IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 20, 2019, accepted December 30, 2019, date of publication January 3, 2020, date of current version January 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963768

Sound Source Localization Based on GCC-PHAT With Diffuseness Mask in Noisy and Reverberant Environments

RAN LEE[1], MIN-SEOK KANG[1], BO-HYUN KIM[1], KANG-HO PARK[2], SUNG Q LEE[2], AND HYUNG-MIN PARK[1], (Senior Member, IEEE)

[1]Department of Electronic Engineering, Sogang University, Seoul 04107, South Korea
[2]Intelligent Sensors Research Section, Electronics Telecommunications Research Institute, Daejeon 34129, South Korea

Corresponding author: Hyung-Min Park (hpark@sogang.ac.kr)

This work was supported in part by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korean Government Ministry of Science and ICT (MSIT) (Development of Human Enhancement Technology for Auditory and Muscle Support) under Grant 2017-0-00050, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant NRF-2017R1A2B4009964.

ABSTRACT Although sound source localization is a desirable technique in many communication systems and intelligence applications, the distortion caused by diffuse noise or reverberation makes the time delay estimation (TDE) between signals acquired by a pair of microphones a complicated and challenging problem. In this paper, we describe a method that can efficiently achieve sound source localization in noisy and reverberant environments. This method is based on the generalized cross-correlation (GCC) function with phase transform (PHAT) weights (GCC-PHAT) to achieve robustness against reverberation. In addition, to estimate the time delay robust to diffuse components and to further improve the robustness of the GCC-PHAT against reverberation, time-frequency(t-f) components of observations directly emitted by a point source are chosen by ''inversed'' diffuseness. The diffuseness that can be estimated from the coherent-to-diffuse power ratio (CDR) based on spatial coherence between two microphones represents the contribution of diffuse components on a scale of zero to one with direct sounds from a source modeled to be fully coherent. In particular, the ''inversed'' diffuseness is binarized with a very rigorous threshold to select highly reliable components for accurate TDE even in noisy and reverberant environments. Experimental results for both simulated and real-recorded data consistently demonstrated the robustness of the presented method against diffuse noise and reverberation.

INDEX TERMS Diffuseness mask, GCC-PHAT, reverberation, sound source localization.

I. INTRODUCTION

Sound source localization is a desirable technique in various communication systems and intelligence applications, including speech enhancement in noisy and reverberant environments by forming a beam toward the target source [1], [2]. Typically, it can be achieved by exploiting the difference among the signals obtained by spatially separated microphones. If many microphones are available, the direction of arrival (DOA) of a sound source can be accurately estimated by numerous approaches combining various information based on the microphone signals (e.g., [3]–[6]), and learning-based approaches such as [7] may provide even better localization performance. They may result in successful localization of multiple sound sources in various real-world situations (e.g., [8]–[10]).

Given a pair of microphones possibly due to a limited resource or a compact size, the DOA of a source is related to the difference between the times taken by the sound from a source to reach the microphones. A lot of methods ranging from exploitation of interaural cues inspired by binaural processing to introduction of various techniques for mathematical analysis, even with learning from data, have been proposed (e.g., [11]–[18]). Among those methods, a class of cross-correlation-based methods is the most intuitive and popular approach, where the frame-wise relative time

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

delay can be estimated by determining the maximum of the cross-correlation function of signals acquired by the pair of microphones [18]. The localization would be easy if the two signals were merely delayed and scaled versions of each other. In real-world situations, however, the acquired signals are frequently prone to contamination by ambient noise. Furthermore, the signals may contain multiple attenuated and delayed replicas of the source signal caused by reverberation. The distortion caused by noise or reverberation makes the time delay estimation (TDE) a complicated and challenging problem [13], [19]. Over the past few decades, researchers have tackled this problem by exploiting different aspects of the observed signals and developed numerous algorithms [2].

Typically, the cross-correlation function is more affected by low-frequency components where most of the natural sound (including speech) energy is concentrated. Therefore, the cross-correlation function may frequently have a flatter peak that can disturb accurate TDE. To overcome this vulnerability, Knapp and Carter [18] introduced the generalized cross-correlation (GCC) function, which results in a cross-correlation function with a frequency weighting, and several weightings have been presented for accurate TDE in considered situations [1]. One of the most commonly used weightings is phase transform (PHAT) weighting, which makes the TDE robust against reverberation. Since the steered-response power (SRP) objective function can be expressed as a sum of the GCCs for different microphone pairs with many microphones available, the GCC function with PHAT weights (GCC-PHAT) is essentially the same as the SRP function with PHAT weights (SRP-PHAT) given a pair of microphones [20], [21]. However, this weighting is known to be sensitive to additive noise, and the sensitiveness can be overcome by the maximum-likelihood (ML) weighting. In order to combine the advantages of both the weightings, Wang and Chu proposed the MLR weighting that is robust to both noise and reverberation [1], [22].

In order to achieve further robustness, masks have been applied to remove time-frequency(t-f) components of observed signals that were harmful for source localization by containing noise or reverberation significantly (e.g., [14], [23]–[27]). Using the non-stationarity of speech, the transition noise masks were estimated [28], or masks were obtained by signal-to-noise ratios (SNRs) computed with stationary noise estimates [29]. Wilson and Darrell exploited cues corresponding to sudden increases in audio energy by finding a mapping from reverberated signal spectrograms to localization precision as a soft mask, which exhibited behavior consistent with the precedence effect [30], [31] from psychoacoustic studies [32]. Since deep-learning-based t-f masking has dramatically improved the performance of monaural speech separation and enhancement, deep neural networks (DNNs) estimated masks to identify t-f components harmful for accurate TDE (e.g., [33]–[36]). Whereas masks estimated by [28] and [29] assumed stationary noise, learning-based methods such as [32]–[36] might not obtain successful masks with insufficient learning data or for an environment unmatched with training data.

A typical acoustic impulse response characterizing the propagation of sound from a source to a microphone consists of the direct impulse and the early and late reflections. Considering that an observed signal can be modelled by the corresponding source signal convolved with an acoustic impulse response, signal components through direct paths provide the DOAs corresponding to sound locations, contrary to the reflected sound. Since a direct path is the shortest path from a source to a microphone, a sound arrives earlier at a microphone through the direct path than others. Therefore, the auditory onset, that is, the start of a discrete event in an acoustic signal, is robust against reverberation because a microphone mainly obtains sound through a direct path from a source during the onset [14], [37], [38]. If two microphone signals contain only direct sounds emitted from a source, they are delayed and scaled versions of each other, and fully coherent, whereas other components caused by diffuse noise or reverberation may be assumed to be diffuse. Since the spatial coherence between two microphones, as an efficient measure to distinguish coherent components from diffuse components, was used for signal enhancement [39], heuristic methods for noise reduction and dereverberation have been proposed [40], [41]. Especially, the spatial coherence (as known as interaural coherence from binaural auditory studies) was used for source localization [42], [43].

In this paper, we describe a method that can efficiently achieve sound source localization in noisy and reverberant environments. This method is based on the GCC-PHAT to achieve robustness against reverberation. In addition, for the TDE robust to diffuse noise or reverberation, t-f components of observed signals directly emitted by a point source are selected by ''inversed'' diffuseness [44]. Instead of using the spatial coherence estimates directly, a diffuseness estimator that represents the contribution of diffuse components on a scale of zero to one is computed from the coherent-to-diffuse power ratio (CDR) based on the spatial coherence estimates, formerly used for dereverberation [45]. Especially, the ''inversed'' diffuseness is binarized with a very rigorous threshold, prior to masking the t-f components of observed signals, in order to use highly reliable components for accurate TDE even in noisy and reverberant environments. Experimental results confirmed the robustness of the proposed method with signals affected by diffuse noise or reverberation.

II. REVIEW OF A DIFFUSENESS ESTIMATOR BASED ON A DOA-INDEPENDENT CDR

In performing the localization of sound sources, direct sound from a source is helpful, but diffuse components interfere with the localization. In the short-time Fourier transform (STFT) domain, let us consider the i-th microphone signal $X_i(m, k)$ composed of a direct signal component $S_i(m, k)$ and an interfering component $N_i(m, k)$ as

$$X_i(m, k) = S_i(m, k) + N_i(m, k), \tag{1}$$

where m and k index time frame and frequency bin, respectively. The short-time complex spatial coherence function of signals acquired by a pair of omnidirectional microphones is estimated by

$$\hat{\Gamma}_X(m,k) = \frac{\hat{\Phi}_{X_1X_2}(m,k)}{\sqrt{\hat{\Phi}_{X_1X_1}(m,k)\hat{\Phi}_{X_2X_2}(m,k)}}, \tag{2}$$

where $\hat{\Phi}_{X_iX_j}(m,k)$ denotes a short-time auto- or cross-power spectral density estimate between $X_i(m,k)$ and $X_j(m,k)$ typically obtained by recursive averaging: [45]

$$\hat{\Phi}_{X_iX_j}(m,k) = \lambda\hat{\Phi}_{X_iX_j}(m-1,k) + (1-\lambda)X_i(m,k)X_j^*(m,k), \tag{3}$$

where λ is a constant between 0 and 1, and $(\cdot)^*$ denotes the complex conjugate.

If two microphone signals contain only direct sounds emitted from a source, they are delayed and scaled versions of each other, and fully coherent, as mentioned in Section I. An estimate of the ratio between the coherent and diffuse components, termed as the CDR [45] can be derived from the spatial coherence functions by [45]

$$\widehat{\mathrm{CDR}}(m,k) = \frac{\hat{\Gamma}_N(m,k) - \hat{\Gamma}_X(m,k)}{\hat{\Gamma}_X(m,k) - \hat{\Gamma}_S(m,k)}, \tag{4}$$

where $\hat{\Gamma}_S(m,k)$ and $\hat{\Gamma}_N(m,k)$ are estimates of the short-time complex spatial coherence functions of direct signal components $S_i(m,k)$ and interfering components $N_i(m,k)$,[1] respectively. Assuming a diffuse or spherically isotropic sound field for $N_i(m,k)$, an ideal coherence function of $N_i(m,k)$ is real-valued time-invariant and given by [46]

$$\tilde{\Gamma}_N(k) = \frac{\sin(2\pi f_k d/c)}{2\pi f_k d/c}, \tag{5}$$

where f_k is the analog frequency at the k-th frequency bin. d and c denote the distance between the two microphones and the speed of sound, respectively.

Without the need for explicit DOA estimation to obtain $\hat{\Gamma}_S(m,k)$, the knowledge that the direct sound from a source is fully coherent derives the DOA-independent CDR estimator given by [45], [47]

$$\widetilde{\mathrm{CDR}}(m,k) = \frac{\tilde{\Gamma}_N(k)\Re\{\hat{\Gamma}_X(m,k)\} - |\hat{\Gamma}_X(m,k)|^2 - \bar{\Gamma}(m,k)}{|\hat{\Gamma}_X(m,k)|^2 - 1}, \tag{6}$$

where

$$\bar{\Gamma}(m,k) = \sqrt{\tilde{\Gamma}_N^2\Re\{\hat{\Gamma}_X\}^2 - \tilde{\Gamma}_N^2|\hat{\Gamma}_X|^2 + \tilde{\Gamma}_N^2 - 2\tilde{\Gamma}_N\Re\{\hat{\Gamma}_X\} + |\hat{\Gamma}_X|^2}. \tag{7}$$

[1] In the context of the CDR, the direct signal and interfering components usually mean coherent and diffuse components, respectively. Background noise including late reverberation belongs to the diffuse components.

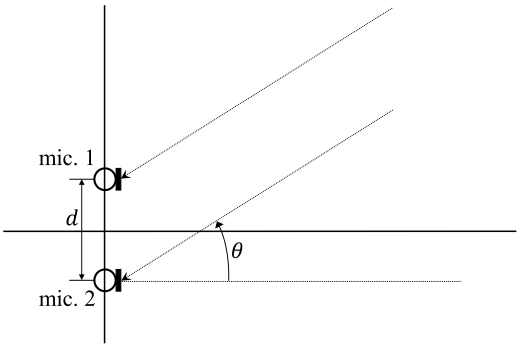

FIGURE 1. DOA estimation of a sound source using TDE for a pair of microphones.

The frame and frequency bin arguments are omitted for brevity. $\Re$ and $|\cdot|$ denote the real part and absolute values of a complex number, respectively.

Then, the contribution of diffuse components in the microphone signals on a scale of zero to one can be obtained by the diffuseness estimator defined as [45]

$$\tilde{D}(m,k) = \frac{1}{\widetilde{\mathrm{CDR}}(m,k) + 1}. \tag{8}$$

III. PROPOSED SOUND SOURCE LOCALIZATION METHOD

In Fig. 1, the DOA of a sound source for a pair of microphones spaced by d, assuming that the distance from the source to the center of microphones is much larger than d, can be given by [1]

$$\theta \approx \arcsin\left(\frac{\tau_{\mathrm{TDE}} \cdot c}{d}\right), \tag{9}$$

where τ_{TDE} denotes the difference between the times taken by the sound from the source to reach the two microphones. Therefore, the source direction can be determined by TDE between signals acquired by the two microphones. The most intuitive and popular approach to obtain τ_{TDE} is to find the time lag corresponding to the maximum of the GCC function of the two microphone signals: [1], [16], [18]

$$\tau_{\mathrm{TDE}} = \frac{1}{f_s} \cdot \arg\max_t R_{12}(t), \tag{10}$$

where f_s and t are the sampling frequency and time sample index, respectively. The GCC function $R_{12}(t)$ of the two signals can be efficiently computed by the inverse DFT of the cross-power spectral density function:

$$R_{12}(t) = \mathcal{F}^{-1}\{\Psi \odot \mathbf{X}_1 \odot \mathbf{X}_2^*\}, \tag{11}$$

where the vector Ψ represents the frequency weighting, and $\mathbf{X}_i$ denotes a vector whose k-th element is $X_i(m,k)$ with the omitted frame index. $\odot$ and * represent the Hadamard product and the element-wise complex conjugate operation on the vector, respectively.

Among several weightings, the PHAT weight at frame m and frequency bin k is given by

$$\Psi(m,k) = \frac{1}{|X_1(m,k)X_2^*(m,k)|}. \tag{12}$$

The weighting normalizes the magnitudes to provide equal weights for all frequency bins to form sharp peaks. Therefore, the GCC-PHAT may provide the TDE robust against reverberation, but it is known to be sensitive to ambient noise as the normalization emphasizes frequency components with small powers.

As mentioned above, direct sound from a source is helpful for sound source localization, but components caused by diffuse noise or reverberation interferes with the localization. The direct sound and other components caused by diffuse noise or reverberation may be assumed to be fully coherent and diffuse, respectively. Because the diffuseness $\tilde{D}(m,k)$ represents the contribution of diffuse components in microphone signals on the scale of zero to one, it can be used to select t-f components of the observed signals directly emitted by the source that are useful to achieve a TDE robust to diffuse components and to further improve the robustness of the GCC-PHAT against reverberation. Especially, in order to choose highly reliable components containing direct sound dominantly for accurate TDE in noisy and reverberant environments, the mask is binarized with a very rigorous threshold as

$$M(m,k) = \begin{cases} 0 & \text{if } \tilde{D}(m,k) > \delta_{\tilde{D}}, \\ 1 & \text{otherwise}, \end{cases} \tag{13}$$

where $\delta_{\tilde{D}}$ denotes the threshold.[2] The GCC-PHAT on the masked observations estimates the DOA θ by (9) after the mask is applied in the STFT domain by replacing (11) with

$$R_{12}(t) = \mathcal{F}^{-1}\{\Psi \odot \mathbf{M} \odot \mathbf{X}_1 \odot \mathbf{X}_2^*\}, \tag{14}$$

where $\mathbf{M}$ denotes a vector whose k-th element is $M(m,k)$ with the omitted frame index.

Furthermore, the time resolution of cross-correlation-based methods including the GCC-PHAT is limited by the sampling interval. To obtain an accurate DOA estimation at a low sampling frequency for two microphones in close proximity, the time resolution at a subsampling level should be investigated. It can be performed by interpolating the GCC function and determining its maximum. The interpolation of the GCC function can be efficiently accomplished by padding zeros to the weighted cross-power spectral density function in (11).

In summary, the overall procedure of the proposed sound source localization method is as follows:

- Begin
  - Step 1 Make a new frame of input data;
  - Step 2 Transform the frame into the frequency domain;
  - Step 3 Compute the DOA-independent CDR estimate by using (6) with (2), (5), and (7);
  - Step 4 Compute the diffuseness estimate by using (8);
  - Step 5 Compute the GCC function by using (14) with (12) and (13);
  - Step 6 Estimate a localization angle by using (9) with (10);
  - Step 7 Go to Step 1.
- End

[2]The threshold is optimized empirically in pilot experiments.

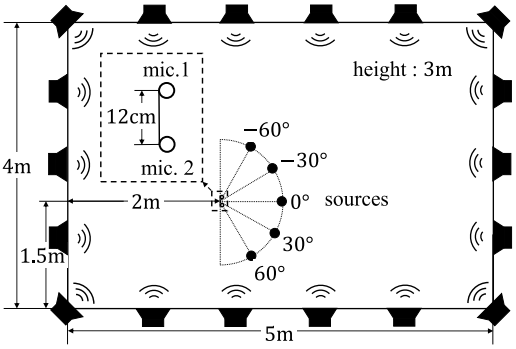

FIGURE 2. Source and microphone positions for experiments on simulated data.

IV. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed sound source localization method in noisy and reverberant environments, we simulated signals observed at two 12-cm-apart microphones[3] from a source in a 5 m × 4 m × 3 m rectangular room. An observed signal was obtained by convolving the source signal with an impulse response that simulates the acoustics from the source to a microphone [48] and by adding diffuse noise. Fig. 2 describes configurations to generate observations. A sound source was placed in one of five different angles at a distance of 1 m from the center of the two microphones. The common height of the source and microphones was 1 m. The reflection coefficients were chosen to provide the reverberation times RT_{60}s of every 0.2 s from 0.2 s to 0.6 s. The source signal was composed of concatenated sentences uttered by a speaker from the TIMIT database [49]. The diffuse noise was simulated by summing up signals convolved with the generated impulse responses from virtual speakers playing randomly selected sections of babble noise from NOISEX-92 [50] that were placed at 1-m-spaced locations along walls with random heights. In the convolving process, the source signal was upsampled to 1024 kHz, convolved with acoustic filters generated at a sampling rate of 1024 kHz, and downsampled back to 16 kHz because the original sampling of 16 kHz was too low to simulate signal delay at the two microphones standing nearby. The length of the signals was fixed as 15 s. Noise components for both the microphones were scaled by the same factor to give a designated SNR at ''mic. 1''. As the distance between the two microphones was 12 cm, the maximum time delay between the microphones was 5.565 samples at the sampling

[3]For a 16-kHz sampling rate, an alias-free microphone spacing is about 2.1 cm that is too close to acquire sufficiently different signals from each other. In a practical situation, acquired signals usually contain noise. If the microphone spacing is too close, the noise can be an important factor in preventing accurate source localization.

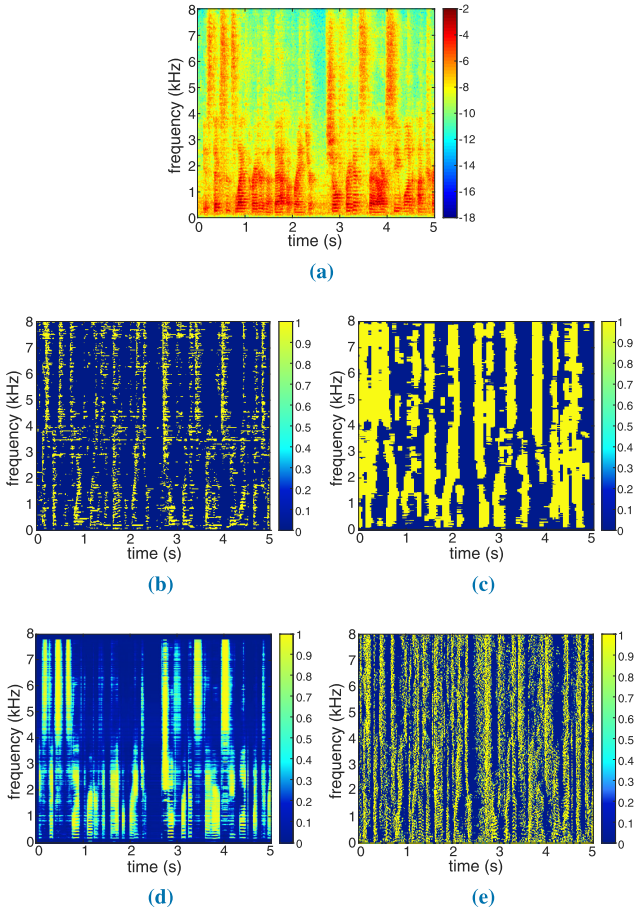

FIGURE 3. Spectrograms of (a) an input signal with an SNR of 20 dB and an RT_{60} of 0.4 s, (b) its binary diffuseness mask estimated when $\lambda = 0.8$ and $\delta_{\bar{D}} = 0.2$, (c) its transition noise mask by [28], (d) its DNN-based masks by [36], and (e) its coherence mask by [42].

rate of 16 kHz. In order to obtain a sufficient time resolution, we padded 32,256 zeros to 512 STFT coefficients such that the maximum and minimum angle resolutions were 0.16° and 4.3°,[4] respectively.

Fig. 3 displays the spectrograms of an input signal with an SNR of 20 dB and an RT_{60} of 0.4 s and its binary diffuseness mask estimated when $\lambda = 0.8$ and $\delta_{\bar{D}} = 0.2$. To compare the diffuseness mask of the proposed method with t-f masks with a range of values between 0 and 1 estimated by other methods, Fig. 3 also shows a transition noise mask by [28], a DNN-based mask by [36], and a coherence mask by [42]. The parameters for source localization in [36] were optimized or trained by 600 utterances uttered by five male and five female speakers (disjoint from speakers for evaluation) at five different angles same as in Fig. 2 for four input SNRs and three RT_{60}s. Although the masks were similar to the mask estimated by the proposed method in that they distinguished t-f components dominated by a target speech signal, it is noteworthy that the mask obtained by the proposed method successfully and rigorously selected t-f components corresponding to auditory onsets that were helpful for robust source localization.

The median value of localization angle estimates obtained for all frames in an utterance was used as the estimated angle for the utterance. Instead of the mean value, the median value was used to remove outliers of the angle estimates at frames where the target utterance was not dominant. Then, the localization performance was evaluated in terms of the mean absolute error (MAE) of estimated angles defined as

$$E_\theta = \frac{\sum_{u=1}^{N_u} |\theta_u^{\text{est}} - \theta_u^{\text{true}}|}{N_u}, \tag{15}$$

where θ_u^{est} and θ_u^{true} denote the estimated and true localization angles for the u-th utterance. N_u is the number of utterances, and we used 30 concatenated sentences uttered by three male and three female speakers at five different angles in this experiment. In addition, we also considered the MAE of estimated time delays to determine source directions because the conversion by (9) is not linear. The MAE of estimated time delays is defined as

$$E_{\tau_{\text{TDE}}} = \frac{\sum_{u=1}^{N_u} |\tau_{\text{TDE}_u}^{\text{est}} - \tau_{\text{TDE}_u}^{\text{true}}|}{N_u}, \tag{16}$$

where $\tau_{\text{TDE}_u}^{\text{est}}$ and $\tau_{\text{TDE}_u}^{\text{true}}$ denote the median value of time delay estimates (with the zero-padding) obtained at all frames and the true time delay for the u-th utterance, respectively.

The MAEs of the proposed localization method are displayed in Fig. 4 for four different input SNRs and three different RT_{60}s. For comparison, we showed the results of the localization methods based on the conventional GCC function with uniform (CC), ML (GCC-ML), MLR (GCC-MLR), and PHAT (GCC-PHAT) weights. We further presented the results of the proposed method using a continuous diffuseness mask without binarization (Prop. w/o bin.). Regardless of the used methods, the MAEs increased in general as RT_{60} increased or the input SNR decreased. This is because an increase in RT_{60} or a decrease in the input SNR increased noisy or reverberant components interfering with the localization. Under all the tested cases, the proposed method considerably reduced the MAEs and consistently showed comparable or lower MAEs than the others. In particular, the proposed method provided comparable or better performance than the method using a diffuseness mask without binarization by choosing only t-f components highly reliable for accurate DOA estimation. In order to compare the proposed method with other methods based on t-f mask estimation, Fig. 5 displays the MAEs of the GCC-PHAT method with no applied masks (No masks), the transition noise masks by [28] (Masks by [28]), localization precision masks by [32] (Masks by [32]), DNN-based masks by [36] (Masks by [36]), coherence masks by [42] (Masks by [42]) and masks by the proposed method. The parameters for source localization in [32] were also optimized or trained by the same data as in [36] as mentioned above. Similar to Fig. 4, an increase in RT_{60} or a decrease in the input SNR generally increased the MAEs. The proposed

[4]Without the zero-padding, the maximum and minimum angle resolutions were about 10° and 35° for each sample delay, respectively.

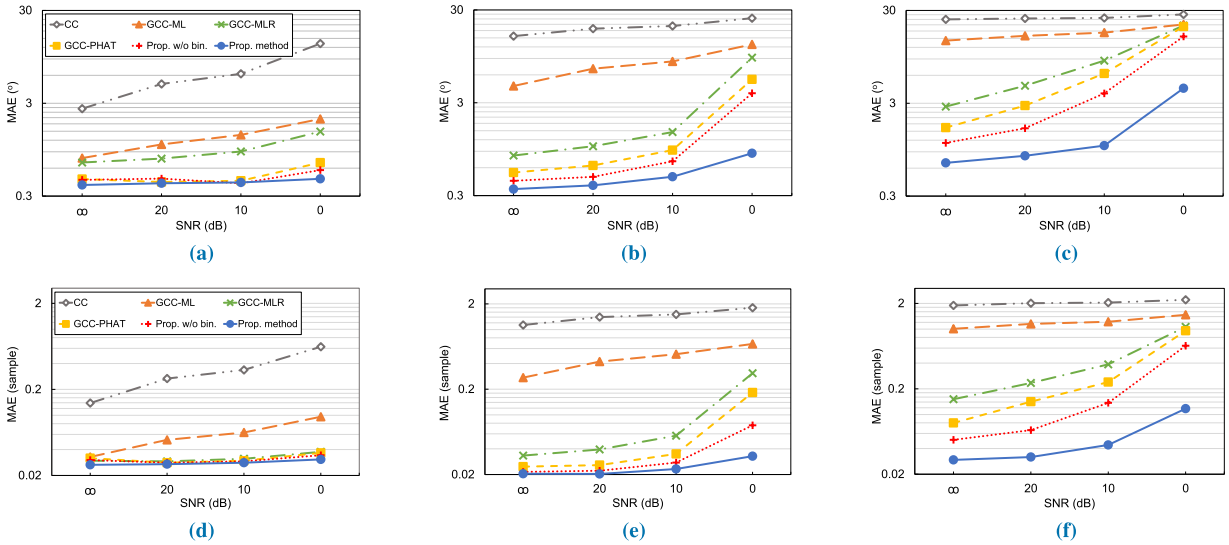

FIGURE 4. MAEs of estimated angles (in the upper row) and time delays (in the lower row) for the localization methods based on the conventional GCC function on 30 signals uttered by three male and three female speakers at five different azimuthal angles with RT_{60}s of 0.2 s (in the left column), 0.4 s (in the center column), and 0.6 s (in the right column). The infinite SNR means that no noise was added. MAEs of estimated angles on data at RT_{60}s of (a) 0.2 s, (b) 0.4 s, and (c) 0.6 s, and MAEs of estimated time delays on data at RT_{60}s of (d) 0.2 s, (e) 0.4 s, and (f) 0.6 s.

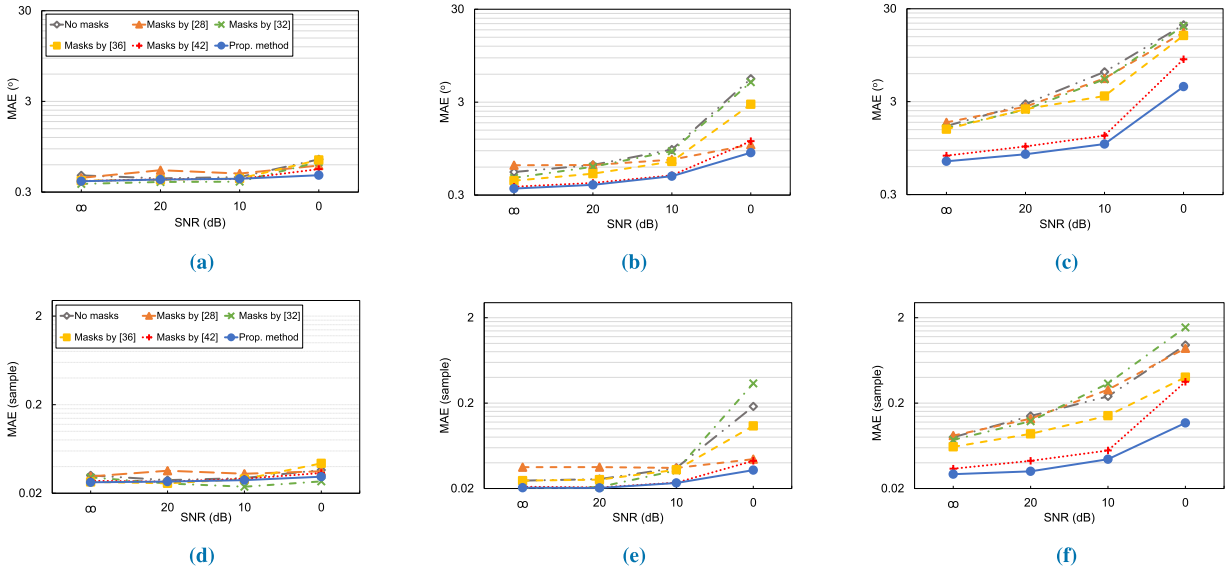

FIGURE 5. MAEs of estimated angles (in the upper row) and time delays (in the lower row) for the GCC-PHAT method with no masks, masks by [28], [32], [36], [42], and the proposed method on 30 signals uttered by three male and three female speakers at five different azimuthal angles with RT_{60}s of 0.2 s (in the left column), 0.4 s (in the center column), and 0.6 s (in the right column). The infinite SNR means that no noise was added. MAEs of estimated angles on data at RT_{60}s of (a) 0.2 s, (b) 0.4 s, and (c) 0.6 s, and MAEs of estimated time delays on data at RT_{60}s of (d) 0.2 s, (e) 0.4 s, and (f) 0.6 s.

method provided comparable or better performance than the others including the learning-based methods of [32] and [36]. Although the learning-based methods might achieve better performance with more various training data, it is worth noting that the proposed method does not require a learning process with training data in advance. Moreover, the results demonstrated that the proposed method successfully selected t-f components that were helpful for sound source localization and the localization using the selected t-f components achieved robustness against noise and reverberation.

To evaluate the probability of successful localization estimates, Figs. 6 and 7 show the rates of localizations averaged over the same 30 concatenated sentences as above. The localization rate is defined as the ratio of the number of frames providing time delay estimates corresponding to successful localization, to the number of all frames, where errors of the time delay estimates (with the zero-padding) were less than or equal to three samples. Regardless of the used methods, the rates of localizations decreased as RT_{60} increased or the input SNR decreased because localization was disturbed by increased noisy or reverberant components.

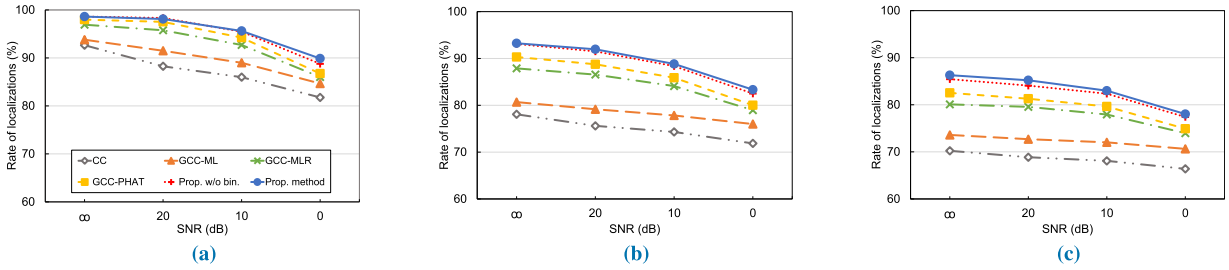

FIGURE 6. Rates of localizations for the localization methods based on the conventional GCC function averaged over 30 signals uttered by three male and three female speakers at five different azimuthal angles with RT_{60}s of (a) 0.2 s, (b) 0.4 s, and (c) 0.6 s. The infinite SNR means that no noise was added.

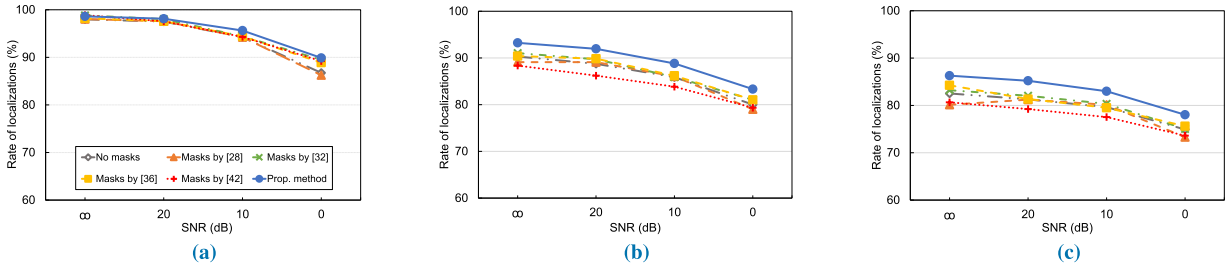

FIGURE 7. Rates of localizations for the GCC-PHAT method with no masks, masks by [28], [32], [36], [42], and the proposed method averaged over 30 signals uttered by three male and three female speakers at five different azimuthal angles with RT_{60}s of (a) 0.2 s, (b) 0.4 s, and (c) 0.6 s. The infinite SNR means that no noise was added.

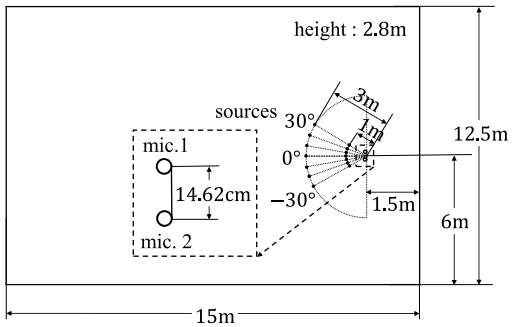

FIGURE 8. Source and microphone positions for experiments on real-recorded data.

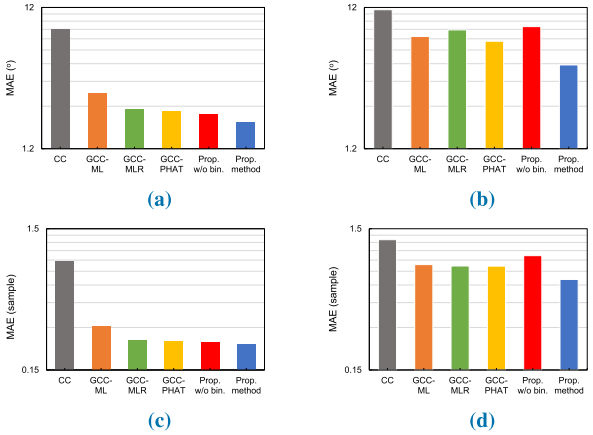

FIGURE 9. MAEs of estimated angles (in the upper row) and time delays (in the lower row) for the localization methods based on the conventional GCC function on 28 real-recorded signals uttered by two male and two female speakers at distances of 1 m (in the left column) and 3 m (in the right column) from the center of the two microphones at seven different azimuthal angles in a hall. MAEs of estimated angles on data at distances of (a) 1 m and (b) 3 m, and MAEs of estimated time delays on data at distances of (c) 1 m and (d) 3 m.

In particular, the proposed method showed comparable or better performance than the others, which demonstrated that the proposed method provided frame-wise localization estimates with a higher probability.

To conduct experiments for real-recorded data, we used four utterances uttered by two male and two female speakers from the TIMIT database [49]. As shown in Fig. 8, 2-s-long data (in the beginning part of each utterance) captured with two microphones at a sampling frequency of 16 kHz in a hall were considered for performing localization. A sound source placed at a distance of 1 m or 3 m from the center of the two microphones at an angle among azimuthal angles of every 10° from −30° and 30°. The heights of the source and microphones were 1.4 m and 1.55 m, respectively. Figs. 9 and 10 show the MAEs on 28 signals uttered by two male and two female speakers at seven different angles for the proposed and other localization methods. In this experiment, the parameters for [32] were optimized by 84 signals uttered by three male and three female speakers (disjoint from speakers for evaluation) at seven angles same as in Fig. 8 for two distances, and the parameters used in the previous experiment for [36] were finely tuned by the same data. Regardless of the used methods, the MAEs increased as the distance between the source and microphones increased because direct sound from the source that was helpful for localization became relatively diminished. Similar to the experiments on simu-

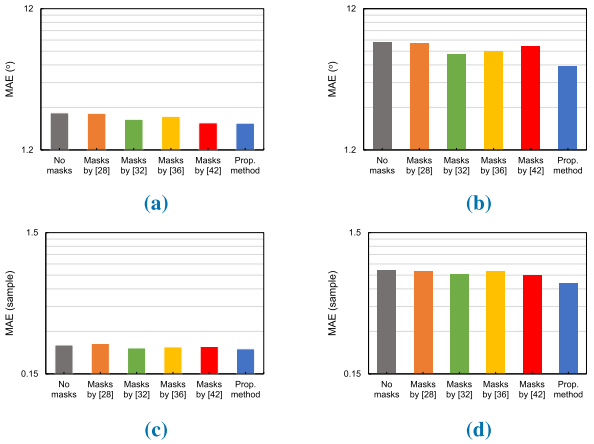

FIGURE 10. MAEs of estimated angles (in the upper row) and time delays (in the lower row) for the GCC-PHAT method with no masks, masks by [28], [32], [36], [42], and the proposed method on 28 real-recorded signals uttered by two male and two female speakers at distances of 1 m (in the left column) and 3 m (in the right column) from the center of the two microphones at seven different azimuthal angles in a hall. MAEs of estimated angles on data at distances of (a) 1 m and (b) 3 m, and MAEs of estimated time delays on data at distances of (c) 1 m and (d) 3 m.

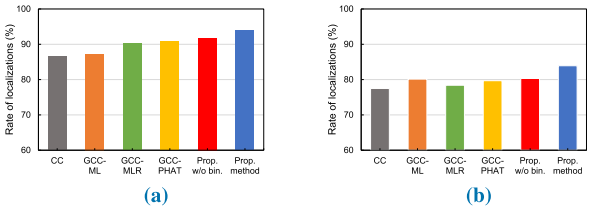

FIGURE 11. Rates of localizations for the localization methods based on the conventional GCC function averaged over 28 real-recorded signals uttered by two male and two female speakers at distances of (a) 1 m and (b) 3 m from the center of the two microphones at seven different azimuthal angles in a hall.

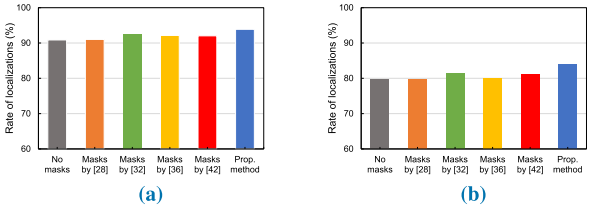

FIGURE 12. Rates of localizations for the GCC-PHAT method with no masks, masks by [28], [32], [36], [42], and the proposed method averaged over 28 real-recorded signals uttered by two male and two female speakers at distances of (a) 1 m and (b) 3 m from the center of the two microphones at seven different azimuthal angles in a hall.

lated data, Figs. 11 and 12 show averaged rates of localizations. Although the rates of localizations decreased as the distance increased, the proposed method provided frame-wise localization estimates with a higher probability than the others. Above all, the proposed method consistently provided comparable or more accurate DOA estimates than the others, which indicated that the proposed method accomplished successful sound source localization even for the real-recorded data in a hall.

V. CONCLUSION

In this paper, we presented a robust sound source localization method based on GCC-PHAT in noisy and reverberant environments. To estimate the DOA robust against diffuse noise and reverberation, t-f components of observations directly emitted by a source were selected by the diffuseness that was computed from the CDR. In particular, the "inversed" diffuseness was binarized with a very rigorous threshold to choose highly reliable components for accurate DOA estimation even in noisy and reverberant environments. The experimental results for both simulated and real-recorded data consistently demonstrated the robustness of the presented method against diffuse noise and reverberation.

REFERENCES

[1] I. J. Tashev, *Sound Capture Processing: Practical Approaches*. Hoboken, NJ, USA: Wiley, 2009.
[2] R. R. Fay, *Sound Source Localization*, vol. 25. New York, NY, USA: Springer, 2006.
[3] A. Bertolino, P. Gandhi, A. Joasil, C. Obi, K. Chandra, and C. Thompson, "Distributed sensing for acoustic source localization in indoor reverberant environments," *J. Acoust. Soc. Amer.*, vol. 145, no. 3, p. 1733, Mar. 2019.
[4] L. Birnie, T. D. Abhayapala, H. Chen, and P. N. Samarasinghe, "Sound source localization in a reverberant room using harmonic based music," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 651–655.
[5] H. He, X. Wang, Y. Zhou, and T. Yang, "A steered response power approach with trade-off prewhitening for acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 143, no. 2, pp. 1003–1007, Feb. 2018.
[6] T. Padois, "Acoustic source localization based on the generalized cross-correlation and the generalized mean with few microphones," *J. Acoust. Soc. Amer.*, vol. 143, no. 5, pp. EL393–EL398, May 2018.
[7] P. Pertila and M. Parviainen, "Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 436–440.
[8] B. Yang, H. Liu, C. Pang, and X. Li, "Multiple sound source counting and localization based on TF–wise spatial spectrum clustering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1241–1255, Aug. 2019.
[9] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 770–774.
[10] H. Liu, B. Yang, and C. Pang, "Multiple sound source localization based on TDOA clustering and multi-path matching pursuit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 3241–3245.
[11] B. R. Hammond and P. J. B. Jackson, "Robust full-sphere binaural sound source localization using interaural and spectral cues," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 421–425.
[12] E. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raumel, and S. Argentieri, "Binaural localization of multiple sound sources by non–negative tensor factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1072–1082, Jun. 2018.
[13] C. Pang, H. Liu, J. Zhang, and X. Li, "Binaural sound localization based on reverberation weighting and generalized parametric mapping," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1618–1632, Aug. 2017.
[14] S.-Y. Lee and H.-M. Park, "Multiple reverberant sound localization based on rigorous zero-crossing-based ITD selection," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 671–674, Jul. 2010.
[15] J.-W. Cho and H.-M. Park, "Imposition of sparse priors in adaptive time delay estimation for speaker localization in reverberant environments," *IEEE Signal Process. Lett.*, vol. 16, no. 3, pp. 180–183, Mar. 2009.
[16] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, pp. 1–19, Dec. 2006.

[17] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.
[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
[19] J. Zhang and H. Liu, "Robust acoustic localization via time–delay compensation and interaural matching filter," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4771–4783, Sep. 2015.
[20] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Dept. Divis. Eng., Brown Univ., Providence, RI, USA, 2000.
[21] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP–PHAT functional for robust real–time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.
[22] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Nov. 2002, pp. 187–190.
[23] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auto. Syst.*, vol. 55, no. 3, pp. 216–228, Mar. 2007.
[24] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Sep. 2004, pp. 133–136.
[25] H.-G. Kang, M. Graczyk, and J. Skoglund, "On pre-filtering strategies for the GCC–PHAT algorithm," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.
[26] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 4, pp. 806–815, Apr. 2013.
[27] H. Liu and J. Zhang, "A binaural sound source localization model based on time-delay compensation and interaural coherence," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1424–1428.
[28] F. Grondin and F. Michaud, "Noise mask for TDOA sound source localization of speech on mobile robots in noisy environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 4530–4535.
[29] F. Grondin and F. Michaud, "Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 6149–6154.
[30] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1633–1654, 1999.
[31] P. M. Zurek, "The precedence effect," in *Directional Hearing*. New York, NY, USA: Springer, 1987, pp. 85–105.
[32] K. Wilson and T. Darrell, "Learning a precedence effect–like weighting function for the generalized cross–correlation framework," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2156–2164, Nov. 2006.
[33] P. Pertila and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 6125–6129.
[34] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.
[35] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2017, pp. 136–140.
[36] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time–frequency masking and deep neural networks," in *Proc. Interspeech*, Aug. 2018, pp. 322–326.
[37] S. Lin, "Reverberation–robust localization of speakers using distinct speech onsets and multichannel cross correlations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2098–2111, Nov. 2018.
[38] J. Huang, N. Ohnishi, and N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, Aug. 1997.
[39] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal–processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, Oct. 1977.
[40] R. La Bouquin-Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 484–487, Sep. 1997.
[41] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histograms," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2767–2777, May 2013.
[42] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.
[43] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2136–2147, Apr. 2008.
[44] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," *J. Acoust. Soc. Amer.*, vol. 131, no. 3, pp. 2141–2151, Mar. 2012.
[45] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.
[46] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.
[47] A. Schwarz and W. Kellermann, "Unbiased coherent-to-diffuse ratio estimation for dereverberation," in *Proc. 14th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 6–10.
[48] E. Habets. (Sep. 2010). *Room Impulse Response (RIR) Generator*. [Online]. Available: http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator
[49] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep. NISTIR 4930, 1993.
[50] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

RAN LEE received the B.S. and M.S. degrees in electronic engineering from Sogang University, Seoul, South Korea, in 2016 and 2018, respectively. Her current research interests include robust speech recognition and natural language understanding.

MIN-SEOK KANG received the B.S. degree in computer science and engineering from Sogang University, Seoul, South Korea, in 2019, where he is currently pursuing the master's degree in electronic engineering. His current research interests include speech processing and deep learning.

BO-HYUN KIM received the B.S. degree in mathematics from Ewha University, Seoul, South Korea, in 2017, and the M.S. degree in electronic engineering from Sogang University, Seoul, in 2019. Her current research interests include deep-learning-based speech recognition and computer vision.

KANG-HO PARK received the B.S., M.S., and Ph.D. degrees in physics from Seoul National University, Seoul, South Korea, in 1987, 1989, and 1994, respectively. Since 1994, he has been working as a Principle Researcher at the Electronics Telecommunications Research Institute (ETRI), Daejeon, South Korea. His current research interests include the development and commercialization of sound field security and safety sensors such as the intrusion and fire detection sensor using sound-field variation method, and the sound source localization and the sound beamforming technology using wearable microphone array module, and the development of the smalls sized piezoelectric vibration actuator for human skin sensory devices. He received the commendation for his contribution to semiconductor technology from the Minister of Ministry of Trade, Industry, Commerce, and Energy, in 2014. He is a member of the American Physical Society and the Acoustical Society of America.

SUNG Q LEE received the B.S., M.S., and Ph.D. degrees in mechanical engineering from the Korea Advanced Institute of Science and Technology, in 1994, 1996, and 2001, respectively. Since 2001, he has worked as a Principal Researcher with the Electronics and Telecommunications Research Institute (ETRI). He has authored more than 80 articles, and more than 50 inventions. As the Director of the Nano Convergence Sensor Research Team, he was awarded ETRI Outstanding Researcher in 2011, and received the Prime Minister' Award in 2018. His main areas of interest are MEMS microphone, sound field security sensor, sound source localization, piezoelectric devices, and implantable neuro engineering device brain and brain computer interface.

HYUNG-MIN PARK (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1997, 1999, and 2003, respectively. From 2003 to early 2005, he held a postdoctoral position at the Department of Biosystems, KAIST. From 2005 to early 2007, he was with the Language Technologies Institute, Carnegie Mellon University. In 2007, he joined the Department of Electronic Engineering, Sogang University, Seoul, South Korea, where he is currently a Professor. His main research interests include multichannel speech processing and robust speech recognition.