Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Method of Estimating Signal-to-Noise Ratio Based on Optimal Design for Sub-band Voice Activity Detection

Shota Morita

Department of Computer Science
Fukuyama University
985-1 Sanzo, Higashimura-cho, Fukuyama-shi, Hiroshima, 729-0292, Japan
morita@fuip.fukuyama-u.ac.jp

Xugang Lu

National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
xugang.lu@nict.go.jp

Masashi Unoki and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi-shi, Ishikawa 923-1292, Japan
unoki@jaist.ac.jp; akagi@jaist.ac.jp

Received February 2017; revised May 2017

Abstract. *The global signal to noise ratio (gSNR) is the ratio of concurrent powers between speech and noise in a noisy speech signal. Its estimates play an important role in power envelope restoration and predictions of speech intelligibility based on the speech transmission index (STI). Here, we propose a gSNR estimation framework that mainly consists of sub-band processing, voice activity detection (VAD), and threshold optimization. This process made the detection of speech and noise much more accurate than that with the global full-band process. In addition, an optimal threshold was designed to detect speech and noise under all testing conditions (e.g., different SNRs) rather than using a fixed decision threshold in VAD under all testing conditions, which has been done in most studies. This optimal threshold was obtained based on minimizing the root mean square (RMS) of the false acceptance rate (FAR) and false rejection rate (FRR) on the receiver operating characteristic (ROC) curves in each sub-band. Global SNR was calculated by summarizing the powers of speech and noise in all sub-bands with the help of the sub-band process and optimal design for VAD decision. Comprehensive evaluations were carried out using various types of noise and gSNR conditions. Classical VAD methods based on G.729B and thresholding using Otsu's method were used in comparative gSNR estimate. The results revealed that the proposed scheme could obtain higher accuracy in estimates of gSNR than the comparative methods.*

Keywords: Global SNR estimation, Optimal threshold, Voice activity detection, Sub-band processing

1. **Introduction.** Noise degrades speech quality, intelligibility, and the performance of applications in speech communication and many speech technologies. Techniques of noise reduction and speech enhancement are intended to mitigate this degradation. Estimates

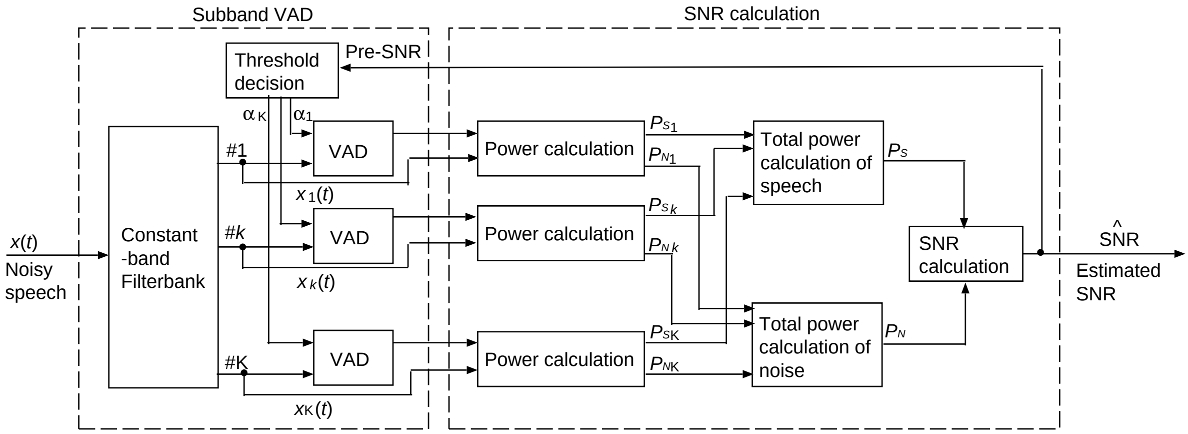


Figure 1. Block diagram of signal flow with proposed method.

of the signal to noise ratio (SNR) from noisy speech signals are necessary steps in these techniques. There are two parameters in the estimation of SNR.

The first is local SNR, which is the sub-band SNR. Local SNR estimation as *a priori* SNR and *a posteriori* SNR are used in Wiener filtering [1] and Minimum mean-square error short-time spectral amplitude (MMSE-STSA) [2] are used to obtain gain functions. The *a priori* SNR is estimated in each frequency. A decision-directed (DD) *a priori* SNR estimation has been used in MMSE-STSA [2]. Recently, *a priori* SNR estimation using voice activity detection (VAD) [3], data-driven *a priori* SNR estimation using neural networks [4], and *a priori* SNR estimation based on a multiple linear regression technique [5] have been proposed. These methods have been used to estimate local SNR in the short-term.

The second parameter is global SNR (gSNR), which is full-band SNR. This gSNR is needed in various speech signal processing, e.g., in speech enhancement [6], power envelope restoration based on the modulation transfer function (MTF) concept [7], the prediction of speech intelligibility, calculating the speech transmission index (STI) [8], and STI estimation [9]. The simplest gSNR estimation has been used in VAD [10]. Some other approaches have been proposed to correctly estimate SNR and noise, such as SNR estimation based on instantaneous amplitude [11], SNR estimation based on statistical distributions with sub-band processing [12], SNR estimation using modulation spectra and neural networks [13], SNR estimation based on Gamma distributions [14], and Computational auditory scene analysis (CASA) based on SNR estimation [15]. The gSNR estimation is more important than local SNR estimation as a base technology because gSNR can be more widely used. However, gSNR estimation does not work well with various types of noise.

This paper proposes gSNR estimation using VAD with a novel feature, which is a reasonable decision strategy based on finding the optimal threshold for VAD on receiver operating characteristic (ROC) curves. The optimal threshold is obtained by using training data with various types of noise. Sub-band processing, optimal design for thresholds of sub-band VAD, and iteration were used to improve the accuracy of gSNR estimation. These details are explained in the following sections.

This paper is organized as follows. Section 2 explains gSNR estimation using VAD and describes the problem setting for gSNR estimation. Section 3 introduces the process for the gSNR estimation we propose and describes how the sub-bands were designed, and how the decision threshold was determined. Section 4 describes comprehensive evaluations we carried out and Section 5 summarizes the conclusions we drew.

2. gSNR Estimation Using VAD.

2.1. gSNR definition.

This subsection defines gSNR using VAD. General gSNR is defined as:

$$\mathrm{gSNR} = 10\log_{10}\left(\frac{P_S}{P_N}\right), \tag{1}$$

where P_S and P_N are the powers of the speech and noise signals. They must be estimated from detected speech and non-speech periods since the powers of speech and noise are not known from observed noisy speech signals. Robust VAD is necessary in these estimates [10]. The gSNR is calculated with VAD as:

$$\mathrm{gS\hat{N}R} = 10\mathrm{log}_{10}\left(\frac{\int_0^T \left(P_{SN}\left(t\right) - \overline{P_N}\right) H_S\left(t\right) dt}{\int_0^T \overline{P_N} H_S\left(t\right) dt}\right), \tag{2}$$

where $P_{SN}\left(t\right)$ is the instantaneous power of both speech and noise at time t, $\overline{P_N}$ is the average of noise power, and T is the time duration of the signal. The $\overline{P_N}$ is generally calculated using the power of non-speech periods. The VAD decision satisfies:

$$H_S\left(t\right) = \begin{cases} 1 \ ; & \text{speech} \\ 0 \ ; & \text{non-speech.} \end{cases} \tag{3}$$

A decision strategy is used in designing VAD to obtain excellent performance by taking into consideration of the false acceptance rate (FAR) and false rejection rate (FRR) of speech. The criterion for the accuracy of detection with FAR and FRR is defined as:

$$\mathrm{FAR} = \frac{N_{FA}}{N_{ns}} \times 100 \ \ (\%), \tag{4}$$

$$\mathrm{FRR} = \frac{N_{FR}}{N_s} \times 100 \ \ (\%), \tag{5}$$

where N_s and N_{ns} correspond to the total numbers of speech and non-speech samples. The N_{FR} is the total number of false rejections, where samples are detected as non-speech, but are actually speech. The N_{FA} is the total number of false acceptances, where samples are detected as speech, but are actually non-speech.

2.2. Problematic issues.

The decision strategy was built based on setting a threshold to classify the signal as speech or non-speech classes, e.g., the likelihood ratio or power level. The simplest decision strategy was comparing the power level with a given threshold for VAD decision. The decision threshold in most VAD is fixed. This fixed threshold is usually determined by considering the FAR and FRR under all testing conditions, which is not reasonable for various noisy environments.

Receiver operating characteristics (ROC) curves are used, which indicate a trade-off between FAR and FRR. Fixing a decision threshold for VAD means setting the condition of performance as one point on an ROC curve for one required noisy condition. Robust VAD needs to adjust different performance conditions on ROC curves under different noisy conditions.

Accurate gSNR estimation needs accurate VAD with high levels of performance for FAR and FRR on ROC. However, the shape of the ROC curve should be convexed downward on the ROC. Both FAR and FRR need to be used in measurements for evaluation under these conditions. Therefore, measurements using both FAR and FRR on the ROC are important to accurately estimate the gSNR.

Therefore, we propose a reasonable decision strategy based on finding the optimal threshold for VAD on ROC curves in this paper. In addition, since noise has a different effect on speech in each sub-band, the optimal threshold is determined in each sub-band by using a training data set in the gSNR estimates. The gSNR is estimated by summarizing the powers of speech and noise from all sub-bands with the help of sub-band VAD.

3. Framework for Proposed Method.

The block diagram in Fig. 1 outlines the concept underlying the proposed method. Noise has a different effect on speech in different frequency bands. Therefore, our processing was designed on sub-bands. The proposed method consisted of two main function blocks (the two dashed rectangles in Fig. 1). The one on the left is VAD for sub-band signals. That on the right is the sub-band power estimation for speech and noise, and final gSNR calculations. Iterations between gSNR and VAD decisions are applied since accurate speech and noise power estimation needs VAD, while accurate VAD needs speech and noise power estimation. The final gSNR is obtained after several iterations. There is an optimal decision threshold in VAD designed based on the criterion of minimizing the root mean square (RMS) of FAR and FRR on the ROC curves of each sub-band. The details are given in the following subsections. It was assumed that noise would be stationary for sub-band processing in the proposed method. Noisy speech was used as additive background noise.

3.1. Filterbank design.

VAD was designed and applied to the sub-band signal shown in Fig. 1. The k is an index of the sub-bands, and K is the total number of sub-bands in this figure. A constant-bandwidth based filterbank (CBFB) was used for sub-band processing in this study. The CBFB filterbank consisted of band-pass filters of constant-bandwidth. The main purpose of using sub-band processing in this research was to separate the noise effect on speech in different frequency bands. One advantage of this sub-band processing is that it was possible to obtain several sub-bands with high SNRs even when the signal had a low SNR in the full band. The detection of speech in the high SNR sub-bands was much easier than that in the low SNR sub-bands. The bandwidth was set to 100 Hz [7] in the research discussed in this paper. There were 40 sub-bands in the CBFB, where the sampling frequency was 8 kHz. The speech and non-speech periods in each sub-band were detected by comparing their power levels with a given threshold in VAD processing.

3.2. Optimization on ROC curves for VAD.

Sub-band signals were obtained after being processed by the CBFB. The speech and non-speech periods were separately detected in each sub-band since sub-band signals had different local SNRs. These periods in each sub-band were detected by VAD processing with a given power level thresholding for each sub-band. At here, a lot of speech and non-speech periods were detected in each sub-band, because these periods were simply detected by power level threshold processing. Then, the final speech periods in each sub-band were decided as first and last sample points of detected speech periods. The other periods were detected as non-speech periods. A band-limited signal by low pass filtering of the cut-off frequency of 50 Hz in each sub-band was used in VAD processing. Different thresholds in VAD should be set on each sub-band for detection. These decision thresholds were designed based on minimizing the RMS of FAR and FRR on the ROC curves. The RMS of FAR and FRR can reasonably represent performance in detecting speech and non-speech periods including the start and end points. The SNR estimation needs speech and non-speech periods to be accurately estimated on both FAR and FRR.

Different pairs of FAR(α) and FRR(α) are obtained in the VAD design based on different decision thresholds α. An ROC curve is obtained with these FAR(α) and FRR(α) pairs. The decision threshold in the k-th sub-band is rewritten as α_k, and FAR(α) and

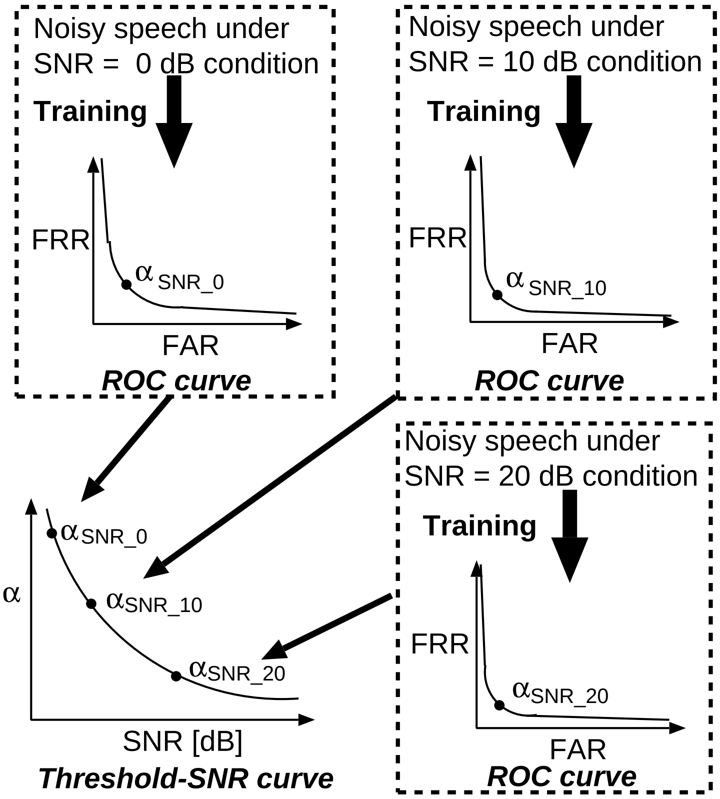


Figure 2. Prediction of decision threshold under any SNR conditions.

FRR(α) are rewritten as FAR(α_k) and FRR(α_k) for the sake of simplicity. There is an Apparent trade-off between FAR(α_k) and FRR(α_k) attained by changing the decision threshold in VAD. One point on the ROC curve means performance has been fixed that concerns FAR(α_k) and FRR(α_k) for VAD. We tried to find an optimal threshold on this ROC curve in our study, rather than fixing performance for FAR and FRR, by using a noisy data corpus for training under various gSNR conditions and noise types.

The FAR(α_k) and FRR(α_k) pairs are calculated by varying the decision threshold in VAD in the k-th sub-band under all SNR conditions to obtain the ROC curves. The objective function in finding an optimal decision threshold is defined by minimizing the RMS of FAR and FRR as:

$$\alpha_k^* = \arg\min_{\alpha_k} \mathrm{RMS}\left(\alpha_k\right), \tag{6}$$

where RMS is defined as:

$$\mathrm{RMS}\left(\alpha_k\right) = \sqrt{\frac{\mathrm{FAR}^2\left(\alpha_k\right) + \mathrm{FRR}^2\left(\alpha_k\right)}{2}}. \tag{7}$$

The definition of RMS in Eq. (7) considers the best trade-off between the FAR and FRR in VAD. Therefore, this is a reasonable criterion to detect speech and non-speech periods in VAD evaluations. Many optimal SNR dependent thresholds have been obtained in each sub-band based on the criterion defined in Eq. (6) by using a large training data set under various SNR conditions and various noise types. However, it has been difficult to cover all SNR conditions since training has only been carried out on data sets under limited SNR conditions. A curve fitting algorithm was applied to the threshold-SNR curves we obtained to acquire a decision threshold under any given SNR conditions. A decision threshold under any testing SNR conditions can be predicted by using this algorithm.

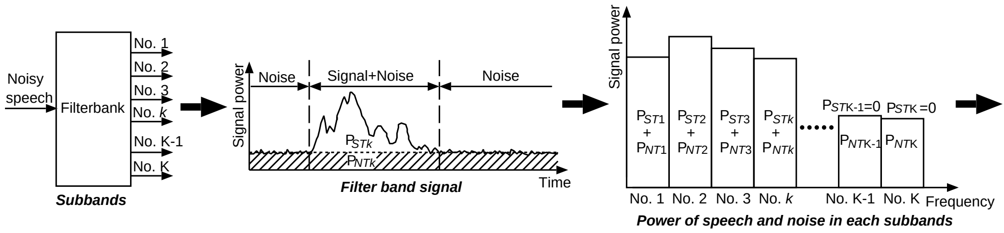


FIGURE 3. Signal flow for sub-band summation for global SNR estimations.

The concept behind this curve fitting algorithm is illustrated in the graphs in Fig. 2. We can implicitly obtain a decision threshold versus an SNR function from these graphs. Here, the optimal decision threshold was obtained under all SNR conditions, and then a fitting curve was used on the threshold-SNR curves. A 4th-order Sigmoid function was used in the curve fitting algorithm in our study. The MMSE of optimal thresholds and true SNR were used as fitting criteria in the fitting function. The main reason for using the Sigmoid function is explained with the results obtained from training on threshold-SNR curves in Subsection 4.1.

3.3. **Power estimation from sub-bands.** The power of a signal in the time domain is calculated based on the Parseval's theorem from the summation of power from all sub-bands. The processing flow for gSNR estimation is outlined in Fig. 3. Since VAD has been separately designed in each sub-band with different decision thresholds, the estimates of noise and speech powers are much more accurate than direct estimates in the time domain. The final gSNR is obtained from the power fusion of all sub-bands as:

$$\hat{\mathrm{SNR}} = 10\log_{10}\left(\frac{\sum_{k=1}^{\mathrm{K}} P_{STk}}{\sum_{k=1}^{\mathrm{K}} P_{NTk}}\right), \tag{8}$$

$$P_{NTk} = \int_0^T \overline{P_{Nk}} H_{Sk}(t)\,dt, \tag{9}$$

$$P_{STk} = \int_0^T P_{SNk}(t)\,H_{Sk}(t)\,dt \quad - \int_0^T \overline{P_{Nk}} H_{Sk}(t)\,dt, \tag{10}$$

where P_{NTk} and P_{STk} are the total power of noise and speech in the k-th sub-band, $P_{SNk}(t)$ is the instantaneous power of both speech and noise at time t in the k-th sub-band, and $\overline{P_{Nk}}$ is the average of noise power in the non-speech periods in the k-th sub-band. The $H_{Sk}(t)$ is the the VAD decision in the k-th sub-band, and is the same as that in Eq. (3).

We can see if one of the signal periods from the k-th sub-band is classified as non-speech by following the flow in Fig. 3. Noise power in this sub-band is calculated as the sum of the signal in the current period (calculated from Eq. (9)). If one of the signal period from the k-th sub-band is classified as speech, this periods is a mixture of speech and noise signals, and then speech power is estimated from the power of noisy speech (calculated from Eq. (10) by subtracting noise power). The noise power in this period is predicted from the average noise power in the detected non-speech periods. The final gSNR is obtained based on this calculation as Eq. (8).

3.4. **Refining estimation with iterations.** There is a feedback loop in Fig. 4 from gSNR estimation to VAD decision through the threshold-SNR curves, i.e., the gSNR estimation needs VAD while VAD requires SNR calculations. The estimated SNR (pre-gSNR) is fed to the threshold decision stage for VAD in this loop. The decision thresholds for VAD are then reset for the estimates of gSNR in the next iteration. This loop is iterated

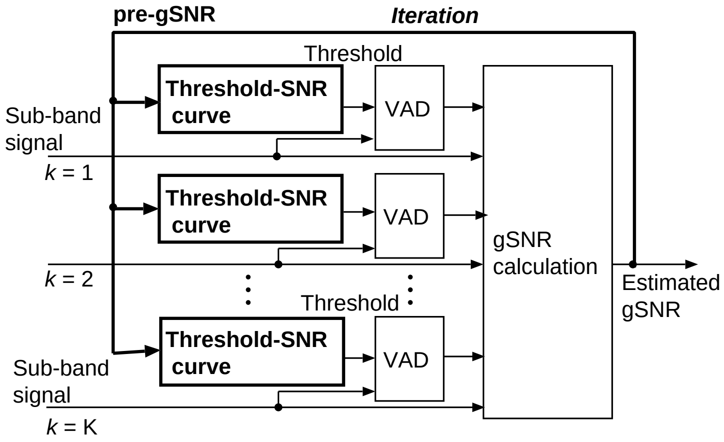


Figure 4. Refining estimation with iterations.

several times to further refine the estimates of gSNR. The estimated gSNR is converged based on the threshold-SNR curves to the convergent point of gSNR for the loop.

4. Evaluations.

We carried out experimental evaluations on the proposed framework, which are described in this section. The training stage that was used for optimal threshold prediction in VAD was first implemented. The effect of iterations on the SNR-VAD loop was then examined. Finally, comprehensive evaluations and comparisons of gSNR estimation were carried out.

4.1. Finding optimal decision thresholds for VAD.

We first trained the functions that were used to find optimal thresholds for VAD, as outlined in Fig. 2. We selected 8440 utterances from the AURORA-2J data set [16] in training as clean speech. White noise, pink noise, and babble noise in NOISEX-92 [17] were used as background noise. Noisy speech signals were artificially created as $y(t) = x(t) + n(t)$, where $x(t)$ is the clean speech signal, and $n(t)$ is the background noise signal. Noisy speech signals with SNRs of $50, 40, 30, 20, 10, 0, -10$, and -20 dB were generated. These clean and noisy speech signals were then used to find the optimal thresholds in each sub-band in VAD design. The sampling frequency was 8 kHz, the bandwidth of sub-bands was 100 Hz, and the number of sub-bands was 40. The optimal threshold in each sub-band under each SNR condition was obtained using all noisy speech under different SNR conditions.

A fitting function with the Sigmoid function was estimated under all SNR conditions (for SNRs from 50 to 0 dB) after the optimal threshold versus SNR pairs in each sub-band had been obtained. The MMSE was used in the fitting function. There is an example of a fitting curve that has been plotted in Fig. 5 with the $k = 10, 20, 30$, and the 40-th sub-bands. The closed circles in Fig. 5 plot the optimal thresholds under all SNR conditions obtained from the results of training. The blue fitting curves are almost reasonably fitted by the Sigmoid function.

4.2. Effect of iterations on SNR-VAD loop.

We first evaluated the effect of iterations on the SNR-VAD loop, as shown in Fig. 1. We selected 1001 utterances for testing from the AURORA-2J data set [16] as clean speech. Five types of noise that were white, pink, babble, factory (factory1), and car (Volvo) noise in NOISEX-92 [17] were used as background noise. These noise signals were added to the speech signals for noisy speech in testing with SNRs of 20, 15, 10, 5, 0, -5, and -10 dB. Five iterations were carried out

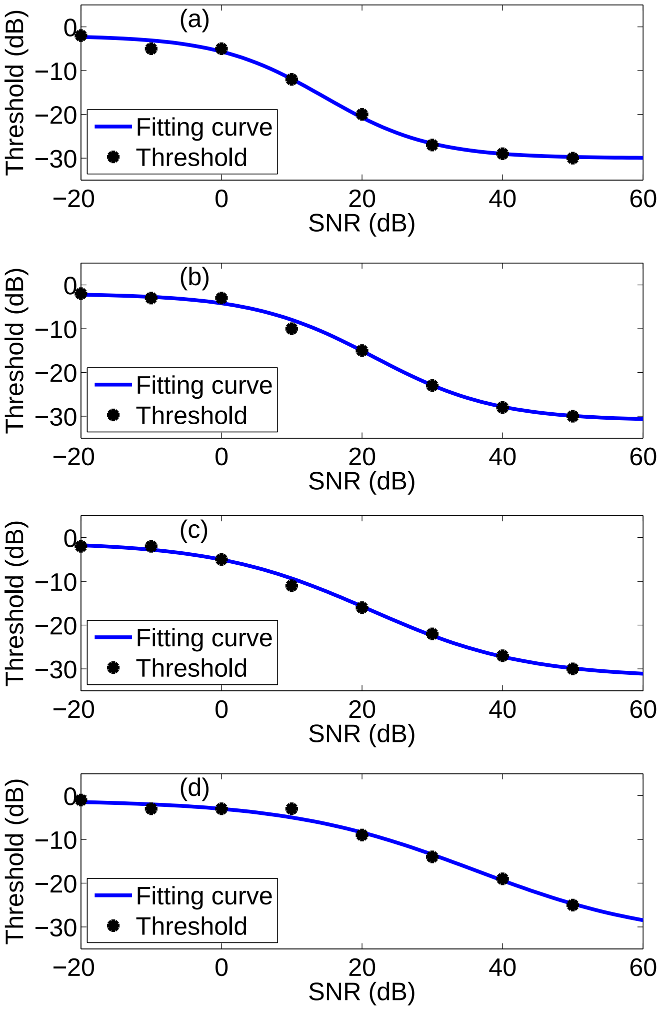
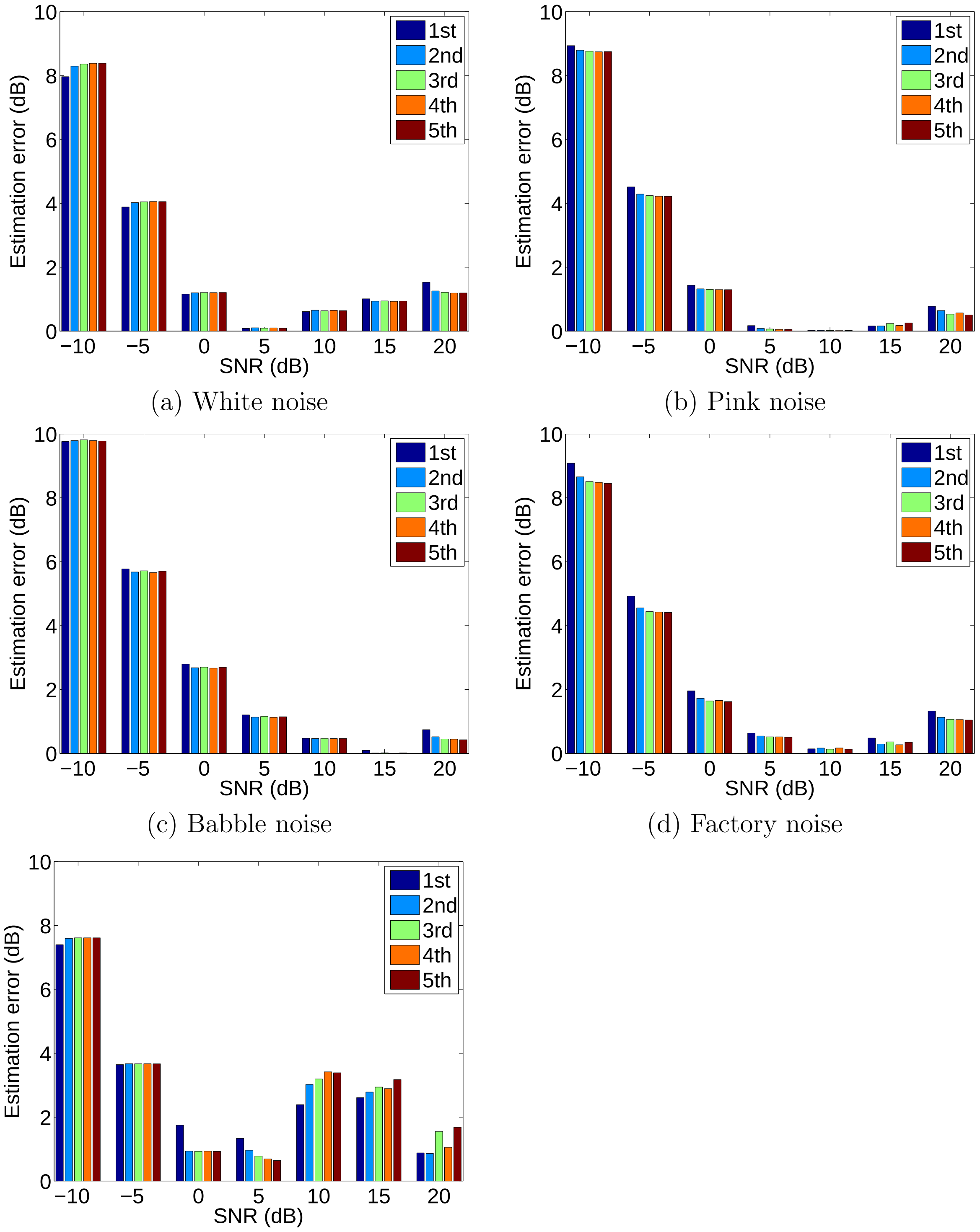


Figure 5. Example of fitting curve for decision threshold vs SNR: (a) $k = 10$th sub-band, (b) $k = 20$th sub-band, (c) $k = 30$th sub-band, and (d) $k = 40$th sub-band. Solid line: fitting curve and points: optimal threshold from training.

in the SNR-VAD loop to estimate gSNR. The RMS error between the estimated gSNR and true gSNR was used in the evaluation.

The results are given in Fig. 6, and we can see that error in gSNR estimation consistently decreased with increasing numbers of iterations, particularly under some SNR conditions, e.g., white noise (SNR = 15 and 20 dB), pink noise (except for SNR = 15 dB), babble noise (except for SNR = -10 and 10 dB), factory noise (except for SNR = 10 dB), and car noise (except for SNR = $10, 15$, and 20 dB). However, under a few conditions, error in gSNR estimation increased due to the effect of the iteration process, e.g., white noise under low SNR conditions. We used three iterations in the evaluations that followed. The number of iterations was empirically determined while taking into account the computational cost and error in SNR.

4.3. **Comprehensive evaluation for SNR estimations.** We evaluated the performance of the proposed method against two other VAD methods in estimates of SNR for comparison (a total of three). The first was VAD used in G.729B [18], which is a speech compression algorithm that has been standardized by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T). It is widely known in



(a) White noise

(b) Pink noise

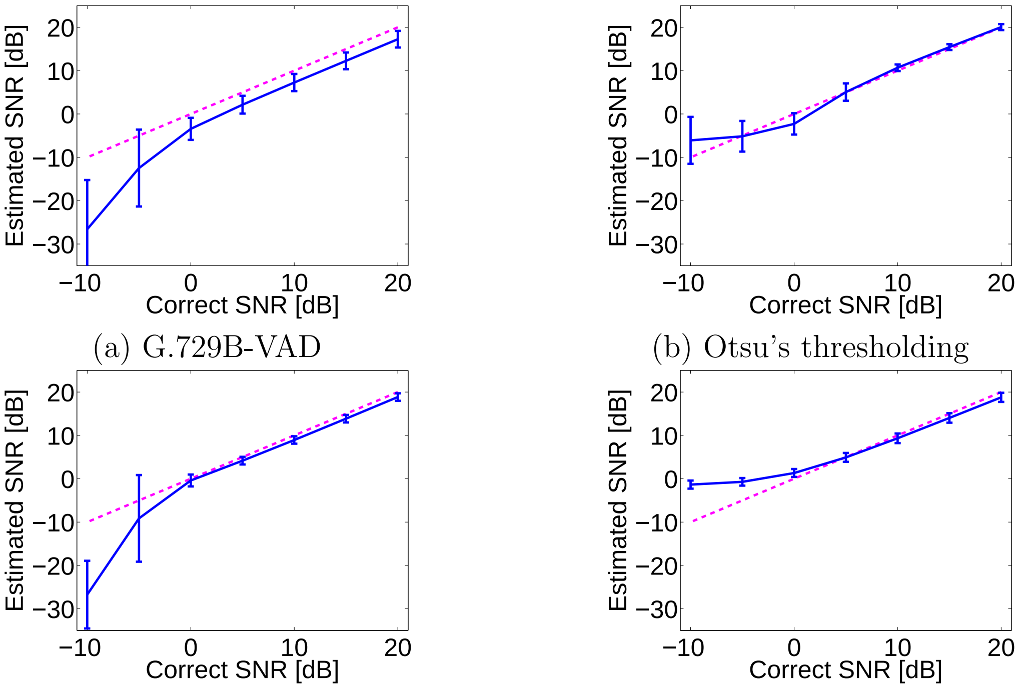
(c) Babble noise

(d) Factory noise

(e) Car noise

Figure 6. Estimation errors and repetitive estimates.

mobile communications. The second approach was VAD using Otsu's method [19]. It adopted Otsu's method [20] to determine a flexible threshold in VAD decision. The third approach was VAD used in adaptive multi-rate option 2 VAD (AMR opt.2-VAD) [21]. It is a robust VAD for noisy environments where its performance is similar to that with improved G.729B for noisy robustness [22]. These three VAD methods were used in the estimates of gSNR. The evaluation conditions were the same as those used in Subsection



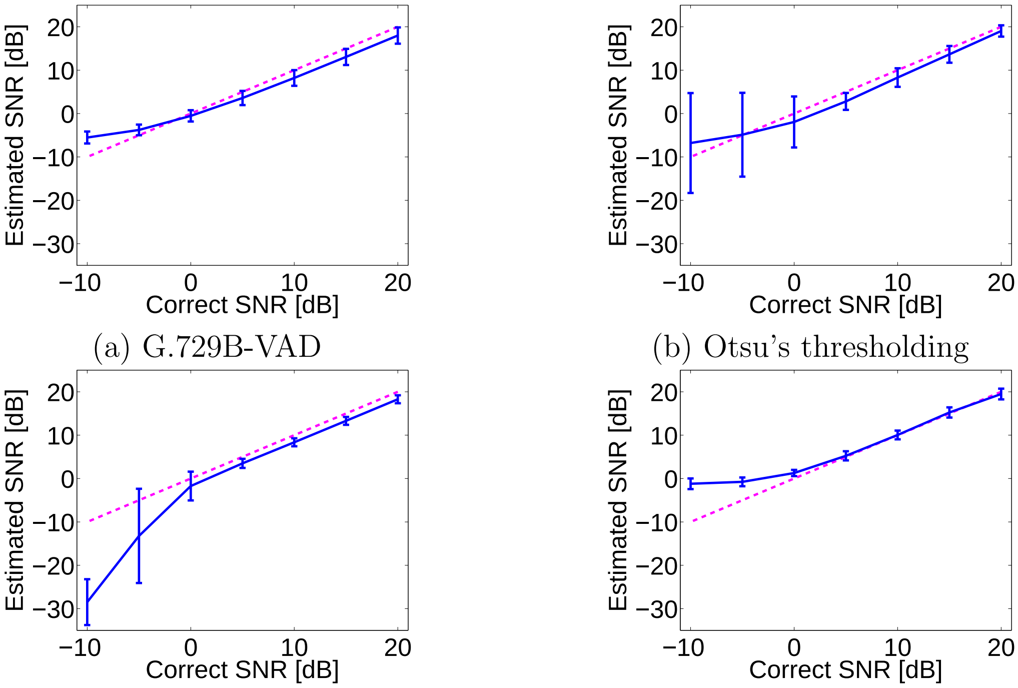
(a) G.729B-VAD

(b) Otsu's thresholding

(c) AMR opt.2-VAD

(d) Proposed method

Figure 7. Results for estimates of SNR under white noise conditions.



(a) G.729B-VAD

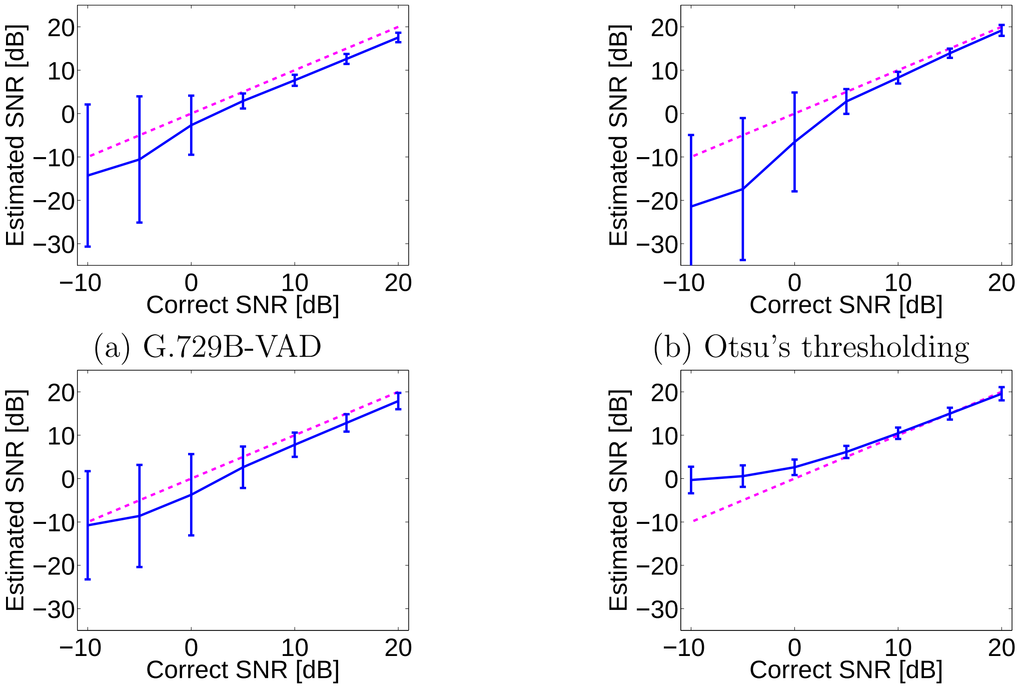
(b) Otsu's thresholding

(c) AMR opt.2-VAD

(d) Proposed method

Figure 8. Results for estimates of SNR under pink noise conditions.

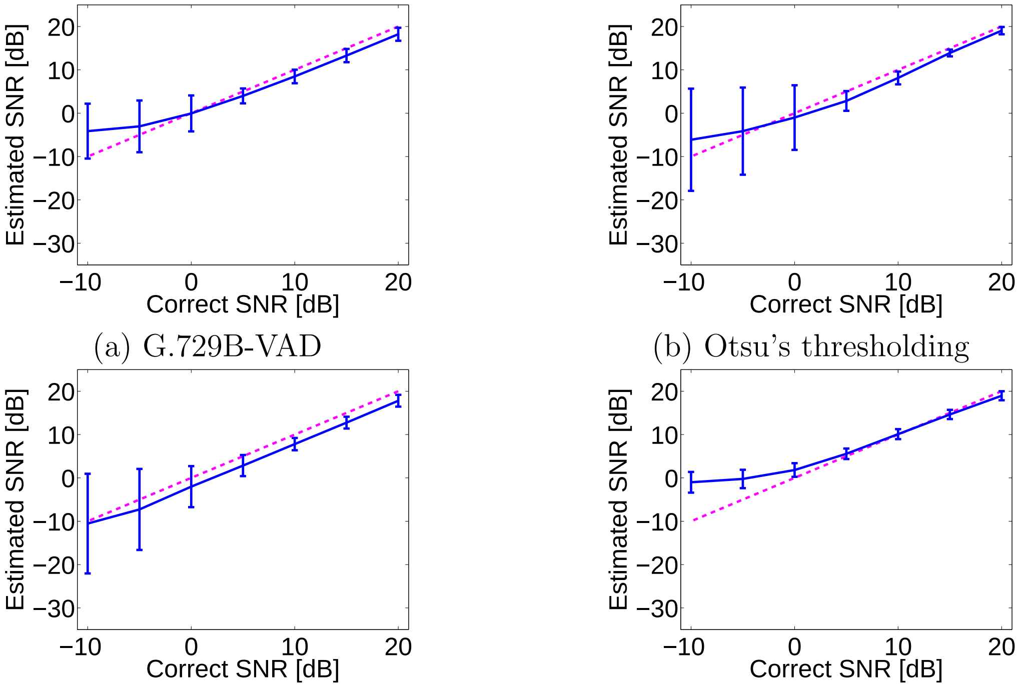
4.2. The estimated gSNR was mandatorily changed to −40 dB under the estimated gSNR up to −40 dB conditions including $-\infty$ dB to statistically evaluate the average.



(a) G.729B-VAD (b) Otsu's thresholding

(c) AMR opt.2-VAD (d) Proposed method

Figure 9. Results for estimates of SNR under babble noise conditions.

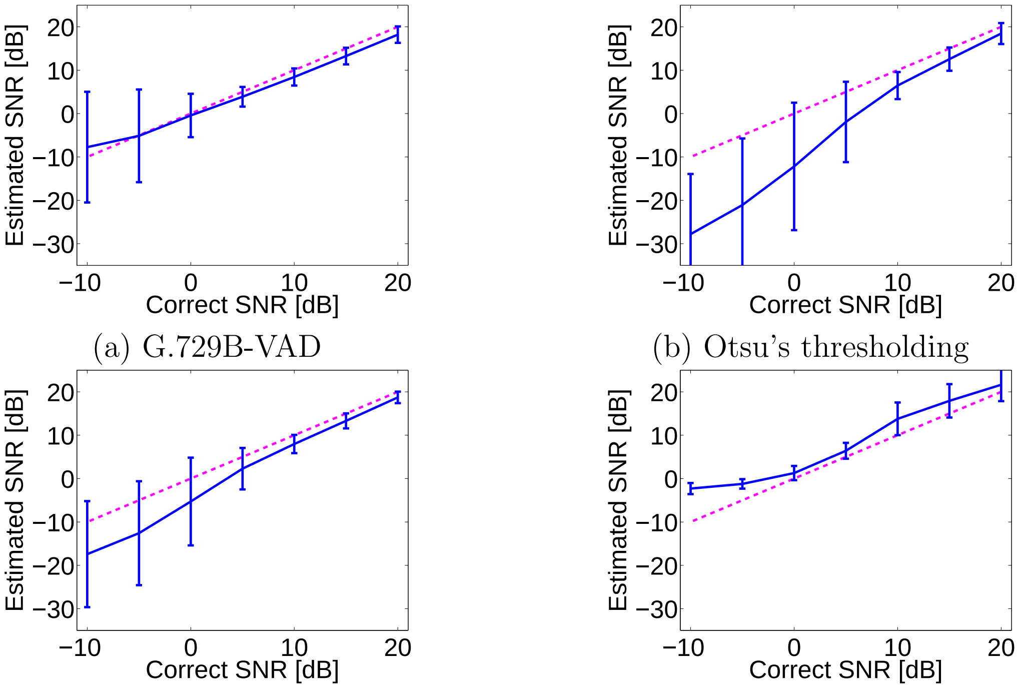


(a) G.729B-VAD (b) Otsu's thresholding

(c) AMR opt.2-VAD (d) Proposed method

Figure 10. Results for estimates of SNR under factory noise conditions.

The results obtrained from the estimates of gSNR are plotted in Figs. 7– 11. The error bars in the figures indicate the standard deviations. The true SNR was calculated according to the definition in Eq. (2). We can see from these results that the estimates deviate from the true values for all methods.



(a) G.729B-VAD

(b) Otsu's thresholding

(c) AMR opt.2-VAD

(d) Proposed method

Figure 11. Results for estimates of SNR under car noise conditions.

4.4. **Discussion.** The optimal decision threshold for VAD in each sub-band was predicted from a model trained using mixed noisy speech signals with white, pink, and babble noises in our experiments. Performance could have been improved if the model had been trained using only one noise type for each. In addition, we found in our study that the sigmoid function was a reasonable choice to predict the decision threshold when SNR was given, as it was in Fig. 5.

The gSNR estimation based on VAD in G.729B had error in SNR of 3 dB under most conditions with all types of noise in the comprehensive evaluation. The gSNR estimation based on VAD using Otsu's method had large errors in SNR except under white noise conditions. The gSNR estimation based on VAD in AMR opt.2 had large errors under most noise conditions. The estimation error with the proposed method was smaller than that for the other methods under conditions from SNR 20 to 0 dB. In addition, variations in estimates (error bars in the figure) with the proposed method were smaller than those for the other methods under almost all conditions. This suggests that our proposed method was much more robust than the other methods from SNR 20 to 0 dB. Variations in comprehensive gSNR estimation were drastically larger than those with the proposed method under less than SNR 0 dB conditions. These results mean that comprehensive gSNR estimation did not work stably and had many errors under the conditions. Although the proposed method had estimation errors, it worked stably under almost all conditions. Comprehensive gSNR estimation using robust VAD (VAD using Otsu's method and VAD in AMR opt. 2) failed during estimates of speech and noisy periods under the conditions, and added error effects to the results of gSNR estimation. However, as our proposed method used sub-band processing and a summation of the sub-band power of noise and speech, the estimation error was absorbed by the process. Therefore, time-frequency information was as important to gSNR estimation as it was with the other speech signal processes.

The standard deviation with our proposed method was smaller than that with the compared methods. Our proposed method could estimate gSNR under all the evaluation conditions. However, the compared methods failed to detect speech and non-speech periods under the lower SNR conditions. The gSNR in that case was estimated as $-\infty$ dB. This affected the averaging procedure and variance. The estimated gSNR discussed in this paper was mandatorily changed to -40 dB for statistical evaluation when the estimated gSNR was less than -40 dB including $-\infty$. This was the main reason that the standard deviation with our proposed method was smaller than that with the compared methods.

The estimated SNR with our proposed method came close to 0 dB under -5 and -10 dB. The noise power level should be greater than the speech power under less than 0 dB conditions to obtain accurate gSNR estimates. However, almost all the noise power in each sub-band was estimated as speech power in each sub-band due to problems with VAD performance in our proposed method. Therefore, the estimated speech periods included most noisy periods and thus the estimated gSNR came close to SNR = 0 dB. Accurate VAD in each sub-band is required in future work to improve the proposed method for SNRs of -5 and -10 dB.

The Results for car noise under SNRs of 10 to 20 dB in Fig. 11 were not satisfactory. Here, the estimated SNRs were overestimated under the conditions. The power spectrum of car noise was drastically different with noise used during training. The power of car noise was concentrated in the low frequency band. Therefore, the power of car noise was included as the speech power on speech/non-speech decisions in low frequency bands.

5. **Conclusions.** We proposed a method of making global SNR estimates. The power of speech and noise were estimated from a sub-band process. This sub-band process separated the noise effect on speech in each sub-band, which made the estimates much more accurate than those in wide-band processing. We designed power-level-based VAD to detect speech and noise periods in each sub-band. Our decision threshold in each sub-band was optimized using a training data set that was composed of noisy speech under various SNR conditions and noise types (three in this paper), which was different from fixing the decision threshold in VAD that has been used in most studies. The optimal thresholds were defined as those that minimized the RMS of FAR and FRR on the ROC curves under each SNR condition in the optimization. The power of speech and noise were accurately estimated in each sub-band based on this strategy. The final gSNR was estimated from the summation of the power of speech and noise in all sub-bands.

We found that the estimates of SNR had large variations under low SNR conditions in our study. The VAD may not work well for discriminating speech and noise that result in large errors in gSNR estimation. Special processing under low SNR conditions (SNR of 0 to -10 dB) should be further considered to improve performance in the future.

Acknowledgments. This work was supported by an A3 foresight program made available by the Japan Society for the Promotion of Science and by the Strategic Information and Communications R & D Promotion Programme (SCOPE: 131205001) of the Ministry of Internal Affairs and Communications (MIC), Japan. This study was also supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and by Secom Science and Technology Foundation.

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, All-pole modeling of degraded speech, *IEEE Trans. ASSP*, vol.26, no.3, pp.197–210, 1978.

[2] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, *IEEE Trans. ASSP*, vol.32, no.6, pp.1109–1121, 1984.

[3] K. Nakayama, S. Higashi, and A. Hirano, A noise estimation method based on improved VAD used in noise spectral suppression under highly non-stationary noise environments, *Proc. EUSIPCO2009*, pp.2494–2498, 2009.

[4] S. Suhadi, C. Last, and T. Fingscheidt, A data-driven approach to *a priori* SNR estimation, *IEEE Trans. Audio, Speech, and Language Processing*, vol.19, no.1, pp.186–195, 2001.

[5] S. Lee, C. Lim, and J. H. Chang, A new *a priori* SNR estimator based on multiple linear regression technique for speech enhancement, *Digigal Signal Processing*, vol.30, pp.154–164, 2014.

[6] X. Shen and L. Deng, A dynamic system approach to speech enhancement using the H_∞ filtering algorithm, *IEEE Trans. Speech and Audio Processing*, vol.7, no.4, pp.391–399, 1999.

[7] S. Morita, X. Lu, M. Unoki, M, Akagi and R, Hoffmann, MTF-based sub-band power-envelope restoration for robust speech recognition in noisy reverberant environments, *Proc. APSIPA2011*, 2011.

[8] IEC 60268-16, Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index, 2003.

[9] A. Miyazaki, S. Morita, and M. Unoki, Study on blind method of estimating speech transmission index from noisy reverberant amplitude modulated-signals, *Journal of Signal Processing*, vol.18, no.4, pp.201–204, 2014.

[10] M. Vondrasek and P. Pollak, Methods for speech SNR estimation: evaluation tool and analysis of VAD dependency, *Radioengineering*, vol.14, no.1, pp.6–11,2005.

[11] R. Martin, An efficient algorithm to estimate the instantaneous SNR of speech signal, *Proc. EUROSPEECH1993*, pp.1093–1096, 1993.

[12] E. Nemer, R. Goubran, and S. Mahmoud, SNR estimation of speech signals using subbands and forth-order statistics, *IEEE signal processing letters*, vol.6, no.7, pp.171–174, 1999.

[13] M. Kleinschmidt and V. Hohmann, Sub-band SNR estimation using auditory feature processing, *Speech Communication*, vol.39, pp.47–63, 2003.

[14] C. Kim and R. M. Stern, Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis, *in Proc. Interspeech2008*, pp.2598–2601, 2008.

[15] A. Narayanan and D. Wang, A CASA-based system for long-term SNR estimation, *IEEE Trans. Audio, Speech, and Language Processing*, vol.20, no.9, pp.2514–2527, 2012.

[16] http://www.slp.cs.tut.ac.jp/CENSREC/en/CENSREC/AURORA-2J/, 2012.

[17] A. Varga and H. J. M. Steeneken, "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.13, pp.247–251, 1993.

[18] A. Benyassine, E. Shlomot, S. Huan-yu, 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 application, *IEEE Commun. Mag.*, vol.35, pp.64–73, 1997.

[19] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda and S. Nakamura, CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments, *Acoust. Sci. & Tech.*, vol.30, no.5, pp.363–371, 2009.

[20] N. Otsu, "A threshold selection method from graylevel histograms, *IEEE Trans. Syst. Man. Cyber.*, SMC-9, pp.61–66, 1979.

[21] ETSI EN 301 v7.1, Digital cellular telecommunications system, Voice Activity Detection (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels, 1999.

[22] H. Farsi, M. A. Mozaffarian, and H. Rahmani, Improving Voice Activity Detection Used in ITU-T G.729.B, *Proc. CISST'09*, pp.11–15, 2009 1999.