Selection of Reliable Likelihood Ratios for Statistical Model-Based Voice Activity Detection

[#]Younggwan Kim, [†]Youngjoo Suh, and [*]Hoirin Kim
Korea Advanced Institute of Science and Technology, 119, Munjiro, Yuseong-gu, Daejeon, Korea
[#]E-mail: cleanthink@kaist.ac.kr Tel: +82-42-350-6221
[†]E-mail: yjsuh@kaist.ac.kr Tel: +82-42-350-6830
[*]E-mail: hrkim@ee.kaist.ac.kr Tel: +82-42-350-6139

Abstract— **A statistical model-based voice activity detection (VAD) is a robust algorithm in noisy condition to detect speech region from input signal by speech and non-speech statistical model such as complex Gaussian probability density function (PDF). The decision rule used in this VAD is based on Bayes' rule and considers likelihood ratios (LRs) in whole frequency region. In this VAD, however, the Bayes' rule may cause a decision error. With the statistical model, we analyze why this problem happens and show how we can decrease the decision error by using the LRs at selected frequency bins having relatively high spectral power in each frame. The performance of this VAD is evaluated by receiver operating characteristic (ROC) curves and summarized in a table, and the results from proposed methods show better performances than those of typical statistical model-based VAD.**

I. Introduction

The purpose of voice activity detection (VAD) is to discriminate speech and silence region from input signal in various noisy conditions. Nowadays, VAD system is used in many fields such as speech recognition, speaker recognition, speech coding, and speech enhancement as a preprocessor because VAD helps to increase the performance of these systems and save computation time efficiently.

In these days, a popular algorithm for VAD has been the statistical model-based method using complex Gaussian probability density function proposed by Sohn *et al*. [1]. To enhance the performance of this VAD, many additional methods, which consider smoothed decision rule [2], noise adaptation [3], decision rules in sequential frame [4], and various statistical models [5], have been proposed.

The statistical model-based VAD can make a decision by arithmetic average of log likelihood ratios (LLRs) in whole frequency region. As simply mentioned above, additional methods for this VAD tried to improve the performance but not to analyze decision rule itself. However, this decision rule holds a possibility which can cause errors with several assumptions and estimations used to constitute a statistical model. This unconsidered problem can debase the accuracy of the VAD system.

In this paper, we first introduce the conventional statistical model-based VAD and present why the Bayes' rule with statistical model and average of LLRs at every frequency bin can make inaccurate decision and how we can reduce the effects of this decision rule by selecting specific LLRs according to spectral powers of frequency bins. Next, based on our analysis, we propose two modified decision rules using selected LLRs only at relatively high-power frequency bins in each frame. Finally, we show that the proposed decision rule enhances performance of the statistical model-based VAD and point our further work.

II. Statistical Model-Based VAD

The statistical model-based VAD starts from two hypotheses H_0 and H_1 which assume that there exists only noise or noisy speech, respectively. The assumptions are described as

$$\begin{aligned} H_0&: \boldsymbol{Y(n)} = \boldsymbol{N(n)} \\ H_1&: \boldsymbol{Y(n)} = \boldsymbol{S(n)} + \boldsymbol{N(n)} \end{aligned}$$

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

1) Clean speech and noise are uncorrelated.
2) All DFT coefficients are independent.
3) The likelihood of $Y_k(n)$ conditioned on each hypothesis can be modeled by zero-mean complex Gaussian pdf.

By those assumptions, the likelihoods of $\boldsymbol{Y(n)}$ are given by

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

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

where $\lambda_{N,k}$ and $\lambda_{S,k}$ denote the variance of noise and clean speech. With these likelihoods, the decision rule by LLRs in entire frequency region can be constituted by

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

$$\phi(n) = \frac{1}{M}\sum_{k=0}^{M-1} \Lambda_k(n) \;\; \gtrless_{H_0}^{H_1} \;\eta \quad (4)$$

where ξ_k is $\lambda_{S,k}/\lambda_{N,k}$ denoting *a priori* signal to noise ratio (SNR) and γ_k is $|Y_k(n)|^2/\lambda_{N,k}$ denoting *a posteriori* SNR. When the result is greater than a threshold value η, we are allowed to determine that the signal in current frame includes speech signal.

III. Analysis of Bayes Decision Rule

Unlike (1) and (2), there is no fixed information about the variance of noise and clean speech. Thus, we have difficulty to constitute the decision rule using *a priori* and *a posteriori* SNR and the well-known method to estimate *a priori* SNR $\hat{\xi}_k$ is the decision-directed (DD) method [6] which is given by

$$\hat{\xi}_k(n) = \alpha\frac{|\hat{S}_k(n-1)|^2}{\hat{\lambda}_{N,k}(n-1)} + (1-\ \alpha)MAX[\hat{\gamma}_k(n) - 1{,}0] \quad (5)$$

where $|\hat{S}_k(n)|^2$ is an estimate for clean speech by minimum mean square error short-time spectral amplitude (MMSE-STSA) estimator, α is a smoothing parameter, $\hat{\lambda}_{N,k}(n)$ is the estimated value by the noise estimation process in [7], and $\hat{\gamma}_k(n) = |Y_k(n)|^2/\hat{\lambda}_{N,k}(n)$.

The complex Gaussian models of input signal which are $p(\boldsymbol{Y}(\boldsymbol{n})|H_0)$ and $p(\boldsymbol{Y}(\boldsymbol{n})|H_1)$ show that the variance of $p(\boldsymbol{Y}(\boldsymbol{n})|H_1)$ is always greater than or equal to the variance of $p(\boldsymbol{Y}(\boldsymbol{n})|H_0)$. In addition, the decision rule assumes that $\phi(n)$ is higher when speech is present than absent. In practice, however, $p(\boldsymbol{Y}(\boldsymbol{n})|H_1)$ cannot always greater than $p(\boldsymbol{Y}(\boldsymbol{n})|H_0)$ even when there is a speech in input signal. So to speak, increased variance does not guarantee increased value of LLR. As shown in Fig. 1, even if there is a speech at a specific frequency bin, LLR is not always greater than 0 because the inaccuracy of noise estimate does not allow $|Y_k(n)|^2$ to be always greater than σ_{th}.

As described in [7], the variance of noise in current frame is estimated in previous frame without any information about current frame by assuming that noises are almost stationary. However, because of imperfect estimation of noise variance, $\hat{\gamma}_k(n)$ is often less than 1, so the LLRs at those frequency bins can possibly decrease the accuracy of decision rule in current frame.

IV. Selection of Frequency Bins Participating in Decision Rule

In practice, the decision rule of VAD considers the LLRs at every frequency bin, but not all of them contribute to make a correct decision. Therefore, we should select the frequency bins having reliable LLR for correct decision. To find the frequency bins, we first need to analyze $\hat{\xi}_k(n)$ and $\hat{\gamma}_k(n)$.

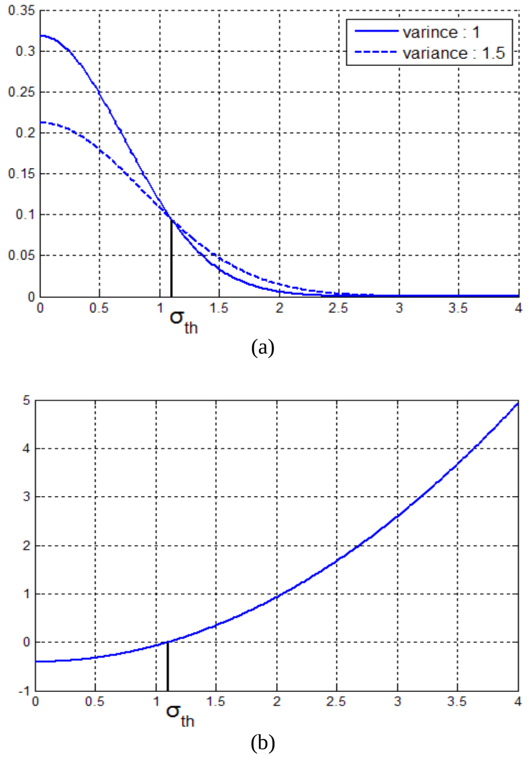


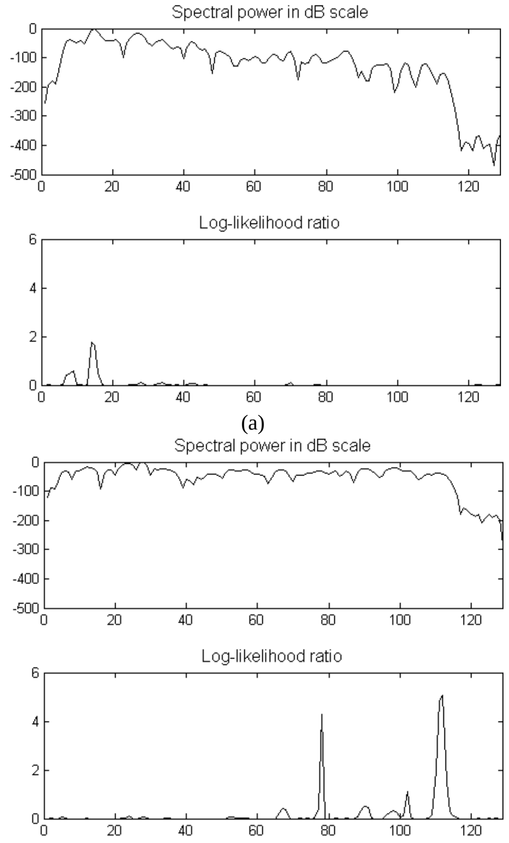
Fig. 1 (a) Two complex Gaussian distributions having same zero-mean and different variances. (b) Log likelihood ratio of distribution of dashed line over distribution of solid line. Note that the x-axis denotes $|Y_k(n)|$ in both figures.

$\hat{\xi}_k(n)$ and $\hat{\gamma}_k(n)$ are represented in terms of not only noise variance but also speech variance and spectral power of input signal. Thus, it can be said that the higher $|\hat{S}_k(n)|^2$ and $|Y_k(n)|^2$ are, the less $\hat{\xi}_k(n)$ and $\hat{\gamma}_k(n)$ are affected by inaccuracy of noise estimation because the difference between actual and estimated values can make influences on those SNRs greater as the value of numerator becomes smaller. Therefore, at low-power frequency bins, LLRs are more severely affected by inaccurate noise estimation. Thus, a possibility of error can be caused more likely at low-power frequency bins because $|Y_k(n)|^2$ at those bins can be often less than σ_{th} and that is the reason why we have to choose high-power frequency bins as the members of decision rule.

As shown in Fig. 2, in case of speech frame, there are two peaks of LLRs at high-power frequency region, and also it is noticeable that the LLR is close to zero at most low-power frequency region. However, in case of noise-only frame, LLRs are relatively high in low-power frequency range although this frame has only noise signal. To explain this situation, we need to simplify (3) as follows:

$$\Lambda_k(n) = \frac{\gamma_k \xi_k}{1+\xi_k} - ln[1+\xi_k] \approx \gamma_k \xi_k\,, \qquad \text{if } \xi_k \ll 1. \;(6)$$

As shown in the third part of (6), LLRs are dominated only by $\gamma_k \xi_k$ when current frame is in sequence of noise frames and



(a)

(b)

Fig. 2 Examples of spectral shape in dB scale and log-likelihood ratio of (a) speech frame and (b) silence frame corrupted by 5dB car noise.

γ_k is very closely related to $|Y_k(n)|^2$ rather than $\lambda_{N,k}$. In practice, when $\boldsymbol{Y(n)}$ is noise-only signal and $|Y_k(n-1)|^2$ is very low, $\hat{\lambda}_{N,k}(n)$ is also estimated as low as $|Y_k(n-1)|^2$ is and changed smoothly. In nth frame, even if $|Y_k(n)|^2$ is increased slightly, it can cause high effect on $\hat{\gamma}_k(n)$ and LLR. That is why LLRs in low-power frequency range can be higher than those in high-power frequency range. As you can also see in Fig.2, if we use (4) as a decision rule, we cannot discriminate this speech frame from input signal because it is apparent that the average of LLRs in whole frequency range is less than those of the noise frame. Thus, choosing the LLRs at high power frequency bins can be helpful to decrease the result of decision rule of noise-only frame and increase those of speech frame.

With these properties, we propose two ways to select the frequency bins having reliable LLRs for decision rule according to spectral power. At first, we reorder the input signal vector in terms of power such as $\boldsymbol{Y^R}(n) = [Y^1(n), Y^2(n), \dots, Y^M(n)]$ where $|Y^r(n)|^2 \geq |Y^s(n)|^2$ when $r > s$ and also define LLR vector, $\boldsymbol{\Lambda(n)} = [\Lambda^1(n), \Lambda^2(n), \dots, \Lambda^M(n)]$ where each element $\Lambda^r(n)$ is related to $Y^r(n)$. With these vectors, the modified decision rule is proposed by

$$\phi_{high-power}(n) = \frac{1}{H} \sum_{r=M-H+1}^{M} \Lambda^r(n) \tag{7}$$

where H denotes the number of likelihood ratios selected by the power of frequency bins. In this decision rule, we only consider the LLRs related to high-power frequency bins. The second approach is to compare the bin-power with average power in each frame. With this consideration, the second modified decision rule is proposed by

$$Y_{avg}(n) = \frac{1}{M} \sum_{k=0}^{M-1} |Y_k(n)|^2 \tag{8}$$

$$\phi_{average-power}(n) = \frac{1}{Q} \sum_{r=1}^{M} F[\Lambda^r(n), Y_{avg}(n)] \tag{9}$$

where $F\left[\Lambda^r(n), Y_{avg}(n)\right] = \Lambda^r(n)$ if $|Y^r(n)|^2 \geq Y_{avg}(n)$, and $F\left[\Lambda^r(n), Y_{avg}(n)\right] = 0$ otherwise, and Q is the number of frequency bins having a power greater than or equal to average power of each frame.

V. Experiments and Results

The proposed decision rules are evaluated by receiver operating characteristic (ROC) curves in Fig. 3 which show performances of VAD in terms of speech detection rate (*SDR*) and false-alarm rate (*FAR*) such that

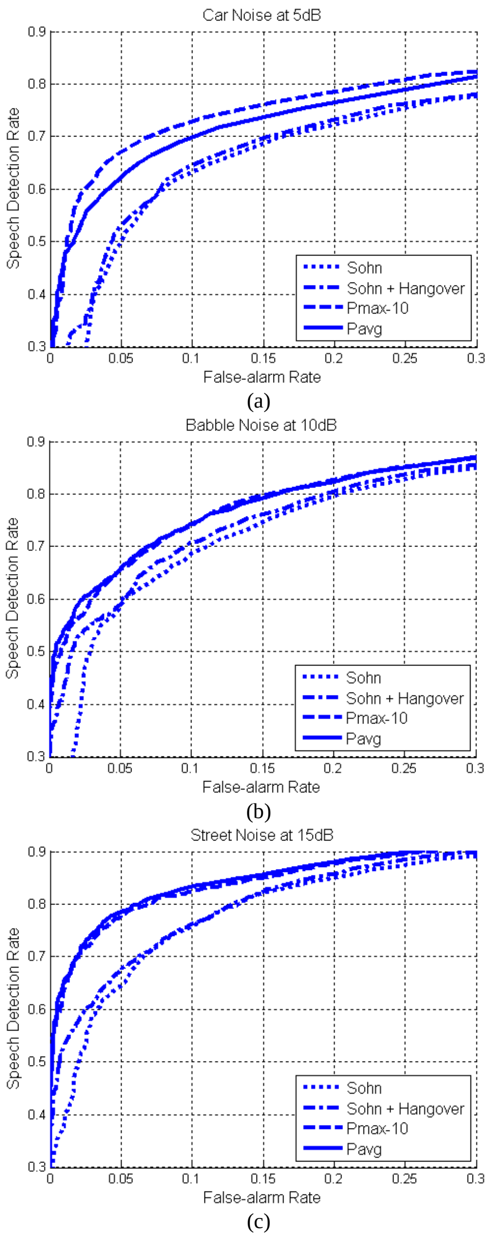
$$SDR = \frac{N_{CS}}{N_{TS}} \tag{10}$$

$$FAR = \frac{N_{FS}}{N_{TN}} \tag{11}$$

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 frames, and total silence frames, respectively. Each curve is plotted as the threshold value η is changed.

The test data were composed of 60 s long speech data from IEEE sentence and noise data from AURORA database. The speech data were spoken by 3 male and 3 female speakers and sampled at 8 kHz. We used 20 ms frame size and shifted it 10 ms as a decision unit. The test material was all hand-labeled and consisted of 67% of speech and 33% of silence frames. For the test, we also used three types of noises such as car noise, babble noise, and street noise. To evaluate proposed algorithms, we compared this algorithm with the VAD proposed by Sohn *et al.* [1], which also includes HMM-based hang-over scheme.

In this experiment, we used H = 10 for (7) which is the first proposed decision rule. As shown in Fig. 3, Pmax–10, Pavg, and Hangover denote the results from (7), (9), and HMM-based hang-over scheme, respectively. The proposed methods show much better performances, especially for $FAR < 0.1$. In



(a)

(b)

(c)

Fig. 3 ROC curves for (a) car noise at 5dB, (b) babble noise at 10dB, and (c) street noise at 15dB

case of stationary noise such as car noise, the result from Pmax-10 shows the best performance. In case of less stationary noises such as babble and street noise, Pavg shows slightly better performance than Pmax-10.

The overall evaluated performances according to types of noises and SNRs are summarized in Table I where P_{SDR} and P_{FAR} denote *SDR* and *FAR* in percentage. To compare the proposed methods with typical methods, we organized Table I with equal or slightly different P_{FAR} in same noise condition. In Table I, Pmax-10 and Pavg show that the proposed methods can save much more speech frames on same P_{FAR} than traditional algorithms can do.

TABLE I
SUMMARY OF PERFORMANCE EVALUATION IN DIFFERENT NOISE CONDITIONS.

	Sohn		Sohn + Hang-over		Pmax-10		Pavg	
SNR (dB)	P_{SDR}	P_{FAR}	P_{SDR}	P_{FAR}	P_{SDR}	P_{FAR}	P_{SDR}	P_{FAR}
Car noise								
5	50.21	5.02	52.94	5.02	67.24	4.98	62.37	4.98
10	64.84	5.02	65.62	5.02	80.10	4.98	78.89	5.09
15	77.22	5.09	78.17	5.09	87.72	5.02	86.14	5.09
Babble noise								
5	42.73	5.06	44.69	5.06	46.64	5.02	48.93	4.94
10	59.31	5.02	59.91	5.02	65.55	5.02	65.61	4.98
15	69.87	5.02	73.14	5.02	76.90	5.02	77.93	5.06
Street noise								
5	44.11	9.92	43.59	10.0	52.86	10.1	52.67	10.0
10	59.87	9.92	62.11	9.96	68.64	9.96	69.81	9.92
15	75.58	9.96	76.22	10.0	82.49	9.96	83.44	9.96

VI. CONCLUSIONS

In this paper, we have proposed a new approach to the statistical model-based VAD through reliable spectral power and modified decision rules which can reduce the error rate caused by Bayes' rule with complex Gaussian distribution. Our analysis showed how inaccurate noise estimation affects LLRs in low-power frequency range. By considering this influence, our modified decision rules have been proved better than typical statistical model-based VAD algorithms. Further work of this study is to find more accurate condition for the selection of frequency bins in each frame.

REFERENCES

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[2] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, pp. 276–278, 2001.

[3] J. Sohn and W. Sung, "A Voice Activity Detector employing soft decision based noise spectrum adaptation," *ICASSP*, vol. 1, pp. 365–368, 1998.

[4] J. Ram´ırez, J. C. Segura, C. Ben´ıtez, L. Garc´ıa, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, 2005.

[5] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, pp. 1965–1976, Jun. 2006.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1190–1121, Dec. 1984.

[7] N. S. Kim and J. -H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Process. Lett.*, vol. 7, no. 5, pp. 108–110, May 2000.