Improved Statistical Model-Based Voice Activity Detection with Noise Reduction for the SIG-2 Humanoid Robot

*Ui-Hyun Kim, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno

Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

1. Introduction

Voice activity detection (VAD) is an essential technique in the robot audition system. VAD can facilitate robot speech processing in which the presence or absence of human speech is detected, and can also be used to deactivate some process during speech-absent period of an audio signal to reduce the computational cost and an unexpected error. Therefore robots can have a better performance with their VAD systems.

The purpose of the VAD is to provide delimiters for the beginning and end of a continuous speech-present period as exactly as possible from background noise such as music or other non-speech signals. For this purpose, it first extracts some features or quantities from the audio signal and compares these observed values with those of estimated noise according to some decision rules.

The most conventional VAD algorithms are based on zero-crossing rate, periodicity estimation, and signal energy level detection. The most well-known algorithm of this kind is the G.729B VAD [1]. However, these conventional VAD algorithms have the weakness that their performance is not good enough in a situation where the background noise level is high. In other words, they cannot work well in the low signal-to-noise ratio (SNR) case. To cope with this problem, many improved VAD algorithms have been designed and proposed but they also have their bad points of using heuristics which makes it difficult to optimize the relevant parameters.

Sohn et al. have proposed a voice activity detector (VAD) based on a statistical model. This statistical model-based VAD algorithm requires fewer parameters for optimization than the G.729B VAD and uses the log likelihood ratio (LLR) of speech and background noise variances of statistics for the low SNR case [2]. However, this statistical model-based VAD algorithm simply uses the power subtraction method to estimate the a priori SNR even if the speech spectrum is more changeable than the noise spectrum and it also assumes that the noise variance is already known through the noise statistic estimation procedure, thus these limit its VAD performance.

The performance problem on the existing statistical model-based VAD algorithm is the insufficient *a priori* SNR estimation by the power subtraction method. In this paper, we improve the *a priori* SNR by utilizing the two-step noise reduction (TSNR) technique [3] with recursive noise adaption instead of the power subtraction method. Then we present an improved statistical model-based VAD algorithm for the SIG-2 humanoid robot which has already included the VAD system [4].

The paper is organized as follows. Section 2 summarizes a statistical model-based VAD algorithm. Section 3 presents an improved statistical model-based VAD algorithm employing the two-step noise reduction technique with recursive noise adaptation. Section 4 evaluates experimental results with discussions. Finally, Section 5 concludes this paper.

2. A statistical model-based voice activity detection algorithm

Assuming that clean speech is degraded by uncorrelated additive noise, the observed signal can be represented with two hypotheses, speech absence H_0 and speech presence H_1, as follows:

$$H_0 : speech\ absent\ : X[f,n] = N[f,n] \quad (1)$$

$$H_1 : speech\ present : X[f,n] = S[f,n] + N[f,n]. \quad (2)$$

where $X[f,n]$, $S[f,n]$, and $N[f,n]$ are f-th elements of the short-time Fourier transform (STFT) of the noisy speech, clean speech, and uncorrelated additive noise, respectively, on the n-th time-frame index in the time-frequency domain. $f \in \{0, fs / T, \ldots, fs\ (T-1) / T\}$ is a frequency, fs is a sampling frequency, and T is a frame size for the STFT.

Adapting the Gaussian statistical model that means the STFT coefficients of clean speech and uncorrelated additive noise are asymptotically independent Gaussian random variables, the probability density functions (PDF) conditioned on two hypotheses H_0 and H_1 are given by

$$p(X[f,n]|H_0) = \prod_{f=0}^{T-1} \frac{1}{\pi \lambda_N[f,n]} \exp\left\{ -\frac{|X[f,n]|^2}{\lambda_N[f,n]} \right\}, \quad (3)$$

$$p(X[f,n]|H_1) = \prod_{f=0}^{T-1} \frac{1}{\pi(\lambda_S[f,n]+\lambda_N[f,n])} \cdot \exp\left\{-\frac{|X[f,n]|^2}{\lambda_S[f,n]+\lambda_N[f,n]}\right\}, \quad (4)$$

where $\lambda_S[f,n]$ and $\lambda_N[f,n]$ is the variances of $S[f,n]$ and $N[f,n]$, respectively. Based on the assumed statistical models, the likelihood ratio (LR) is

$$\Lambda[f,n] = \frac{p(X[f,n]|H_1)}{p(X[f,n]|H_0)} = \frac{1}{1+\xi[f,n]} \exp\left\{\frac{\gamma[f,n]\xi[f,n]}{1+\xi[f,n]}\right\}, \quad (5)$$

where $\xi[f,n]=\lambda_S[f,n]/\lambda_N[f,n]$ is the *a priori* SNR and $\gamma[f,n]=|X[f,n]|^2/\lambda_N[f,n]$ is the *a posteriori* SNR.

The *a priori* SNR $\xi[f,n]$ and the noise variance $\lambda_N[f,n]$ are unknown in advance as the noisy speech $X[f,n]$ alone is available. Therefore, $\xi[f,n]$ and $\lambda_N[f,n]$ need to be estimated by some procedure. This statistical model-based VAD algorithm assumes that $\lambda_N[f,n]$ is already known through the noise statistic estimation procedure and $\xi[f,n]$ can be derived by the power subtraction method as follows:

$$\hat{\xi}[f,n] = \frac{|X[f,n]|^2 - \lambda_N[f,n]}{\lambda_N[f,n]} = \gamma[f,n] - 1. \quad (6)$$

The VAD decision rule is derived from substituting Equation (6) into Equation (5) and the mean of the LLR for individual frequency bins:

$$\text{if } \hat{\Lambda}[n] > \eta \text{ then } n = \text{speech present}; \quad \text{else } n = \text{speech absent} \quad (7)$$

where

$$\hat{\Lambda}[n] = \frac{1}{T}\sum_{f=0}^{f_s(T-1)/T} \log\Lambda[f,n] = \frac{1}{T}\sum_{f=0}^{f_s(T-1)/T} \log\left\{\frac{1}{\gamma[f,n]}\exp(\gamma[f,n]-1)\right\} = \frac{1}{T}\sum_{f=0}^{f_s(T-1)/T} \{\gamma[f,n] - \log\gamma[f,n] - 1\}, \quad (8)$$

η is a threshold.

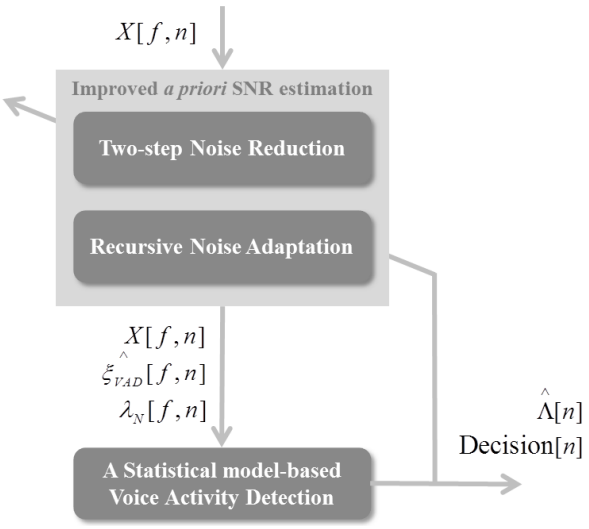

Fig.1 Block diagram of an improved statistical model-based VAD algorithm employing the TSNR technique and recursive noise adaptation.

3. Proposed voice activity detection algorithm

This section gives the improved statistical model based VAD algorithm employing the two-step noise reduction technique with recursive noise adaptation. The performance problem on the existing statistical model-based VAD algorithm is the insufficient *a priori* SNR estimation by the power subtraction method. The TSNR technique is utilized to improve the insufficient *a priori* SNR estimation instead of the power subtraction method.

Figure 1 shows the block diagram of the proposed VAD system. In the proposed VAD system, the *a priori* SNR is optimized after the TSNR technique with reconceive noise adaptation, and then it is used in the existing statistical model-based VAD process which is described in Section 2.

3.1 A priori SNR estimation with the two-step noise reduction technique

In the first step in the TSNR technique, the *a priori* SNR is computed with the decision-directed (DD) estimation approach to reduce the bias of an estimator [5] as follows:

$$\hat{\xi}_{DD}[f,n] = \alpha\frac{|\hat{S}[f,n]|^2}{\lambda_N[f,n]} + (1-\alpha)P\{\gamma[f,n]-1\}, \quad (9)$$

where $P[\cdot]$ is the half-wave rectification which is defined by $P[x]=x$ if $x\geq 0$ and $P[x]=0$ otherwise, and α is the forgetting factor whose value is typically chosen as 0.98

($0 < \alpha < 1$). Then, the spectral gain $G_{DD}[f,n]$ is obtained by applying Equation (9) to the Wiener amplitude estimator as follows:

$$G_{DD}[f,n] = \frac{\hat{\xi}_{DD}[f,n]}{1+\hat{\xi}_{DD}[f,n]}. \quad (10)$$

In the second step, $G_{DD}[f,n]$ is used for estimation of the TSNR *a priori* SNR as follows:

$$\hat{\xi}_{TSNR}[f,n] = \frac{|G_{DD}[f,n]X[f,n]|^2}{\lambda_N[f,n]}. \quad (11)$$

Finally, the spectral gain $G_{TSNR}[f,n]$ to enhance speech is obtained by applying Equation (11) to the Wiener amplitude estimator again:

$$G_{TSNR}[f,n] = \frac{\hat{\xi}_{TSNR}[f,n]}{1+\hat{\xi}_{TSNR}[f,n]}, \quad (12)$$

and the enhanced speech can be obtained by applying $G_{TSNR}[f,n]$ to the noisy signal as the following equation:

$$\hat{S}_{TSNR}[f,n] = G_{TSNR}[f,n]X[f,n]. \quad (13)$$

3.2 Noise adaptation

To compensate for fluctuations of noise power level, the noise variance $\lambda_N[f,n]$ is updated in a recursive way as follows:

$$\lambda_N[f,n] = \beta\lambda_N[f,n-1] + (1-\beta)\left\{|X[f,n]|^2 - |\hat{S}_{TSNR}[f,n]|^2\right\}, \quad (14)$$

where β is the forgetting factor ($0 < \beta < 1$). This noise adaptation function is performed on the speech-absent frames determined by the VAD decision rule.

3.3 Improved statistical model-based VAD algorithm

We use the speech spectrum enhanced by the TSNR technique in Equation (13) to improve the *a priori SNR* estimation for the statistical model-based VAD algorithm instead of the existing power subtraction method in Equation (6) as follows:

$$\hat{\xi}_{VAD}[f,n] = \frac{\hat{S}_{TSNR}[f,n]}{\lambda_N[f,n]}. \quad (15)$$

An improved VAD decision rule to be substituted for Equation (8) can be derived by substituting Equation (15) into Equation (5) and the mean of the LLR:

$$\hat{\Lambda}[n] = \sum_{f=0}^{f_s(T-1)/T} \log\left\{\frac{1}{1+\hat{\xi}_{VAD}[f,n]}\right\} + \frac{\gamma[f,n]\hat{\xi}_{VAD}[f,n]}{1+\hat{\xi}_{VAD}[f,n]}, \quad (16)$$

where the noise variance is recursively updated by Equation (14).

4. Experimental results

We constructed an improved statistical model-based VAD system employing the TSNR technique with recursive noise adaptation for audition system of the SIG-2 humanoid robot.

An objective test was conducted to evaluate the performance of the proposed statistical model-based VAD algorithm. The results are shown in Fig. 2. Speech signals of a male and a female who speak Japanese were recorded for 7 seconds (see Fig. 2(a).) with a 16kHz sampling frequency and then mixed with background music as additive noise (see Fig. 2(b).). The sentences of the speech signals, which were used in the conversation, are “kono uwagiwa Suzuki-san nodesuka?” (female) and “ie, sono uwagiwa Lee-san nodesu.” (male). The speech signal enhanced by Equation (13) in the TSNR technique is shown in Fig. 2(c). The *a posteriori* SNR and the *a priori* SNRs estimated by the power subtraction method in Equation (6) and the TSNR technique in Equation (15) are shown in green, blue, and red, respectively (see Fig. 2(d).). The VAD decisions by Equation (8) in the existing statistical model-based VAD algorithm and Equation (16) in the proposed VAD algorithm are also shown in blue and red (see Fig. 2(e).).

As the results of the test show, the proposed VAD algorithm improves *a priori* SNRs by 5.31 dB from those of the existing statistical model-based VAD algorithm on the speech-present period. These improved *a priori* SNRs make LLR values for the VAD decision more affluent and it reduces the detection errors of speech-present period as the final outcome. This means that the *a priori* SNR estimation is a key function in improving the performance of the existing statistical model-based VAD algorithm. The proposed statistical model-based VAD algorithm employing the TSNR technique with recursive noise adaptation could distinguish the speech-present and speech-absent frames with 12.28% higher accuracy comparing to the existing statistical model-based VAD process.

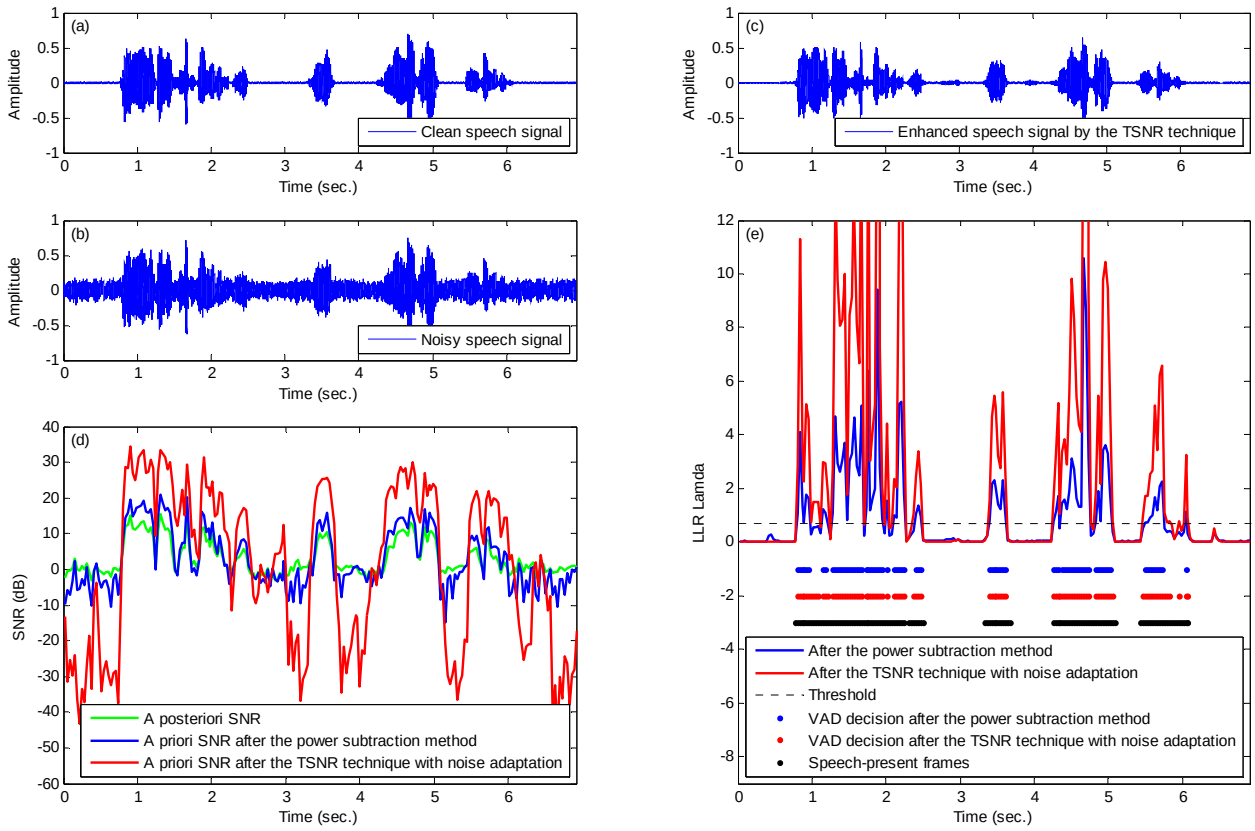

Fig.2 Examples of VAD using the power subtraction method and the TSNR technique with recursive noise adaptation. (a) Clean speech signal. (b) Noisy signal with music. (c) Enhanced speech signal by the TSNR technique. (d) *A posteriori* SNR and *A priori* SNRs. (e) LLR Λ and VAD decisions.

5. Conclusion

In this paper, we presented an improved statistical model-based VAD algorithm employing the TSNR technique with recursive noise adaptation for the SIG-2 humanoid robot. To obtain a better performance of the existing statistical model-based VAD algorithm, we improved the *a priori* SNR estimation by the TSNR technique with recursive noise adaptation instead of the power subtraction method. We could improve *a priori* SNRs with the TSNR technique by 5.31 dB on average during speech-present period. Experimental results demonstrated that the proposed statistical model-based VAD algorithm can indicate the presence and absence of speech with 12.28% higher accuracy than the existing statistical model-based VAD algorithm.

6. Acknowledgment

This research was partially supported by JSPS Grant-in-Aid for Scientific Research (S), JST Japan-France Cooperative Research Project BINAAHR, and the Global COE Program of Graduate School of Informatics, Kyoto University.

REFERENCES

[1] A. Benyassine, E. Shlomot, H.Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, “ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,” IEEE Communications Magazine, vol. 35, pp. 64–73, Sept. 1997.

[2] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 365–368, 1998.

[3] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Transactions on Audio, Speech & Language Processing, pp.2098-2108, 2006.

[4] U. H. Kim, T. Mizumoto, T. Ogata, and H. G. Okuno, "Improvement of Speaker Localization by Considering Multipath Interference of Sound Wave for Binaural Robot Audition," IEEE/RSJ International Conference on Intelligent Robots and Systems, September 25-30, 2011.

[5] Y. Ephraïm and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech, Signal Processing, vol. no. 6, pp. 1109–1121, Dec. 1984.