an observation vector \( \hat{y} \) to be classified into two classes (\( \omega_0 \) or \( \omega_1 \)), the optimal decision rule is the likelihood ratio test based on comparison of the statistics of two classes to the fixed threshold named LRT threshold

\[
L(\hat{y}) = \frac{p_{\omega_0}(\hat{y} | \omega_0)}{p_{\omega_1}(\hat{y} | \omega_1)},
\]

where \( p_{\omega_i}(\cdot) \) is the probability distribution function over classes \( \omega_i \) for \( i = 0,1 \). The above decision rule assumes that the probability distributions of the observation vector are exactly known over two classes. However, the distributions of speech and noise in the practical VAD problem can only be estimated approximately. Consequently, the inaccuracy of the statistical models results in an increasing in the detection error when the LR is simply compared with the fixed threshold.

* This work was sponsored by the National Basic Research Program of China (973 Program, No. 2007CB311100).
Although multiple vectors, which are assumed independent, are used in MO-LRT, the above problem is not yet addressed. In order to overcome the above drawback of current methods, we consider the LR from a vector as a random variable and make a decision according to the statistical characteristics of the LRs from multiple vectors under the two hypotheses. However, the distribution of LR is impractical to be described by a statistical model. To address this problem, we incorporate sign test into the decision problem based on the robustness of the sign test, which only requires the coarse information of the statistics of the random variable. We refer to this method as the likelihood ratio sign test (LRST), which is described as follows.

Given $n$ independent observation vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ to be available in the LRST, the log LR equivalent to the LR in Eq. (13) can be written as follows for each observation

$$
\ell_k = \log \frac{P_{y_k|H_0} (\mathbf{y}_k | a_k)}{P_{y_k|H_1} (\mathbf{y}_k | a_k)} k = 1, \ldots, n.
$$

Equivalently, considering the log likelihood ratio $\ell_k$, also named LR, as a random variable, two hypotheses ($H_0$ and $H_1$) are to be tested. The null hypothesis $H_0$ is intended to describe the situation that speech is absent and the alternative hypothesis $H_1$ is to describe that speech is present. Although the exact probability distribution of LR is unknown, a simple information about the statistics of the LR can also be obtained, that is, the probability of the LR is greater than the LR threshold under the two hypotheses. Consequently, the two probabilities can be represented as

$$
\Pr(\ell_k \geq \eta | H_0) = q_k,
$$

$$
\Pr(\ell_k > \eta | H_1) = p_k,
$$

where $\eta$ defines the LR threshold. The LRST is defined as

$$
T_n = \sum_{k=1}^{n} u(\ell_k - \eta) \geq \tau_n
$$

where $\tau_n$ defines the LRST threshold and $u(\cdot)$ is a unit step function as

$$
u(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}
$$

Moreover, according to Eq. (3) and (4), the false alarm probability $P_f$ and the detection probability $P_d$ of the LRST can be written as follows

$$
P_f = \sum_{k=1}^{n} \binom{n}{k} q^k (1-q)^{n-k}
$$

$$
P_d = \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k}
$$

In terms of Eq. (6) and Eq. (7), the global error can be represented as

$$
P_e = P_f \Pr(H_0) + (1-P_d) \Pr(H_1)
$$

where $\Pr(H_0)$ and $\Pr(H_1)$ are the a priori probabilities of speech absence and presence, respectively. Furthermore, let $P(H_0) = a, P(H_1) = 1 - a$, and the LRST threshold can be obtained by minimizing the global error

$$
t_{\text{optimal}} = \arg \min_{t_l} \{ P_f \}
$$

$$
= \arg \min_{t_l} \left[ a \sum_{k=1}^{n} \binom{n}{k} q^k (1-q)^{n-k} + (1-a) \left( 1 - \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \right) \right]
$$

Since the a priori probabilities of speech absence and presence have little effect on the optimal LRST threshold, a suitable approximation of it can be obtained by $P(H_0) = P(H_1) = a = 0.5$, the details as shown in section 4.

3. Likelihood ratio sign test for voice activity detection

This section addresses the application of the proposed LRST to the problem of voice activity detection. The use of the LRST for VAD is mainly motivated by two factors: 1) The statistical models can not exactly model the distribution of the clean speech and noise, and this reflects on the LR to cause detection error; 2) The available information about the statistical characteristics of LRs is very scarce. The VAD based on LRST rule is described as follows.

The LRST VAD is defined over a sliding window consisting of $n = 2m + 1$ observation vectors $\{\mathbf{y}_{-m}, \ldots, \mathbf{y}_0, \ldots, \mathbf{y}_m\}$ around the $l$th observation vector for which the decision is made. In term of Eq. (4), the LRST VAD for the $l$th observation vector over the sliding window can be defined as

$$
T_{i,n} = \sum_{k=i-n}^{i} u(\ell_k - \eta),
$$

where $\eta$ defines the LRST threshold.

By defining $S(l) = u(\ell_l - \eta)$, the LRST VAD can be recursively computed by

$$
T_{i+1,n} = T_{i,n} - S(l-m) + S(l+m+1),
$$

where the sliding widow length is $n = 2m + 1$ and its hop size is one. And the decision rule is defined as

$$
T_{i,n} \geq \tau_n, \text{ accept } H_1, \mathbf{y}_l \text{ is classified as speech}
$$

$$
T_{i,n} < \tau_n, \text{ accept } H_0, \mathbf{y}_l \text{ is classified as nonspeech}
$$

In order to implement the proposed LRST VAD on an incoming signal, an adequate statistical model for the feature vectors in presence and absence of speech needs to be selected. The model selected is similar to that used by Sohn [6] that assumes the discrete Fourier transform (DFT) coefficients of the clean speech ($S_k$) and the noise ($N_k$) to be asymptotically independent Gaussian random variables

$$
P_{y_k|\theta_k} (\mathbf{y}_k | a_k) = \prod_{j=0}^{J} \frac{1}{\sqrt{\lambda_k(j)}} \exp \left[ -\frac{|X_k(j)|^2}{\lambda_k(j)} \right]
$$
are the DFT bin, and  \( \lambda_n(j) \) and  \( \lambda_j(j) \) denote the variances of  \( N_n \) and  \( S_j \), respectively. By evaluating the log LR for the  \( j \)th DFT bin and averaging all  \( J \) the DFT bins, the log LR defined in Eq. (2) can be obtained by

\[
\ell_j = \frac{1}{J} \sum_{m=1}^{J} \left( \frac{\gamma_n(j)}{\xi(j)} \frac{\xi(j)}{\gamma_n(j)} \log (1 + \xi(j)) \right)
\]

where  \( \gamma_n(j) \) and  \( \xi(j) \) are the a priori and a posteriori SNRs that are estimated using the minimum mean-square error (MMSE) estimator [14]. The estimation is also obtained under the assumption of Gaussian model for DFT coefficients of the noisy speech and noise. And this means that the estimations of SNRs are also inexact due to the issue as mentioned before.

The LRST VAD working on a sliding window with window length  \( n=2m+1 \), imposes an m-frame delay to the algorithm, but this is not a serious implementation obstacle in several applications [10].

4. Experiments

The selection of the relevant parameters used in the LRST VAD and experimental results are shown in this section.

4.1. Selection of relevant parameters

The total number of the observation vectors  \( n \), namely the sliding window length, has an optimum value according to the conclusion of [10]. The similar method is used in the LRST and the selected window length is 13, and the difference of the optimum value in different noise environment is ignored in this paper.

The LRT threshold  \( \eta \) in Eq. (3) can be obtained by training data, which is similar to that used by Sohn [6]. The LRT threshold is chosen following the policy of best tradeoff between speech and nonspeech classification error in LRT, because it causes the minimum global classification error. When the LRT threshold is chosen, the parameters  \( p \) and  \( q \) in Eq. (3) are determined by it. This is helpful to choose the optimal LRST threshold and to improve the performance of the LRST.

In order to calculate the LRST threshold by Eq. (9), the a priori knowledge of the probabilities of occurrence of each hypothesis,  \( Pr(H_0) \) and  \( Pr(H_1) \) , are required. In a practical implementation, we find that the priori probabilities have little effect on the optimal LRST threshold. The optimal LRST thresholds are calculated and shown in Table 1 when the probability of nonspeech  \( a \) varies from 0.1 to 0.9 for the fixed values of  \( p \) and  \( q \). It can be seen that the optimal thresholds vary in a small range for some fixed values of  \( p \) and  \( q \). For example, when  \( p = 0.80, q = 0.10 \) , the optimal LRST thresholds calculated by Eq. (9) are between 5 and 6 when  \( a \) varies from 0.1 to 0.9. Therefore, this means that a suitable approximation for the optimal LRST threshold can be yielded when  \( P(H_0) = P(H_1) = a = 0.5 \). Moreover, the latter

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |

experimental results show that the approximation can satisfy the practical requirements as shown in Fig 1.

4.2. Experimental results

In this subsection, we analyze the proposed LRST VAD and compare its performance to the MO-LRT VAD without hang-over scheme. The analysis is based on the receiver operating characteristics (ROC) curves. The ROC curves is a frequently used methodology to describe the performance of the VAD based on the false alarm probability  \( (P_f) \) and detection probability  \( (P_d) \). For training, we made reference decisions on a clean speech material of 125 s by manually labeling the active and inactive duration every 10 ms frame. The percentage of the hand-marked frame was 71.4%, which consisted of 43.08% voice and 28.4% unvoiced frame. The white, babble, and vehicle noise from NOISEX-92 [15] database at 5 dB SNR were added to the clean speech to create noisy environments, respectively. Furthermore, the LRT threshold was trained and the probabilities,  \( p \) and  \( q \) , were also obtained corresponding to the LRT threshold. The parameters were stored as the parameter for LRST based on the training data.

For testing, a different speech of 130 s from TIMIT database was used as the test material and also manually labeled using every 10 ms frames for the evaluation purposes. Similarly as above, the white, babble, and vehicle noise from NOISEX-92 database were added to the clean speech to simulate the noisy conditions, respectively. The noisy speech was segmented into frames and the frame length is 20 ms, and the number of observation frames  \( n \) is 13. Considering that this paper only focuses on the issues of the statistical model and decision rule, the hang-over scheme for VAD is not used in this experiment for a fair comparison. When the LRST threshold  \( \tau_n \) varied from 1 to 13, we got the receiver operating characteristics (ROC) curve which shows the trade-off between the false alarm probability  \( (P_f) \) and detection probability  \( (P_d) \). The ROCs of LRST and MO-LRT as well as the optimal points obtained by using Eq. (9) in black square, are shown in Fig. 1 (a-c) in white, babble, and vehicle noise environments at 5 dB SNR, respectively. These figures show that the performance of the LRST VAD has a good enhancement comparing with the MO-LRT VAD and the approximation of the optimal LRST thresholds are suitable to the detection task as well. The detail test results are summarized in Table 2 where the false alarm probability  \( (P_f) \) and the detection probability  \( (P_d) \) of MO LRT and LRST in different type noises at 5, 10, 15 dB SNR are listed respectively. The test results confirm that the proposed LRST method effectively enhances the performance of the statistical
5. Conclusions

This paper shows a new VAD for improving speech detection robustness in noisy environments. The proposed method is developed on the basis of signal nonparametric detection theory, in which the exact statistical model of signal is unnecessary. The statistical likelihood ratio is incorporated into the sign test to provide the new decision rule. Furthermore, the optimal threshold is derived and the corresponding parameters are analyzed as well. Moreover, experimental results show that the proposed VAD algorithm outperforms the conventional statistical model-based VAD.

6. References