Direction-of-Arrival Based SNR Estimation for Dual-Microphone Speech Enhancement

Seon Man Kim and Hong Kook Kim, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a method for estimating target speech by exploring the spatial cues in adverse noise environments. This method is able to reliably estimate the signal-to-noise ratio (SNR) using the phase difference obtained from dual-microphone signals. To this end, spatial cues such as the phase difference are used to estimate the target-to-non-target directional signal ratio (TNR). Based on the estimated TNR, a direction-of-arrival (DOA)-based SNR is then estimated by using a statistical model-based log-likelihood ratio test for the target speech activity decision followed by a decision-directed approach. The estimate is then incorporated into a Wiener filter in order to obtain a spectral-gain attenuator. The perceptual evaluation of speech quality shows that the performance of a dual-microphone speech enhancement system employing the proposed estimation method outperforms single- and dual-microphone speech enhancement systems that use conventional methods such as Wiener filtering, beamforming, or phase-error-based filtering under noise conditions whose SNR ranges from 0 to 20 dB.**

***Index Terms*—Direction-of-arrival, dual-microphone signal, signal-to-noise ratio, spatial cue, speech enhancement, target-to-non-target directional signal ratio.**

I. Introduction

RECENTLY, speech enhancement in noisy environments has attracted a great deal of research interest, especially in applications such as mobile voice communications, automatic speech recognition, and hearing aids [1]–[4]. The goal of speech enhancement is to suppress additive background noise components while maintaining the quality and intelligibility of speech [1]. This task is usually accomplished by preserving the characteristics of speech using the short-term spectral amplitude (STSA), for which the reliable estimation of the signal-to-noise ratio (SNR) is crucial in noisy environments [1], [2].

The SNR estimation generally involves two steps: a noise variance estimation and an SNR computation that is based on the estimated noise variance [1], [3]. For the SNR computation, it has been reported that a decision-directed (DD) approach, in which the decision was based on the *a priori* and *a posteriori* SNR estimates, provided a simple but effective solution with a reasonable computational cost [2]. However, since the *a posteriori* SNR estimate is directly obtained from the noisy speech and the estimated noise variance, an unreliable noise variance estimate can affect the SNR estimation, which can then distort the estimated clean speech in severely adverse noise environments [2]. Thus, accurate noise variance estimation plays an important role in reliable SNR estimation; however, it may also be unreliable in adverse noise environments [2], [3].

To improve the speech enhancement performance in adverse noise environments, multiple-microphone systems are increasingly becoming popular despite their additional cost since they result in a better performance than can be obtained by single-microphone speech enhancement algorithms [4]. As opposed to only time information from a single microphone signal, multiple-microphone signals allow additional spatial cues to be exploited because the direction-of-arrival (DOA) of target speech is strongly linked to the phase difference between multiple-microphone signals [4]–[6]. Thus, by assuming that only one sound source power is dominant in each time-frequency (T-F) bin, the dominance of the target speech's presence or absence in each T-F bin can be determined using the phase difference [5]–[10]. This binary decision has become quite well-known for achieving a good performance [5]–[10]. However, enhanced speech continues to suffer from musical noise, because the inherent discontinuous zero-paddings result in artifact distortions in the estimated target signal. This problem is even worse in real world environments where the sparseness assumption cannot quite be satisfied.

On one hand, under non-sparse conditions, the phase difference from a directional target speech signal would be variable due to non-target-directional noises. Thus, the binary decision for the dominance of the target speech in each T-F bin would be inappropriate [10]–[19]. For example, a beamformer was utilized to construct the directional sensitivity—referred to as the spatial directivity pattern (SDP)—and attenuate spatially-unwanted noises arising from non-target directions [14]–[16]. A post-filtering technique was also utilized to further improve the noise reduction performance of the beamformer [4], [17], and an independent component analysis (ICA) based blind source separation (BSS) was used to estimate the acoustic paths from the sound sources to each microphone [18]. In addition, a soft-masking method, which was referred to as a phase-error-based filter (PEF) [19], was employed as the spectral-gain function of phase differences. This approach was motivated by the fact that phase difference errors between the dual-microphone signals were related to the SNR of the observed noisy speech signal.

Manuscript received January 28, 2014; revised May 07, 2014; accepted September 22, 2014. Date of publication September 26, 2014; date of current version November 05, 2014. This work was supported in part by a National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT and Future Planning (MSIP) (No. 2012-010636), and ICT R&D program of MSIP/IITP [2014-044-055-002, Loudness Based Broadcasting Loudness and Stress Assessment of Indoor Environment Noises]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Man-Wai Mak.

S. M. Kim is with the Institute of Sound and Vibration Research (ISVR), University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: kobem30002@gmail.com).

H. K. Kim is with the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea (e-mail: hongkook@gist.ac.kr).

Digital Object Identifier 10.1109/TASLP.2014.2360646

However, since the phase difference of the target speech is estimated based on target DOA information, an unreliable DOA estimate can distort the estimated clean speech in a PEF or a beamformer [4], [19]. Thus, the accurate DOA estimation of the target speech plays an important role in the reliability of the target speech estimation. However, it may also be unreliable in adverse noise environments [4]. Since the accurate estimation of the target DOA is beyond the scope of this paper, this paper assumes that the target DOA is known *a priori*.

Besides the DOA problem, the strategies of using beamformers, ICA-based BSS, PEF, and post-filtering techniques limit the performance of dual-microphone speech enhancements due to the following reasons. First, the SDPs of a beamformer were constrained by the number of microphones. In other words, the performance of beamformer-based speech enhancement systems was found to be highly dependent on the number of microphones used [4]. Thus, the performance obtained via dual-microphones might not be satisfactory, compared to the masking-based methods such as PEF [19]. Second, for the acoustic path estimation in an ICA-based BSS, the number of sound sources should be smaller than or equal to that of the microphones, which might be impractical [18], [20]. Third, PEF was effective at reducing non-target directional noise under low SNR conditions, but it could distort the target-directional speech at high SNRs [19]. Moreover, a beamformer-based post-filtering technique estimates the spectral-gain attenuator by utilizing the power ratio between the input and output signals of the beamformer, and the estimated spectral-gain attenuator was then applied to the beamformer output signal to further reduce the noise components. However, since such a post-filtering technique was similar to PEF [19], the performance at a high SNR was poorer than those at low SNRs. Nevertheless, when the DOA of target speech is uncertain, the performance of non-target directional signal attenuation by the PEF could be worse than that by a single-microphone technique using a temporal cue. This is because PEF works only on the basis of a DOA cue represented by phase differences [19]. Accordingly, it could be more beneficial for better spectral estimate to use both the temporal and DOA cues than to use either of them.

As motivated above, this paper proposes a method for estimating the target speech in a dual-microphone speech enhancement system by exploring the spatial cues in adverse noise environments in order to reliably estimate the SNR using the phase difference obtained from dual-microphone signals. First, spatial cues such as the phase difference are derived as a form of target-to-non-target directional signal ratio (TNR), which is defined as the power ratio between the target-directional enhanced and rejected signals [21]. Next, a DOA-based SNR is defined using the TNR, from which a DOA-based SNR estimation method is also proposed. Then, the estimated DOA-based SNR is incorporated into a Wiener filter in order to obtain a spectral-gain attenuator. Finally, the performance of a dual-microphone speech enhancement system employing the proposed method is evaluated by measuring the source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) [22] under interference speech conditions whose SNR ranges from 0 to 20 dB. In addition, we evaluate the performance by measuring the perceptual evaluation of speech quality (PESQ) scores [23] under four different noise conditions such as interference speech, factory, vacuum cleaner, and white noise. Finally, we compare the performance of a dual-microphone speech enhancement system employing the proposed method with that of a conventional one using either a super-directive beamformer (SDB) [4], [14], [15], a generalized sidelobe canceller (GSC) [16] followed by a post Wiener filter (GSC-PW), PEF [19], or an angular spectrum-based masking method (ASBM) [24].[1]

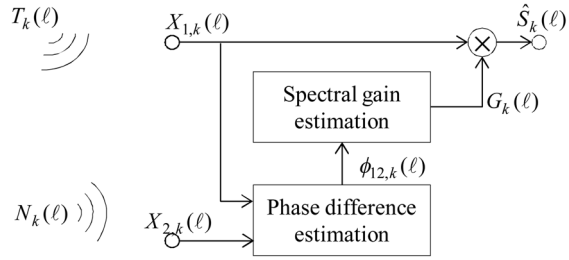

Fig. 1. Block diagram of a dual-microphone speech enhancement system that utilizes the phase differences as spatial cues.

The remainder of this paper is organized as follows. Section II briefly reviews a dual-microphone speech enhancement system using a spatial cue such as the phase differences between dual-microphone signals. Section III proposes a DOA-based SNR estimation method to improve the performance of a dual-microphone speech enhancement system. Section IV evaluates the performance of a dual-microphone speech enhancement system employing the proposed method and compares it with those of conventional single- and dual-microphone speech enhancement systems. Finally, our findings are summarized in Section V.

II. Dual-Microphone Speech Enhancement Using Phase Differences

In this section, we briefly describe a dual-microphone speech enhancement system that utilizes spatial cues estimated from dual-microphone signals. As shown in Fig. 1, spatial cues such as the phase differences between dual-microphone signals are first estimated. Then, the spectral gain for estimating target speech is obtained from the estimated phase differences.

In order to model dual-microphone signals, we assume that sound sources can be classified into target and non-target directional sources, and that the target directional speech source is located far enough from the microphone array, which is referred to as the acoustic far field condition. In other words, the target-directional speech is free from any room reverberation effects since only the direct path is considered in modeling dual-microphone input signals. Then, the k-th spectral component of the m-th microphone signal at the ℓ-th frame, $X_{m,k}(\ell)$, can be represented as

$$X_{1,k}(\ell) = T_k(\ell) + N_{1,k}(\ell), \tag{1}$$

$$X_{2,k}(\ell) = T_k(\ell)e^{-j\omega_k\tau_{12}} + N_{2,k}(\ell) \tag{2}$$

where $T_k(\ell)$ $(k = 0, 1, \cdots, K-1)$ is the k-th spectral component representing the target directional source, and $N_{m,k}(\ell)(m = 1, 2)$ is each microphone-recorded version of the non-target directional source $N_k(\ell)$ [4], [19]. Note here that K denotes the total number of frequency bins. In addition, τ_{12} is the time difference-of-arrival (TDOA) between the two

[1]All abbreviations in this paper were listed in Appendix I.

microphones, and $\omega_k \in [-\pi, \pi]$ is the angular frequency in radians at the k-th frequency bin. Here, we can also assume that the DOA of the desired signal source is either easily predictable or known *a priori*, such as for an interlocutor-facing hearing-aid wearer [4] or a driver in a car telematics system [2], [25]. Moreover, even when it is difficult to predict the DOA of the desired signal source, the TDOA can be estimated before applying a speech enhancement algorithm. To this end, we can apply popular localization algorithms such as the generalized cross correlation (GCC) or steered response power (SRP) with phase transform (PHAT) weighting $W_k^{PHAT}(\ell)$ [4], [19]. For example, the TDOA estimate by GCC-PHAT, $\hat{\tau}_{12}$, is represented as [19]

$$\hat{\tau}_{12} = \arg\max_{\tau} \sum_{\ell=0}^{L-1} \sum_{k=-K}^{K} W_k^{PHAT}(\ell) X_{1,k}(\ell) X_{2,k}^*(\ell) e^{-j\omega_k \tau} \tag{3}$$

where $W_k^{PHAT}(\ell) = 1/|X_{1,k}(\ell)X_{2,k}^*(\ell)|$ and $*$ is a complex conjugate operator. Next, by multiplying $\exp(j\omega_k\hat{\tau}_{12})$ into $X_{2,k}(\ell)$ in (2), $X_{2,k}(\ell)$ can be rewritten as $X'_{2,k}(\ell) = X_{2,k}(\ell) \cdot \exp(j\omega_k\hat{\tau}_{12}) = T_k(\ell) \cdot \exp(-j\omega_k(\tau_{12} - \hat{\tau}_{12})) + N_{2,k}(\ell) \cdot \exp(j\omega_k\hat{\tau}_{12})$. Then, assuming $\tau_{12} = \hat{\tau}_{12}$, we have

$$X'_{2,k}(\ell) = T_k(\ell) + N'_{2,k}(\ell) \tag{4}$$

where $N'_{2,k}(\ell) = N_{2,k}(\ell) \cdot \exp(j\omega_k\hat{\tau}_{12})$ also becomes a non-target directional source.

As shown in Fig. 1, dual-microphone speech enhancement algorithms such as binary masking, beamformers, and PEF attempt to estimate the spectrum of the target directional signal, $\hat{T}_k(\ell)$, in the form $\hat{T}_k(\ell) = G_k(\ell)X_{1,k}(\ell)$ or $\hat{T}_k(\ell) = G_k(\ell)X_{2,k}(\ell)$, where $G_k(\ell)$ is the spectral-gain attenuator estimated from spatial cues such as the phase differences between dual-microphone signals. For example, when the dual-microphone beamforming output from (1) and (4) is denoted as $B_k(\ell)$, the transfer function of the beamformer, $H_{B,k}(\ell) = B_k(\ell)/X_{1,k}(\ell)$, for the target-directional source is given by

$$G_k^{BF}(\ell) = W_{1,k}^*(\ell) + W_{2,k}^*(\ell) \cdot \exp(j\phi_{12,k}(\ell)) \tag{5}$$

where $W_{m,k}(m = 1, 2)$ denotes a beamformer weight of the m-th microphone signal [4], [14], [21], and $\phi_{12,k}(\ell)$ is a phase difference between $X_{1,k}(\ell)$ and $X'_{2,k}(\ell)$.

Depending on the process how the beamformer weights are chosen, beamformers are classified as either data-independent or data-dependent [14]. To be specific, the weights in data-independent beamformers are designed so that the spatial directivity approximates a desired pattern, which is independent of the array input data; a delay-and-sum beamformer (DSB) or blocking matrix (BM) belongs to such a category. Conversely, the weights in data-dependent beamformers are determined based on the statistics of the input audio signals. As an example of a data-dependent beamformer, a super-directive beamformer (SDB) has been designed to obtain an optimal performance on the diffuse noise field, in which the beamformer weights $\mathbf{W}_k^{SDB} = [W_{1,k}^{SDB}, W_{2,k}^{SDB}]^T$ are derived as $\mathbf{W}_k^{SDB} = \mathbf{\Gamma}^{-1}\mathbf{d}_k/(\mathbf{d}_k^H\mathbf{\Gamma}^{-1}\mathbf{d}_k)$ [4], [15]. Here, H represents the Hermitian operator, $\mathbf{d}_k = [1, \exp(-j\omega_k\tau_{12})]^T$, and $\mathbf{\Gamma}$ denotes the coherence matrix of the diffuse noise sound field, whose element $\Gamma_{uv} = \mathrm{sinc}(\omega_k d_{uv}/c)$ represents the coherence between the u-th and v-th microphones [15], [17] where d_{uv} is the distance between the u-th and v-th microphones and c is the speed of sound (≈ 340 m/s in dry air). The performance of beamformer-based speech enhancement systems are highly dependent on the number of microphones used [4]. For example, the performance of an SDB implemented with dual microphones might be worse than that of a PEF [19].

As another example, PEF is motivated by the fact that the error in phase difference between dual-microphone signals varies depending on the power ratio between the target and non-target directional signals, i.e., TNR. The TNR at the k-th spectral component and the ℓ-th frame, $\eta_k(\ell)$, is defined as [21]

$$\eta_k(\ell) \triangleq \frac{|T_k(\ell)|^2}{|N_k(\ell)|^2} \tag{6}$$

where $T_k(\ell)$ and $N_k(\ell)$ are respectively the k-th spectral components of the target directional source signal and non-target directional signal at the ℓ-th frame. Then, the TNR-based spectral-gain attenuator, $G_k^{TNR}(\ell)$, can be derived by employing the Wiener filter formulation as

$$G_k^{TNR}(\ell) = \frac{\eta_k(\ell)}{\eta_k(\ell) + \gamma} \tag{7}$$

where γ is a constant used to control the degree of noise attenuation, in which a higher value of γ results in better performance at low input TNRs but worse performance at high input TNRs [19], [21].

PEF attempts to find a corresponding TNR, $\eta_k(\ell)$, to determine the phase difference $\phi_{12,k}(\ell)$ between $X_{1,k}(\ell)$ and $X'_{2,k}(\ell)$, which is then utilized to estimate $G_k^{TNR}(\ell)$. It has been reported in [19] that $\eta_k(\ell)$ is related to the phase difference, and the inverse of the squared phase difference, $1/|\phi_{12,k}(\ell)|^2$, can be used to approximate $\eta_k(\ell)$. Thus, PEF uses $1/|\phi_{12,k}(\ell)|^2$ instead of $\eta_k(\ell)$ in order to obtain the spectral-gain attenuator, which is then denoted as $G_k^{PEF}(\ell)$ [19], i.e.,

$$G_k^{PEF}(\ell) = \frac{1}{1 + \gamma \cdot \phi_{12,k}^2(\ell)}. \tag{8}$$

It has been shown that the PEF provided a higher digit recognition accuracy than both the dual-microphone SDB and the beamformer with a post-filter [19]. Thus, it was postulated that the PEF could be effective in non-target directional noise suppression, potentially improving the desired signal performance. However, though the PEF was effective at reducing non-target directional signals under low TNR conditions, it could distort the target-directional desired signal at high TNRs [19]. This could be a major drawback with the PEF since the phase difference-based TNR estimation approach may be unreliable, although it gives better performance than other dual-microphone speech enhancement algorithms.

Consequently, the phase difference is seen to play a crucial role in estimating the spectral-gain attenuator using spatial cues. Therefore, it is important how phase differences are utilized for the reliable enhancement of target noisy speech components; this paper attempts to provide a solution to deal with this problem.

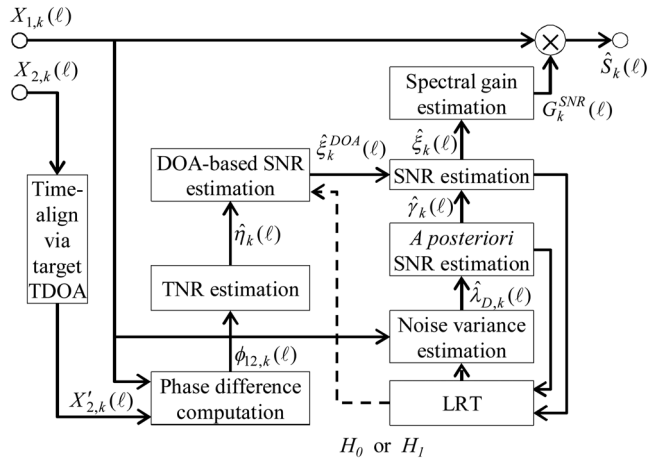

Fig. 2. Block diagram of a dual-microphone speech enhancement system employing the proposed SNR estimation method.

III. Proposed DOA-Based SNR Estimation Using Dual-Microphone Signals

As mentioned in Section I, the performance of a speech enhancement method using a single-microphone signal such as the Wiener filter may be poorer than that of a speech enhancement method using multiple-microphone signals; however, the better *a priori* SNR estimation can provide a more accurate spectral-gain attenuation [1]. Thus, in this section, an SNR-estimation method based on dual-microphone phase differences is proposed in order to reliably utilize spatial cues for the target speech estimation.

Fig. 2 shows a block diagram of the proposed method. First, the phase difference, $\phi_{12,k}(\ell)$, between $X_{1,k}(\ell)$ and a time-aligned version of $X_{2,k}(\ell)$, $X'_{2,k}(\ell)$, is estimated and utilized to obtain a TNR estimate, $\hat{\eta}_k(\ell)$. Next, $\hat{\eta}_k(\ell)$ is used to estimate the DOA-based SNR, $\hat{\xi}_k^{DOA}(\ell)$, which is also utilized to reliably obtain an SNR estimate, $\hat{\xi}_k(\ell)$. Finally, a speech enhancement spectral-gain attenuator, $G_k^{SNR}(\ell)$, is obtained on the basis of $\hat{\xi}_k(\ell)$, which is subsequently applied to the first or second microphone noisy speech signal.[2]

A. TNR Estimation

As shown in (7) and (8), the spectral-gain attenuator realized by $\hat{\eta}_k(\ell) = 1/|\phi_{12,k}(\ell)|^2$ was found to provide better speech enhancement performance than a beamforming algorithm [19]. In this subsection, we explain the TNR estimation method that is based on the phase differences.

For this task, we use the spectral-gain function of phase difference that is developed for a beamformer as in (5). Fortunately, this spectral-gain function makes it possible to improve the performance of the beamformer by replacing the phase difference, $\phi_{12,k}(\ell)$, with its frequency-normalized version, $\phi_{12,k}^{norm}(\ell) = \phi_{12,k}(\ell) \cdot c/(\omega_k \cdot d)$. This is because we understand that a beamforming function employing $\phi_{12,k}^{norm}(\ell)$ enables a better performance than employing $\phi_{12,k}(\ell)$; the performance benefit from using $\phi_{12,k}^{norm}(\ell)$ instead of $\phi_{12,k}(\ell)$ was also reported in [5] and [10]. Thus, we can obtain the DSB transfer function to estimate the target-directional speech, $G_k^{DSB}(\ell)$, where the beamformer weights for DSB are expressed as $W_{1,k}(\ell) = 0.5$ and $W_{2,k}(\ell) = 0.5\exp(j\omega_k\tau_{12})$ [4]. That is,

$$G_k^{DSB}(\ell) \approx 0.5(1 + \exp(j\phi_{12,k}^{norm}(\ell))). \tag{9}$$

On the other hand, since $W_{1,k}(\ell) = 1$ and $W_{2,k}(\ell) = -\exp(j\omega_k\tau_{12})$, the transfer function of a BM, $G_k^{BM}(\ell)$, can be obtained to reject the target-directional speech [4] as

$$G_k^{BM}(\ell) \approx 1 - \exp(j\phi_{12,k}^{norm}(\ell)). \tag{10}$$

By multiplying $X_{1,k}(\ell)$ by $G_k^{DSB}(\ell)$ and $G_k^{BM}(\ell)$, the estimates of $T_k(\ell)$ and $N_k(\ell)$ can be obtained as $\hat{T}_k(\ell) = G_k^{DSB}(\ell) \cdot X_{1,k}(\ell)$ and $\hat{N}_k(\ell) = G_k^{BM}(\ell) \cdot X_{1,k}(\ell)$, respectively. Finally, as derived in Appendix II of this paper, the TNR estimate, $\hat{\eta}(\ell)(= \hat{T}(\ell)/\hat{N}(\ell))$, can be obtained as the ratio between the two transfer functions of $G^{DSB}(\ell)$ and $G^{BM}(\ell)$, such that[3]

$$\begin{aligned}\hat{\eta}(\ell) &\approx \left|\frac{1}{2} \cdot \frac{1 + \exp(j\phi_{12}^{norm}(\ell))}{1 - \exp(j\phi_{12}^{norm}(\ell))}\right|^2 \\ &= \frac{1}{4} \cdot \frac{1 + \cos(\phi_{12}^{norm}(\ell))}{1 - \cos(\phi_{12}^{norm}(\ell))}\end{aligned}. \tag{11}$$

Consequently, as shown in (11), the TNR can be estimated by only using the phase difference term. In the next subsection, it is explored how the DOA-based SNR can be estimated with this TNR estimate.

B. DOA-Based SNR Estimation

For the dual-microphone array given in (1) and (4), we assume that each target-directional signal, $T(\ell)$, or non-target signal, $N(\ell)$, can be decomposed into the target speech, $S(\ell)$, and noise, $D(\ell)$. In other words, we assume that $D(\ell)$ can be represented as $D(\ell) = D^t(\ell) + D^n(\ell)$, where $D^t(\ell)$ and $D^n(\ell)$ denote the target- and non-target-directional noise spectral components, respectively. Similar to $D(\ell)$, $S(\ell)$ can be represented as $S(\ell) = S^t(\ell) + S^n(\ell)$, where $S^t(\ell)$ and $S^n(\ell)$ also denote the target- and non-target-directional speech spectral components, respectively. Even if $S^n(\ell)$ can be affected by reverberation effects in real environments, $S^n(\ell)$ can be negligible when the target speech comes close to the microphones. Consequently, we can assume $S(\ell) = S^t(\ell)$. Therefore, two hypotheses pertaining to target speech activity, H_0 and H_1, can then be represented as

$$\begin{aligned}H_0 &: |X(\ell)| = |D^t(\ell)| + |D^n(\ell)| \\ H_1 &: |X(\ell)| = |D^t(\ell)| + |D^n(\ell)| + |S^t(\ell)|\end{aligned}. \tag{12}$$

The above equation can subsequently be extended as

$$\begin{aligned}H_0 &: \frac{|X(\ell)|}{|D^n(\ell)|} - 1 = \frac{|D^t(\ell)|}{|D^n|} \\ H_1 &: \frac{|X(\ell)|}{|D^n(\ell)|} - 1 = \frac{|D^t(\ell)|}{|D^n(\ell)|} + \frac{|S^t(\ell)|}{|D^n(\ell)|}\end{aligned} \tag{13}$$

Since $|X(\ell)|/|D^n(\ell)| - 1 = (|X(\ell)| - |D^n(\ell)|)/|D^n(\ell)| = (|S^t(\ell)| + |D^t(\ell)|)/|D^n(\ell)| = |T(\ell)|/|N(\ell)| = \eta^{1/2}(\ell)$, the two hypotheses above can be represented as

$$\begin{aligned}H_0 &: \eta^{1/2}(\ell) = \eta_D^{1/2}(\ell) \\ H_1 &: \eta^{1/2}(\ell) = \eta_D^{1/2}(\ell) + \eta_S^{1/2}(\ell)\end{aligned} \tag{14}$$

[2]In this paper, the spectral gain is applied to the first microphone signal, $X_{1,k}(\ell)$.

[3]Hereafter within this section, we remove the frequency bin index k in equations to improve the readability.

where $\eta_D^{1/2}(\ell)$ and $\eta_S^{1/2}(\ell)$ denote $|D^t(\ell)|/|D^n(\ell)|$ and $|S^t(\ell)|/|D^n(\ell)|$, respectively. Next, we can define the DOA-based SNR, $\xi^{DOA}(\ell)$, as

$$\xi^{DOA}(\ell) = \frac{E(|S^t(\ell)|^2)}{E(|D^t(\ell)|^2) + E(|D^n(\ell)|^2)} = \frac{E(\eta_S(\ell))}{E(\eta_D(\ell)) + 1} = \frac{\lambda_{\eta_S}(\ell)}{\lambda_{\eta_D}(\ell) + 1} \quad (15)$$

where $\lambda_{\eta_S}(\ell)$ and $\lambda_{\eta_D}(\ell)$ denote the expected values of $\eta_S(\ell)$ and $\eta_D(\ell)$, which are defined as $\lambda_{\eta_S}(\ell) = E(\eta_S(\ell))$ and $\lambda_{\eta_D}(\ell) = E(\eta_D(\ell))$, respectively. To obtain an estimate of $\xi^{DOA}(\ell)$, $\lambda_{\eta_S}(\ell)$ and $\lambda_{\eta_D}(\ell)$ should be estimated. First, the estimate of $\lambda_{\eta_D}(\ell)$, $\hat{\lambda}_{\eta_D}(\ell)$, can be obtained through a recursive procedure with a smoothing parameter β_η, such that

$$\hat{\lambda}_{\eta_D}(\ell) = \beta_\eta \cdot \hat{\lambda}_{\eta_D}(\ell - 1) + (1 - \beta_\eta) \cdot \hat{\eta}(\ell). \quad (16)$$

Note that (16) is performed only when H_0 is true, in which the decision is based on the statistical model-based log-likelihood ratio test (LRT) for the target speech activity decision [26] used in this paper.

On the other hand, to obtain the estimate of $\lambda_{\eta_S}(\ell)$ in (15), i.e., $\hat{\lambda}_{\eta_S}(\ell)$, a Wiener filtering procedure is employed as

$$\hat{\lambda}_{\eta_S}(\ell) = \frac{\hat{\phi}_\eta(\ell)}{\hat{\phi}_\eta(\ell) + 1} \cdot \hat{\eta}(\ell) \quad (17)$$

where $\hat{\phi}_\eta(\ell)$ is estimated using a DD approach [1], [2], such that

$$\hat{\phi}_\eta(\ell) = \alpha \cdot \frac{\hat{\lambda}_{\eta_S}(\ell - 1)}{\hat{\lambda}_{\eta_D}(\ell - 1)} + (1 - \alpha) \cdot \max\left(\frac{\hat{\eta}(\ell)}{\hat{\lambda}_{\eta_D}(\ell)} - 1, 0\right) \quad (18)$$

where α is the smoothing parameter. Consequently, we can obtain the DOA-based SNR estimate as $\hat{\xi}^{DOA}(\ell) = \hat{\lambda}_{\eta_S}(\ell)/(\hat{\lambda}_{\eta_D}(\ell) + 1)$ by using $\hat{\lambda}_{\eta_S}(\ell)$ in (17) and $\hat{\lambda}_{\eta_D}(\ell)$ in (16).

C. SNR Update by a Posteriori SNR

By using the estimate of the DOA-based SNR, $\hat{\xi}^{DOA}(\ell)$, the SNR, $\hat{\xi}(\ell)$, can also be estimated using a DD approach as

$$\hat{\xi}(\ell) = \beta \cdot \hat{\xi}^{DOA}(\ell) + (1 - \beta) \cdot \max(\hat{\gamma}(\ell) - 1, 0) \quad (19)$$

where $\beta (0 \le \beta < 1)$ is the smoothing parameter. In (19), $\hat{\gamma}(\ell)$ is the estimate of the *a posteriori* SNR, which is expressed as

$$\hat{\gamma}(\ell) = \frac{|X(\ell)|^2}{\hat{\lambda}_D(\ell)} \quad (20)$$

where $\hat{\lambda}_D(\ell)$ is the noise variance estimate, and can be estimated via a recursive procedure with a smoothing parameter β_D, such that

$$\hat{\lambda}_D(\ell) = \beta_D \cdot \hat{\lambda}_D(\ell - 1) + (1 - \beta_D) \cdot |X_1(\ell)|^2 \text{ if } H_0 \text{ is true.} \quad (21)$$

D. Target Speech Reconstruction

The proposed noise attenuating spectral gain for the speech enhancement based on the Wiener filter, $G^{SNR}(\ell)$, is represented as

$$G^{SNR}(\ell) = \frac{\hat{\xi}(\ell)}{\hat{\xi}(\ell) + 1} \quad (22)$$

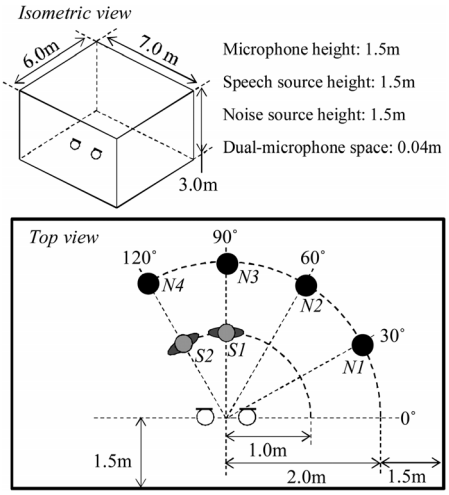

Fig. 3. The geometric configuration of a room environment simulation for obtaining dual-microphone noisy speech signals.

where $\hat{\xi}(\ell)$ is obtained in (19). Thus, the target speech spectral estimate, $\hat{S}(\ell)$, can be obtained by multiplying $G^{SNR}(\ell)$ by $X_1(\ell)$, which is transformed back into a time domain signal.

IV. Performance Evaluation

In this section, we evaluated the performance of a speech enhancement system employing the proposed DOA-based SNR estimation approach. Fig. 3 illustrates the geometric configuration for the simulation setup to obtain dual-microphone noisy speech signals in a reverberant room. In this paper, even though the proposed method was developed without considering any reverberant effects of the target speech signals, we attempted to evaluate the performance of the proposed method in reverberation environments. To this end, we assumed that a reverberant room had dimensions of 7 m × 6 m × 3 m and a dual-microphone system with 4-cm spacing was located in the room. Here, four reverberation times (RT60s) of 0, 100, 200, and 300 ms were modeled using the image source method (ISM) [27], [28]. For the ISM, early and lateral reflections were all considered by the diffused decay model in [28].

For an intensive analysis, we considered two target speech source locations and four noise source locations, as shown in Fig. 3. All the heights of sources and microphones were set to 1.5 m. In this simulation, we utilized ten target speech utterances (five males and five females) from the TIMIT database [29]. The target speech sources were located along the azimuth directions of 90° and 120° on a circle of 1 m radius, which were referred to as $S1$ and $S2$, respectively. In order to test the robustness against speech interference, four interfering speech sources from the TIMIT database [29] were positioned along the azimuth directions of 30°, 60°, 90°, and 120° on a circle at a radius of 2 m; these were referred to as *N1, N2, N3*, and *N4*.

We combined two speech-source and four noise-source locations to construct six different scenarios, as described in Table I. They were *S2-N1, S2-N2, S2-N3, S1-N3, S1-N4-N1*, and *S2-N3-N1*. Among them, *S2-N1, S2-N2*, and *S2-N3,* which were respectively referred to as *Case 1, Case 2*, and *Case 3*, were comparably easier scenarios than others, because the sources were located on the opposite side. On the other hand, *S1-N3* (referred to as *Case 4*) was the most difficult because the sources were located close to one another and they had very

TABLE I
SIMULATION SCENARIOS ACCORDING TO SPEECH AND NOISE SOURCE LOCATIONS

Scenarios	Notation	Description
Case 1	*S2-N1*	*S2* and *N1*
Case 2	*S2-N2*	*S2* and *N2*
Case 3	*S2-N3*	*S2* and *N3*
Case 4	*S1-N3*	*S1* and *N3*
Case 5	*S1-N1-N4*	*S1* and *N1* and *N4*
Case 6	*S2-N3-N1*	*S2* and *N3* and *N1*

similar DOAs. Furthermore, two-noise mixture scenarios such as *S1-N1-N4* and *S2-N3-N1* were respectively referred to as *Case 5* and *Case 6*. Consequently, to evaluate the noise-robust speech enhancement performance, 50 test noisy speech signals (10 signals at different SNRs of 0, 5, 10, 15, or 20 dB) were prepared for each scenario (*Cases 1, 2, 3, 4, 5*, and *6*) under each reverberation time of 0, 100, 200, and 300 ms.

A. Algorithm Implementation

For the simulation, it was assumed that all TDOAs τ_{12}'s were known *a priori*, i.e. $\bar{\tau}_{12}$, instead of using the TDOA estimate $\hat{\tau}_{12}$ in (3). Furthermore, each test signal was down-sampled from 16 kHz to 8 kHz, then segmented into consecutive frames by half-overlapping a cosine window with a length of 32 ms, which corresponded to 256 samples at a sampling rate of 8 kHz. In this paper, SIR, SAR, and SDR were measured to evaluate the performance of the speech enhancement algorithms in noisy environments [22].

The true clean target signal, $s_{\text{target}}(n)$, and its estimate, $\hat{s}(n)$, were related by $\hat{s}(n) = s_{\text{target}}(n) + e_{\text{interf}}(n) + e_{\text{noise}}(n) + e_{\text{artif}}(n)$, where $e_{\text{interf}}(n)$, $e_{\text{noise}}(n)$, and $e_{\text{artif}}(n)$ were the errors associated with the interference, noise, and artifacts, respectively; they were obtained through a least-square projection [22]. By using those errors, SIR, SAR, and SDR were defined as [22]

$$\text{SIR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2}, \tag{23}$$

$$\text{SAR} = 10\log_{10}\frac{||s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}||^2}{||e_{\text{artif}}||^2}, \tag{24}$$

$$\text{SDR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2} \tag{25}$$

where $\|\cdot\|$ is the norm operator.

As shown in (23)–(25), the SIR could mainly measure how well the algorithm suppressed interfering sources, while SAR measured how many artifacts remained in the separated (target) source. SDR was a global performance index, which might give a better assessment of the overall performance of the algorithms under comparison by considering both the interference and artifact components. Note that the left (or the first) microphone signal of the dual-microphone system was used as the reference for measuring SDR, SIR, and SAR in this paper.

In order to select proper values for β_η in (16), α in (18), β in (19), and β_D in (21), we measured the SDR values according to different values of β_η, α, β, and β_D for all SNRs, RT60s, and scenarios. Consequently, it was found that setting $\beta_\eta = 0.94$ in (16), $\alpha = 0.2$ in (18), $\beta = 0.99$ in (19), and $\beta_D = 0.98$ in (21) provided the maximized SDR value. Thus, we set those parameters as fixed for all the simulations.

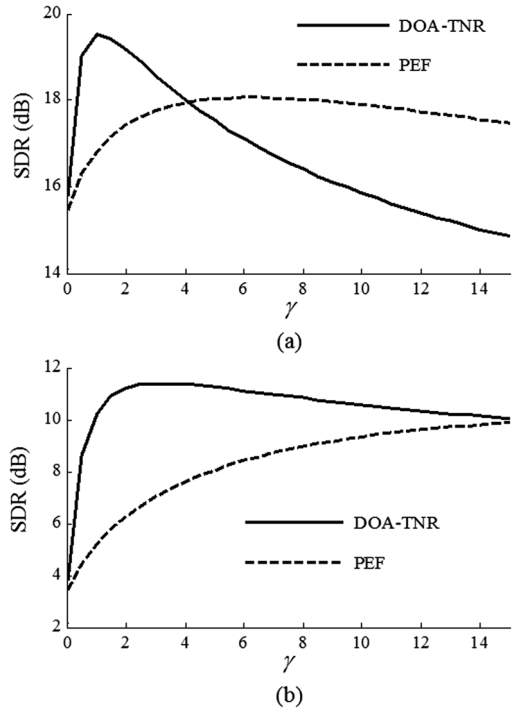

Fig. 4. Comparison of SDR values averaged over (a) high SNRs (15 and 20 dB) and (b) low SNRs (0, 5, and 10 dB) between PEF and the proposed TNR method for *Case 1* with no reverberation ($\text{RT}60 = 0$ ms), according to the different values of γ.

B. Objective Quality Test and Results

We compared the speech enhancement performance of $G_k^{TNR}(\ell)$ in (11) with that of $G_k^{PEF}(\ell)$ in (8) according to the parameter γ that controls the degree of noise attenuation. To this end, we implemented the TNR-based spectral gain $G_k^{TNR}(\ell)$ using $\hat{\eta}_k(\ell)$. Fig. 4 shows the comparison of SDR values of the signals processed by the PEF method in (8) and the proposed method in (11), where γ varied from 0.1 to 15. Note that the SDR values for *Case 1* were averaged over high SNRs (15 and 20 dB SNRs) or low SNRs (0, 5, and 10 dB SNRs). It was shown from the figure that the proposed TNR and PEF methods provided the highest SDR values when γ was about 1 and 5, respectively, under high SNR conditions. On the other hand, at low SNRs, the proposed TNR achieved the highest SDR value when γ was about 3, and higher SDR values than PEF for all γ's. This implies that the proposed TNR method can provide better speech enhancement performance than the conventional PEF by properly setting γ.

Figs. 5(a) and (b) show the spectrograms of the target-directional desired speech and its noise-contaminated version at 5 dB SNR for *Case 1*, respectively, under no reverberation condition, i.e., $RT60 = 0$ ms. In addition, Fig. 5(c) was obtained by applying the spectral gain $G_k^{TNR}(\ell)$ with $\gamma = 3$, based on the TNR estimate $\hat{\eta}_k(\ell)$ to Fig. 5(b). In addition, Fig. 5(d) was obtained by applying the proposed $G_k^{SNR}(\ell)$ in (22) to Fig. 5(b). It was shown from the figure that the proposed TNR method effectively suppressed the non-target-directional noise components. Moreover, it was shown by comparing Figs. 5(c) and 5(d)

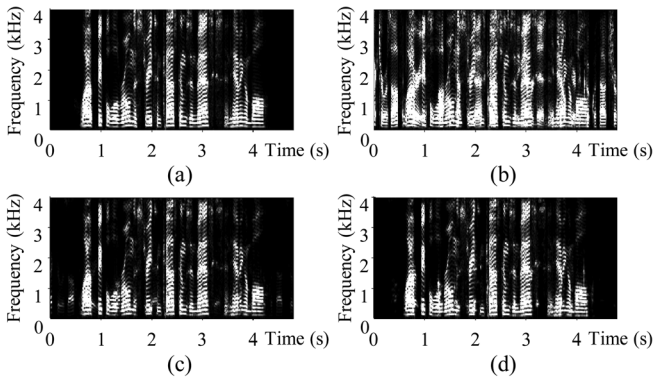

Fig. 5. Step-by-step illustrations of: spectrograms of (a) clean male speech, (b) interference speech contaminated by noisy speech at 5 dB SNR in *Case 1* (target speech at 120° and noise at 30°) under no reverberation conditions, (c) signal processed by $G_k^{TNR}(\ell)$, and (d) signal processed by $G_k^{SNR}(\ell)$.

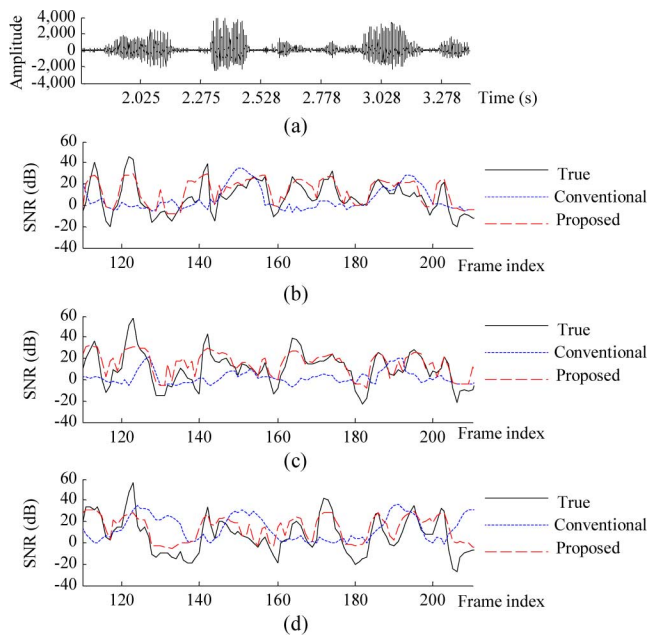

Fig. 6. Illustration of (a) the noisy speech signal at 5 dB SNR of *Case 1* with no reverberation, and the SNRs estimated by the conventional and proposed methods at (b) 1 kHz, (c) 2 kHz, and (d) 3 kHz.

that $G_k^{SNR}(\ell)$ could provide more suppressed non-target-directional noise components and more-refined target speech than $G_k^{TNR}(\ell)$.

To demonstrate the effectiveness of the proposed method further, the proposed SNR estimate, $\hat{\xi}_k(\ell)$ in (19), was compared to the conventional DD-based estimate in [3], where the true SNR was used as a reference. Fig. 6 shows the performance comparison of the SNR estimate at specific frequencies of 1, 2, and 3 kHz for the methods applied to the noisy speech signal in Fig. 5(b). It was shown from the figure that the proposed method gave an SNR estimate that was more similar to the true value than the conventional single-microphone method. To quantify the similarity, Table II compares the root mean square (RMS) errors between the true and estimated SNR values at specific frequencies from 500 Hz to 3 kHz at a step of 500 Hz. As shown in the table, the proposed method achieved much smaller RMS errors than the conventional method. These results imply that the

TABLE II
RMS ERROR BETWEEN TRUE AND ESTIMATED SNR AT DIFFERENT FREQUENCIES

Method	Frequency (kHz)					
	0.5	1	1.5	2	2.5	3
Conventional	22.66	16.06	18.55	18.84	21.98	22.27
Proposed	10.45	9.32	8.92	9.97	10.11	11.26

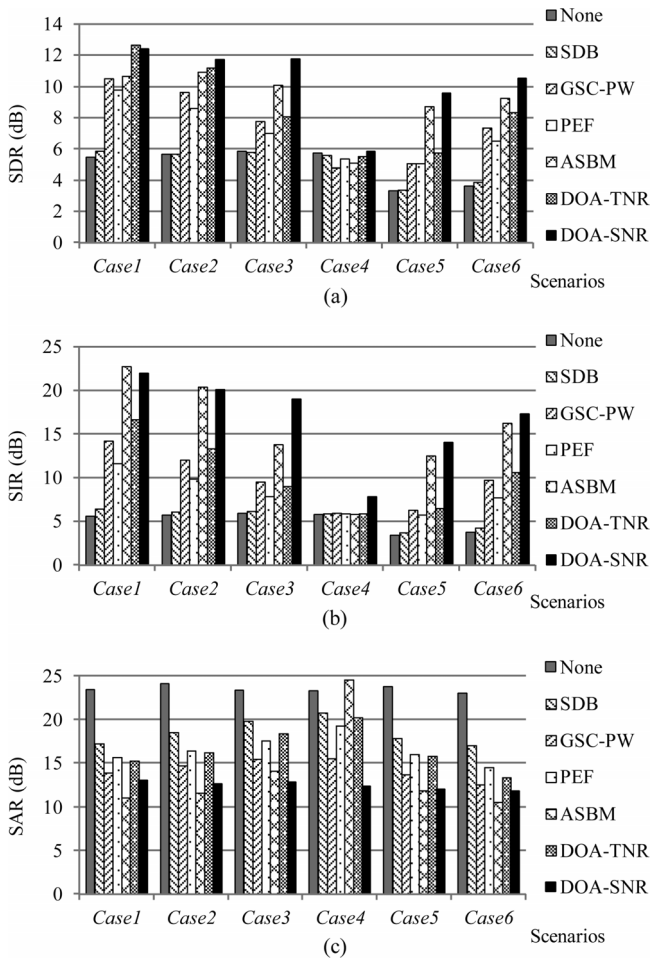

Fig. 7. Comparison of (a) SDR, (b) SIR, and (c) SAR values according to different approaches for all scenarios.

SNR estimate based on the DOA information from the dual-microphone signals can provide more reliable information for estimating the SNR.

Fig. 7 compares the SDR, SIR, and SAR values averaged over all SNRs and RT60s of different methods according to different scenarios. In this figure, we labeled no speech enhancement as 'None,' and we compared the proposed dual-microphone TNR and SNR-based methods with four different conventional dual-microphone speech enhancement methods (SDB, GSC followed by post-filter (GSC-PW), PEF, and angular spectrum-based masking (ASBM)).[4] As shown in the figure, the SAR values did not seem to be dependent on the scenario; however, the SDR and SIR values in one-noise mixture scenarios (*Cases 1, 2*, and *3*) seemed to be higher than those in two-noise mixture scenarios (*Cases 5* and *6*). Note that the SDR values were little varied according to the different dual-microphone

[4]Some speech samples can be found inhttp://hucom.gist.ac.kr/TASL-2014/sample.html.

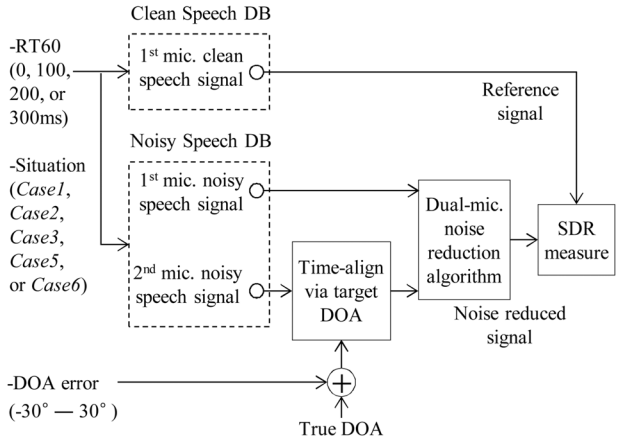

Fig. 8. Conceptual diagram for a DOA error simulation.

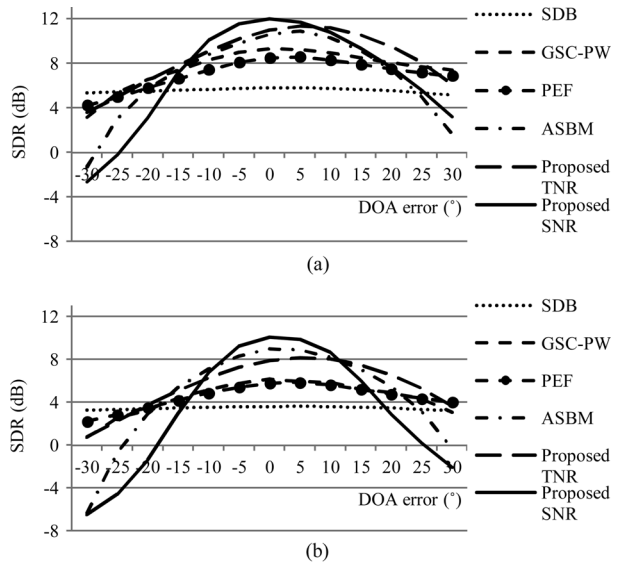

Fig. 9. Comparison of SDR averaged over all SNRs and RT60s for different speech enhancement methods under DOA errors ranging from -30° to 30°; (a) SDR values averaged over *Cases 1, 2*, and *3*, (b) SDR values averaged over *Cases 5* and *6*.

methods in *Case 4*; this was because in *Case 4* the target speech and noise were in the same direction, which made it difficult to discriminate between the target and non-target signals using DOA cues. These DOA cue ambiguities present a challenging problem for our future work. In *Cases 1* and *2*, the proposed method did not provide a higher SIR value than the conventional methods. However, the proposed method achieved the highest SDR value, which implies that a speech enhancement system employing the proposed SNR estimation method could provide a better speech enhancement performance than conventional methods, since the SDR was an assessment measure that represented the overall speech enhancement performance [22]. Note that the proposed TNR-based method provided a better SDR performance than SDB, GSC-PW, and PEF in all scenarios except *Case 4*; however, its performance did not seem to be better than that of ASBM or the proposed DOA-based SNR methods.

As mentioned earlier, we assumed that the target DOA was known *a priori*; however, in a real environment, the target DOA estimation might not be robust, thus the performance against the target DOA error was also investigated in an interference speech environment. As depicted in Fig. 8, in order to simulate the DOA error, the clean speech signal from the left microphone was fixed as the reference signal for measuring the SDR over all DOA error conditions. Fig. 9 compares the SDR values of a noisy target speech according to the target DOA error, ranging from -30° to 30°, which denotes the relative difference to the true DOA value. It should be noted that the SDR values were averaged over all SNRs. In the figure, the SDB provided nearly constant SDR values, regardless of the DOA error under both one-noise mixture situations (*Cases 1, 2*, and *3*) and two-noise mixture situations (*Cases 5* and *6*). This was because the SDB using a dual-microphone broadside array had almost the same directional sensitivity over all the target directions of errors. On the other hand, in terms of SDR values, the GSC-PW, PEF, ASBM, and proposed TNR/SNR estimation methods provided the speech enhancement performance that was dependent on the DOA error. In particular, it was shown that the best performance of the proposed SNR estimation method had a DOA error of around 0°, denoting an accurate target DOA estimate. Note that although the proposed TNR and SNR methods provided the highest SDR values at DOA errors ranging from -15° to 15°, they achieved lower SDR values than the conventional methods beyond this range. Nevertheless, the proposed method was deemed more effective for speech enhancement than the convention methods, because the conventional GCC or SRP approach for the DOA estimation exhibited a sufficiently reliable localization performance in terms of allowable DOA errors for the range of -15° to 15°.

Table III compares the relative improvements of the SDR values of the noise-reduced signals over their corresponding original noisy speech signals. The improved SDR were averaged over the scenarios of *Cases 1, 2, 3, 5*, and *6*, according to the different approaches under various SNR and reverberation conditions. It was shown from the table that the improved SDR values of all the methods decreased when the reverberation was severed (i.e., RT60 was increased). Among all compared methods, SDB provided the lowest improvement in SDR values for all RT60s, but the proposed TNR method provided more highly improved SDR values than SDB, PEF, and GSC-PW, despite its simplicity. Moreover, ASBM seemed to be the most effective for speech enhancement in terms of SDR values among the conventional methods. However, the proposed DOA-based SNR estimation method yielded the highest improvement in SDR values among all the considered speech enhancement approaches for all SNRs when RT60 is below 300 ms. Note here that the proposed DOA-based SNR method outperformed the proposed DOA-based TNR method. This was because the proposed DOA-based SNR method well utilized spatial cues, even under reverberation conditions, while the proposed TNR method was sensitive to the reverberation conditions. Consequently, it could be concluded here that a speech enhancement system employing the proposed DOA-based SNR estimation method outperformed those using the conventional and TNR-based estimation methods regardless of reverberation conditions.

Table IV subsequently compares the improved SDR values of the speech signals processed by the different speech enhancement methods under three different noise conditions such as factory, vacuum cleaner, and white noise [30]. It was shown

TABLE III
COMPARISON OF IMPROVED SDR VALUES ACCORDING TO DIFFERENT METHODS UNDER VARIOUS SNR AND REVERBERATION CONDITIONS

SNR (dB)	RT60 = 0 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
20	0.24	4.15	3.42	5.20	5.08	5.86
15	0.27	5.15	3.89	6.80	5.99	7.85
10	0.28	5.85	4.28	8.85	6.97	9.47
5	0.28	6.42	4.54	10.72	7.86	11.23
0	0.29	6.65	4.59	12.41	8.52	12.65
Avg.	0.27	5.64	4.14	8.80	6.88	9.41

SNR (dB)	RT60 = 100 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
20	0.21	2.54	2.61	3.66	4.12	5.17
15	0.27	4.05	3.26	5.58	5.02	6.88
10	0.29	4.88	3.61	7.52	5.72	8.20
5	0.30	5.33	3.80	8.95	6.27	9.88
0	0.29	5.51	3.82	10.15	6.58	11.53
Avg.	0.27	4.46	3.42	7.17	5.54	8.33

SNR (dB)	RT60 = 200 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
20	-0.09	-0.10	0.94	-0.67	1.94	1.91
15	0.16	2.20	2.10	2.55	3.35	4.76
10	0.24	3.41	2.64	5.14	4.20	6.34
5	0.26	3.83	2.91	7.08	4.75	8.16
0	0.27	4.19	2.95	8.14	4.98	9.49
Avg.	0.17	2.70	2.31	4.45	3.84	6.13

SNR (dB)	RT60 = 300 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
20	-1.39	-4.29	-2.20	-6.76	-2.15	-3.40
15	-0.35	-0.75	0.11	-2.55	0.55	0.48
10	0.04	1.24	1.27	1.03	2.13	3.05
5	0.17	2.38	1.76	3.68	2.94	5.20
0	0.22	2.76	1.90	5.29	3.26	6.55
Avg.	-0.26	0.27	0.57	0.14	1.35	2.38

TABLE IV
COMPARISON OF IMPROVED SDR VALUES AVERAGED OVER LOW SNRS OF 0, 5, AND 10 dB (OR HIGH SNRS OF 15 AND 20 dB) ACCORDING TO DIFFERENT METHODS UNDER VARIOUS NOISE TYPES

Noise Type	RT60 = 0 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
Factory	0.21 (0.01)	2.03 (0.50)	1.20 (0.48)	7.52 (3.53)	6.27 (3.49)	8.29 (4.18)
Vacuum Cleaner	0.53 (0.00)	2.83 (-0.46)	2.21 (0.47)	5.74 (1.84)	5.58 (2.42)	6.40 (2.62)
White	2.24 (2.21)	7.72 (4.91)	6.65 (5.01)	9.79 (5.87)	8.59 (6.16)	11.07 (6.95)
Avg.	0.99 (0.74)	4.19 (1.65)	3.35 (1.99)	7.68 (3.75)	6.81 (4.03)	8.59 (4.59)

Noise Type	RT60 = 100 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
Factory	0.23 (-0.21)	1.43 (-1.94)	1.02 (-0.56)	6.22 (-0.23)	5.66 (2.23)	7.78 (3.23)
Vacuum Cleaner	0.44 (-0.73)	1.21 (-5.11)	1.50 (-2.05)	3.96 (-3.86)	4.76 (0.08)	5.98 (0.89)
White	2.07 (1.99)	7.28 (3.91)	6.32 (4.46)	8.64 (4.35)	7.66 (5.31)	10.17 (6.30)
Avg.	0.92 (0.35)	3.31 (-1.05)	2.95 (0.62)	6.27 (0.09)	6.03 (2.54)	7.97 (3.47)

Noise Type	RT60 = 200 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
Factory	0.15 (-1.26)	0.56 (-4.96)	0.62 (-2.81)	3.53 (-6.12)	4.31 (-1.32)	6.23 (-0.50)
Vacuum Cleaner	0.04 (-3.38)	-0.79 (-9.35)	0.15 (-6.30)	-0.14 (-11.15)	2.68 (-5.42)	3.80 (-4.46)
White	1.93 (1.56)	6.24 (1.65)	5.74 (3.04)	7.19 (0.63)	6.39 (3.40)	8.77 (4.34)
Avg.	0.70 (-1.03)	2.00 (-4.22)	2.17 (-2.03)	3.52 (-5.54)	4.46 (-1.12)	6.27 (-0.21)

Noise Type	RT60 = 300 ms					
	SDB	GSC -PW	PEF	ASBM	Proposed	
					TNR	SNR
Factory	-0.30 (-4.38)	-1.36 (-9.17)	-0.58 (-6.80)	-1.62 (-13.00)	1.48 (-7.02)	2.41 (-6.40)
Vacuum Cleaner	-1.32 (-8.52)	-3.84 (-14.25)	-2.34 (-11.62)	-6.39 (-18.46)	-1.52 (-12.18)	-0.89 (-11.18)
White	1.77 (0.25)	5.08 (-1.63)	4.62 (0.04)	3.77 (-5.39)	4.35 (-0.57)	5.86 (-0.15)
Avg.	0.05 (-4.22)	-0.04 (-8.35)	0.57 (-6.13)	-1.41 (-12.28)	1.43 (-6.59)	2.46 (-5.91)

from the table that the proposed DOA-based SNR estimation method yielded the most highly improved SDR values of all the considered methods under each noise condition, which also outperformed the proposed TNR-based method under all SNR and RT60 conditions.

In addition to the SDR, the PESQ [23] score was measured in order to compare the quality of the target speech reconstructed by the different speech enhancement methods. Note here that the simulation settings of the SDR measurements were maintained in this evaluation. Table V compares the averaged PESQ scores averaged over all RT60s according to all scenarios. As shown in the table, ASBM provided higher PESQ scores than the other conventional methods under all noise conditions. However, the proposed method achieved higher PESQ scores than ASBM. Consequently, it could be concluded that the speech enhancement system employing the proposed

TABLE V
COMPARISON OF PESQ SCORES AVERAGED OVER LOW SNRS OF 0, 5, AND 10 dB (OR HIGH SNRS OF 15 AND 20 dB) ACCORDING TO DIFFERENT METHODS UNDER VARIOUS NOISE TYPES

Noise Type	None	SDB	GSC -PW	PEF	ASBM	Proposed	
						TNR	SNR
Interference Speech	2.241 (2.976)	2.313 (3.035)	2.474 (3.257)	2.439 (3.229)	2.594 (3.272)	2.509 (3.264)	2.677 (3.442)
Factory	2.417 (3.307)	2.511 (3.393)	2.794 (3.559)	2.658 (3.458)	2.954 (3.748)	2.774 (3.450)	3.125 (3.803)
Vacuum Cleaner	2.661 (3.586)	2.772 (3.676)	3.142 (3.841)	2.904 (3.769)	3.238 (3.947)	2.982 (3.592)	3.386 (3.964)
White	1.749 (2.407)	1.842 (2.572)	2.369 (3.116)	2.299 (3.057)	2.325 (3.057)	2.140 (3.150)	2.392 (3.277)
Avg.	2.267 (1.581)	2.360 (3.169)	2.695 (3.443)	2.575 (3.378)	2.778 (3.506)	2.601 (3.364)	2.895 (3.622)

DOA-based SNR method outperformed the conventional and proposed TNR-based methods, regardless of the noise type.

V. Conclusion

In this paper, we proposed a new SNR estimation method based on the phase differences for dual-microphone speech enhancement. For this task, the DOA-based TNR was first estimated by using the phase difference between the dual-microphone signals, which was then utilized to estimate the DOA-based SNR under target speech presence uncertainty. Next, the estimated DOA-based SNR was incorporated into a Wiener filter in order to obtain a spectral-gain attenuator. To investigate the effect of the proposed method on the performance of a speech enhancement system, test noisy speech signals were artificially prepared to simulate interference speech environments at different SNRs of 0, 5, 10, 15, and 20 dB under reverberant conditions (RT60s) of 0, 100, 200, and 300 ms. By using the test noisy speech signals, we evaluated the speech enhancement performance of the proposed method in terms of SDR, SIR, and SAR. Consequently, it was confirmed from SDR, SIR, and SAR comparisons that the Wiener filter employing the proposed DOA-based SNR estimation method outperformed the method that used a single-microphone Wiener filter, in addition to the dual-microphone methods such as SDB, GSC-PW, PEF, and ASBM under reverberant noise conditions. It was also shown from PESQ comparison that the Wiener filter using the proposed method provided a better speech quality than the other speech enhancement systems under comparison.

Appendix I

List of Abbreviations and their Meanings

SNR	Signal-to-Noise Ratio
TNR	Target-to-Non-target directional signal Ratio
DOA	Direction-Of-Arrival
STSA	Short-Term Spectral Amplitude
DD	Decision-Directed
T-F	Time-Frequency
SDP	Spatial Directivity Pattern
ICA	Independent Component Analysis
BSS	Blind Source Separation
PEF	Phase Error-based Filter
SDR	Source-to-Distortion Ratio
SIR	Source-to-Interference Ratio
SAR	Source-to-Artifacts Ratio
PESQ	Perceptual Evaluation of Speech Quality
SDB	Super-Directive Beamformer
GSC	Generalized Sidelobe Canceller
GSC-PW	GSC followed by a Post Wiener filter
ASBM	Angular Spectrum-Based Masking
TDOA	Time Difference-Of-Arrival
GCC	Generalized Cross-Correlation
SRP	Steered Response Power
PHAT	Phase Transform
DSB	Delay-and-Sum Beamformer
BM	Blocking Matrix
LRT	Log-likelihood Ratio Test
ISM	Image Source Method

Appendix II
TNR Estimation Based on Phase Difference

In this Appendix, we derive (11) from (9) and (10). First, $\hat{T}_k(\ell)/\hat{N}_k(\ell)$ is obtained by using the relationships of (9) and (10) as

$$\begin{aligned}\frac{\hat{T}_k(\ell)}{\hat{N}_k(\ell)} &\approx \frac{G_{DSB,k}(\ell)}{G_{BM,k}(\ell)} = \frac{1}{2}\cdot\frac{1+\exp(j\phi_{12,k}(\ell))}{1-\exp(j\phi_{12,k}(\ell))} \\ &= \frac{1}{2}\cdot\left(\frac{1+\exp(j\phi_{12,k}(\ell))}{1-\exp(j\phi_{12,k}(\ell))}\right) \\ &\quad\cdot\left(\frac{1+\exp(-j\phi_{12,k}(\ell))}{1+\exp(-j\phi_{12,k}(\ell))}\right) \\ &= \frac{1}{2}\cdot\frac{1+\cos(\phi_{12,k}(\ell))}{-j\sin(\phi_{12,k}(\ell))} \end{aligned} \tag{26}$$

since $(1-\exp(j\phi_{12,k}(\ell)))(1+\exp(-j\phi_{12,k}(\ell)) = -2j\sin(\phi_{12,k}(\ell))$ and $(1+\exp(j\phi_{12,k}(\ell)))(1+\exp(-j\phi_{12,k}(\ell)) = 2+2\cos(\phi_{12,k}(\ell))$. Then, the proposed TNR estimate $\hat{\eta}_k(\ell) = |\hat{T}_k(\ell)/\hat{N}_k(\ell)|^2$ can be finally obtained from (26) as

$$\begin{aligned}\hat{\eta}_k(\ell) &= \left|\frac{\hat{T}_k(\ell)}{\hat{N}_k(\ell)}\right|^2 = \frac{1}{4}\cdot\frac{(1+\cos(\phi_{12,k}(\ell)))^2}{\sin^2(\phi_{12,k}(\ell))} \\ &= \frac{1}{4}\cdot\frac{(1+\cos(\phi_{12,k}(\ell)))^2}{1-\cos^2(\phi_{12,k}(\ell))} \\ &= \frac{1}{4}\cdot\frac{1+\cos(\phi_{12,k}(\ell))}{1-\cos(\phi_{12,k}(\ell))}. \end{aligned} \tag{27}$$

References

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. New York, NY, USA: Springer-Verlag, 2005.

[2] S. J. Lee, B. O. Kang, H. Jung, Y. Lee, and H. S. Kim, "Statistical model-based noise reduction approach for car interior applications to speech recognition," *ETRI J.*, vol. 32, no. 5, pp. 801–809, Oct. 2010.

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

[4] P. M. Brandstein and D. Ward, *Microphone Arrays*. New York, NY, USA: Springer-Verlag, 2001.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation of arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[6] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[8] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. Technol.*, vol. 22, no. 2, pp. 149–157, May 2001.

[9] B. Loesch and B. Yang, "Online blind source separation based on time-frequency sparseness," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 117–120.

[10] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-and-add method applied to source separation using a time frequency mask," in *Proc. ICASSP*, Philadelphia, PA, USA, Mar. 2005, vol. 3, pp. 81–84.

[11] J. H. Park, J. S. Yoon, and H. K. Kim, "HMM-based mask estimation for a speech recognition front-end using computational auditory scene analysis," *IEICE Trans. Inf. Syst.*, vol. E91-D, no. 9, pp. 2360–2364, Sep. 2008.

[12] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, Beijing, China, Oct. 2000, vol. 1, pp. 373–376.
[13] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501, Nov. 2006.
[14] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
[15] S. Dolco and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 617–631, Feb. 2007.
[16] S. Han, J. Hong, S. Jeong, and M. Hahn, "Probabilistic adaption mode control algorithm for GSC-based noise reduction," *IEICE Trans. Fundam. Electron., Commun., Comput. Sci.*, vol. E93-A, no. 3, pp. 627–630, Mar. 2010.
[17] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, May 1998.
[18] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech*, J. Benesty, Y. Huang, and M. Sondhi, Eds. New York, NY, USA: Springer, Nov. 2007, pp. 1–34.
[19] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Syst., Man, Cybern.-Part B: Cybern.*, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.
[20] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundam. Electron., Commun., Comput. Sci.*, vol. E87-A, no. 8, pp. 1941–1948, Aug. 2004.
[21] S. M. Kim and H. K. Kim, "Target-to-non-target directional ratio estimation based on dual-microphone phase differences for target-directional speech enhancement," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 3254–3258.
[22] R. Gribonval, C. Févotte, and E. Vincent, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[23] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech coders," ITU-T Recommendation P.862, Feb. 2001.
[24] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, Aug. 2012.
[25] S. M. Kim and H. K. Kim, "Hybrid probabilistic adaptation mode controller for generalized sidelobe canceller-based target-directional speech enhancement," in *Proc. ICASSP*, Prague, Czech Republic, May 2011, pp. 2532–2535.
[26] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[27] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, Jul. 2008.
[28] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1429–1439, Aug. 2010.
[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
[30] A. Varga, H. J. M. Steenneken, M. Tomilson, and D. Jones, "The NOISEX–92 study on the effect of additive noise on automatic speech recognition," *Documentation on the NOISEX–92 CD–ROMs*, Jun. 1992.

Seon Man Kim is a researcher at the Institute of Sound and Vibration Research (ISVR), University of Southampton. He received a B.S. degree in mechanical design engineering from Chonbuk National University, Korea, in 2003 and an M.S. degree in mechatronics engineering from Gwangju Institute of Science and Technology (GIST), Korea, in 2005. From 2005 to 2007, he was a researcher at the Samsung Reciprocating Compressor RND Group, Gwangju, Korea. He received a Ph.D. degree in information and communications engineering from GIST in 2013. His current research interests involve audio and speech signal processing using multi-microphone techniques, non-negative matrix factorization, and auditory image models.

Hong Kook Kim (M'99–SM'01) received a B.S. degree in control and instrumentation engineering from Seoul National University, Korea, in 1988. He then received both M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea in 1990 and 1994, respectively. He was a senior researcher at the Samsung Advanced Institute of Technology (SAIT), Korea, from 1990 to 1998. During 1998–2003, he was a senior technical staff member with the Voice Enabled Services Research Lab at AT&T Labs-Research, Florham Park, New Jersey. Since August 2003, he has been with the School of Information and Communications at Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, as a Professor. From July 2014, he is a visiting professor at the City University of New York, New York, USA. He is currently an IEEE Senior Member and an affiliate member of IEEE Speech and Language Technical Committee. Since 2012, he has served as an editorial committee member and area editor of Digital Signal Processing. His current research interests include large vocabulary speech recognition, audio coding and speech/audio source separation, and embedded algorithms and solutions for speech and audio processing for handheld devices.