AN IMPROVED STATISTICAL MODEL-BASED VAD ALGORITHM WITH AN ADAPTIVE THRESHOLD

Sang-Sik Ahn and Yoon-Chang Lee*

ABSTRACT

A voice activity detection (VAD) algorithm with a fixed threshold cannot follow fluctuations of signal and noise power level in a time-varying environment and the inability to adapt to a time-varying environment severely limits the VAD performance. Therefore, we need to employ an adaptive threshold with which the VAD will have enhanced performance even in a time-varying SNR environment. In this paper, we propose an improved statistical model-based VAD algorithm employing preprocessing and an adaptive threshold. We also perform extensive computer simulations to demonstrate performance improvement of the proposed algorithm under various noise environments, when compared to other algorithms.

Key Words: voice activity detection, preprocessing, adaptive threshold.

I. INTRODUCTION

As a means to reduce the average bit rate for speech storage and transmission, silence suppression techniques have been proposed (Srinivasan and Gersho, 1993). By assigning fewer or no bits while a speaker is in a silent state, a speech coder employing a voice activity detection (VAD) can reduce the required transmission or storage capacity significantly. In a typical conversation, a speaker talks for less than 40% of the time and he or she is in silence for the remaining period. When the VAD algorithm is employed and discontinuous transmission is in operation, the transmitter is switched off if silence is detected. Utilizing this technique, mobile communication systems, for example, can increase the system capacity by reducing the required bandwidth and transmitting power.

In systems where the background noise level is very low, a simple energy level detecting algorithm can be used to detect the silence period. The most well known VAD algorithm of this kind is the G.729B VAD (ITU-T Recommendation, 1996). On the other hand, in systems where a large background noise is present, it is impossible to distinguish noisy speech from background noise by using the simple energy level detection. Yang (1993) proposed an improved energy level-based VAD utilizing the average subband a priori SNR. However, its performance is not good enough in the low SNR case. Therefore, a more intelligent algorithm is required. To this end, a statistical model-based algorithm has been proposed (Sohn and Sung, 1998) and further optimized (Sohn *et al.*, 1999) by employing the decision-directed method to estimate the unknown parameters.

With the fixed threshold level, however, the VAD cannot follow fluctuations of signal and noise power level in a time-varying environment and the inability to adapt to a time-varying environment severely limits the VAD performance. Therefore, we need to employ an adaptive threshold with which the VAD will have enhanced performance even in a time-varying SNR environment. In this paper, we propose an improved statistical model-based VAD algorithm employing preprocessing and an adaptive threshold.

This paper is organized as follows. In section II, we summarize a statistical model-based VAD algorithm. In section III, we present a preprocessing algorithm employing power subtraction and matched filter, and propose an improved statistical model-based VAD algorithm with an adaptive threshold. Finally, we evaluate the performance of the proposed

*Corresponding author. (Tel: 82-41-860-1792; Fax: 82-41-866-0401; Email: ychlee@korea.ac.kr)

The authors are with the Department of Electronics and Information Engineering, Korea University, 208, Seochang, Jochiwon, Yonki, Chungnam, 339-700, Korea.

algorithm by extensive computer simulations in section IV, and discuss experimental results in the conclusion.

II. EXISTING STATISTICAL MODEL-BASED VAD ALGORITHMS

For each frame n, assuming that the clean speech $\boldsymbol{s}_n$ is degraded by uncorrelated additive noise $\boldsymbol{v}_n$, the measured signal $\boldsymbol{x}_n$ on the two hypotheses H^0 and H^1 can be represented as follows:

$$H^0 \text{ (speech absence): } \boldsymbol{x}_n = \boldsymbol{v}_n$$

$$H^1 \text{ (speech presence): } \boldsymbol{x}_n = \boldsymbol{s}_n + \boldsymbol{v}_n. \quad (1)$$

While a simple energy level-based algorithm can be used to detect the silence period when background noise level is very low, it is impossible to distinguish noisy speech from background noise by using the simple energy level detection when a large background noise is present. Therefore, a more intelligent algorithm is required. To this end, a statistical model-based algorithm has been proposed.

Under the assumption that the DFT coefficients of noisy speech signals are asymptotically independent Gaussian random variables (Ephraim and Malah, 1984), conditional probability density functions on two hypotheses H^0 and H^1 are given by

$$p(X_{k,n}|H_n^0) = \frac{1}{\pi|V_{k,n}|^2}\exp\{-\frac{|X_{k,n}|^2}{|V_{k,n}|^2}\}, \quad (2)$$

$$p(X_{k,n}|H_n^1) = \frac{1}{\pi(|S_{k,n}|^2+|V_{k,n}|^2)}\exp\{-\frac{|X_{k,n}|^2}{|S_{k,n}|^2+|V_{k,n}|^2}\}, \quad (3)$$

where $S_{k,n}$, $V_{k,n}$, and $X_{k,n}$ are k^{th} element of M point DFT coefficient vectors of speech, noise, and noisy speech at frame n, respectively. Sohn and Sung (1998) proposed a statistical model-based VAD algorithm using a log likelihood ratio (LLR):

$$\Lambda_n = \log\frac{p(X_n|H^1)}{p(X_n|H^0)} = \sum_{k=0}^{M-1}\{\frac{|X_{k,n}|^2}{|\hat{V}_{k,n-1}|^2} - \log\frac{|X_{k,n}|^2}{|\hat{V}_{k,n-1}|^2} - 1\} \underset{H^0}{\overset{H^1}{\lessgtr}} \eta. \quad (4)$$

The noise power of each frequency bin $|\hat{V}_{k,n}|^2$ is estimated by the following recursive equation

$$|\hat{V}_{k,n}|^2 = \frac{1}{1+\varepsilon\Lambda_n}|X_{k,n}|^2 + \frac{\varepsilon\Lambda_n}{1+\varepsilon\Lambda_n}|\hat{V}_{k,n-1}|^2 \quad (5)$$

where $\varepsilon = P(H^1)/P(H^0)$ and $P(H^1)$ is the probability that the measured signal x_n is in state H^1. Thus, the noise power can be updated in every frame without a secondary VAD. However, performance degradation occurs in the speech offset regions. To overcome this problem, Cho and Kondoz (2001) proposed the following smoothed likelihood ratio (SLR)

$$\Psi_{k,n} = \exp\{\kappa\log\Psi_{k,n-1} + (1-\kappa)\log\Lambda_{k,n}\}, \quad 0 \le \kappa \le 1 \quad (6)$$

where κ is a smoothing factor, and a likelihood ratio for each frequency bin k and frame n is given by $\Lambda_{k,n} = p(X_{k,n}|H^1)/p(X_{k,n}|H^0)$, where $p(X_{k,n}|H^1)$ is the conditional probability density function that $X_{k,n}$ is in state H^1. Then, the decision on the voice activity is carried out by

$$\Psi_n = \{\prod_{k=0}^{M-1}\Psi_{k,n}\}^{1/M} \underset{H^0}{\overset{H^1}{\lessgtr}} \eta. \quad (7)$$

While this method can reduce detection error in the speech offset regions, it shows increased false-alarm probability and may have stability problems.

Despite all the efforts at improving decision statistics, performance is still heavily dependent on the threshold level η since the fixed threshold cannot follow fluctuations of signal and noise power level. Therefore, we need to employ an adaptive threshold with which VAD shows good performance even in a time-varying SNR environment.

III. PROPOSED VAD ALGORITHM

1. Preprocessing Using Power Subtraction and Matched Filter

The simplest way of enhancing speech signal in an additive noise environment is to perform a spectral decomposition of a frame of noisy speech signal and to attenuate particular spectral lines. Under the same hypothesis as in (1), the well known method is the spectral decomposition using DFT and enhancing the speech signal by power subtraction (Mcaulay and Malpass, 1980):

$$\hat{s}_{m,n} = \frac{1}{M}\sum_{k=0}^{M-1}\hat{S}_{k,n}\exp\{j\frac{2\pi km}{M}\} \quad (8)$$

where

$$\hat{S}_{k,n} = \sqrt{|X_{k,n}|^2 - |\hat{V}_{k,n}|^2}\,\frac{X_{k,n}}{|X_{k,n}|}, \quad (9)$$

where $|\hat{V}_{k,n}|^2$ is the estimated noise power and $X_{k,n}$ is the k^{th} element of M point DFT coefficient vector of noisy speech, and k, m, and n stand for frequency bin, time, and frame index, respectively.

After the power subtraction speech enhancement preprocessing, we consider $\hat{S}_{k,n}$ as the sum of the clean speech signal $S_{k,n}$ and residual noise $\omega_{k,n}$, i.e. $\hat{S}_{k,n} = S_{k,n} + \omega_{k,n}$. Then, to further enhance the SNR of $\hat{S}_{k,n}$ we utilize a matched filter. The filtered signal ζ_n is denoted by

$$\zeta_n = \sum_{k=0}^{M-1} \alpha_{k,n} \hat{S}_{k,n}, \tag{10}$$

where $\alpha_{k,n}$ is the frequency response of the matched filter. To find the $\alpha_{k,n}$ that maximizes the SNR of ζ_n, we assume that the residual noise $\omega_{k,n}$ in each frequency bin is zero mean with power $|\overline{V}_{k,n}|^2$ and they are uncorrelated with each other, then the noise power σ_n^2 in ζ_n becomes

$$\sigma_n^2 = \sum_{k=0}^{M-1} |\alpha_k|^2 |\overline{V}_{k,n}|^2 . \tag{11}$$

and the SNR of ζ_n, $|\zeta_n|^2/\sigma_n^2$ is maximized by Schwarz inequality when

$$\sigma_{k,n} = \frac{\hat{S}_{k,n}^*}{|\overline{V}_{k,n}|^2}, \tag{12}$$

where * stands for complex conjugation.

Therefore, the matched filtered signal $\tilde{S}_{k,n}$ of each frequency bin k becomes

$$\tilde{S}_{k,n} = \frac{\left| |X_{k,n}|^2 - |\hat{V}_{k,n}|^2 \right|}{|\overline{V}_{k,n}|^2} \frac{X_{k,n}}{|X_{k,n}|} . \tag{13}$$

Finally, we notice that the SNR of $\tilde{S}_{k,n}$ is much larger than that of $\hat{S}_{k,n}$. We also note that preprocessing was employed not for speech coding but for generating enhanced speech signal to help the following VAD decide whether the speech component exists or not. Therefore, the matched filtered signal is not recognizable.

2. Proposed VAD Algorithm with an Adaptive Threshold

When we employ the statistical model-based VAD algorithm with the preprocessed signal (13), we get the following modified LLR test Λ_n (Lee and Ahn, 2001)

$$\Lambda_n = \sum_{k=0}^{M-1} \{ \frac{|\tilde{S}_{k,n}|^2}{|\tilde{V}_{k,n-1}|^2} - \log \frac{|\tilde{S}_{k,n}|^2}{|\tilde{V}_{k,n-1}|^2} - 1 \} \underset{H^0}{\overset{H^1}{\lessgtr}} \eta \tag{14}$$

where $\tilde{S}_{k,n}$ denotes k^{th} element of M-point discrete Fourier transform coefficient vector of the preprocessed speech signal $\tilde{s}_n$ and the noise power $|\tilde{V}_{k,n}|^2$ is updated in the speech-absence frame with the help of the secondary VAD. It was shown that the VAD with (14) as a decision statistic had better performance than one with (4) (Lee and Ahn, 2001). However, with a fixed threshold η, (14) still cannot follow fluctuations of noise power level and thus performance is limited. To make matters worse, performance is heavily dependent on that of the secondary VAD. Therefore, to achieve satisfactory performance in a time-varying SNR environment, we need to employ an adaptive threshold such that the threshold will be proportional to the noise level. In this paper, we propose a new statistical model-based VAD algorithm with an adaptive threshold η_n:

$$\Lambda_n = \sum_{k=0}^{M-1} \{ \frac{|\tilde{S}_{k,n}|^2}{|\tilde{V}_{k,n-1}|^2} - \log \frac{|\tilde{S}_{k,n}|^2}{|\tilde{V}_{k,n-1}|^2} - 1 \} \underset{H^0}{\overset{H^1}{\lessgtr}} \eta_n , \tag{15}$$

where

$$|\tilde{S}_{k,n}|^2 = \left(\frac{|X_{k,n}|^2 - |\hat{V}_{k,n}|^2}{|\overline{V}_{k,n}|^2} \right)^2 \tag{16}$$

and the noise power $|\tilde{V}_{k,n}|^2$ is estimated by

$$|\tilde{V}_{k,n}|^2 = \mu |\tilde{V}_{k,n-1}|^2 + (1-\mu) |\tilde{S}_{k,n-1}|^2 ,$$
$$0 < \mu < 1 \tag{17}$$

during speech-absence frame without secondary VAD. We don't need the secondary VAD, since with an adaptive threshold η_n, the performance of the VAD is good enough to estimate the noise power correctly and the effect of the occasional decision errors will be attenuated by the forgetting factor μ. Generally, forgetting factor μ is selected between 0.9 and 0.95 in a stationary environment such that the current estimate is affected by about 10 recent frames. In this paper, considering a non-stationary situation and simulation results shown in Fig. 1, we choose $\mu = 0.82$ to track well the varying noise power and get the best performance. The forgetting factor μ can be decreased more for highly non-stationary noise environments, but this would make the VAD performance more sensitive to the noise power estimation error. In Fig. 1, detection probability P_D is the probability that speech or noise frames are correctly detected [$P(H^1/H^1)$ + $P(H^0/H^0)$] and total error probability P_E is the sum of clipping error probability [$P(H^0/H^1)$] and false-alarm

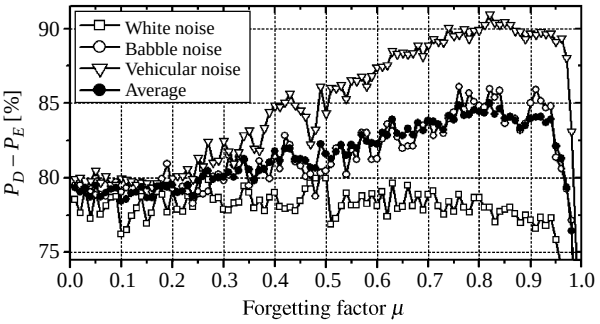

Fig. 1 Finding an optimum forgetting factor μ

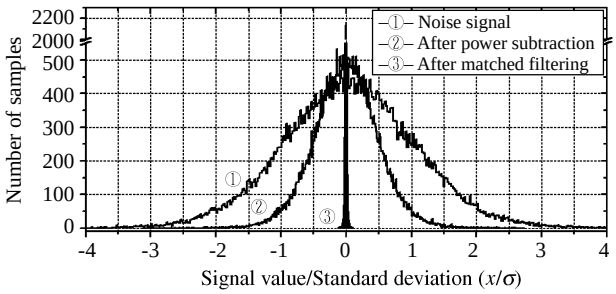

Fig. 2 Histograms of the noise signal

probability [$P(H^1/H^0)$].

The proposed adaptive threshold updating algorithm is working as follows. Within speech-absence frames, we store the LLR Λ_n and calculate the mean $\bar{\Lambda}$ and standard deviation σ_Λ. Then, the threshold is adaptively updated by the following recursive equation:

$$\eta_{n+1} = \mu\eta_n + (1-\mu)(\bar{\Lambda} + \beta_n\sigma_\Lambda),\ 0 < \mu < 1 \tag{18}$$

where μ is a forgetting factor and β_n is a weighting factor and updated by the following procedure:

STEP 1: increase frame index n
STEP 2: verify $\eta_n < (1 + 2INC)\Lambda_n$
- if yes, increase weighting factor $\beta_n = \beta_n(1 + INC)$
- if no, a) backup weighting factor $\beta_{temp} = \beta_n$
  - b) decrease weighting factor $\beta_n = \beta_n(1 - DEC)$
  - c) calculate temporary threshold η_{temp}
  - d) verify $\eta_{temp} < (1 + 2INC)\Lambda_n$
    - if yes, restore weighting factor $\beta_n = \beta_{temp}$
    - if no, use decreased weighting factor β_n

STEP 3: verify n is final frame
- if yes, END
- if no, go to STEP 1,

where INC and DEC are an increasing factor and a decreasing factor, respectively. Different values are employed to track the varying noise power more precisely at the onset and offset region of speech signal. In general, the power of the speech signal is increasing fast at onset regions and decreasing slowly at the offset regions. Therefore, INC must be larger than DEC to track the varying noise power more precisely at transitional regions.

IV. COMPUTER SIMULATIONS

Speech and noise samples are drawn from Telecom DB (Si Pro Laboratory, 1995) and NOISEX-92 DB (Varga and Steeneken, 1993), respectively. Speech signals of two males and two females who speak American English are selected from Telecom DB, and white, babble, and vehicular (Volvo car) noises are selected from NOISEX-92 DB. We arrange 25 second-long speech signals by concatenating 4 different speaker's signals and prepare total 125 second-long speech signals by concatenating five 25 second-long speech signal for the simulations. Then, to produce noisy signals at specific SNR, 125 second-long noise signals are added to the 125 second-long clean speech signals. All signals are digitized at 8 KHz sampling frequency with 16-bit resolution. Finally, we take 128 point fast Fourier transforms with 240 point-long Hamming windowed signal.

During the initial 128 frames, we store the LLR Λ_n, estimate the mean $\bar{\Lambda}$ and standard deviation σ_Λ, and update the weighting factor β_n and η_n in every frame since the VAD decisions are not reliable yet in this period. After that, we store the LLR Λ_n, estimate the mean $\bar{\Lambda}$ and standard deviation σ_Λ, and update the weighting factor β_n and η_n only in speech-absence frames.

1. Effect of the Preprocessing

To see the effect of the preprocessing, we first plotted histograms of the noise signal in Fig. 2.

We can observe the effect of the preprocessing from Fig. 2 and Fig. 3, which confirm that the preprocessing reduces noise power and increases the output SNR. Then we plotted LLR Λ_n of a noisy signal in Fig. 4.

We also notice that from Fig. 4, after preprocessing, Λ_n of a speech-presence frame is much larger than that of a speech-absence frame and thus the noise margin is increased. Noise margin is the difference between the LLR Λ_n and the threshold η_n i.e. $|\Lambda_n -$

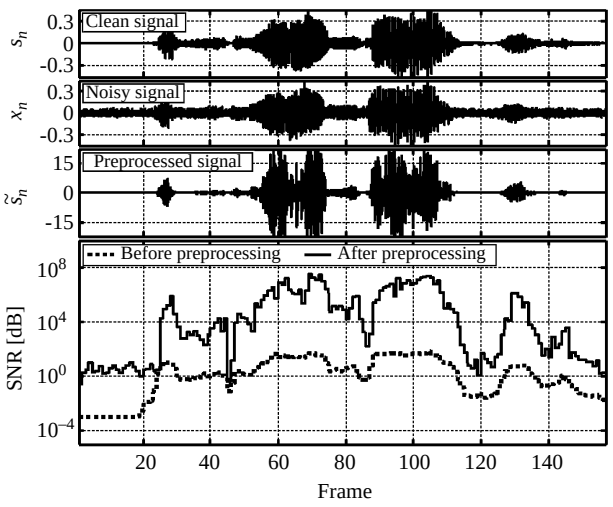

Fig. 3 SNR comparison between before and after preprocessing

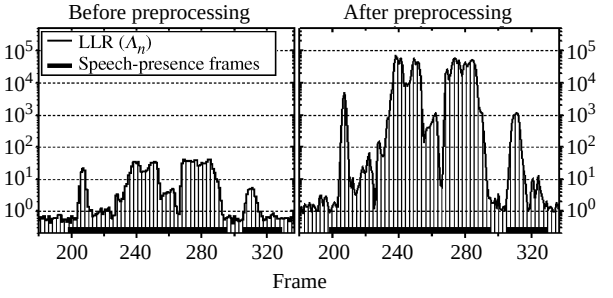

Fig. 4 LLR comparison between before and after preprocessing

$\eta_n|$. This assures that the preprocessing can help VAD correctly decide whether the speech component exists or not.

2. Effect of the Weighting Factor β_n

Figure 5 shows the adaptation of the threshold η_n when the weighting factor β_n is fixed.

We observe that when β_n is small the threshold is updated too slowly to follow the LLR Λ_n variation, and the false alarm rate will increase. On the other hand, when β_n is large the threshold is updated so fast that clipping error probability will increase. Furthermore, when clipping error occurs, the threshold will be updated in the speech-presence frame as well and, as a result, noise power will be estimated incorrectly. Therefore, we need to update β_n adaptively to track the variation of the noise power level more precisely. On the basis of Kondoz (1999) and simulation results shown in Fig. 6, we choose 1/8 and 1/32 as an increasing factor (INC) and a decreasing factor (DEC), respectively.

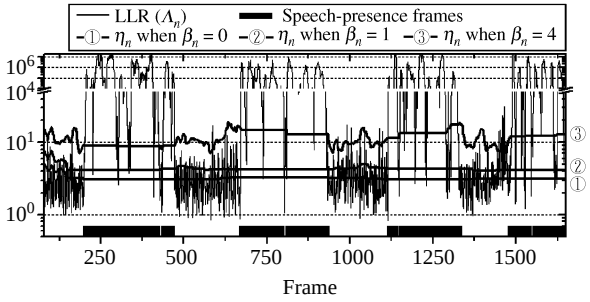

Fig. 5 Adaptation of the η_n with a fixed weighting factor β_n

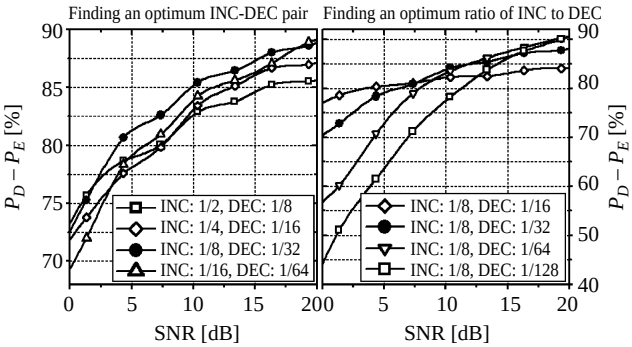

Fig. 6 Finding the optimum INC-DEC pair and their INC/DEC ratio

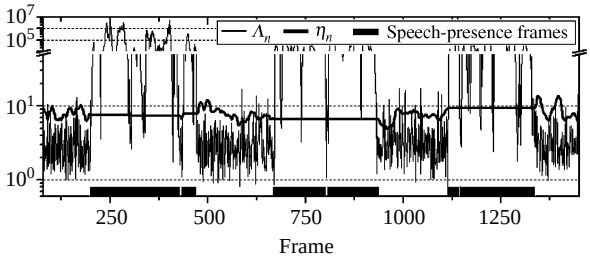

Fig. 7 Threshold adaptation with a speech-absence initial period

3. Threshold Adaptation

To demonstrate the performance of the adaptive threshold update algorithm, we present the simulation results in two cases separately. The first case, when the initial period is composed of speech-absence frames, is shown in Fig. 7 and the second case, when the initial period is composed of speech-presence frames, is shown in Fig. 8. We can see that in both cases the threshold is adapting well to the fluctuation of noise level in a short time period.

4. Performance Comparison

It is common in modern VAD algorithms to use

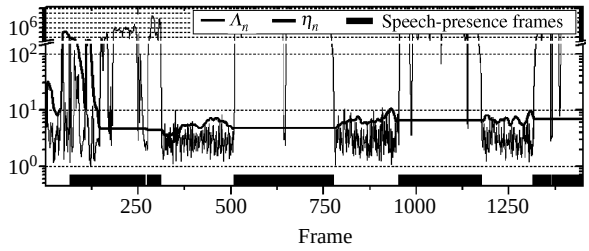

Fig. 8 Threshold adaptation with a speech-presence initial period

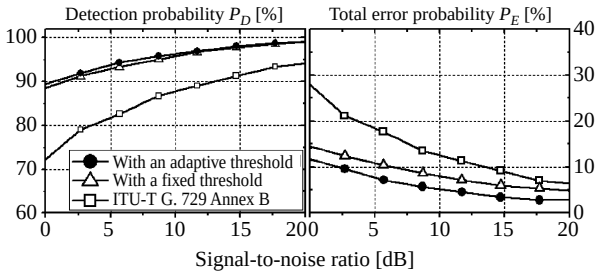

Fig. 10 Performance comparison in vehicular noise environment

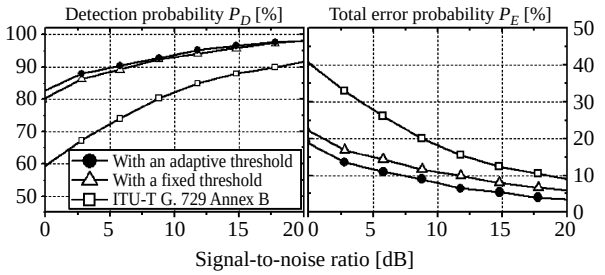

Fig. 9 Performance comparison in white noise environment

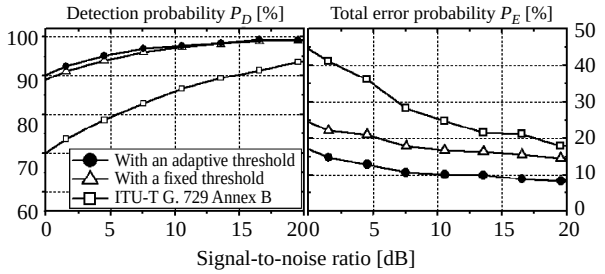

Fig. 11 Performance comparison in babble noise environment

a hangover period of a few frames to delay any premature transition from speech to noise. This is to minimize the clipping error probability for low-power unvoiced speech signals. In this simulation, we assigned 4 frames as a hangover period. We performed extensive computer simulations to demonstrate the performance improvement of the proposed algorithm under various noise environments and the results are summarized in Fig. 9, Fig. 10, and Fig. 11.

V. CONCLUSIONS

To obtain good performance of VAD in a time-varying SNR environment, we proposed an improved VAD algorithm employing preprocessing and an adaptive threshold. We first derived the preprocessing procedure, which is necessary to enhance the SNR and to make the LLR of a speech-presence frame much larger than that of a speech-absence frame, so that the adaptive threshold update algorithm can be applied. Then we proposed an adaptive threshold update algorithm using the mean and variance of the modified LLR which are stored within speech-absence frames.

After preprocessing, we first showed that the threshold is adapting well to a time-varying SNR environment in a short time period even with a speech-presence initial period. The simulation results summarized in Fig. 9, Fig. 10, and Fig. 11 confirm that the proposed statistical model-based VAD algorithm with an adaptive threshold has better performance than G.729B and statistical model-based algorithms with fixed thresholds. Especially, the proposed algorithm has lower clipping error probability than other algorithms, which is important for decoded speech quality. Finally, with an adaptive update of the parameters such as decision threshold and weighting factor, we hope that the proposed VAD algorithm can be applied in practical environments.

REFERENCES

Cho, Y. D., and Kondoz, A. M., 2001, "Analysis and Improvement of a Statistical Model-based Voice Activity Detector," *IEEE Signal Processing Letters*, Vol. 8, Issue 10, pp. 276-278.

Ephraim, Y., and Malah, D., 1984, "Speech Enhancement Using a Minimum Mean-square Error Short-time Spectral Amplitude Estimator," *IEEE Transactions Acoustics, Speech, Signal Processing*, Vol. ASSP-32, pp. 1109-1121.

ITU-T Recommendation, 1996, G.729 Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70.

Kondoz, A. M., 1999, *Digital Speech: Coding for Low Bit Rate Communication Systems*, John Wiley & Sons, New York, USA, pp. 337-341.

Lee, Y. C., and Ahn, S. S., 2001, "An Improved Voice Activity Detection Algorithm Employing Speech Enhancement Preprocessing," *The Institute of*

Electronics, Information and Communication Engineers, Transactions on Fundamentals, Vol. E84-A, No. 6, pp. 1401-1405.

Mcaulay, J., and Malpass, L., 1980, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-28, No. 2, pp. 137-145.

Sohn, J., and Sung, W., 1998, "A Voice Activity Detector Employing Soft Decision-based Noise Spectrum Adaptation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 365-368.

Sohn, J., Kim, N. S., and Sung, W., 1999, "A Statistical Model-based Voice Activity Detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1-3.

Srinivasan, K., and Gersho, A., 1993, "Voice Activity Detection for Cellular Networks," *IEEE Workshop on Speech Coding for Telecommunications*, St. Jovite, Quebec, Canada, pp. 85-86.

Telecom D. B., 1995, ftp://ftp.sipro.com, Telecom Inc.

Varga, A., and Steeneken, H. J. M., 1993, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, Vol. 12, No. 3, pp. 247-251.

Yang, J., 1993, "Frequency Domain Noise Suppression Approaches in Mobile Telephone System," *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, MN, USA, Vol. 2, pp. 363-366.

Manuscript Received: Aug. 31, 2005
Revision Received: Nov. 15, 2005
and Accepted: Dec. 08, 2005