*International Journal of Innovative*
*Computing, Information and Control*
Volume **9**, Number **6**, June **2013**

ICIC International ©2013 ISSN 1349-4198
pp. **2533–2542**

# DUAL-MICROPHONE VOICE ACTIVITY DETECTION INCORPORATING GAUSSIAN MIXTURE MODELS WITH AN ERROR CORRECTION SCHEME IN NON-STATIONARY NOISE ENVIRONMENTS

Ji Hun Park and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology
1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea
{ jh_park; hongkook }@gist.ac.kr

Received March 2012; revised July 2012

Abstract. *In this paper, a voice activity detection (VAD) method is proposed based on Gaussian mixture models (GMMs) by exploiting the spatial selectivity in dual-microphone environments. In other words, each GMM is constructed according to the direction-of-arrival (DOA) to detect speech intervals. Based on the assumption that the target speech is located in front of dual-microphones, the VAD is performed by comparing the likelihood obtained from the GMM constructed for the front of the microphones with those obtained from GMMs for other DOAs. In addition, to mitigate false rejection errors of VAD arising from the low spatial correlation in unvoiced intervals of target speech, VAD results are refined by employing a VAD error correction scheme. The error correction scheme analyzes the ratio between the energy of high and low frequency bands (HILO) to discriminate between an unvoiced interval of speech and a non-speech interval. The performance of the proposed GMM-based VAD method with the HILO-based error correction scheme is evaluated by measuring the false alarm rate (FAR) and false rejection rate (FRR) and comparing them with those of conventional dual-microphone VAD methods, where the FAR and FRR are measured by comparing the VAD results of each VAD method with those of manual segmentation. It is shown from the evaluation that the proposed GMM-based VAD method with the HILO-based VAD error correction outperforms a Gaussian kernel density-based VAD method and a GMM-based VAD method without VAD correction.*

**Keywords:** Voice activity detection (VAD), End-point detection, Gaussian mixture model (GMM), Spatial selectivity, VAD error correction

1. **Introduction.** Voice activity detection (VAD) is a technique for detecting the presence or absence of desired speech. VAD is used in various speech-based applications such as speech recognition [1] and speech coding since it deactivates some processes during non-speech intervals to reduce the number of computations and amount of network bandwidth usage [2]. Many VAD methods have been proposed to discriminate speech intervals from non-speech intervals. Among the potential methods, techniques based on energy levels and zero-crossing rates are the most common. They can detect speech intervals effectively with low complexity, but their performance is degraded due to the reduced discrimination capability of features such as energy levels and zero-crossing rates under low signal-to-noise ratio (SNR) conditions [3]. To overcome this problem, noise-robust VAD features such as a periodicity measure [4], cepstral features [5], and a long-term spectral divergence [6] have been proposed. In particular, Davis *et al.* incorporated Welch's method [7] into a VAD method to obtain a low-variance spectrum estimate [8], where both the estimate of the power spectral density of noise and the variance of SNRs were estimated

from non-speech intervals and used as VAD features. This method provided a stable VAD performance under different SNR conditions when noise is stationary. However, the VAD method compared the VAD features with thresholds that were updated during non-speech intervals. Thus, its performance is apt to be degraded in non-stationary noise environments because it was very hard to estimate the reliable thresholds according to the unexpected fluctuation of non-stationary noise.

Recently, there have been a number of research works reported to improve the VAD performance in non-stationary noise environments by exploiting the spatial selectivity of multiple microphones [9, 10, 11]. Most of them employed a microphone array to extract spatial features in order to be beneficial in non-stationary noise environments. For example, Kim and Cho proposed a multi-microphone VAD method using a phase vector as a VAD feature [11]. However, the performance of such a microphone array-based VAD approach relied strongly upon the number of microphones. This was because the microphone array techniques required a considerable number of microphones to get improved VAD performance, which caused high implementation costs [12]. Thus, a statistical model-based VAD method using two microphones was proposed which classified speech as either active or inactive intervals by comparing the log likelihoods obtained from Gaussian kernel density-based statistical models [13]. This approach reduced implementation costs by using a smaller number of microphones. However, the performance of the VAD method proposed in [13] deteriorated owing to low spatial correlations during unvoiced intervals of the target speech. In other words, the VAD method extracted spatial features on the basis of a cross-correlation technique. Generally, the relatively reliable cross-correlation distribution could be obtained between periodic signals rather than non-periodic signals. Thus, the unvoiced intervals of speech signals had low spatial correlation owing to their non-periodic characteristics. Therefore, a speech frame might be classified as a non-speech frame when an unvoiced interval of the target speech overlapped with an interval of noise signals that were strongly spatially correlated.

In this paper, a Gaussian mixture model (GMM)-based dual-microphone VAD method is proposed to improve the VAD performance in non-stationary environments. Compared with the VAD methods proposed in [8, 11], spatial cues and a logarithmic root mean squared energy (log-RMSE) are extracted and used as features for the proposed VAD method. To this end, speech intervals are detected via the GMMs, where each GMM represents the probabilistic distributions of the spatial cues and the log-RMSE according to the direction-of-arrival (DOA). In addition, to mitigate the VAD errors caused by low spatial correlation during the unvoiced intervals, a VAD correction scheme is incorporated by using the ratio between the energies of the high and low frequency bands (HILO).

Following this Introduction, Section 2 describes how the proposed VAD method with an error correction scheme can be constructed by incorporating GMM. Section 3 describes the detailed procedure of the HILO-based error correction scheme, and Section 4 evaluates the performance of the proposed VAD method in terms of false alarm errors and false detection errors. Finally, we conclude our findings in Section 5.

## 2. Proposed GMM-Based VAD.

Figure 1 shows a block diagram of the proposed GMM-based VAD with an HILO-based error correction scheme. As shown in the figure, the proposed method extracts feature parameters, such as spatial cues and a log-RMSE, from the binaural auditory signals. Next, the VAD is performed by comparing the log likelihoods estimated by applying the feature parameters to the GMMs, in which one GMM is trained for each DOA. The VAD results are then refined using the HILO-based correction scheme. The following subsections describe the processing steps of the proposed GMM-based VAD with the HILO-based correction scheme.

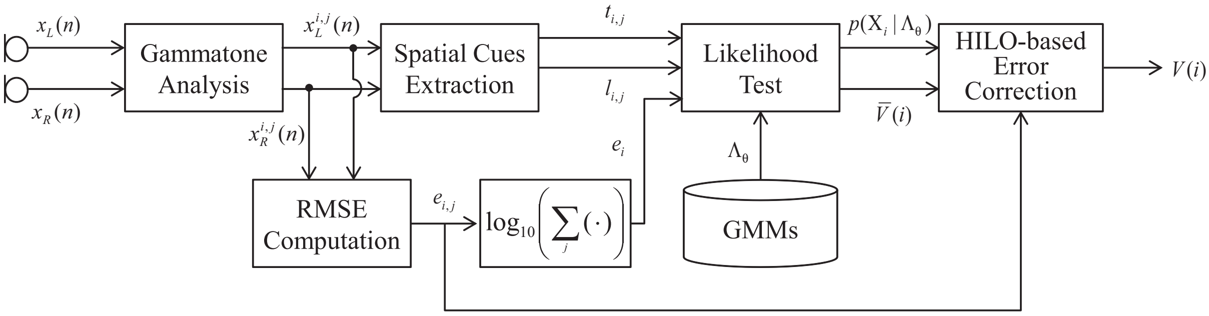


FIGURE 1. Block diagram of the proposed GMM-based dual-microphone VAD method with an HILO-based error correction scheme

2.1. **Gammatone analysis.** As shown in Figure 1, $x_L(n)$ and $x_R(n)$ are the left and right input signals and are sampled at a rate of 16 kHz. They are decomposed into auditory signals by the gammatone filterbanks [14] whose center frequencies are linearly spaced on an equivalent rectangular bandwidth (ERB) scale [15]. Note that the gammatone filters for all frequency bands have different group delays from each other, resulting in a phase shift between the frequency bands. In order to compensate for such a phase shift, a phase correction term is applied to the impulse responses of the gammatone filterbanks [16]. Auditory signals decomposed by the gammatone filterbanks are then windowed at a frame rate of 100 Hz using a Hamming window with a time resolution of 20 ms. Thus, the left and right auditory signals for the $i$-th frame and $j$-th gammatone frequency band are obtained, which are denoted as $x_L^{i,j}(n)$ and $x_R^{i,j}(n)$, respectively.

2.2. **Extraction of spatial cues.** In order to construct feature vectors for the GMMs, spatial cues, such as the interaural time difference (ITD) and the interaural level difference (ILD), are extracted for each time-frequency (T-F) segment. First of all, to extract the ITD for the $ij$-th T-F segment, a normalized cross-correlation (CC) coefficient between the left and right auditory signals of the $ij$-th T-F segment is computed as

$$C_{i,j}(\tau) = \frac{\sum_{n=0}^{N-1} x_L^{i,j}(n) x_R^{i,j}(n-\tau)}{\sqrt{\sum_{n=0}^{N-1} \left(x_L^{i,j}(n)\right)^2} \sqrt{\sum_{n=0}^{N-1} \left(x_R^{i,j}(n)\right)^2}} \tag{1}$$

where $N$ represents the number of speech samples per frame. In this paper, $N$ is set to 320, which is identical to the length of the Hamming window. In addition, $\tau$ is a time lag, and its range corresponds to an angle range for the spots where sound sources can be located. Assuming that a sound source is located from $-90°$ to $90°$, $\tau_{\min}$ and $\tau_{\max}$ can be defined as

$$\tau_{\min} = \left\lfloor \frac{d \cdot \sin(\theta_{\min}) \cdot f_s}{c} \right\rfloor \tag{2}$$

and

$$\tau_{\max} = \left\lceil \frac{d \cdot \sin(\theta_{\max}) \cdot f_s}{c} \right\rceil \tag{3}$$

where $c$ is the speed of sound, $d$ is the distance between two microphones, and $f_s$ is the sampling frequency. $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ indicate the flooring and ceiling operators, respectively. $\theta_{\min}$ and $\theta_{\max}$ denote the minimum and maximum angles for the location of the sound source on a radian frequency scale and are set to $-90°$ to $90°$, respectively. Therefore, $\tau$ ranges from $-8$ to 8 at a sampling rate of 16 kHz. Thus, the ITD for the $ij$-th T-F segment can be defined as the time lag at which the normalized CC coefficient is maximized. In other words, the ITD for the $ij$-th T-F segment is represented as

$$t_{i,j} = \left| \arg \max_{\tau_{\min} \le \tau \le \tau_{\max}} C_{i,j}(\tau) \right|. \tag{4}$$

In addition to ITD extraction, the ILD for the $ij$-th T-F segment is estimated as the ratio of energies obtained from the left and right auditory spectral signals using the equation of

$$l_{i,j} = \left| 10 \log_{10} \left( \frac{\sum_{n=0}^{N-1} \left( x_L^{i,j}(n) \right)^2}{\sum_{n=0}^{N-1} \left( x_R^{i,j}(n) \right)^2} \right) \right|. \tag{5}$$

2.3. **Extraction of root mean squared energy.** The proposed VAD method assumes that speech and noise are all point sources. In addition, the speech source is assumed to be located directly in front of dual microphones, i.e., at an angle of 0°, but noise sources can be located in arbitrary directions except in front of the dual microphones. When the signal comes from the front of the microphones, i.e., 0°, there are no time and level differences between the signals from the left and right microphones, which results in an ITD of 0 and ILD of 0. Thus, the distributions of ITD and ILD extracted from the silent intervals are similar to those extracted from the target speech. To discriminate speech intervals from silent intervals, a log-RMSE is also extracted and incorporated into the GMMs. The log-RMSE for the $i$-th analysis frame, $e_i$, is defined as

$$e_i = \log_{10} \sum_{j=0}^{J-1} e_{i,j} \tag{6}$$

where $J$ is the number of frequency bands and $e_{i,j}$ is the RMSE for the $ij$-th T-F segment, defined as

$$e_{ij} = \frac{1}{2} \sum_{j=0}^{J-1} \left( \sqrt{\sum_{n=0}^{N-1} \left( x_L^{i,j}(n) \right)^2} + \sqrt{\sum_{n=0}^{N-1} \left( x_R^{i,j}(n) \right)^2} \right). \tag{7}$$

As described in Equation (6), $e_i$ is obtained by taking a logarithm of the sum of $e_{i,j}$ over all frequency bands, where $J$ is set to 32 in this paper.

2.4. **GMM-based VAD classification.** As described in Section 2.3, the speech source is assumed to be located directly in front of the dual-microphone. Thus, the speech intervals can be discriminated by comparing the likelihood estimated from GMM for DOA = 0° with those from GMMs for other DOAs. Since the difference in ITD corresponding to an angular step of 10° is about one at a sample rate of 16 kHz, as described in Equations (2) and (3), ten GMMs are constructed for once every 10° DOA from 0° to 90° and one GMM is for silent intervals, resulting in eleven GMMs in total. Note here that the GMMs from −10° to −90° are same to GMMs from 10° to 90° because the ITD and ILD are always a positive value as expressed in Equations (4) and (5). Each GMM is trained for each frequency band based on a maximum likelihood estimation criterion. In other words, for a given training observation sequence, the optimal GMM for the $j$-th frequency band and the DOA of $\theta$, $\Lambda_{\theta,j} = \left( \mu_{\theta,j}^m, \Sigma_{\theta,j}^m, w_{\theta,j}^m \right)$, is estimated by running the expectation-maximization (EM) iteration, where $\mu_{\theta,j}^m$, $\Sigma_{\theta,j}^m$, and $w_{\theta,j}^m$ represent the mean vector, covariance matrix, and weight vector of the $m$-th Gaussian mixture of $\Lambda_{\theta,j}$, respectively. Note that the number of Gaussian mixtures of $\Lambda_{\theta,j}$ is set as $M = 4$ from the preliminary experiments.

For a given set of trained GMMs, $\Lambda_\theta = \{\Lambda_{\theta,j} | 0 \leq j < J\}$, and observations for the $i$-th frame, $X_i = \{t_{i,j}, l_{i,j}, e_i | 0 \leq j < J\}$, the likelihood for each GMM, $p(X_i|\Lambda_\theta)$ is defined by

$$p(X_i|\Lambda_\theta) = \sum_{j=0}^{J-1} \log\left(p(X_{i,j}|\Lambda_{\theta,j})\right) \tag{8}$$

where

$$p(X_{i,j}|\Lambda_{\theta,j}) = \sum_{m=0}^{M-1} w_{\theta,j}^m N(X_{i,j}; \mu_{\theta,j}^m, \Sigma_{\theta,j}^m), \quad \theta \in \Theta. \tag{9}$$

In Equation (9), $X_{i,j} = \{t_{i,j}, l_{i,j}, e_i\}$ indicates the observation for the $ij$-th T-F segment and $N(\cdot)$ denotes the operation of the Gaussian distribution. In addition, $\Theta = \{0°, 10°, \cdots, 90°, \mathit{silence}\}$ represents all eleven GMMs. Next, the GMM having the maximum *a posteriori* probability is obtained as

$$\theta_i^* = \arg\max_\theta p(X_i|\Lambda_\theta). \tag{10}$$

By using Equation (10), the $i$-th analysis frame is declared as either a speech or non-speech interval by the following equation of

$$\overline{V}(i) = \begin{cases} 1, & \text{if } \theta_i^* = 0° \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where 0 and 1 represent speech and non-speech intervals, respectively.

The proposed VAD method described above discriminates speech intervals from non-speech intervals by exploiting spatial selectivity. Thus, speech intervals with weak spatial correlations, such as unvoiced intervals, might be classified as a non-speech frame. In order to remedy such a problem, the VAD result, $\overline{V}(i)$ is refined by employing the following VAD error correction scheme.

## 3. HILO-Based VAD Error Correction Scheme.

In general, the ITD and ILD for each T-F segment are mainly affected by a dominant sound source. Thus, if the noise signal is stronger than the target speech signal for a given analysis frame, the frame may be declared as a non-speech interval, resulting in an erroneous VAD result. This comes from a typical situation in which the unvoiced interval of the target speech overlaps with an interval of noise signals. To mitigate this problem, the analysis frame declared as a non-speech interval, i.e., $\overline{V}(i) = 0$ is subsequently analyzed to correct the GMM-based VAD result.

The correction scheme starts by examining a frame when the likelihood for DOA $= 0°$ is greater than the average likelihood for the other DOAs with the exception of $\theta_i^*$. Since the energy above $f_H$ ( $= 4$ kHz) is greater than that below $f_L$ ( $= 2$ kHz) in an unvoiced interval, HILO is useful for classifying unvoiced intervals [17]. The HILO is defined as

$$HILO(i) = \log\left(\sum_{j=J_H}^{J-1} e_{i,j}\right) - \log\left(\sum_{j=0}^{J_L} e_{i,j}\right) \tag{12}$$

where $J_H$ and $J_L$ denote the frequency bands corresponding to $f_H$ and $f_L$, respectively, and they are set to $J_H = 25$ and $J_L = 18$ in this paper. By using HILO, the VAD result can be refined as

$$V(i) = \begin{cases} 1, & \text{if } \overline{V}(i) = 1 \\ 1, & \text{if } (\overline{V}(i) = 0 \text{ and } HILO(i) > 0) \text{ and} \\ & (p(X_i|\Lambda_\theta)|_{\theta=0°} > \alpha \cdot p_{avg}(X_i|\Lambda_\theta)|_{\theta \in \overline{\Theta}_i}) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

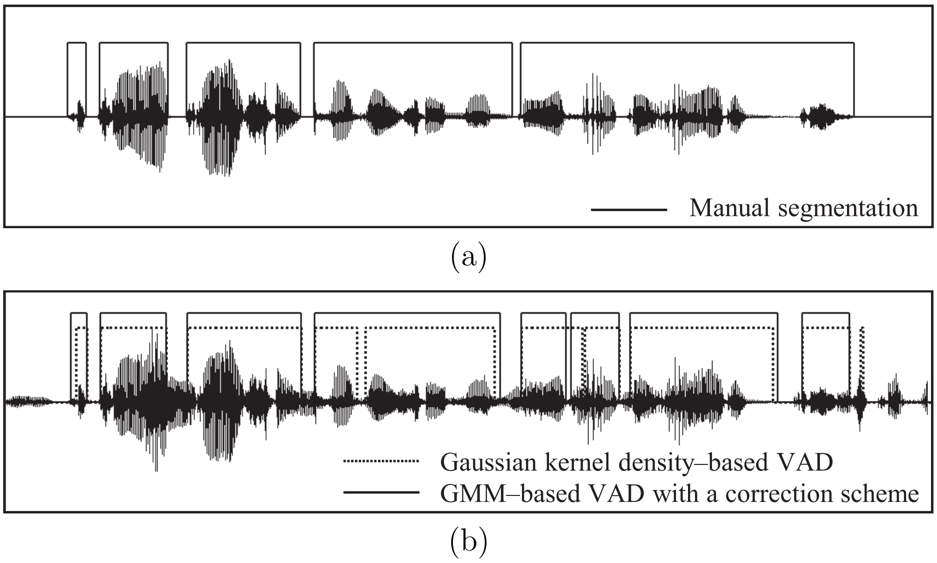


Figure 2. Illustrations of VAD results obtained from different VAD methods under a competing speech noise condition at 10 dB SNR; (a) clean speech and the VAD result by manual segmentation, and (b) noisy speech and the VAD results of the Gaussian kernel density-based method and the proposed GMM-based VAD method with the HILO-based error correction scheme.

where $\overline{\Theta}_i = \{\theta | \theta \in \{\Theta - \{0^\circ, \theta_i^*\}\}\}$ represents all GMMs except for $\theta_i^*$ and $0^\circ$. In addition, $p_{avg}(X_i|\Lambda_\theta)|_{\theta \in \overline{\Theta}_i}$ and $\alpha$ indicate the average likelihoods for $\theta \in \overline{\Theta}_i$ and a weighting factor to control the degree of the unvoiced interval of the target speech, respectively.

Figure 2 shows the waveform of the input speech and the VAD results of different VAD methods, such as Gaussian kernel density-based VAD [13] and the proposed GMM-based VAD with the HILO-based error correction scheme. As shown in the figure, the VAD result of the proposed GMM-based VAD with the HILO-based correction scheme has a relatively small number of VAD errors compared to the Gaussian kernel density-based VAD method. In particular, the proposed VAD method more reduces misclassification errors of a speech interval being declared a non-speech interval.

## 4. **Performance Evaluation.**

The performance of the proposed GMM-based VAD method with the HILO-based error correction scheme was evaluated in terms of the false rejection rate (FRR) and false alarm rate (FAR), and the results were compared with those of the Gaussian kernel density-based VAD method [13] and the GMM-based VAD method without any correction scheme. Note that the Gaussian kernel density-based VAD method [13] discriminated speech intervals from non-speech intervals by comparing the likelihood obtained from the Gaussian kernel model for 0 with those from other Gaussian kernel models, where the Gaussian kernel models were trained under the same conditions as the GMMs were trained.

In order to additionally demonstrate the effectiveness of the proposed VAD method, the performance of the proposed VAD method was also compared with those of the different single-channel VAD method [8] and the microphone array-based VAD method [11]. The former used the spectral density of noise and the variance of SNRs as VAD features [8], and the latter utilized the phase vector obtained between the signals from different microphones as a VAD feature [11]. In this paper, the microphone array-based VAD method was implemented using only two microphones.

4.1. **Speech database.** To estimate GMMs for the proposed VAD method, a target speech database was artificially constructed using speech data spoken by 10 males and 10 females taken from the TIMIT corpus [18]. In addition, speech material spoken by 4 males and 4 females in the TIMIT corpus was utilized as the interfering speech signals, i.e., noise signals, which were not identical to anyone in the target speech database. Note that the length of all the target speech data was 280 second long in total. The percentage of speech frames in the target speech database was 64.3%, while 35.7% of the target speech material was non-speech frames, i.e., silent frames. Then, each speech signal from the database and noise signals were convolved with a head-related impulse response (HRIR) modeled from a KEMAR dummy head [19]. In other words, speech signals were filtered using an HRIR with an angle of 0°, while the noise signals were convolved with HRIRs at an angle ranging from 10° to 90° in a step of 10°. Finally, the speech and noise signals were combined with different SNRs of 5, 10, 15, and 20 dB.

4.2. **Performance analysis.** In this subsection, we measured the FRR and FAR of different VAD methods. Fortunately, the TIMIT database provided manual segmentation results for each utterance, which allowed for evaluation of the FRRs and FARs. In other words, by comparing the VAD result of each VAD method with that of the manual segmentation given by the TIMIT database, the FRR and FAR were measured as

$$FRR(\%) = N_n/N_s^{Ref} \times 100 \tag{14}$$

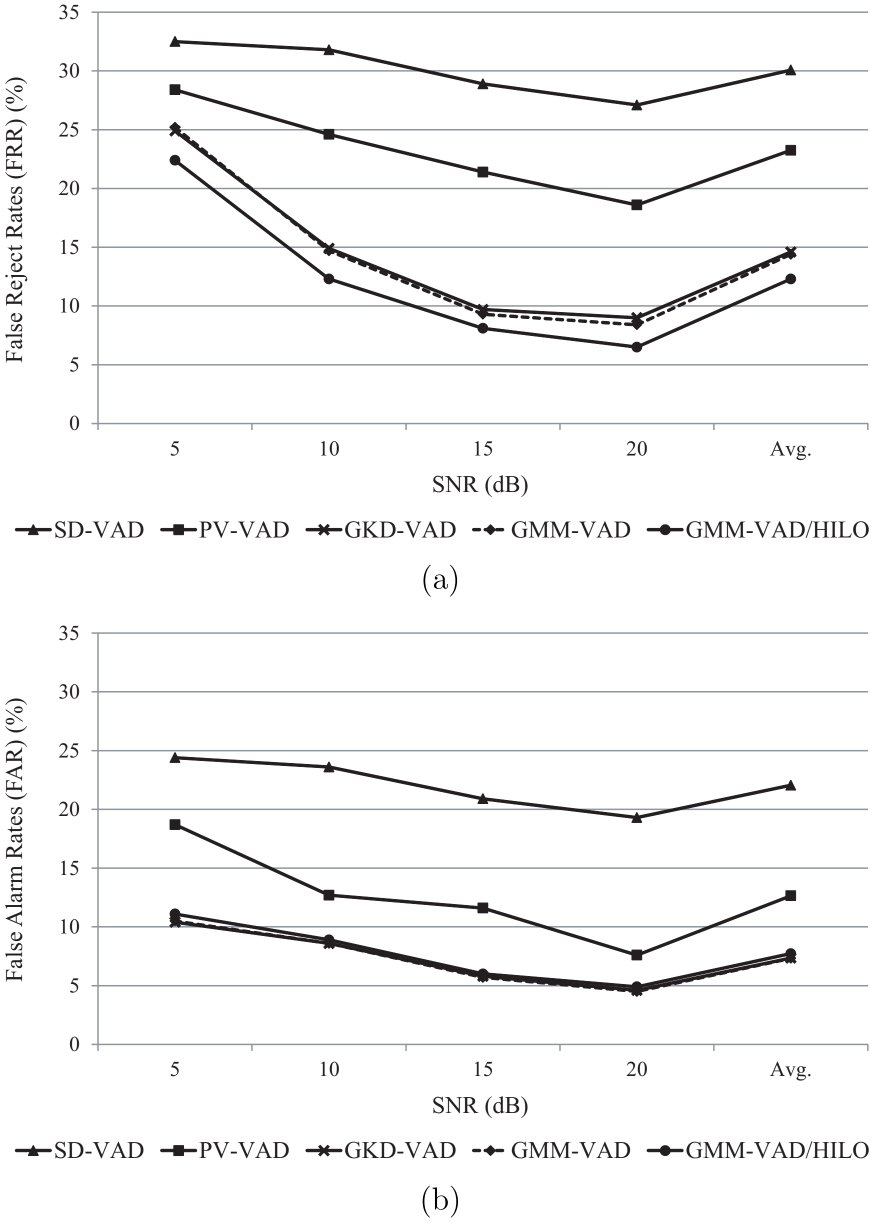
and

$$FAR(\%) = N_s/N_n^{Ref} \times 100 \tag{15}$$

where $N_s^{Ref}$ and $N_n^{Ref}$ are the total numbers of speech and non-speech frames as labeled by the manual segmentation, respectively. $N_s$ and $N_n$ are the numbers of incorrectly detected speech and non-speech frames, respectively.

Figure 3 shows the FRRs and FARs of several VAD methods under different SNRs. In the figure, SD-VAD, PV-VAD, GKD-VAD, GMM-VAD, and GMM-VAD/HILO refer to the spectral density-based single-channel VAD method [8], the phase vector-based VAD method using a microphone array [11], the Gaussian kernel density-based VAD method [13], the GMM-based VAD method without any correction scheme, and the proposed GMM-based VAD method with the HILO-based VAD error correction scheme, respectively. As shown in Figures 3(a) and 3(b), SD-VAD provided the highest FRRs and FARs because the noise signals were non-stationary interfering speech signals. In addition, PV-VAD gave higher FRRs and FARs than GMM-VAD, GKD-VAD, and GMM-VAD/HILO. Here, PV-VAD parameterized the distribution of phase vectors for speech present and absent intervals as a Gaussian function. In particular, the Gaussian parameters of speech absent intervals were derived from frames initially assumed to be speech-absent, i.e., the initial period of target speech utterance. Thus, when the noise signals were also absent during the initial period, the parameters of PV-VAD were not correctly derived, which resulted in degradation of the performance of PV-VAD.

Figure 3(a) shows that GMM-VAD provided slightly lower FRRs than GKD-VAD due to finer expression on the likelihoods for speech presence or absence using GMMs. On the other hand, GMM-VAD/HILO achieved a significant reduction in FRRs for all SNR conditions, compared with GKD-VAD and GMM-VAD. This was because GMM-VAD/HILO was sort of a refined version of GMM-VAD regarding the VAD result. In contrast to the FRRs, GMM-VAD/HILO provided FARs comparable to GKD-VAD and GMM-VAD. This also implies that GMM-VAD/HILO could correct the false rejection errors of GMM-VAD in unvoiced intervals while maintaining the false alarm errors.



(a)

(b)

Figure 3. Comparison of VAD performance of different VAD methods according to different SNRs; (a) FRRs, and (b) FARs of the spectral density-based method (SD-VAD), the phase vector-based method (PV-VAD), the Gaussian kernel density-based method (GKD-VAD), the GMM-based method without error correction (GMM-VAD), and the proposed GMM-based method with the HILO-based error correction (GMM-VAD/HILO).

5. **Conclusions.** In this paper, we proposed a GMM-based dual-microphone VAD method incorporating an HILO-based error correction scheme to improve VAD performance in non-stationary noisy environments. In particular, speech intervals were detected by searching for the GMM with the maximum likelihood among all GMMs trained for both different DOAs and silence. In addition, an HILO-based error correction scheme was applied to reduce the VAD errors during unvoiced intervals of the target speech. To evaluate the performance of the proposed VAD method, the FRRs and FARs of the

proposed VAD method were measured and compared with those of other VAD methods. It was shown from the evaluation that the proposed GMM-based VAD method with the HILO-based correction reduced average FRR by 17.8%, 10.9%, 2.3%, and 2.1%, compared with the spectral density-based VAD method, phase vector-based VAD method, Gaussian kernel density-based VAD method, and GMM-based VAD method without any correction scheme, respectively.

**Acknowledgment.** This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-010636).

## REFERENCES

[1] M. Nakayama and S. Ishimitsu, Speech support system using body-conducted speech recognition for disorders, *International Journal of Innovative Computing, Information and Control*, vol.5, no.11(B), pp.4255-4265, 2009.

[2] J. C. Junqua, B. Mak and B. Reaves, A robust algorithm for word boundary detection in the presence of noise, *IEEE Transactions on Speech and Audio Processing*, vol.2, no.3, pp.406-412, 1994.

[3] L. R. Rabiner and M. R. Sambur, An algorithm for determining the endpoints of isolated utterances, *Bell System Technical Journal*, vol.54, no.2, pp.297-315, 1975.

[4] R. Tuker, Voice activity detection using a periodicity measure, *IEE Proceedings-I, Communications, Speech, and Vision*, vol.139, no.4, pp.377-380, 1992.

[5] J. A. Haigh and J. S. Mason, Robust voice activity detection using cepstral features, *Proc. of IEEE Region 10 Conference (TENCON)*, Beijing, China, pp.321-324, 1993.

[6] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre and A. Rubio, Efficient voice activity detection algorithms using long-term speech information, *Speech Communication*, vol.42, no.3-4, pp.271-287, 2004.

[7] P. D. Welch, The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, *IEEE Transactions on Audio Electroacoustics*, vol.15, no.2, pp.70-73, 1967.

[8] A. Davis, S. Nordholm and R. Tognery, Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold, *IEEE Transactions on Audio, Speech, and Language Processing*, vol.14, no.2, pp.412-424, 2006.

[9] J. Chen and W. Ser, Speech detection using microphone array, *Electronic Letters*, vol.36, no.2, pp.181-182, 2000.

[10] I. Potamitis, Estimation of speech presence probability in the field of microphone array, *IEEE Signal Processing Letters*, vol.11, no.12, pp.956-959, 2004.

[11] G. Kim and N. I. Cho, Voice activity detection using phase vector in microphone array, *Electronics Letters*, vol.43, no.14, pp.783-784, 2007.

[12] J. E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani and M. Fujimoto, Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, vol.4, pp.385-388, 2007.

[13] J. H. Park, M. H. Shin and H. K. Kim, Statistical model-based voice activity detection using spatial cues and log energy for dual-channel noisy speech recognition, *Communications in Computer and Information Science*, vol.120, pp.172-179, 2010.

[14] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth and P. Rice, An efficient auditory Filterbank based on the Gammatone functions, *APU Report 2341, MRC*, Applied Psychology Unit, Cambridge, U.K., 1988.

[15] B. R. Glasberg and B. C. J. Moore, Derivation of auditory filter shapes from notched-noise data, *Hearing Research*, vol.47, no.1-2, pp.103-138, 1990.

[16] G. J. Brown, *Modelling Auditory Processing and Organization*, Cambridge University Press, U.K., 1993.

[17] L. Siegel and A. Bessey, Voiced/unvoiced/mixed excitation classification of speech, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.30, no.3, pp.451-460, 1982.

[18] W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, The DARPA speech recognition research database: Specifications and status, *Proc. of DARPA Speech Recognition Workshop*, Palo Alto, CA, pp.93-99, 1986.

[19] W. G. Gardner and K. D. Martin, HRTF measurements of a KEMAR, *Journal of the Acoustical Society of America*, vol.97, no.6, pp.3907-3908, 1995.