A Statistical Model-Based Voice Activity Detection Employing Minimum Classification Error Technique

Sang-Ick Kang, Ji-Hyun Song, Kye-Hwan Lee, Yun-Sik Park, Joon-Hyuk Chang

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

Abstract

In this paper, we apply a discriminative weight training to a statistical model-based voice activity detection (VAD). In our approach, the VAD decision rule is expressed as the geometric mean of optimally weighted likelihood ratios (LRs) based on a minimum classification error (MCE) method. That approach is different from that of previous works in that different weights are assigned to each frequency bin and is considered to be more realistic. According to the experimental results, the proposed approach is found to be effective for the statistical model-based VAD using the LR test.

1. Introduction

Voice activity detection (VAD) algorithms are essential part of variable rate speech coding, providing effective means of enhancing the capacity and coverage of communication bandwidth. Among the various VAD algorithms, we consider a statistical model-based VAD approach originating from work on speech enhancement reported by Ephraim and Malah [1]. Sohn et al. applied a Gaussian statistical model to the VAD employing the decision-directed (DD) method-based parameter estimation and reported high detection accuracy [2], [3]. The novelty of the statistical model-based VAD has been recognized in many studies, which employ the decision rule derived from application of the likelihood ratio (LR) test to a set of hypotheses [2]-[8]. From an investigation of the VAD schemes, however, it is observed that no weights can be applied to LRs in all frequency components without taking full consideration of the spectral characteristics of the speech signal, and using the geometric mean of the LRs for the final VAD decision [2], [3]. For this reason, we present a novel VAD technique incorporating optimally weighted LRs based on a minimum error classification (MCE) scheme, an approach that is well known as a discriminative weight training [9]. Based on a number of experiments, the proposed VAD technique is found to yield a better performance than the conventional approaches adopting equally weighted LRs.

2. A statistical model-based VAD

We assume that the noise signal \( n(t) \) is added to the speech signal \( x(t) \), with their sum being denoted by \( y(t) \) in the time domain. \( y(t) \) is transformed by the discrete Fourier transform (DFT) as follows:

\[
Y(t) = X(t) + N(t)
\]

where \( Y(t) = [Y_1(t), Y_2(t), \cdots, Y_M(t)]\), \( X(t) = [X_1(t), X_2(t), \cdots, X_M(t)]\), and \( N(t) = [N_1(t), N_2(t), \cdots, N_M(t)]\) denote the DFT coefficients of the noisy speech signal, clean speech, and the added noise. Given two classes, \( H_0 \) and \( H_1 \), which respectively indicate speech presence and absence, it is assumed that

\[
\begin{align*}
H_0: & \text{speech absent; } Y_k(t) = N_k(t) \\
H_1: & \text{speech present; } Y_k(t) = X_k(t) + N_k(t)
\end{align*}
\]

With the Gaussian pdf assumption, the distributions of the noisy spectral components conditioned on both hypotheses are given by

\[
\begin{align*}
p(Y_k|H_0) &= \frac{1}{\pi \lambda_{n,k}} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{n,k}} \right\} \\
p(Y_k|H_1) &= \frac{1}{\pi \lambda_{a,k} + \lambda_{n,k}} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{a,k} + \lambda_{n,k}} \right\}
\end{align*}
\]

where \( \lambda_{a,k} \) and \( \lambda_{n,k} \) denote the variances of noise and speech for the individual frequency band, respectively. The LR of the \( k \)th frequency band is given by

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

where \( \xi_k = \lambda_{a,k}/\lambda_{n,k} \) and \( \gamma_k = Y_k/\lambda_{a,k} \) denote the a priori signal-to-noise ratio (SNR) and the a posteriori SNR, respectively [1]. The a posteriori SNR \( \gamma_k \) is estimated using \( \lambda_{a,k} \) and the a priori SNR \( \xi_k \) is estimated by the well-known DD method as follows [2]:

\[
\hat{\xi}_k(t) = \alpha \frac{|\hat{X}_k(t-1)|^2}{\lambda_{n,k}(t-1)} + (1 + \alpha) P[\gamma_k(t-1) - 1]
\]

where \( |\hat{X}_k(t-1)|^2 \) is the speech spectral amplitude estimate of the previous frame obtained using the minimum mean-square error (MMSE) estimator [3]. Also, \( \alpha \) is a weight that is usually determined in the range (0.95, 0.99) [1]. 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 established from the geometric mean of the LRs computed for the individual frequency bins [2]-[8] and is obtained by

\[
\log \Lambda(t) = \frac{1}{M} \sum_{k=1}^{M} \log \Lambda_k(t) \geq \eta
\]

where an input frame is classified as speech presence if the geometric mean of the LRs is greater than a certain threshold value \( \eta \) and speech absent otherwise.
3. Weight optimization using MCE training

The previous section notes that the usual procedure to derive a statistical model-based VAD decision is to form the geometric mean of the LRs. In this section, we propose a technique to adopt various weights for the LRs such as \( w_k \) and \( \Lambda_k \), as we speculate that incorporation of the different contributions of the LRs will enhance the performance of the VAD. The weights \( \{ w_k \} \) must satisfy the following conditions:

\[
\sum_{k=1}^{M} w_k = 1, \quad w_k \geq 0. \tag{9}
\]

Let \( \Lambda_w = \{ w_1 \log \Lambda_1, w_2 \log \Lambda_2, \ldots, w_M \log \Lambda_M \} \) represent the optimally weighted LR vector and \( \Lambda_w \) be the 
\[
\frac{1}{M} \sum_{k=1}^{M} w_k \log \Lambda_k.
\]

Two discriminant functions of speech \((g_s)\) and noise \((g_n)\) are prepared to decide whether each frame is classified into speech or noise as follows:

\[
g_s(\Lambda_w) = w_0 - \theta, \quad g_n(\Lambda_w) = \theta - \Lambda_w
\]

where \( \theta \) denotes a threshold value of the combined score. If the discriminant function of \( g_s(\Lambda_w) \) is greater than that of \( g_n(\Lambda_w) \), each frame of \( \Lambda_w \) is classified into the speech frame. Actually, this judgement can be made simply by comparing \( \Lambda_w \) and \( \theta \). However, since the MCE training requires a discriminative function for each cluster, the two functions are prepared. In our approach, estimation of the weights is performed under the discriminative training framework in which generalized probabilistic descent (GPD) technique is applied [9]. Let \( D \) denote the misclassification measure of training data \( \Lambda_w(t) \). Then,

\[
D(\Lambda_w(t)) = \left\{ \begin{array}{ll}
g_s(\Lambda_w(t)) + g_n(\Lambda_w(t)) & \text{if } C_m(Y(t)) = H_1 \\
g_n(\Lambda_w(t)) + g_s(\Lambda_w(t)) & \text{if } C_m(Y(t)) = H_0
\end{array} \right.
\]

where \( C_m() \) denotes a VAD decision and is obtained by manual labeling every frame. When (13) is negative, the classification is considered to be correct. The GPD approach approximates the empirical classification error by a smooth objective function, which is the 0–1 step loss function defined by

\[
L(t) = \frac{1}{1 + \exp(-\gamma D(\Lambda_w(t)))}, \quad \gamma > 0
\]

where \( \gamma \) denotes the gradient of the sigmoid function. Once the parameter \( \gamma \) is specified, the weights are trained according to the following criterion:

\[
\{ \tilde{w}_k \} = \arg \min_{\{ w_k \}} L.
\]

(14)

The steepest descent method is considered the easiest way to optimize the weights according to the above criterion.

On the other hand, direct adoption of the steepest descent technique is found to be difficult due to the constraints on the weights as given by (10). We therefore adopt the following parameter transformation.

\[
\tilde{w}_k = \log w_k \quad (k = 1, \ldots, M).
\]

(15)

Let \( \{ \tilde{w}_k(t) \} \) denote the set of estimates for the transformed weights at time \( t \). Then, it is updated based on the steepest descent algorithms as follows:

\[
\tilde{w}_k(t+1) = \tilde{w}_k(t) - \epsilon \frac{\partial L(t)}{\partial \tilde{w}_k} |_{\tilde{w}_k \rightarrow \tilde{w}_k(t)}
\]

(16)

where \( \epsilon(>0) \) is a step size. The gradient of (16) is obtained as follows [10]:

\[
\frac{\partial L(t)}{\partial \tilde{w}_k} = \frac{\partial L(t)}{\partial D(\Lambda_w(t))} \frac{\partial D(\Lambda_w(t))}{\partial g} \frac{\partial g}{\partial \tilde{w}_k}
\]

(17)

where

\[
\frac{\partial L(t)}{\partial D(\Lambda_w(t))} = \gamma \cdot L(t)(1 - L(t)).
\]

(18)

\[
\frac{\partial D(\Lambda_w(t))}{\partial g} = \left\{ \begin{array}{ll}
-1 & \text{if } D < 0 \\
1 & \text{if } D > 0
\end{array} \right.
\]

(19)

\[
\frac{\partial g}{\partial \tilde{w}_k} = w_k \log \Lambda_k(t).
\]

(20)

After \( \tilde{w}_k \) is updated, \( \tilde{w}_k \) is inversely transformed to \( w_k \) using the following rule:

\[
w_k = \frac{\exp(\tilde{w}_k)}{\sum_{i=1}^{M} \exp(\tilde{w}_i)}
\]

(21)

where (21) includes normalization of the weights to satisfy the constraint in (9). Comparing \( \Lambda_w \) and a given threshold value finally reveals the proposed VAD method which is the optimally weighted LR-based test as follows:

\[
\Lambda_w(t) = \frac{1}{M} \sum_{k=1}^{M} w_k \log \Lambda_k(t) \quad \frac{H_1}{H_0} \eta.
\]

(22)

4. Experimental results

Performance of the proposed algorithm was evaluated on the NTT database that consists of a number of speech materials [6]. All the training data used for the MCE technique were recorded from 4 male speakers and 4 female speakers.

For training, we made reference decisions on the clean speech material of 230 s long by manually labeling the active and inactive regions of the speech signal every 10 ms frame. The percentage of the hand-marked active speech frames was 57.1%, which consisted of 44.0% voiced sounds and 13.1% unvoiced sounds. In order to create noisy environments, we added the car and street noises to the clean speech data at 5 and 15 dB SNR. The parameters used for defining the objective function \( L \) were selected such that \( \gamma = 1 \) and the step size for parameter update was set to \( \epsilon = 1 - \frac{40000}{40000} \). In practice, a threshold value of the combined score was set to 0 as the experimentally chosen boundary in the middle of \( \Lambda_w \) stemming from speech and \( \Lambda_w \) stemming from noise. Subsequently, the weights were optimized on the separate training set in all the training conditions. Finally, among the different sets of the weights, we selected only a single set of the weights as a representative case which is obtained based on an observation that the weights under each training condition seem to be quite similar.

For testing, we used different speech material (220 s in duration) from the NTT database. For evaluation purposes, we manually labeled the test material using 10 ms frames. To simulate noisy conditions, car and street noises are again added to the clean speech data at 5 and 15 dB SNR. The receiver operating characteristics (ROCs), showing the trade-off between \( P_0 \) and \( P_1 \) of clean, car and street noise environments are shown in Figs. 1-5. In addition, the performance improvement was investigated for voiced sounds and unvoiced sounds, respectively. While the performance improvement is not observed in
the clean speech condition as shown in Fig. 1, the proposed algorithm performs better than the conventional VAD in all noisy conditions as illustrated in Figs. 2-5. The test results confirm that the proposed MCE method effectively enhance the performance of the statistical model-based VAD. In particular, it is obvious form the results that the detection accuracy with the proposed scheme is considerably improved for unvoiced sounds while preserving performance for voiced sounds.

5. Conclusions

In this paper, we have proposed a novel VAD technique based on the MCE algorithm in which optimally weighted LRs are integrated into the geometric mean for a robust VAD decision.

The proposed approach yields better performance than the conventional method in stationary and non-stationary noise environments.

6. Acknowledgements

This work was supported in part by IITA through IT Leading R&D Support Project.

7. References


