EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Reliable likelihood ratios for statistical model-based voice activity detector with low false-alarm rate

Younggwan Kim, Youngjoo Suh* and Hoirin Kim

Abstract

The role of the statistical model-based voice activity detector (SMVAD) is to detect speech regions from input signals using the statistical models of noise and noisy speech. The decision rule of SMVAD is based on the likelihood ratio test (LRT). The LRT-based decision rule may cause detection errors because of statistical properties of noise and speech signals. In this article, we first analyze the reasons why the detection errors occur and then propose two modified decision rules using reliable likelihood ratios (LRs). We also propose an effective weighting scheme considering spectral characteristics of noise and speech signals. In the experiments proposed in this study, with almost no additional computations, the proposed methods show significant performance improvement in various noise conditions. Experimental results also show that the proposed weighting scheme provides additional performance improvement over the two proposed SMVADs.

Keywords: voice activity detector, statistical model, reliability of likelihood ratio

1. Introduction

The purpose of a voice activity detector (VAD) is to discriminate between speech and non-speech regions from the input signals in various noisy conditions. VAD techniques have widely been used in many speech applicable fields, such as speech recognition, speaker recognition, speech coding, and speech enhancement as a preprocessor because they can help us to improve the performance of those recognition systems and enhance the channel efficiency of the speech coding system. In general, most of the conventional VAD systems assume that the statistical property of noise is stationary over longer period than that of speech, which makes it possible to estimate noise statistics in spite of the occasional presence of speech [1]. By comparing estimated noise and speech statistics, we can detect speech regions from the unknown input signals.

As the demands for more accurate VADs in noisy conditions increase, a lot of efforts have been made to enhance the performance of VAD [2-14]. One successful approach is the statistical model-based VAD (SMVAD) proposed by Sohn et al. [2]. It utilizes the complex Gaussian probability density function (PDF). More recently, various efforts have been made to optimize SMVAD by modifying the decision rule originally derived from the likelihood ratio test (LRT). To decrease detection errors at speech offset regions, Sohn et al. [3] proposed an effective hang-over scheme based on the hidden Markov model (HMM), and Cho and Kondoz [4] proposed smoothed likelihood ratios (SLRs) in the decision rule. Other approaches have involved various statistical models for noise and noisy speech [5], and discriminative weight training (DWT) scheme [6]. The DWT scheme is a good approach in the name of optimizing frequency weights, but it does not yet consider temporal variations of input signal statistics because the optimized weights can be calculated only once through the whole training data. Also, the DWT scheme is not very practical since the weights need to be optimized differently in various noise conditions according to noise types and signal-to-noise ratio levels. Another technique is the moving-averaged decision rule over a certain number of neighboring frames applied for performance improvement of SMVAD [7]. However, to our knowledge, it seems that there has been no study about the reliability of likelihood ratios (LR).

* Correspondence: yjsuh@kaist.ac.kr
Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, 335 Gwahangno, Yuseong-Gu, Daejeon 305-701, Korea

In this article, we analyze the problem of the LRT-based decision rule and various properties of noise spectra in terms of the signal power and the related SNRs. Based on our analysis, we propose modified decision rules by selecting reliable LRs and a weighting scheme which can well take into account the difference between noise and noisy speech. The main advantage of these methods is that each proposed method shows significant performance improvement with almost no additional computational cost.

This article is organized as follows. In Section 2, we introduce the modeling concept of noise and noisy speech used to constitute the decision rule of SMVAD. In addition, we demonstrate estimation techniques for related parameters such as *a priori* and *a posteriori* SNRs, noise variance, and speech absence probability (SAP). In Section 3, we analyze the LRT-based decision rule of SMVAD and explain overlooked phenomena produced by noise or noisy speech. In Section 4, we propose new decision rules by selecting reliable LRs and a weighting scheme applied to every LR to reduce detection errors. In Section 5, we show test environments and demonstrate the significantly improved performance of proposed methods, compared with the conventional SMVAD methods in various noise conditions.

2. Statistical model-based voice activity detector

In this section, we briefly review the overall process of SMVAD using the complex Gaussian PDF to detect speech regions in the adverse noise environment. Basically, the PDF used for SMVAD assumes that there is no correlation between the real and imaginary parts of spectral components.

2.1. Noise and noisy speech modeling

The SMVAD is based on two hypotheses H_0 and H_1 which assume the only two cases, noise or noisy speech, respectively,

H_0 - Speech absence: $Y(n) = N(n)$

H_1 - Speech presence: $Y(n) = S(n) + N(n)$

where $Y(n) = [Y_0(n), Y_1(n),...,Y_{M-1}(n)]$, $N(n) = [N_0(n), N_1(n),...,N_{M-1}(n)]$, and $S(n) = [S_0(n), S_1(n),...,S_{M-1}(n)]$ represent M-dimensional discrete Fourier transform (DFT) coefficient vectors of the input signal, noise, and clean speech at the nth frame, respectively. In the SMVAD, the following assumptions are given:

1. Noise is additive and its statistics is uncorrelated with speech.
2. All DFT coefficients are independent of each other.
3. The likelihood of $Y_k(n)$ conditioned on each hypothesis can be modeled by the zero-mean complex Gaussian PDF.

Under these assumptions, the PDFs of $Y(n)$ conditioned on each hypotheses are given by

$$p(Y(n)|H_0)\prod_{k=0}^{M-1}\frac{1}{\pi\lambda_{N,k}}\exp\left[\frac{|Y_k(n)|^2}{\lambda_{N,k}}\right] \quad (1)$$

$$p(Y(n)|H_1)\prod_{k=0}^{M-1}\frac{1}{\pi(\lambda_{N,k}+\lambda_{S,k})}\exp\left[-\frac{|Y_k(n)|^2}{\lambda_{N,k}+\lambda_{S,k}}\right] \quad (2)$$

where k is the frequency bin index, and $\lambda_{N,\,k}$ and $\lambda_{S,\,k}$ denote the variances of noise and speech, respectively.

2.2. Decision rule based on LRT

The decision rule of the SMVAD can be derived from log likelihood ratios (LLRs) at every frequency bin which is given by

$$\Lambda_k(n) = ln\left[\frac{p(Y_k(n)|H_1)}{p(Y_k(n)|H_0)}\right] = \frac{\gamma_k(n)\xi_k(n)}{1+\xi_k(n)} - ln[1+\xi_k(n)] \quad (3)$$

where ξ_k(n) is $\lambda_{S,\,k}/\lambda_{N,\,k}$ representing the *a priori* signal-to-noise ratio (SNR) and $\gamma_k(n)$ is $|Y_k(n)|^2/\lambda_{N,\,k}$ denoting the *a posteriori* SNR. $\lambda_{S,\,k}$ and $\lambda_{N,\,k}$ should be estimated and the well-known method for estimating the *a priori* SNR is the decision-directed (DD) method [15] which is given as

$$\hat{\xi}_k(n) = \alpha\frac{|S_k(n-1)|^2}{\hat{\lambda}_{N,k}(n-1)} + (1-\alpha)\max\left[\hat{\gamma}_k(n) - 1.0\right] \quad (4)$$

where α is the weighting term, e.g., 0.98, $\hat{\gamma}_k(n) = |Y_k(n)|^2/\hat{\lambda}_{N,k}(n)$ is an estimate for the short-time power spectrum of clean speech derived from the minimum mean square error short-time spectral amplitude (MMSE-STSA) estimator [15], $\hat{\gamma}_k(n) = |Y_k(n)|^2/\hat{\lambda}_{N,k}(n)$, and $\hat{\lambda}_{N,k}(n)$ is the estimated noise variance which is given by [16] as

$$\hat{\lambda}_{N,k}(n) = \zeta_N\hat{\lambda}_{N,k}(n-1) + (1-\zeta_N)E\left[|N_k(n)|^2|Y_k(n)\right] \quad (5)$$

where $0 < \zeta_N < 1$ is the smoothing parameter, $E[\cdot]$ the expectation operator, and $|N_k(n)|^2$ the noise power spectrum. In Equation 5, the expectation term is also given by

$$E\left[|N_k(n)|^2|Y_k(n)\right] = E\left[|N_k(n)|^2|Y_k(n), H_0\right]p\left(H_0|Y_k(n)\right) + E\left[|N_k(n)|^2|Y_k(n), H_1\right]p\left(H_1|Y_k(n)\right) \quad (6)$$

Where

$$E\left[|N_k(n)|^2|Y_k(n), H_0\right] = \left|Y_k(n)\right|^2 \quad (7)$$

$$E\left[|N_k(n)|^2|Y_k(n), H_1\right] = \left(\frac{\hat{\xi}_k(n)}{1+\hat{\xi}_k(n)}\right)\hat{\lambda}_{N,k}(n) + \left(\frac{1}{1+\hat{\xi}_k(n)}\right)^2 |Y_k(n)|^2 \quad (8)$$

In Equation 6, $p(H_0|Y_k(n))$ is the SAP at the kth frequency bin and derived from the Bayes' rule such that [8]

$$p\left(H_0|Y_k(n)\right) = \frac{p(Y_k(n)|H_0)p(H_0)}{p(Y_k(n)|H_0)p(H_0) + p(Y_k(n)|H_1)p(H_1)} = \frac{1}{1+(p(H_1)/p(H_0))\exp(\Lambda_k(n))} \quad (9)$$

with $p(H_0)$ representing the *a priori* probability of speech absence which is set to 0.2 in our case.

With the estimated parameters, the decision rule of SMVAD is given by

$$\phi(n) = \frac{1}{M}\sum_{k=0}^{M-1}\hat{\Lambda}_k(n) \underset{H_0}{\overset{H_1}{\gtrless}} n \quad (10)$$

where $\hat{\Lambda}_k(n)$ is the LLR utilizing $\hat{\xi}_k(n)$ and $\hat{\gamma}_k(n)$ and η is the decision threshold.

3. Analysis of LRT-based decision rule

In general, the LRT-based decision rule of SMVAD overlooks two undesirable problems. The first is that LRs cannot always show high values even if the input signal contains speech. Because of the basic assumption that noise is uncorrelated with speech, the complex Gaussian models of the input signal must satisfy the following condition:

$$\lambda_{N,k} + \lambda_{S,k} \geq \lambda_{S,k}. \quad (11)$$

With condition (11), the peak of $p(Y(n)|H_1)$ is always lower than or equal to that of $p(Y(n)|H_0)$. Thus, $p(Y(n)|H_1)$ cannot always larger than $p(Y(n)|H_0)$ even in the case of speech presence. Therefore, an increased variance does not guarantee an increased LLR values. Figure 1 shows an example of this case with the three complex Gaussian PDFs having different variances. In Figure 1, the dotted line indicates the PDF only with noise variance, and the dashed and the solid lines represent the PDFs for which the *a priori* SNRs are -10 and 5 dB with the given noise variance, respectively.

Figure 2 shows two LLR curves related to Figure 1 with respect to the spectral amplitude of input signals where the dashed line indicates the LLR with the lower *a priori* SNR and the solid line represents the LLR with the higher *a priori* SNR, respectively. In Figure 2, the dashed circles show the difference between two LLRs for a low- and high-powered spectra at the given frequency bin. In the case of the left circle, there is a small difference between the solid and the dashed lines, and the solid line may be rather lower than the dashed line,

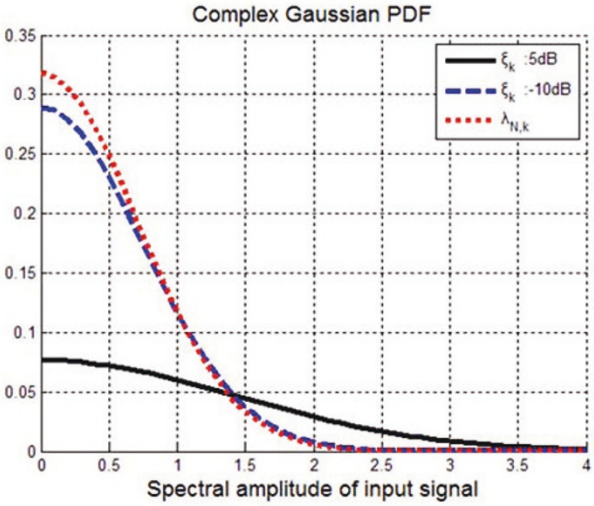


Figure 1 Three complex Gaussian PDFs with different variances.

even though the given *a priori* SNRs show a substantial difference. On the other hand, the right circle shows a large difference between the two LLR lines.

By inspecting the two cases, it is observed that the spectral power of input signals plays an important role in making the decision rule have a better discriminative property, because the conventional decision rule of the SMVAD was the average of all LLRs. In other words, the accuracy of the decision rule may be degraded by the LLRs derived from low-powered input spectrum.

Figure 3 shows an example of the undesirable cases of LLRs in a speech frame caused by both high *a priori* SNR and low-spectral power. In Figure 3a, x-axis represents the frequency bin index, the solid line indicates

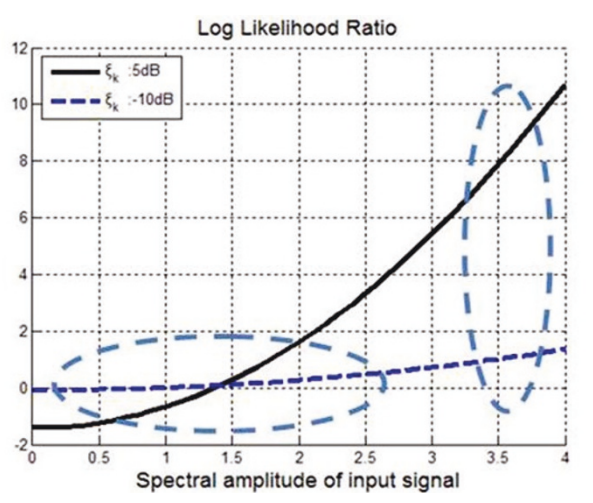


Figure 2 Two LLRs versus amplitude of input signal.

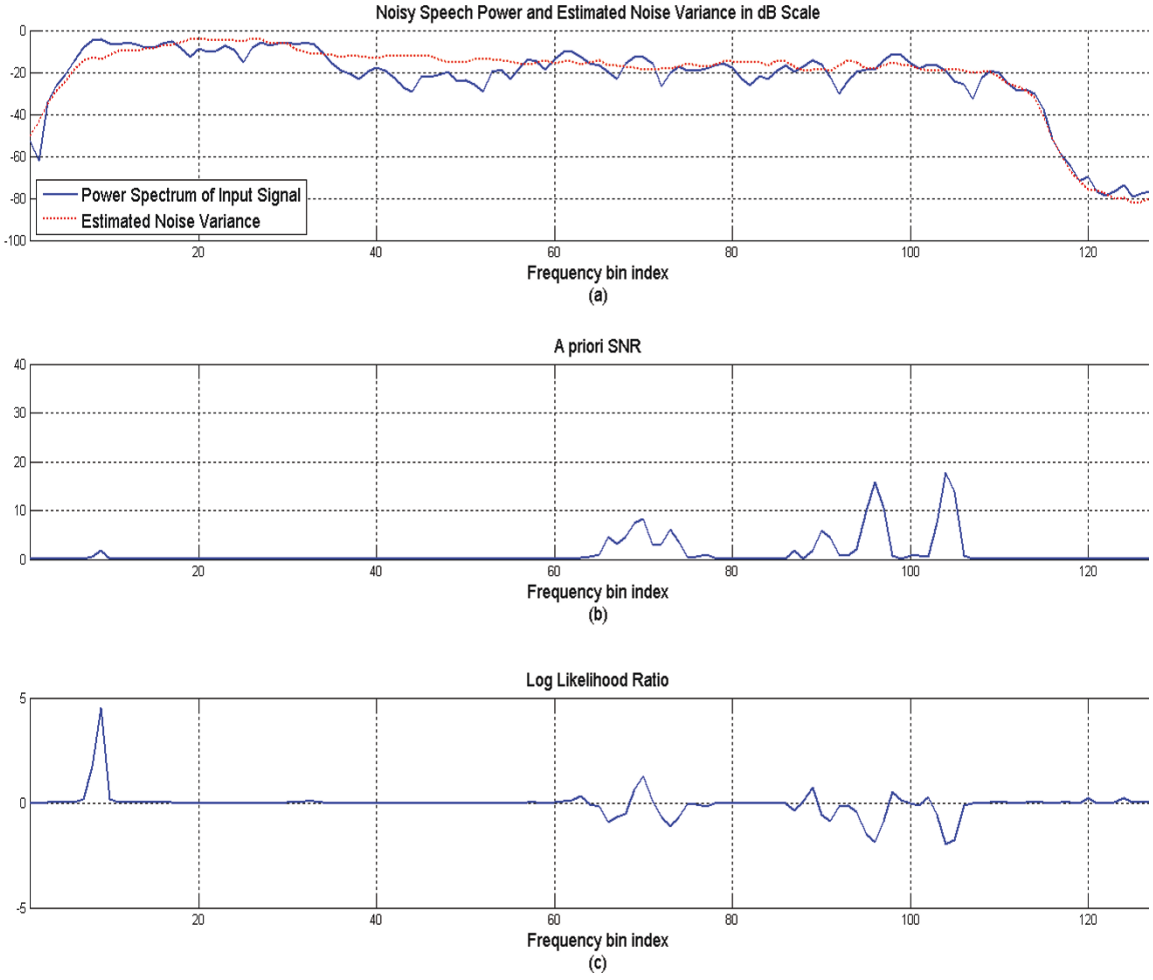


Figure 3 Example of LLR according to input signal power, noise variance, and *a priori* SNR in a speech frame in car noise (5dB SNR).

the spectral power of noisy speech, and the dotted line represents the estimated noise variance. As shown in Figure 3b, even though the *a priori* SNR is estimated to be high in this speech frame, it causes low LLR values when the input signal power is lower than the estimated noise variance. If the LLRs shown in Figure 3c are employed for decisions in SMVAD, this speech frame could not be detected as a speech frame. In Figure 3c, it is also observed that most of LLRs are close to 0 and high LLRs are located on the most high-powered frequency region. From the investigation of the first problem, it is concluded that the decision rule of SMVAD uses LLR at every frequency bin but not all of them contribute to a correct decision in case of the components in the low-powered frequency region.

The second problem of SMVAD also occurs on the low-powered frequency region of the noise. As mentioned in Section 1, the basic assumption on SMVAD is that noise is stationary over a long period of time, but, in practice, most of real noise powers tend to change slightly frame-by-frame. To accommodate this phenomenon, the estimated noise variance, $\hat{\lambda}_{N,k}(n)$, needs to be re-considered. Since the fixed smoothing parameter ζ_N is generally chosen to be very close to 1, estimated noise variance $\hat{\lambda}_{N,k}(n)$ changes very smoothly. By this effect, estimated noise variance $\hat{\lambda}_{N,k}(n)$ can keep the *a priori* SNR $\hat{\xi}_k(n)$ very low along the noise-only frames. As a result, the LLR with estimated SNRs can be simplified by

$$\hat{\Lambda}_k(n) = \frac{\hat{\gamma}_k(n)\hat{\xi}_k(n)}{1+\hat{\xi}_k(n)} - \ln\left[1+\hat{\xi}_k(n)\right] \approx \hat{\gamma}_k(n)\hat{\xi}_k(n), \quad \text{if } \hat{\xi}_k(n) \ll 1 \quad (12)$$

Because of the smoothing operation, the *a priori* SNR $\hat{\xi}_k(n)$ can be kept low or change very smoothly over the non-speech region. However, the *a posteriori* SNR $\hat{\gamma}_k(n)$ and is greatly influenced by the unstable property

of noise when noise is not stationary. Actually, the variations in the noise spectrum are not so serious in terms of the actual absolute value of noise and this phenomenon is good for estimating the reliable noise variance. However, since the *a posteriori* SNR is the ratio of the varying input spectrum $|Y_k(n)|^2$ to the almost fixed noise variance $\hat{\lambda}_{N,k}(n)$, the *a posteriori* SNRs in the low-powered frequency regions are more difficult to be settled on fixed low values. Figure 4 shows an example of the transition of *a posteriori* SNR with two different noise variances where *y*-axis means the ratio, (Noise variance + Varying Range)/Noise variance, which corresponds to the *a posteriori* SNR.

As shown in Figure 4, when the noise variance is very small, the transition of the *a posteriori* SNR is rapidly increased over the same varying range against the transition of the high noise variance. In the noise-only frames, the average of LLRs has to be close to 0. However, the *a posteriori* SNRs in low-powered frequency region can be high and possibly make certain LLR values as high as the levels in the speech frame. Using these LLRs in the decision rule, some of these noise frames can be detected as speech frames. Figure 5, where *x*-axis represents frequency bin, shows an actual case of the average car noise spectrum and the variance of *a posteriori* SNR according to the noise variance estimated on each frame. As already mentioned, in most high-powered frequency regions, the variances of *a posteriori* SNR are very close to 0, which means that there are low possibilities of high LLR values. In a low-powered region, on the contrary, higher variances may be shown. In case of (12), this effect brings about high *a posteriori* SNR which causes high LLRs.

In summary, if the input signal includes the speech signal, the LLRs in the low powered region could not be reliable because there is no way to judge whether the LLRs are caused by the speech signal or varying noise spectral components. Therefore, LLRs in the high-powered region are more important to let decision rule have an enhanced discriminative property. Figure 6 shows an actual case that SMVAD can save a speech frame when the decision rule of SMVAD only uses the LLRs in the high-powered region. In Figure 6, if we average all LLRs for decision, we could never detect the speech frame plotted in Figure 6a. On the contrary, based on our analysis, if we select or properly weight the LLRs for the decision rule, we can detect the speech frame because all of LLRs in low-powered region in Figure 6b can be excluded from the decision or reduced by the proper weights which can attenuate the effects of unreliable LLRs.

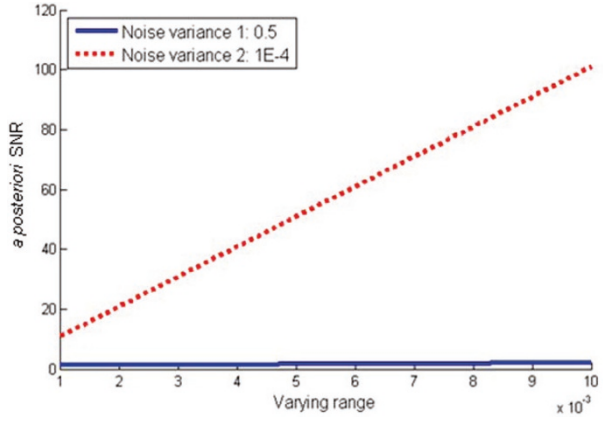


Figure 4 Transition of the *a posteriori* SNR with different noise variances as noise is varied.

4. Modified decision rules

4.1. Selection of reliable LRs

By considering two undesirable phenomena analyzed in Section 3, it is discovered that LLRs in the high-powered frequency region are more reliable than those in the low-powered region. However, the concept for the high-powered region is still ambiguous for the time-varying input frames. In case of additive noise, since the average noise spectrum is almost fixed, the high-powered region may also be fixed for every noise frame, but in case of speech frames, the high-powered region can be moved by speech signals. Therefore, we need to find the high-powered region independently from the neighboring frames and only consider the total power of the current frame. Here, three assumptions for the selection of reliable LLRs are used:

1. The property of noise is mainly dependent on high-powered but less varying frequency components for which LLRs can be kept low.
2. Most of noise spectral power is concentrated on the high-powered region, irrespective of the existence of speech component.
3. When speech component exists in the current frame, the LLRs in the high-powered frequency region obtained because of the speech component may show high value.

Therefore, we propose two modified decision rules by selecting the frequency bins with reliable LLRs on the basis of the spectral power. At first, we reorder the input signal vector in terms of the spectral power such as $Y^{(R)}(n) = [Y^{(1)}(n), Y^{(2)}(n),...,Y^{(M)}(n)]$ where $|Y^{(r)}(n)|^2 \geq |Y^{(s)}(n)|^2$ for $r > s$ and we also define LLR vector, $\hat{\Lambda}^{(R)}(n) = [\hat{\Lambda}^{(1)}(n), \hat{\Lambda}^{(2)}(n), \ldots, \hat{\Lambda}^{(M)}(n)]$ where each element $\hat{\Lambda}^{(r)}(n)$ is related to its corresponding $Y^{(r)}(n)$. With this vector, the first modified decision rule is

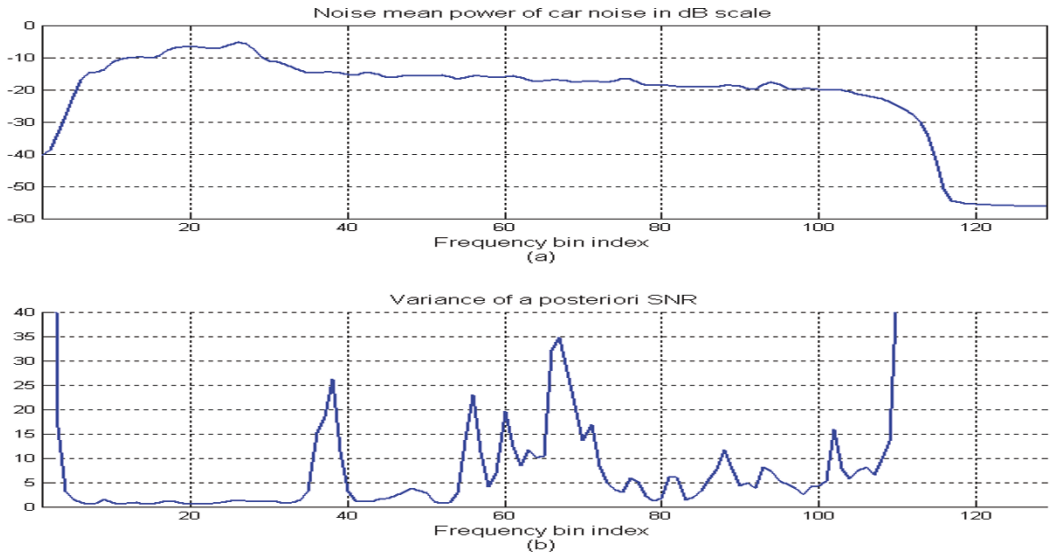


Figure 5 Average power of car noise compared with variance of the *a posteriori* SNR in noise frames.

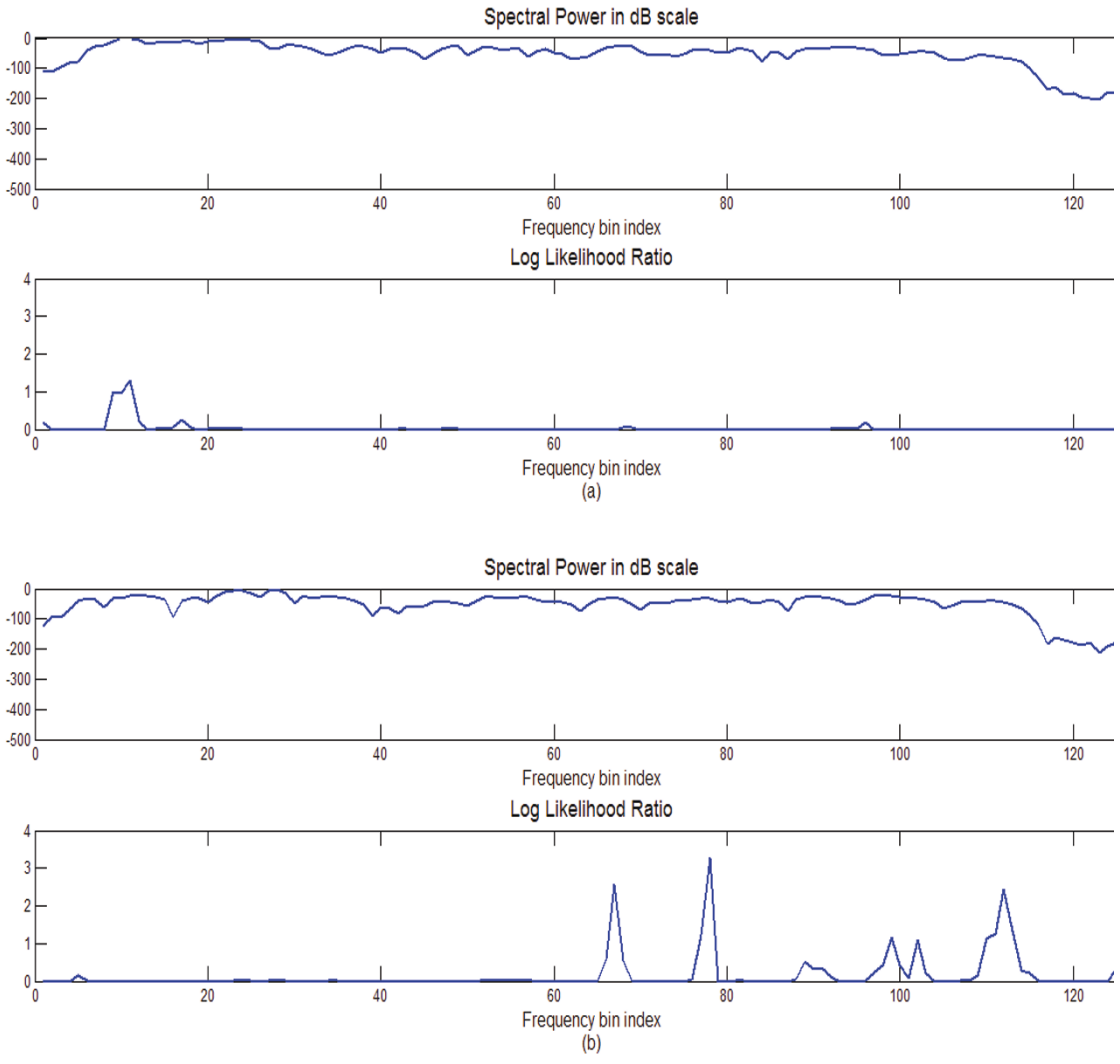


Figure 6 Input signal power and LLR in car noise (5dB SNR). **(a)** a speech frame. **(b)** a noise frame.

defined as

$$\hat{\phi}_{\text{High - power}}(n) = \frac{1}{N_H} \sum_{r=M-N_H+1}^{M} \hat{\Lambda}^{(r)}(n) \underset{<H_0}{\overset{>H_1}{}} \eta \tag{13}$$

where N_H denotes the number of LR selected by the spectral power of frequency bins. By this decision rule, we can only consider the LLRs related to high-power frequency bins and N_H is determined, empirically.

The second method is to compare the bin-power with the average power in each frame. Based on this idea, the second modified decision rule is given by

$$\hat{\phi}_{\text{Average - power}}(n) = \frac{1}{N_A} \sum_{r=1}^{M} f\left[\hat{\Lambda}^{(r)}(n),\, Y_{\text{avg}}(n)\right] \underset{<H_0}{\overset{>H_1}{}} \eta \tag{14}$$

$$Y_{\text{avg}}(n) = \frac{1}{M} \sum_{k=0}^{M-1} \left|Y_k(n)\right|^2 \tag{15}$$

where $f[\hat{\Lambda}^{(r)}(n), Y_{\text{avg}}(n)] = \hat{\Lambda}^{(r)}(n)$ if $|Y^{(r)}(n)|^2 \geq Y_{\text{avg}}(n)$, and $f[\Lambda^{(r)}(n), Y_{\text{avg}}(n)] = 0$ otherwise, and N_A is the number of spectral components greater than or equal to the average power of each frame. In this method, we assumed that the spectral power in the high-powered region of noise is always greater than the frame average power.

4.2. Weighting scheme considering reliability of LRs

With the analysis of LRT-based decision rule, we also propose a weighting scheme to reflect the reliability of each LLR. As mentioned in Section 3, since the LLRs in the low-powered region of noise are not reliable because of the variation of the *a posteriori* SNR, it is desirable to consider more importantly the spectral powers of noisy speech which are much higher than the noise variance. In addition, as the noise variance becomes closer to the highest value of the noise variances at the current frame, the LLRs derived from the *a posteriori* SNRs would be reliable. Thus, the weights applied to each LLR are defined by

$$|Y^{(r)}(n)|^2 \geq Y_{\text{avg}}(n), \tag{16}$$

where $\text{MAX}[\hat{\lambda}_N(n)]$ is equal to the highest variance of the variance vector $\hat{\lambda}_N(n)$ which is composed of $\hat{\lambda}_{N,k}(n)$ s at all kth bins. In (16), each $w_k(n)$ can reduce the effects of the unstable *a posteriori* SNRs and cause LLRs in the high-powered region to remain on their own values or to increase. Thus, the new decision rule with this weight is given by

$$\hat{\phi}_{\text{Weight}}(n) = \frac{1}{M} \sum_{k=0}^{M-1} w_k(n) \hat{\Lambda}_k(n) \tag{17}$$

5. Experiments

In the experiments, test data were composed of 60-s long speech data from the IEEE sentence listed in Table 1 and noise data from the AURORA database. The speech data were spoken by three male and three female speakers and sampled at 8 kHz. We used 20 ms frame size and 10 ms frame shift size. VAD decision was made every frame. The test material was all hand-labeled and consisted of 67% of speech and 33% of silence frames. For these experiments, we also used three types of noises, such as car, babble, and street noises at 5, 10, and 15 dB SNRs, respectively.

To compare with the proposed methods, we used four conventional methods as baseline systems. The first method is the typical SMVAD proposed by Sohn et al.

Table 1 List of sentences for experiments

1. The birch canoe slid on the smooth planks	16. The stray cat gave birth to kittens
2. He knew the skill of the great young actress	17. The lazy cow lay in the cool grass
3. Her purse was full of useless trash	18. The friendly gang left the drug store
4. Read verse out loud for pleasure	19. We talked of the side show in the circus
5. Wipe the grease off his dirty face	20. The set of china hit the floor with a crash
6. Men strive but seldom get rich	21. Clams are small, round, soft, and tasty
7. We find joy in the simplest things	22. The line where the edges join was clean
8. Hedge apples may stain your hands green	23. Stop whistling and watch the boys march
9. Hurdle the pit with the aid of a long pole	24. A cruise in warm waters in a sleek yacht is fun
10. The sky that morning was clear and bright blue	25. A good book informs of what we ought to know
11. He wrote down a long list of items	26. She has a smart way of wearing clothes
12. The drip of the rain made a pleasant sound	27. Bring your best compass to the third class
13. Smoke poured out of every crack	28. The club rented the rink for the fifth night
14. Hats are worn to tea and not to dinner	29. The flint sputtered and lit a pine torch
15. The clothes dried on a thin wooden rack	30. Let's all join as we sing the last chorus

[2] as described in Section 2, and the second method is SMVAD with the HMM hangover scheme which is specified in [3]. The third conventional method is the DWT scheme proposed in [6]. For training and testing of the DWT method, we used same parameters specified in [6]. The experiment of DWT scheme was performed with six sets of 10-s long data used for testing and the remaining data used for training by the round-robin

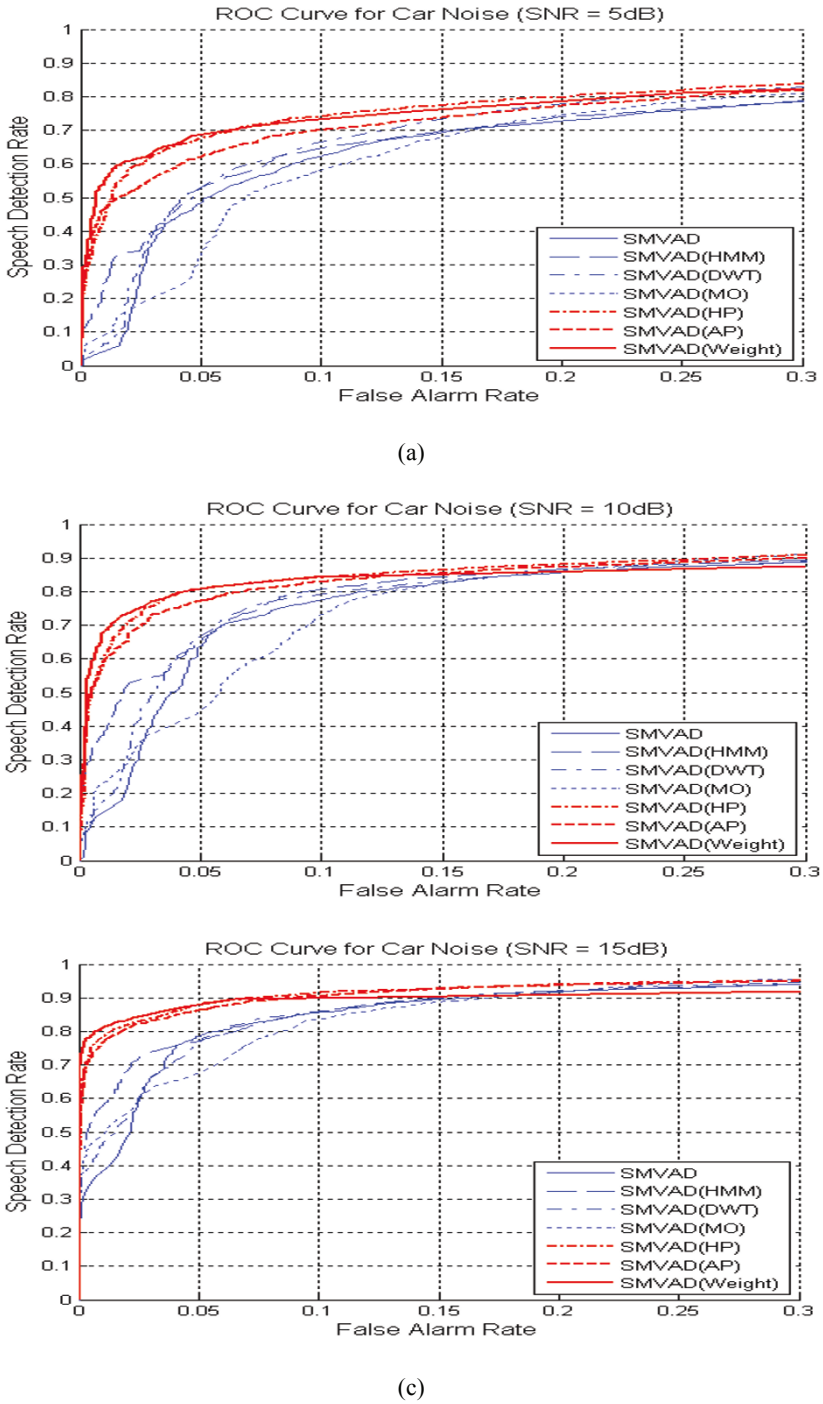
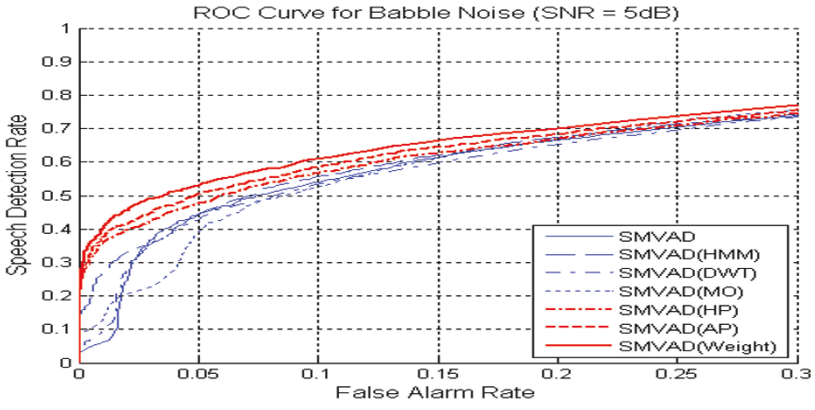


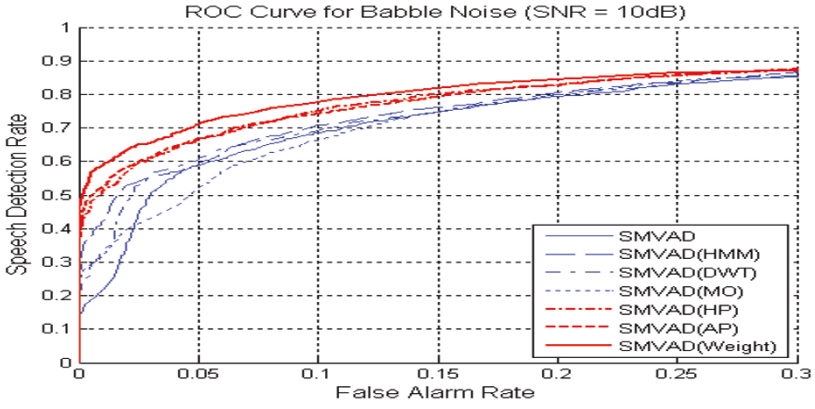
Figure 7 ROC curves for car noise. **(a)** 5dB SNR. **(b)** 10dB SNR. **(c)** 15dB SNR.

test. The fourth method is SMVAD using a multiple observation LRT (MO-LRT) proposed in [7]. For the fourth method, we used one frame, which was experimentally chosen, before and after the current frame which is going to be determined as the speech or the non-speech.

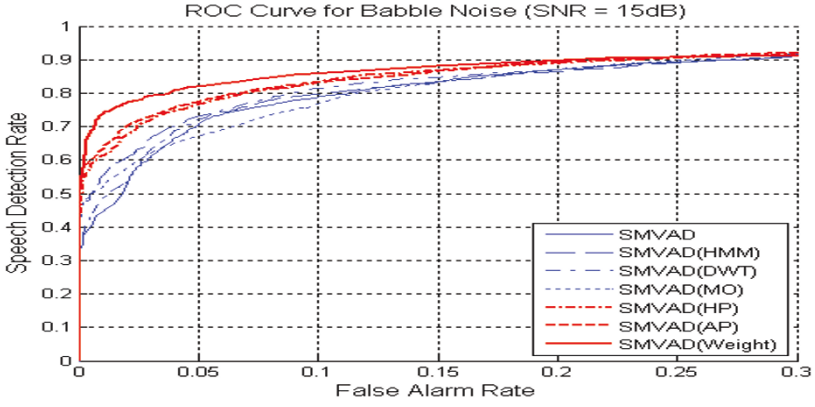
The proposed methods are all evaluated by receiver operating characteristic (ROC) curves which show



(a)

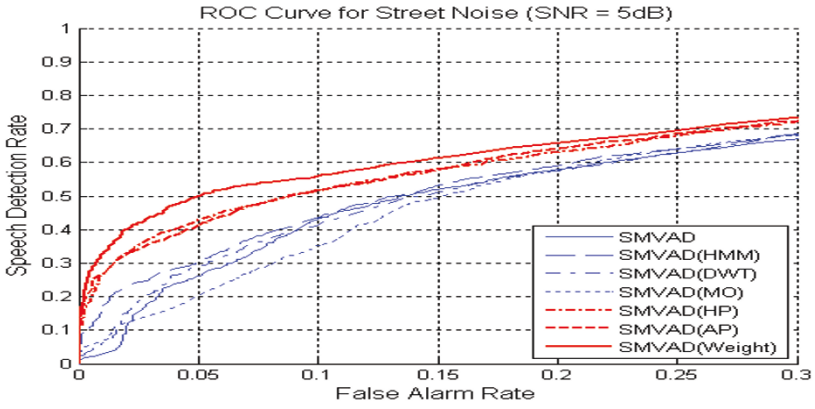


(b)

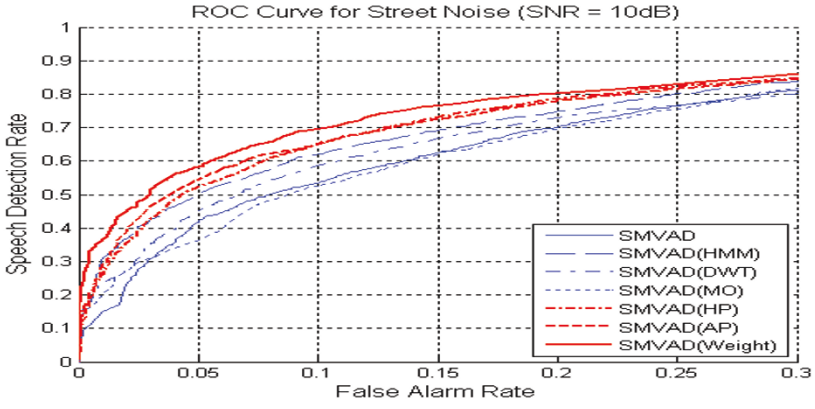


(c)

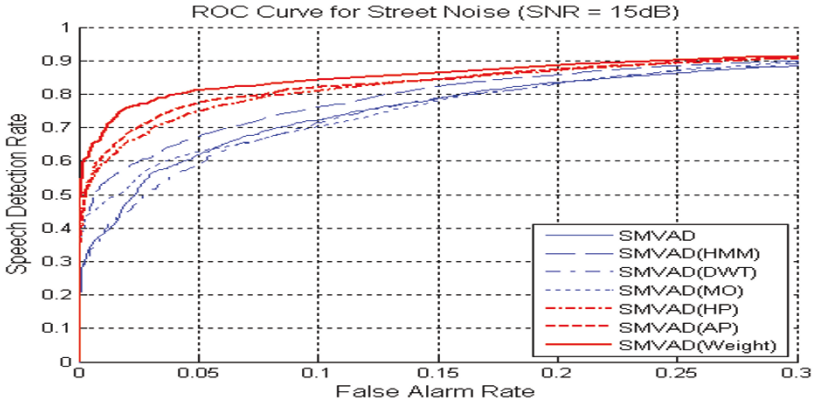
Figure 8 ROC curves for babble noise. **(a)** 5dB SNR. **(b)** 10dB SNR. **(c)** 15dB SNR.



(a)



(b)



(c)

Figure 9 ROC curves for street noise. **(a)** 5dB SNR. **(b)** 10dB SNR. **(c)** 15dB SNR.

discriminative properties of VAD between noise-only and noisy speech frames in terms of the speech detection rate (SDR) and false-alarm rate (FAR) such that

$$\mathrm{SDR} = \frac{N_{\mathrm{CS}}}{N_{\mathrm{TS}}} \tag{18}$$

$$\text{FAR} = \frac{N_{\text{FS}}}{N_{\text{TN}}} \tag{19}$$

where N_{CS}, N_{TS}, N_{FS}, and N_{TN} denote the number of correctly detected speech frames, total speech frames, falsely detected speech frames in silence regions, and total silence frames, respectively. In these experiments, we set $N_H = 10$ for $\hat{\phi}_{\text{High - power}}(n)$ in (13).

In every ROC curve, HMM, DWT, MO, HP, AP, and weight in the parenthesis denote the results from the HMM hang-over scheme, the DWT scheme, the MO-LRT scheme, the first proposed decision rule in (13), the second proposed decision rule in (14), and the decision rule with the proposed weighting scheme, respectively. For practical comparison of performances, we focus on SDRs of the methods when FAR is low. We consider that the VAD can show a reasonable discriminating property when FAR < 0.2. In the ROC curves, the red lines represent the results of the proposed methods and the blue lines indicate the results of the conventional methods.

In the car noise environment of Figure 7, all of the proposed methods show better performance than the conventional methods do. In addition, the SDRs of SMVAD(HP) and SMVAD(Weight) are at least 0.1 higher than those of SMVAD when FAR < 0.05. Especially, SMVAD(Weight) keeps the highest SDR at the extremely low FAR.

In babble noise environment of Figure 8, every proposed method also shows better performance than the conventional methods do as in the car noise environment, but the difference is that the performance improvement of SMVAD(HP) and (AP) are not noticeable. However, we can also observe that the performance improvement of SMVAD(Weight) is kept constant as the SNR becomes higher and can be still considered to be significant.

In street noise environment of Figure 9, SMVAD (Weight) is effective on improving the performance of SMVAD and shows significantly higher SDR at FAR = 0.05 with 5 dB SNR than SMVAD(HMM). In case of 10 dB SNR, the performance of SMVAD(HMM) is almost the same as SMVAD(HP) or SMVAD(AP), but it is still not better than that of SMVAD(Weight).

From the investigation of the experimental results, it is observed that SMVAD(Weight) shows the highest and the most consistent performance improvement in all noise conditions. In addition, SMVAD(DWT) did not show better performance than SMVAD(Weight) does although the complexity of SMVAD(Weight) is almost the same as that of SMVAD. By these results, we can conclude that the variation of input signal statistics has a large influence on the accuracy of VAD, and if we are not sure to know about specific noise type, it would be better to use SMVAD(Weight) for stable performance improvement under unexpected noise environments.

6. Conclusion

In this article, we introduced SMVAD, and analyzed the averaged LRT-based decision rule and its undesirable phenomena which can possibly happen in various noise environments. To reduce the undesirable phenomena, we proposed two types of modified decision rules based on the selection of reliable LRs and a weighting scheme applied to LLRs used in the decision rule. Compared with the conventional methods, it was proved that the proposed methods are much more robust in various noise environments without any training procedure and additional computational complexity. Among the proposed methods, SMVAD(Weight) showed the most reliable performance improvement under various conditions.

For further studies, the properties of speech and noise, which can be applied to the weights for LLRs, are needed to be analyzed in more details.

Abbreviations

DWT: discriminative weight training; FAR: false-alarm rate; HMM: hidden Markov model; LRs: likelihood ratios; LRT: likelihood ratio test; MMSE-STSA: minimum mean square error short-time spectral amplitude; MO-LRT: multiple observation LRT; PDF: probability density function; ROC: receiver operating characteristic; SAP: speech absence probability; SDR: speech detection rate; SLRs: smoothed likelihood ratios; SMVAD: statistical model-based voice activity detector; SNR: signal-to-noise ratio; VAD: voice activity detector.

Acknowledgements

This study was supported by the NRF grant funded by the Korean government (MEST) (No. 2011-0017967).

Competing interests

The authors declare that they have no competing interests.

Received: 10 January 2011 Accepted: 27 July 2011
Published: 27 July 2011

References

1. K Srinivasan, A Gersho, Voice activity detection for cellular networks, in *Proc IEEE Speech Coding Workshop*, 85–86 (October 1993)
2. J Sohn, W Sung, A voice activity detector employing soft decision based noise spectrum adaptation, in *Proc Int Conf Acoustics, Speech, and Signal Processing*, 365–368 (1998)
3. J Sohn, NS Kim, W Sung, A statistical model-based voice activity detection. IEEE Signal Process Lett. **6**(1), 1–3 (1999). doi:10.1109/97.736233
4. YD Cho, A Kondoz, Analysis and improvement of a statistical model-based voice activity detector. IEEE Signal Process Lett. **8**(10), 276–278 (2001). doi:10.1109/97.957270
5. J-H Chang, NS Kim, SK Mitra, Voice activity detection based on multiple statistical models. IEEE Trans Signal Process. **54**(6), 1965–1976 (2006)
6. S-I Kang, Q-H Jo, J-H Chang, Discriminative weight training for a statistical model-based voice activity detection. IEEE Signal Process Lett. **15**, 170–173 (2008)
7. J Ramírez, JC Segura, C Benítez, L García, A Rubio, Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Process Lett. **12**(10), 689–692 (2005)

8. Y-G Kim, Y-J Suh, H-R Kim, Selection of reliable likelihood ratios for statistical model-based voice activity detection in *Proc of Asia Pacific Signal and Information Processing Association Annual Summit and Conf. 2009*, CD-ROM, (October 2009)
9. J Ramírez, JC Segura, C Benítez, Á Torre, AJ Rubio, A new Kullback-Leibler VAD for speech recognition in noise. IEEE Signal Process Lett. **11**(2), 266–269 (2004). doi:10.1109/LSP.2003.821762
10. PN Garner, T Fukada, Y Komori, A differential spectral voice activity detector. in *Proc Int Conf Acoustics, Speech, and Signal Processing*, 597–600 (2004)
11. A Davis, S Nordholm, R Togneri, Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. IEEE Trans Audio Speech Lang Process. **14**(2), 412–424 (2006)
12. JW Shin, HJ Kwon, SH Jin, NS Kim, Voice activity detection based on conditional MAP criterion. IEEE Signal Process Lett. **15**, 257–260 (2008)
13. J Ramírez, JM Górriz, JC Segura, CG Puntonet, AJ Rubio, Speech/non-speech discrimination based on contextual information integrated bispectrum LRT. IEEE Signal Process Lett. **13**(8), 497–500 (2006)
14. J-H Chang, JW Shin, NS Kim, Voice activity detector employing generalised Gaussian distribution. Electron Lett. **40**(24), 1561–1563 (2004). doi:10.1049/el:20047090
15. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process. **ASSP-32**(6), 1109–1121 (1984)
16. NS Kim, J-H Chang, Spectral enhancement based on global soft decision. IEEE Signal Process Lett. **7**(5), 108–110 (2000). doi:10.1109/97.841154

doi:10.1186/1687-6180-2011-31
Cite this article as: Kim *et al.*: Reliable likelihood ratios for statistical model-based voice activity detector with low false-alarm rate. *EURASIP Journal on Advances in Signal Processing* 2011 **2011**:31.