





Article

Improved Speech Spatial Covariance Matrix Estimation for Online Multi-Microphone Speech Enhancement

Minseung Kim , Sein Cheong , Hyungchan Song  and Jong Won Shin * 

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Buk-gu, Gwangju 61005, Republic of Korea

* Correspondence: jwshin@gist.ac.kr

Abstract: Online multi-microphone speech enhancement aims to extract target speech from multiple noisy inputs by exploiting the spatial information as well as the spectro-temporal characteristics with low latency. Acoustic parameters such as the acoustic transfer function and speech and noise spatial covariance matrices (SCMs) should be estimated in a causal manner to enable the online estimation of the clean speech spectra. In this paper, we propose an improved estimator for the speech SCM, which can be parameterized with the speech power spectral density (PSD) and relative transfer function (RTF). Specifically, we adopt the temporal cepstrum smoothing (TCS) scheme to estimate the speech PSD, which is conventionally estimated with temporal smoothing. Furthermore, we propose a novel RTF estimator based on a time difference of arrival (TDoA) estimate obtained by the cross-correlation method. Furthermore, we propose refining the initial estimate of speech SCM by utilizing the estimates for the clean speech spectrum and clean speech power spectrum. The proposed approach showed superior performance in terms of the perceptual evaluation of speech quality (PESQ) scores, extended short-time objective intelligibility (eSTOI), and scale-invariant signal-to-distortion ratio (SISDR) in our experiments on the CHiME-4 database.

Keywords: multi-microphone speech enhancement; speech spatial covariance matrix estimation; temporal cepstrum smoothing; speech PSD estimation; RTF estimation



Citation: Kim, M.; Cheong, S.; Song, H.; Shin, J.W. Improved Speech Spatial Covariance Matrix Estimation for Online Multi-Microphone Speech Enhancement. *Sensors* **2023**, *23*, 111. <https://doi.org/10.3390/s23010111>

Academic Editor: Alberto Bernardini

Received: 10 October 2022

Revised: 10 December 2022

Accepted: 19 December 2022

Published: 22 December 2022



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Speech enhancement is essential to ensure the satisfactory perceptual quality and intelligibility of speech signals in many speech applications, such as hearing aids and speech communication with mobile phones and hands-free systems [1–43]. Currently, devices with multiple microphones are popular, which has enabled multi-microphone speech enhancement, exploiting spatial information as well as spectro-temporal characteristics of the input signals [6–48]. One of the most popular approaches to multi-microphone speech enhancement may be spatial filtering in the time–frequency domain, which aims to extract a target speech signal from multiple microphone signals contaminated by background noise and reverberation by suppressing sounds from directions other than the target direction [6–11].

There have been various types of spatial filters with different optimization criteria [6–10]. Among them, the minimum mean square error (MMSE) criterion for speech spectra estimation led to the multi-channel Wiener filter (MWF), which has shown decent performance [10,12,21,22]. It has been shown that the MWF solution can be decomposed into the concatenation of the minimum-variance distortionless-response (MVDR) beamformer and the single-channel postfilter [11,12]. Spatial filters often require the estimation of acoustic parameters such as the relative transfer function (RTF) between the microphones and the speech and noise spatial covariance matrices (SCMs), which should be estimated from the noisy observations.

For applications such as speech communication and hearing aids, time delays are crucial, and thus an online algorithm is required for multi-microphone speech enhancement. The work in [25] extended the single-channel minima controlled recursive averaging

(MCRA) framework [49,50] for noise estimation in the multi-channel case by introducing the multi-channel speech presence probability (SPP) [24]. In [26], a coherence-to-diffusion ratio (CDR) based *a priori* SPP estimator under the expectation and maximization (EM) framework was proposed to improve the robustness in nonstationary noise scenarios. In [25,26], the speech SCM was estimated with the maximum likelihood (ML) approach, while the multi-channel decision-directed (DD) estimator was proposed in [29]. In [27], the recursive EM (REM) algorithm, which performs an iterative estimation for the latent variables and model parameters in the current frame, was exploited by defining the exponentially weighted log-likelihood of the data sequence. The speech SCM was decomposed into the speech power spectral density (PSD) and RTF under the rank-1 approximation, and these components were estimated by an ML approach using the EM algorithm in [27].

In this paper, we propose an improved speech SCM estimation for online multi-microphone speech enhancement. First, we adopt the temporal cepstrum smoothing (TCS) approach [51] to estimate the speech PSD, which has not yet been tried in multi-channel cases. Furthermore, we propose an RTF estimator based on time difference of arrival (TDoA) estimation using the cross-correlation method. Finally, we propose refining the acoustic parameters by exploiting the clean speech spectrum and clean speech power spectrum estimated in the first pass. The experimental results show that the proposed speech enhancement framework exhibited improved performance in terms of the perceptual evaluation of speech quality (PESQ) scores, extended short-time objective intelligibility (eSTOI), and scale-invariant signal to distortion ratio (SISDR) for the CHiME-4 database. Additionally, we performed an ablation study to understand how each sub-module contributed to the performance improvement.

The remainder of this paper is organized as follows. Section 2 briefly introduces the previous work on multi-microphone speech enhancement depending on various classes of approaches and then summarizes the main contributions of our proposal. Section 3 reviews the previous MMSE multi-channel speech enhancement approach and explains the conventional speech and noise SCM estimation. Section 4 presents the proposed speech SCM estimation based on the novel speech PSD and RTF estimators. Section 5 outlines the experimental results that demonstrate the superiority of the proposed method compared with the baseline in terms of speech quality and intelligibility. Finally, a conclusion is provided in Section 6.

2. Previous Work and Contributions

Recently, many approaches to multi-microphone speech enhancement have been proposed. In [33], the estimation of the speech PSD reduces to seek a unitary matrix and the square roots of PSDs based on the factorization of the speech SCM. The RTF estimate was recursively updated based on these estimates. They also proposed a desmoothing of the generalized eigenvalues to maintain the non-stationarities of estimated PSDs. Furthermore, these parameter estimates were then exploited for a Kalman filter-based speech separation algorithm [35]. In the context of sound field analysis, ref. [34] proposed a masking scheme under the non-negative tensor factorization model and [36] exploited the sparse representation in a spherical harmonic domain. The work in [37] proposed a multi-channel non-negative factorization algorithm in the ray space transform domain.

Deep-learning-based approaches have also been proposed, which can be categorized into several types. One is the combination of deep learning with conventional beamforming methods, in which the deep neural networks (DNNs) are employed to implement beamforming [38,39]. In [38], the complex spectral mapping approach was proposed to estimate the speech and noise SCMs. In contrast, ref. [39] reformulated the MVDR beamformer as a factorized form associated with two complex components and estimated them using a DNN, instead of estimating the parameters of the MVDR beamformer. The other approach is neural beamforming, in which a DNN directly learns the relationship between multiple noisy inputs and outputs in an end-to-end way [40–43]. In [40], they defined spatial regions and proposed a non-linear filter that suppresses signals from the undesired region while

preserving signals from the desired region. In [41], the authors proposed an end-to-end system to estimate the time-domain filter-and-sum beamformer coefficient using a DNN. This approach was later replaced with implicit filtering in latent space [42]. In [43], they built a causal neural filter comprising modules for fixed beamforming, beam filtering, and residual refinement in the beamspace domain.

One of the popular approaches that adapt the spatial filter according to the dynamic acoustic condition is the informed filter, which is computed by utilizing the instantaneous acoustic parametric information [15–18]. Refs. [15,16] exploited the instantaneous direction of arrival (DoA) estimates to find the time-varying RTF used to construct the spatial filter, and [18] formulated a Bayesian framework under the DoA uncertainty. In [19], the eigenvector decomposition was applied to the estimated speech SCM to extract the steering vectors, which were used for the MVDR beamformer. The aforementioned approaches often adopted classical techniques such as ESPRIT [52] or MUSIC [53] for DoA estimation, which may be improved by incorporating more sophisticated sound localization [47,48].

Another set of studies focus on the estimation of the acoustic parameters. An EM algorithm [14] was employed to perform a joint estimation of the signals and acoustic parameters. While clean speech signals were obtained in the E-step, the PSDs of signals, RTF, and SCMs were estimated in the M-step. As the previous EM algorithm processed all of the signal samples at once, REM algorithms [27,28] overcame these issues by carrying out frame-wise iterative processing to handle online scenarios. For the speech PSD estimation, ref. [32] proposed an instantaneous PSD estimation method based on generalized principal components to preserve the non-stationarity of speech signals. For the RTF estimation, previous approaches mainly exploited the sample SCMs [46]. The covariance subtraction (CS) approaches [44,45] estimated the RTF by taking the normalized first column of the SCM obtained by the subtraction of the noisy speech SCM and noise SCM, assuming that the rank of the speech SCM was one. On the other hand, the covariance whitening (CW) approaches [30,54] normalized the dewhitened principal eigenvector of the whitened noisy input SCM to obtain the RTF.

In this paper, we propose an improved speech SCM estimation method for the online multi-microphone speech enhancement system based on the MVDR beamformer–Wiener filter factorization. The main contributions of our proposals are as follows:

1. A speech PSD estimator based on the TCS scheme to take the knowledge on the speech signal in the cepstral domain into account;
2. An RTF estimator based on the TDoA estimate to take advantage of the information from all frequency bins, especially when the signal-to-noise ratio (SNR) is low;
3. The refinement of the acoustic parameter estimates by exploiting the clean speech spectrum and clean speech power spectrum estimated in the first pass.

3. MMSE Multi-Microphone Speech Enhancement

3.1. Signal Model

Suppose that there is an array of M microphones in a noisy and reverberant room. Assuming that a single speech source and noises are additive, the observed microphone signals are given as

$$\begin{aligned}\mathbf{y}(l, k) &= \mathbf{g}(l, k)S_1(l, k) + \mathbf{v}(l, k) \\ &= \mathbf{s}(l, k) + \mathbf{v}(l, k)\end{aligned}\quad (1)$$

where $\mathbf{y}(l, k) = [Y_1(l, k), Y_2(l, k), \dots, Y_M(l, k)]^T$, $\mathbf{s}(l, k) = [S_1(l, k), S_2(l, k), \dots, S_M(l, k)]^T$, and $\mathbf{v}(l, k) = [V_1(l, k), V_2(l, k), \dots, V_M(l, k)]^T$, in which $Y_m(l, k)$, $S_m(l, k)$, and $V_m(l, k)$ are the short-time Fourier transform (STFT) coefficients of the microphone signal, clean speech, and background noises, including reverberations at the m th microphone, respectively, and $\mathbf{g}(l, k) = [1, g_2(l, k), \dots, g_M(l, k)]^T$ is the RTF vector for the direct path from the desired speech source to the microphones. We assume that $S_m(l, k)$ and $V_m(l, k)$ are uncorrelated

as in [16], although early reflections may disrupt this assumption. The SCM for the input signal $\mathbf{y}(l, k)$, $\Phi_{\mathbf{y}}(l, k)$, is given by

$$\begin{aligned}\Phi_{\mathbf{y}}(l, k) &= E[\mathbf{y}(l, k)\mathbf{y}^H(l, k)] \\ &= \Phi_{\mathbf{s}}(l, k) + \Phi_{\mathbf{v}}(l, k),\end{aligned}\quad (2)$$

where $E[\cdot]$ denotes mathematical expectation, and $\Phi_{\mathbf{s}}(l, k) = E[\mathbf{s}(l, k)\mathbf{s}^H(l, k)]$ and $\Phi_{\mathbf{v}}(l, k) = E[\mathbf{v}(l, k)\mathbf{v}^H(l, k)]$ are the SCMs of $\mathbf{s}(l, k)$ and $\mathbf{v}(l, k)$, respectively.

3.2. MWF and MVDR–Wiener Filter Factorization

The objective of multi-microphone speech enhancement is to estimate clean speech $S_1(l, k)$ from the noisy observation $\mathbf{y}(l, k)$, and we assume that prior knowledge on the location of the source or RTF is not available. One of the popular approaches is the MWF, which is a linear MMSE estimator for clean speech $S_1(l, k)$, i.e.,

$$\hat{S}_1(l, k) = \mathbf{w}_{mwf}^H(l, k)\mathbf{y}(l, k), \quad (3)$$

where $\mathbf{w}_{mwf}(l, k)$ denotes the MWF described as [6]

$$\mathbf{w}_{mwf}(l, k) = \frac{\Phi_{\mathbf{v}}^{-1}(l, k)\Phi_{\mathbf{s}}(l, k)}{1 + \text{tr}(\Phi_{\mathbf{v}}^{-1}(l, k)\Phi_{\mathbf{s}}(l, k))}\mathbf{e}_1, \quad (4)$$

where $\mathbf{e}_1 = [1 \ \mathbf{0}_{1 \times M-1}]^T$, in which $\mathbf{0}$ is a zero vector, and $\text{tr}[\cdot]$ denotes the trace of a matrix. It is noted that only the noise and speech SCM, $\Phi_{\mathbf{v}}(l, k)$ and $\Phi_{\mathbf{s}}(l, k)$, need to be estimated to implement the MWF. Previous work often adopted the multi-channel MCRA approach for noise SCM estimation, whereas ML estimation was employed for speech SCM estimation [25,26].

The MWF can be decomposed into the MVDR beamformer, \mathbf{w}_{mvdr} , and a single-channel Wiener postfilter, w_{wiener} , as [11,12]

$$\mathbf{w}_{mwf} = \mathbf{w}_{mvdr} \cdot w_{wiener}, \quad (5)$$

which makes it possible to consider the spatial filtering depending on the RTF \mathbf{g} and the energy-based postfiltering w_{wiener} separately. Note that the frame and frequency indices are omitted for notational convenience.

Let the output of the MVDR beamformer be Z , i.e.,

$$Z = \mathbf{w}_{mvdr}^H \mathbf{y}, \quad (6)$$

where the MVDR beamformer is given as

$$\mathbf{w}_{mvdr} = \frac{\Phi_{\mathbf{v}}^{-1} \mathbf{g}}{\mathbf{g}^H \Phi_{\mathbf{v}}^{-1} \mathbf{g}}. \quad (7)$$

With the distortionless constraint of the MVDR beamformer, the beamformer output can be expressed as [27]

$$Z = S_1 + O, \quad (8)$$

where O is assumed to follow the Gaussian distribution with variance

$$\phi_o = (\mathbf{g}^H \Phi_{\mathbf{v}}^{-1} \mathbf{g})^{-1}. \quad (9)$$

The clean speech spectrum can be obtained by applying the single-channel Wiener filter to the beamformer output Z , as

$$\hat{S}_1 = \frac{\phi_s}{\phi_s + \phi_o} \cdot Z, \quad (10)$$

where $\phi_s = E[|S_1|^2]$ is the speech PSD at the first microphone.

Figure 1 illustrates the block diagram of the multi-microphone speech enhancement system based on the MVDR–Wiener filter factorization. The noisy speech \mathbf{y} is processed by the MVDR beamformer and the Wiener filter sequentially, for which acoustic parameters Φ_v , \mathbf{g} , ϕ_s , and ϕ_o need to be estimated. Existing methods for parameter estimation are present in the next subsection.

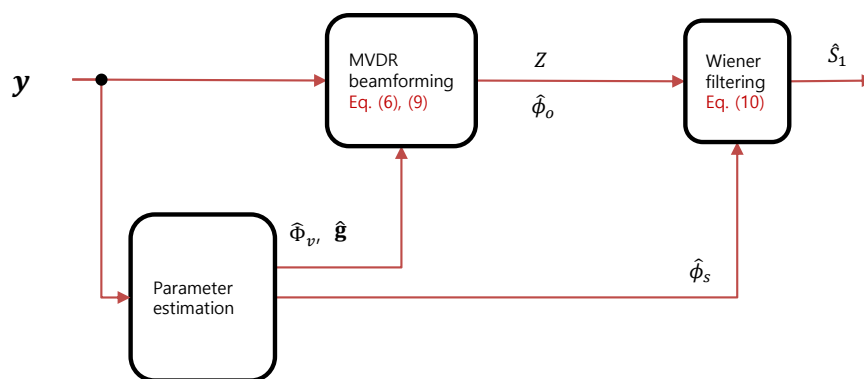


Figure 1. Block diagram of the multi-microphone speech enhancement system based on the MVDR–Wiener filter factorization.

3.3. Speech and Noise SCM Estimation

As for the estimation of the SCM of noise, Φ_v , the multi-channel MCRA approach [25] is widely used, which is given as

$$\hat{\Phi}_v(l, k) = \tilde{\alpha}_v(l, k) \hat{\Phi}_v(l-1, k) + (1 - \tilde{\alpha}_v(l, k)) \mathbf{y}(l, k) \mathbf{y}^H(l, k), \quad (11)$$

where $\tilde{\alpha}_v(l, k) = \lambda + p(H_1(l, k) | \mathbf{y}(l, k)) (1 - \lambda)$ is an SPP-dependent smoothing parameter with a constant $0 < \lambda < 1$. This method updates Φ_v more when the SPP is low and vice versa. The *a posteriori* SPP $p(H_1 | \mathbf{y})$ can be obtained using Bayes' rule as

$$p(H_1 | \mathbf{y}) = \frac{p(H_1) p(\mathbf{y} | H_1)}{p(H_0) p(\mathbf{y} | H_0) + p(H_1) p(\mathbf{y} | H_1)}, \quad (12)$$

where H_0 and H_1 denote the hypotheses for speech absence and presence, respectively, and $p(\mathbf{y} | H_0)$ and $p(\mathbf{y} | H_1)$ are modeled as complex multivariate Gaussian distributions, as follows:

$$p(\mathbf{y} | H_0) = \frac{1}{\pi^M \det[\Phi_v]} \exp\left\{-\mathbf{y}^H \Phi_v^{-1} \mathbf{y}\right\} \quad (13)$$

$$p(\mathbf{y} | H_1) = \frac{1}{\pi^M \det[\Phi_v + \Phi_s]} \exp\left\{-\mathbf{y}^H [\Phi_v + \Phi_s]^{-1} \mathbf{y}\right\}, \quad (14)$$

in which $\det[\cdot]$ denotes the determinant of a matrix. Then, $p(H_1 | \mathbf{y})$ becomes [24]

$$p(H_1 | \mathbf{y}) = \left(1 + \frac{p(H_0)}{p(H_1)} (1 + \xi) \exp\left\{-\frac{\beta}{1 + \xi}\right\}\right)^{-1}, \quad (15)$$

where $\xi = \text{tr}(\Phi_v^{-1} \Phi_s)$, $\beta = \mathbf{y}^H \Phi_v^{-1} \Phi_s \Phi_v^{-1} \mathbf{y}$, and $p(H_1) = 1 - p(H_0)$ is the *a priori* SPP, which can be estimated using the CDR-based [26] or DNN-based [27] method.

The speech SCM is usually estimated with the ML approach, which is defined as [25,26]

$$\widehat{\Phi}_s^{ml} = \widehat{\Phi}_y - \widehat{\Phi}_v, \quad (16)$$

where $\widehat{\Phi}_y$ is obtained by recursive smoothing as

$$\widehat{\Phi}_y(l, k) = \lambda \widehat{\Phi}_y(l-1, k) + (1-\lambda) \mathbf{y}(l, k) \mathbf{y}^H(l, k). \quad (17)$$

Under the rank-1 approximation for the clean speech SCM, $\widehat{\Phi}_s$ can be further refined using the decomposition of Φ_s , with the speech PSD and RTF given by [8]

$$\Phi_s = \phi_s \mathbf{g} \mathbf{g}^H. \quad (18)$$

Adopting the covariance subtraction (CS) approach, which extracts the normalized first column vector of the ML estimator of the speech SCM $\widehat{\Phi}_s^{ml}$, the estimator for the RTF is given as [46]

$$\widehat{\mathbf{g}}^{cs} = \frac{\widehat{\Phi}_s^{ml} \mathbf{e}_1}{\mathbf{e}_1^T \widehat{\Phi}_s^{ml} \mathbf{e}_1}, \quad (19)$$

where the denominator represents the speech PSD, i.e.,

$$\widehat{\phi}_s^{cs} = \mathbf{e}_1^T \widehat{\Phi}_s^{ml} \mathbf{e}_1, \quad (20)$$

in which the superscript ^{cs} indicates the CS approach.

In the REM framework [27], the ML estimator for the RTF based on the observed noisy speech is obtained as [27]

$$\widehat{\mathbf{g}}^{ml}(l, k) = \frac{\sum_{\tau=1}^l \lambda^{l-\tau} p(H_1(\tau, k) | \mathbf{y}(\tau, k)) \mathbf{y}(\tau, k) \widehat{S}_1^*(\tau, k)}{\sum_{\tau=1}^l \lambda^{l-\tau} p(H_1(\tau, k) | \mathbf{y}(\tau, k)) \widehat{S}_1(\tau, k) \widehat{S}_1^*(\tau, k)}, \quad (21)$$

where the summations in the numerator and the denominator can be computed with recursive averaging. The numerator can be thought of as the estimate of the cross-correlation between \mathbf{y} and \widehat{S}_1 in (10), $\widehat{\mathbf{r}}_{ys}$, given by

$$\widehat{\mathbf{r}}_{ys}(l, k) = \lambda \widehat{\mathbf{r}}_{ys}(l-1, k) + (1-\lambda) p(H_1(l, k) | \mathbf{y}(l, k)) \mathbf{y}(l, k) \widehat{S}_1^*(l, k), \quad (22)$$

and the denominator can be considered to be the estimate of the speech PSD obtained by the recursive smoothing of the estimated clean speech power spectrum,

$$\widehat{\phi}_s^{ts}(l, k) = \lambda \widehat{\phi}_s^{ts}(l-1, k) + (1-\lambda) \widehat{|S_1|^2}(l, k), \quad (23)$$

where the superscript ^{ts} indicates it is a temporally smoothed estimate, and $\widehat{|S_1|^2}(l, k)$ is the MMSE estimator of $|S_1|^2$ under the speech presence uncertainty given by

$$\begin{aligned} \widehat{|S_1|^2} &= E[|S_1|^2 | Z] \\ &= p(H_1 | Z) \cdot E[|S_1|^2 | Z, H_1] \\ &= p(H_1 | \mathbf{y}) \cdot \left(\left(\frac{\phi_s}{\phi_s + \phi_o} \right)^2 \cdot |Z|^2 + \frac{\phi_s \phi_o}{\phi_s + \phi_o} \right), \end{aligned} \quad (24)$$

where we let $p(H_1 | Z) = p(H_1 | \mathbf{y})$, as in [27]. With $\widehat{\mathbf{r}}_{ys}$ in (22) and $\widehat{\phi}_s^{ts}$ in (23), $\widehat{\mathbf{g}}^{ml}$ in (21) can be expressed as

$$\widehat{\mathbf{g}}^{ml} = \frac{\widehat{\mathbf{r}}_{ys}}{\widehat{\phi}_s^{ts}}. \quad (25)$$

4. Proposed Speech SCM Estimation

Figure 2 illustrates the block diagram of the proposed speech enhancement system. As in [25–27], the estimation of the speech and relevant statistical parameters is performed twice for each frame, which was shown to be effective for online speech enhancement. In this paper, we propose an improved method for speech SCM estimation, i.e., speech PSD estimation and RTF estimation with a rank-1 approximation, using the speech enhancement system described in Figure 2. Note that the proposed modules are highlighted with red boxes.

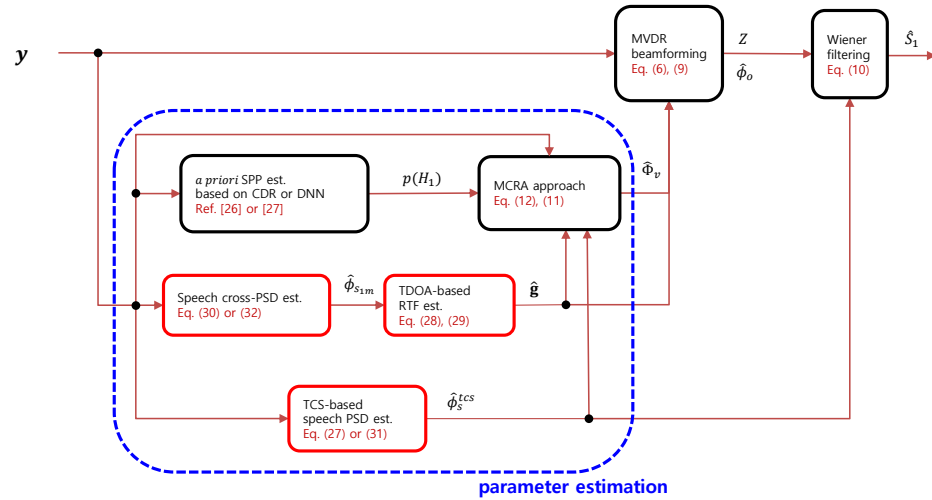


Figure 2. Block diagram of the proposed multi-microphone speech enhancement system.

In the first pass, we exploit the noisy input $\mathbf{y}(l)$ in the current frame and the noise SCM estimate $\hat{\Phi}_v(l-1)$ obtained in the previous frame to estimate the acoustic parameters in the current frame and perform beamforming and postfiltering, as explained in Section 3.2.

The ML estimate of the speech PSD at the first microphone using an instantaneous estimate of the PSD of input noisy signal can be obtained as

$$\hat{\phi}_s^{ml}(l, k) = \max\left(|Y_1(l, k)|^2 - \hat{\phi}_v(l-1, k), \hat{\phi}_s^{min}\right), \quad (26)$$

where $\hat{\phi}_v(l-1, k)$ is the $(1, 1)$ th component of $\hat{\Phi}_v(l-1, k)$, and $\hat{\phi}_s^{min}$ is a certain minimum value for the speech PSD estimate, which is set as $\zeta^{min}\hat{\phi}_v(l-1, k)$ with a tunable parameter ζ^{min} . To estimate the speech PSDs, the ML estimation with temporal smoothing has been commonly used as described in (16) and (17) [25–27]. However, this approach occasionally results in undesired temporal smearing of speech [51]. In this paper, we propose to apply TCS [51] to $\hat{\phi}_s^{ml}$ in (26). TCS is a selective temporal smoothing technique in the cepstral domain motivated by the observation that, although the excitation component resides in a limited number of cepstral coefficients dependent on the pitch frequency, the speech spectral envelope is well-represented by the cepstral bins with low indices [55]. Specifically, the TCS consists of the following procedure: First, the cepstrum of ML speech PSD estimate $\hat{\phi}_s^{ml,ceps}(l, q)$ is computed by the inverse discrete Fourier transform (IDFT) of $\hat{\phi}_s^{ml}$. Next, the selective smoothing is applied to $\hat{\phi}_s^{ml,ceps}(l, q)$, in which the cepstral bins that are less relevant to speech are smoothed more and those representing the spectral envelope and fundamental frequency are less smoothed. Finally, the discrete Fourier transform is used to convert $\hat{\phi}_s^{ml,ceps}(l, q)$ into the TCS-based speech PSD estimate in the spectral domain $\hat{\phi}_s^{tcs}(l, k)$. The bias compensation for the reduced variance due to the cepstral smoothing can be found in [56], and a detailed description of the adaptation of the smoothing parameters

and the fundamental frequency estimation is given in [51]. In this paper, we denote the aforementioned procedure of TCS as an operation:

$$\widehat{\phi}_s^{tcs,f}(l) = \text{TCS}(\widehat{\phi}_s^{ml}(l)), \quad (27)$$

in which the superscript f indicates that this is the estimate in the first pass.

In this paper, we model the RTF vector \mathbf{g} as a relative array propagation vector, which depends on the DoA [16]. Note that the conventional approaches in [27,44] estimate the RTF for each frequency using the input statistics in the frequency bin, ignoring the inter-frequency dependencies. In the presence of heavy noise, the accurate estimation of the RTF may become difficult, and thus it would be beneficial to estimate TDoA by utilizing the input signal in all frequency bins and to reconstruct the RTF using the simplest model. The TDoA for the desired speech can be obtained from the estimate of the cross-PSD of the desired speech, $\phi_{s_{1m}}(l, k) = E[S_1(l, k)S_m^*(l, k)]$, using the cross-correlation method [57]. The TDoA estimate τ_m between the first and the m th microphones is given by

$$\begin{aligned} \tau_m(l) &= \arg \max_{\tau} \gamma_{1m}(\tau), \\ \gamma_{1m}(\tau) &\triangleq \sum_{k=1}^K \widehat{\phi}_{s_{1m}}(l, k) \cdot e^{j2\pi k\tau/K}, \end{aligned} \quad (28)$$

in which $\widehat{\phi}_{s_{1m}}(l, k)$ is the estimate of $\phi_{s_{1m}}(l, k)$. Then, the TDoA-based RTF estimator can be obtained as

$$\widehat{\mathbf{g}}^{tdoa}(l, k) = \exp\{j2\pi k\bar{\tau}(l)/K\}, \quad (29)$$

where $\bar{\tau} = [\tau_1, \tau_2, \dots, \tau_M]^T$.

In the first pass, the cross-PSD estimate $\widehat{\phi}_{s_{1m}}$ can be obtained by taking the $(1, m)$ element of the ML speech SCM estimate $\widehat{\Phi}_s^{ml}(l, k)$ in (16) as

$$\widehat{\phi}_{s_{1m}}^f(l, k) = \mathbf{e}_m^T \widehat{\Phi}_s^{ml}(l, k) \mathbf{e}_1, \quad (30)$$

where $\mathbf{e}_m = [\mathbf{0}_{(m-1)} \quad 1 \quad \mathbf{0}_{(M-m)}]^T$ in which $\mathbf{0}_n$ is an all-zero vector of length n ; $\widehat{\mathbf{g}}^{tdoa,f}$ can be computed using (28) and (29) with $\widehat{\phi}_{s_{1m}}^f$, and $\widehat{\Phi}_s^f$ can be obtained as in (18) using $\widehat{\phi}_s^{tcs,f}(l)$ in (27) and $\widehat{\mathbf{g}}^{tdoa,f}$. The noise SCM is estimated with the multi-channel MCRA approach in (11) utilizing $p(H_1|\mathbf{y})$ in (15) computed with $\widehat{\Phi}_s^f$ and $\widehat{\Phi}_v$. Then, we can compute the beamformer output Z in (6) and ϕ_o in (9), and the estimate for the speech spectrum, \widehat{S}_1 , can be obtained as in (10).

In the second pass, we estimate the acoustic parameters again by additionally utilizing the estimates for the clean speech spectrum, clean speech power spectrum, and a *posteriori* SPP, computed in the first pass. These refined parameters are in turn used to estimate the clean speech once again.

To refine the estimate of the speech PSD, we apply the TCS to the clean speech power spectrum estimate $\widehat{|S_1|^2}$ in (24) as

$$\widehat{\phi}_s^{tcs,r}(l) = \text{TCS}(\widehat{|S_1|^2}(l)), \quad (31)$$

in which the superscript r indicates it is the refined estimate in the second pass. As $\widehat{|S_1|^2}$ would be less affected by the noise compared with the $\widehat{\phi}_s^{ml}$ by virtue of beamforming and the MMSE estimation, $\widehat{\phi}_s^{tcs,r}(l)$ would be more accurate than $\widehat{\phi}_s^{tcs,f}(l)$. As for the RTF estimation, $\widehat{\mathbf{r}}_{ys}$ in (22) is evaluated with \widehat{S}_1 in (10), as in [27]. Instead of using $\widehat{\mathbf{r}}_{ys}$ divided by the estimate of the speech PSD in the first microphone to obtain the RTE, as in [27], we

again estimate the RTF based on the TDoA; $\hat{\phi}_{s_{1m}}^r$ can be computed by extracting the m th element of $\hat{\mathbf{r}}_{ys}$ as

$$\hat{\phi}_{s_{1m}}^r(l, k) = \mathbf{e}_m^T \hat{\mathbf{r}}_{ys}(l, k), \quad (32)$$

in contrast to (30). The TDoA-based RTF estimate in the second pass, $\hat{\mathbf{g}}^{tdoa,r}$, can be obtained through (28) and (29) with $\hat{\phi}_{s_{1m}}^r$. As in the first pass, $\hat{\Phi}_s^r$ is computed with $\hat{\phi}_s^{tcs,r}$ in (31) and $\hat{\mathbf{g}}^{tdoa,r}$, and $p(H_1|\mathbf{y})$ in (15) is updated with $\hat{\Phi}_s^r$. Then, $p(H_1|\mathbf{y})$ and $\hat{\Phi}_v$ are obtained again using (15) and (11), and then the beamformer output Z and ϕ_o are updated using (6) in (9). The final clean speech estimate \hat{S}_1 is obtained by (10) using $\hat{\mathbf{g}}^{tdoa,r}$, $\hat{\Phi}_v$, and $\hat{\phi}_s^{tcs,r}$. The whole procedure of the proposed online multi-microphone speech enhancement method is summarized in Algorithm 1.

Algorithm 1 Proposed multi-microphone speech enhancement algorithm with improved speech SCM estimation.

- 1: **Inputs:** \mathbf{y} for all frames
 - 2: **Output:** \hat{S}_1 for all frames
 - 3: **Initialize** variables and parameters
 - 4: **for** each frame **do**
 - 5: Compute $p(H_1)$ using CDR-based [26] or DNN-based [27] method
 - 6: **(First pass)**
 - 7: Compute $\hat{\phi}_s^{tcs,f}$ via (26) and (27)
 - 8: Compute $\hat{\mathbf{g}}^{tdoa,f}$ via (16), (28)–(30)
 - 9: Estimate $p(H_1|\mathbf{y})$ and $\hat{\Phi}_v$ via (11), (15) and (18)
 - 10: Beamformer: Compute Z and ϕ_o via (6) and (9)
 - 11: Postfilter: Compute \hat{S}_1 via (10)
 - 12: **(Second pass)**
 - 13: Compute $\hat{\phi}_s^{tcs,r}$ via (24) and (31)
 - 14: Compute $\hat{\mathbf{g}}^{tdoa,r}$ via (22), (28), (29) and (32)
 - 15: Estimate $p(H_1|\mathbf{y})$ and $\hat{\Phi}_v$ via (11), (15) and (18)
 - 16: Beamformer: Compute Z and ϕ_o via (6) and (9)
 - 17: Postfilter: Compute \hat{S}_1 via (10)
 - 18: **end for**
-

5. Experiments

5.1. Experimental Settings

To demonstrate the superiority of the proposed algorithm, we conducted a set of experiments to evaluate the performance of the multi-microphone speech enhancement on the simulated set in the CHiME-4 database [58]. In this database, a mobile tablet device with six microphones was used for recording, of which the three microphones numbered 1, 2, and 3 were located in the top left, center, and right with an inter-microphone distance of approximately 10 cm each, while the other three microphones numbered 4, 5, and 6 were placed in the bottom left, center, and right, respectively [58]. The vertical distance between pairs of microphones was approximately 19 cm [58]. All microphones were located on the frontal surface, except for microphone 2. The bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR) types of noise were used, and the SNR was between 0 and 15 dB. The training set consisted of 7138 utterances spoken by 83 speakers, whereas the development and evaluation sets were 1640 utterances and 1320 utterances, respectively, from 4 different speakers. The sampling rate for the signals used in the experiments was 16 kHz, and the square-root Hann window was applied to a 32 ms signal with a 16 ms frame shift. The 512-point DFT was applied to the windowed signal. The reference channel for the algorithms and evaluations was microphone 5, located at the bottom center of the device.

We set the *a posteriori* SPP, $p(H_1|\mathbf{y})$, to zero for the first 10 frames instead of computing it using (15) based on the assumption that the speech would be absent in the initial periods,

which helped the fast stabilization of the algorithm, as in [27]. To mitigate speech distortion at the expense of increased residual noise [59], the lower bounds for the $p(H_1|\mathbf{y})$ to compute $|\widehat{S}_1|^2$ in (24) and the Wiener gain in (10) were configured to 0.5 and -18 dB, respectively. The parameter values for λ and ζ^{min} were set to be 0.9 and -10 dB, respectively. For the TCS schemes in (27) and (31), we followed the procedure in [51], employing the same parameter values except for the constant smoothing parameter, $\bar{\alpha}^{const}$, which was determined empirically as

$$\bar{\alpha}^{const}(q) = \begin{cases} 0.1 & \text{if } q \in \{0, \dots, 2\} \\ 0.5 & \text{if } q \in \{3, \dots, 19\} \\ 0.95 & \text{if } q \in \{20, \dots, 256\}, \end{cases} \quad (33)$$

in which q is the quefrency index.

For the DNN-based *a priori* SPP estimation, we adopted the DNN architecture in [27], which consisted of a uni-directional long short-term memory (LSTM) layer of 512 dimensions, followed by three fully-connected layers of 256 dimensions. The activation functions were the rectified linear unit (ReLU) for the first three layers and sigmoidal activation for the last layer, which produced a 257-dimensional output vector. The number of dimensions of the DNN output is 257. The input for the DNN was the noisy log magnitude spectrum at the reference microphone, and the training target was binary for each bin, which was set by thresholding the instantaneous SNR [13].

5.2. Experimental Results

To demonstrate the superior performance of the proposed speech enhancement method, we evaluated the wideband PESQ score [60], eSTOI [61], and SISDR [62]. As we focused on the online framework in which the algorithm only uses the current and previous audio samples for frame-wise processing, the online algorithms designed in this way were chosen for the baseline methods. Depending on the *a priori* SPP estimator, we compared the performance of the proposed method using the ML framework with the MWF in [26] when the CDR-based *a priori* SPP estimator [26] was adopted, whereas the REM approach [27] was used for performance comparison when the DNN-based *a priori* SPP estimator [27] was employed for the proposed algorithm. As in [27], two versions of the REM approach using the Wiener postfilter and Kalman postfilter, denoted by DNN-REMWF and DNN-REMKF, were included in the experiment. The configuration parameters for the compared methods were set as in the original papers.

Tables 1–3 show the average PESQ score, eSTOI, and SISDR for each method depending on the noise type, respectively. The proposed method with the CDR-based *a priori* SPP estimator, CDR-Proposed, outperformed the previous approach in [26] by 0.39 in terms of the average PESQ score, 0.022 in terms of the eSTOI, and 4.3 dB in terms of the SISDR on average, respectively. With the DNN-based *a priori* SPP estimator, the proposed method, DNN-Proposed, improved the performance of DNN-REMKF by 0.21 in terms of the average PESQ score, 0.017 in terms of the eSTOI, and 1.2 dB in terms of the SISDR on average, respectively. Table 4 shows the PESQ scores, eSTOIs, and SISDRs for the baselines and the proposed method depending on the SNR. As the SNRs for the utterances in the evaluation set of the CHiME-4 database are distributed as in Figure 3, we divided the evaluation set into three groups depending on the SNR: low SNR less than 6.5 dB, medium SNR between 6.5 and 8.5 dB, and high SNR over 8.5 dB. It can be seen that all the measures were improved in all SNR ranges, and the performance improvements were more pronounced in low SNRs. From the results, we may conclude that the proposed speech SCM estimation approach could improve the performance of the multi-microphone speech enhancement method, regardless of the adoption of the DNN for the *a priori* SPP estimation.

Table 1. The average PESQ scores for the different algorithms depending on the noise type.

Method	Noise Type				Avg.
	BUS	CAF	PED	STR	
Noisy	1.32	1.24	1.26	1.28	1.27
CDR-MWF [26]	1.93	1.74	1.86	1.74	1.82
CDR-Proposed	2.35	2.15	2.18	2.17	2.21
DNN-REMWF [27]	2.10	2.01	2.12	2.03	2.07
DNN-REMKF [27]	2.13	2.03	2.15	2.07	2.10
DNN-Proposed	2.43	2.24	2.28	2.29	2.31

Table 2. The average eSTOIs (x100) for the different algorithms depending on the noise type.

Method	Noise Type				Avg.
	BUS	CAF	PED	STR	
Noisy	71.0	66.0	68.8	67.2	68.2
CDR-MWF [26]	83.1	79.7	82.4	80.2	81.4
CDR-Proposed	85.1	83.4	83.7	82.4	83.6
DNN-REMWF [27]	83.7	82.0	83.9	83.2	83.2
DNN-REMKF [27]	84.3	82.6	84.5	83.9	83.8
DNN-Proposed	86.0	85.3	85.9	85.0	85.5

Table 3. The average SISDRs (in dB) for the different algorithms depending on the noise type.

Method	Noise Type				Avg.
	BUS	CAF	PED	STR	
Noisy	6.79	7.77	8.60	6.85	7.51
CDR-MWF [26]	9.68	9.44	10.69	9.75	9.89
CDR-Proposed	14.25	14.55	14.56	13.40	14.19
DNN-REMWF [27]	14.40	14.18	14.96	14.64	14.54
DNN-REMKF [27]	14.71	14.42	15.25	14.99	14.84
DNN-Proposed	15.90	16.07	16.34	15.83	16.04

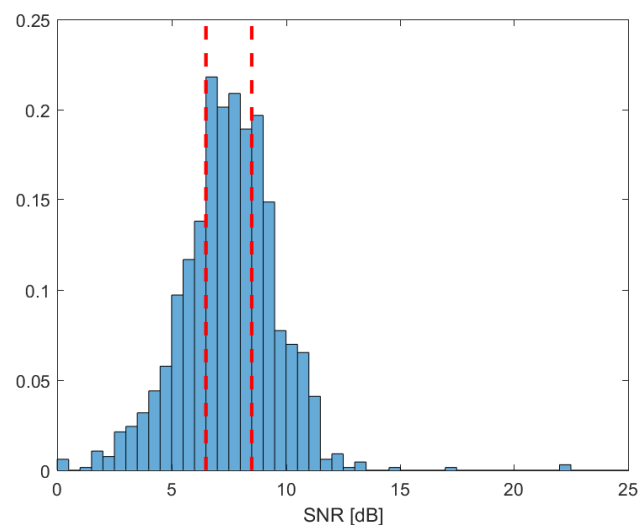
**Figure 3.** Normalized histogram of the SNRs in dB scale for the utterances in the evaluation set of the CHiME-4 database.

Table 4. The PESQ scores, eSTOIs, and SISDRs for the baselines and the proposed method depending on the SNR.

Method	PESQ Score			eSTOI ($\times 100$)			SISDR (in dB)		
	($-\infty, 6.5$)	(6.5, 8.5)	(8.5, ∞)	($-\infty, 5$)	(6.5, 8.5)	(8.5, ∞)	($-\infty, 6.5$)	(6.5, 8.5)	(8.5, ∞)
Noisy	1.22	1.26	1.33	63.3	67.1	74.2	5.04	7.48	9.74
CDR-MWF [26]	1.66	1.83	1.95	76.7	80.7	86.5	8.81	9.80	10.97
CDR-Proposed	2.07	2.22	2.33	80.6	83.2	86.9	12.35	13.95	16.14
DNN-REMWF [27]	1.84	2.07	2.26	79.3	82.8	87.2	12.80	14.41	16.28
DNN-REMKF [27]	1.87	2.10	2.29	80.0	83.4	87.7	13.07	14.70	16.60
DNN-Proposed	2.15	2.32	2.44	82.7	85.2	88.5	14.38	15.83	17.78

5.3. Ablation Study

Additionally, we carried out an ablation study to analyze how much each module in the proposed system contributed to the performance improvement. We propose the speech PSD estimator, $\hat{\phi}_s^{tcs,s}$ in (31), and the RTF estimator, $\hat{\mathbf{g}}^{tdoa,s}$ in (29). The previous approaches were the speech PSD estimator using recursive smoothing, $\hat{\phi}_s^{ts}$ in (23), and the ML estimator of the RTF $\hat{\mathbf{g}}^{ml}$ in (25). The performances of the systems replacing the proposed modules one by one with conventional modules are summarized in Table 5. DNN-REMWF is also included, which uses $\hat{\phi}_s^{ts}$, $\hat{\mathbf{g}}^{ml}$, and the Wiener postfilter, but adopts a different noise SCM estimator, derived from the EM framework.

Table 5. The PESQ scores, eSTOIs, and SISDRs averaged over all noise types for the proposed method with DNN-based *a priori* SPP estimation by replacing the proposed sub-modules with conventional ones, one by one.

Method	PESQ	eSTOI	SISDR
DNN-REMWF [27]	2.07	83.2	14.54
$\hat{\phi}_s^{ts} + \hat{\mathbf{g}}^{ml}$	2.07	84.5	15.79
$\hat{\phi}_s^{ts} + \hat{\mathbf{g}}^{tdoa,s}$	2.12	85.4	16.07
$\hat{\phi}_s^{tcs,s} + \hat{\mathbf{g}}^{ml}$	2.19	83.8	15.52
$\hat{\phi}_s^{tcs,s} + \hat{\mathbf{g}}^{tdoa,s}$ (DNN-Proposed)	2.31	85.5	16.04

The proposed system with conventional speech PSD and RTF estimators, $\hat{\phi}_s^{ts} + \hat{\mathbf{g}}^{ml}$, showed the same average PESQ score and improved eSTOI and SISDR compared with DNN-REMWF [27]. Among the systems in the same framework, the introduction of the proposed speech PSD estimator improved the average PESQ scores by relatively large differences of 0.12 and 0.19, whereas it did not result in increased eSTOIs and SISDRs. On the other hand, employing the proposed RTF estimator improved all three metrics. From the results, we may conclude that both the proposed speech PSD and the RTF estimators contributed to the performance improvement.

5.4. Computational Complexity

Additionally, we have compared the computational complexity of the baseline and proposed methods in terms of the normalized processing time for the MATLAB implementation of the methods. The processing times for each algorithm, normalized by the processing time of the proposed algorithm, are given in Table 6. In this experiment, the *a priori* SPP was estimated by a DNN for all cases. As they depend on implementation details and settings such as the number of microphones, sampling frequency, and the dimensions of the DFT, the numbers given in the table should only be used as a rough indication. To see how much the refinement in the second pass incurred additional computational burden, the proposed method without the second pass (denoted as woSP) is included. From the table, it can be seen that the computational complexity of the proposed method was higher than those for MWF [26] and REMWF [27], but less than that of REMKF [27].

Table 6. Comparison of the normalized processing time when the *a priori* SPP was obtained by a DNN.

	MWF [26]	REMWF [27]	REMKF [27]	Proposed (woSP)	Proposed
Process. time	0.614	0.692	1.132	0.704	1.000

6. Conclusions

Multi-microphone speech enhancement exploits spatial information and spectro-temporal characteristics to reduce noise from the input. The online algorithms are required for the applications sensitive to time delays such as speech communication and hearing aids. In this paper, we propose an improved estimator of the speech SCM for online multi-microphone speech enhancement. Using the decomposition of the speech SCM under a rank-1 approximation, we propose an improved estimator for the speech PSD and RTF. For speech PSD estimation, we adopt the TCS scheme, which exploits knowledge on the speech signal in the cepstral domain to provide a better estimate of the speech PSD compared with the ML estimate. The RTF is estimated based on the TDoA estimate summarizing the information from all frequency bins. These estimators are evaluated once with input statistics and refined with an estimated clean speech spectrum and power spectrum obtained in the first pass. Our proposed speech enhancement method showed an improved speech enhancement performance in terms of the PESQ score, eSTOI, and SISDR in various noise environments for the CHiME-4 dataset, compared with other online multi-microphone speech enhancement algorithms.

Future work may include the incorporation of other spatial cues such as the inter-channel level differences on top of the inter-channel phase differences [47] into the RTF estimation without resorting to the far-field assumption. We may also investigate a deep learning approach to estimate acoustic parameters such as the speech and noise PSDs and RTF in a causal manner in the MVDR–Wiener filter factorization framework.

Author Contributions: Conceptualization, M.K., S.C. and J.W.S.; methodology, M.K. and J.W.S.; software, M.K. and H.S.; validation, M.K., S.C., H.S. and J.W.S.; formal analysis, J.W.S.; investigation, M.K. and S.C.; resources, J.W.S.; data curation, M.K. and H.S.; writing—original draft preparation, M.K.; writing—review and editing, J.W.S.; visualization, M.K.; supervision, J.W.S.; project administration, J.W.S.; funding acquisition, J.W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of Korea, grant number NRF-2019R1A2C2089324, and the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-0-01835) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vary, P.; Martin, R. *Digital Speech Transmission: Enhancement, Coding and Error Concealment*; John Wiley & Sons: Chichester, UK, 2006.
2. Kates, J.M. *Digital Hearing Aids*; Plural Publishing: San Diego, CA, USA, 2008.
3. Rabiner, L.; Juang, B.H. *Fundamentals of Speech Recognition*; Prentice-Hall, Inc.: Hoboken, NJ, USA, 1993.
4. Kim, M.; Shin, J.W. Improved Speech Enhancement Considering Speech PSD Uncertainty. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1939–1951. [[CrossRef](#)]
5. Kim, M.; Song, H.; Cheong, S.; Shin, J.W. iDeepMMSE: An improved deep learning approach to MMSE speech and noise power spectrum estimation for speech enhancement. *Proc. Interspeech* **2022**, *2022*, 181–185.
6. Benesty, J.; Chen, J.; Huang, Y. *Microphone Array Signal Processing*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.

7. Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 692–730. [[CrossRef](#)]
8. Souden, M.; Benesty, J.; Affes, S. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 260–276. [[CrossRef](#)]
9. Markovich-Golan, S.; Gannot, S.; Cohen, I. A weighted multichannel Wiener filter for multiple sources scenarios. In Proceedings of the 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel, 14–17 November 2012; pp. 1–5.
10. Doclo, S.; Spriet, A.; Wouters, J.; Moonen, M. Speech distortion weighted multichannel Wiener filtering techniques for noise reduction. In *Speech Enhancement*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 199–228.
11. Balan, R.; Rosca, J. Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase. In Proceedings of the Sensor Array and Multichannel Signal Processing Workshop Proceedings, Rosslyn, VA, USA, 6 August 2002; pp. 209–213.
12. Thüne, P.; Enzner, G. Maximum-likelihood approach with Bayesian refinement for multichannel-Wiener postfiltering. *IEEE Trans. Signal Process.* **2017**, *65*, 3399–3413. [[CrossRef](#)]
13. Heymann, J.; Drude, L.; Haeb-Umbach, R. A generic neural acoustic beamforming architecture for robust multi-channel speech processing. *Comput. Speech Lang.* **2017**, *46*, 374–385. [[CrossRef](#)]
14. Schwartz, O.; Gannot, S.; Habets, E.A. An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1495–1510. [[CrossRef](#)]
15. Thiergart, O.; Taseska, M.; Habets, E.A. An informed MMSE filter based on multiple instantaneous direction-of-arrival estimates. In Proceedings of the 21st IEEE European Signal Processing Conference (EUSIPCO 2013), Marrakech, Morocco, 9–13 September 2013; pp. 1–5.
16. Thiergart, O.; Taseska, M.; Habets, E.A. An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 2182–2196. [[CrossRef](#)]
17. Taseska, M.; Habets, E.A. Informed spatial filtering for sound extraction using distributed microphone arrays. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1195–1207. [[CrossRef](#)]
18. Chakrabarty, S.; Habets, E.A. A Bayesian approach to informed spatial filtering with robustness against DOA estimation errors. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 145–160. [[CrossRef](#)]
19. Higuchi, T.; Ito, N.; Araki, S.; Yoshioka, T.; Delcroix, M.; Nakatani, T. Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 780–793. [[CrossRef](#)]
20. Jin, Y.G.; Shin, J.W.; Kim, N.S. Spectro-temporal filtering for multichannel speech enhancement in short-time Fourier transform domain. *IEEE Signal Process. Lett.* **2014**, *21*, 352–355. [[CrossRef](#)]
21. Serizel, R.; Moonen, M.; Van Dijk, B.; Wouters, J. Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 785–799. [[CrossRef](#)]
22. Wang, Z.; Vincent, E.; Serizel, R.; Yan, Y. Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments. *Comput. Speech Lang.* **2018**, *49*, 37–51. [[CrossRef](#)]
23. Schwartz, O.; Gannot, S.; Habets, E.A. Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 151–155.
24. Souden, M.; Chen, J.; Benesty, J.; Affes, S. Gaussian model-based multichannel speech presence probability. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 1072–1077. [[CrossRef](#)]
25. Souden, M.; Chen, J.; Benesty, J.; Affes, S. An integrated solution for online multichannel noise tracking and reduction. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2159–2169. [[CrossRef](#)]
26. Taseska, M.; Habets, E.A. Nonstationary noise PSD matrix estimation for multichannel blind speech extraction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2223–2236.
27. Martín-Doñas, J.M.; Jensen, J.; Tan, Z.H.; Gomez, A.M.; Peinado, A.M. Online Multichannel Speech Enhancement Based on Recursive EM and DNN-Based Speech Presence Estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 3080–3094. [[CrossRef](#)]
28. Schwartz, O.; Gannot, S. A recursive expectation-maximization algorithm for online multi-microphone noise reduction. In Proceedings of the 2018 IEEE 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 1542–1546.
29. Jin, Y.G.; Shin, J.W.; Kim, N.S. Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement. *J. Acoust. Soc. Am.* **2017**, *141*, EL228–EL233. [[CrossRef](#)]
30. Markovich, S.; Gannot, S.; Cohen, I. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1071–1086. [[CrossRef](#)]
31. Hwang, S.; Kim, M.; Shin, J.W. Dual microphone speech enhancement based on statistical modeling of interchannel phase difference. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2865–2874. [[CrossRef](#)]
32. Dietzen, T.; Moonen, M.; van Waterschoot, T. Instantaneous PSD Estimation for Speech Enhancement based on Generalized Principal Components. In Proceedings of the 2020 IEEE 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 191–195.

33. Dietzen, T.; Doclo, S.; Moonen, M.; van Waterschoot, T. Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 755–769. [[CrossRef](#)]
34. Mitsufuji, Y.; Takamune, N.; Koyama, S.; Saruwatari, H. Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 607–617. [[CrossRef](#)]
35. Dietzen, T.; Doclo, S.; Moonen, M.; van Waterschoot, T. Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 740–754. [[CrossRef](#)]
36. Pezzoli, M.; Cobos, M.; Antonacci, F.; Sarti, A. Sparsity-Based Sound Field Separation in The Spherical Harmonics Domain. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 1051–1055.
37. Pezzoli, M.; Carabias-Orti, J.J.; Cobos, M.; Antonacci, F.; Sarti, A. Ray-space-based multichannel nonnegative matrix factorization for audio source separation. *IEEE Signal Process. Lett.* **2021**, *28*, 369–373. [[CrossRef](#)]
38. Wang, Z.Q.; Wang, P.; Wang, D. Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2020**, *28*, 1778–1787. [[CrossRef](#)] [[PubMed](#)]
39. Kim, H.; Kang, K.; Shin, J.W. Factorized MVDR Deep Beamforming for Multi-Channel Speech Enhancement. *IEEE Signal Process. Lett.* **2022**, *29*, 1898–1902. [[CrossRef](#)]
40. Markovic, D.; Defossez, A.; Richard, A. Implicit Neural Spatial Filtering for Multichannel Source Separation in the Waveform Domain. *arXiv* **2022**, arXiv:2206.15423.
41. Luo, Y.; Chen, Z.; Mesgarani, N.; Yoshioka, T. End-to-end microphone permutation and number invariant multi-channel speech separation. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6394–6398.
42. Luo, Y.; Mesgarani, N. Implicit Filter-and-Sum Network for End-to-End Multi-Channel Speech Separation. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 3071–3075.
43. Liu, W.; Li, A.; Wang, X.; Yuan, M.; Chen, Y.; Zheng, C.; Li, X. A Neural Beamspace-Domain Filter for Real-Time Multi-Channel Speech Enhancement. *Symmetry* **2022**, *14*, 1081. [[CrossRef](#)]
44. Cohen, I. Relative transfer function identification using speech signals. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 451–459. [[CrossRef](#)]
45. Varzandeh, R.; Taseska, M.; Habets, E.A. An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation. In Proceedings of the 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 1–3 March 2017; pp. 11–15.
46. Zhang, J.; Heusdens, R.; Hendriks, R.C. Relative acoustic transfer function estimation in wireless acoustic sensor networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1507–1519. [[CrossRef](#)]
47. Pak, J.; Shin, J.W. Sound localization based on phase difference enhancement using deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2019**, *27*, 1335–1345. [[CrossRef](#)]
48. Song, H.; Shin, J.W. Multiple Sound Source Localization Based on Interchannel Phase Differences in All Frequencies with Spectral Masks. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 671–675.
49. Cohen, I.; Berdugo, B. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Process. Lett.* **2002**, *9*, 12–15. [[CrossRef](#)]
50. Cohen, I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 466–475. [[CrossRef](#)]
51. Breithaupt, C.; Gerkmann, T.; Martin, R. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 4897–4900.
52. Roy, R.; Kailath, T. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 984–995. [[CrossRef](#)]
53. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [[CrossRef](#)]
54. Markovich-Golan, S.; Gannot, S.; Kellermann, W. Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 2499–2503. [[CrossRef](#)]
55. Noll, A.M. Cepstrum pitch determination. *J. Acoust. Soc. Am.* **1967**, *41*, 293–309. [[CrossRef](#)]
56. Gerkmann, T.; Martin, R. On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling. *IEEE Trans. Signal Process.* **2009**, *57*, 4165–4174. [[CrossRef](#)]
57. Knapp, C.; Carter, G. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 320–327. [[CrossRef](#)]
58. Vincent, E.; Watanabe, S.; Nugraha, A.A.; Barker, J.; Marxer, R. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Comput. Speech Lang.* **2017**, *46*, 535–557. [[CrossRef](#)]

59. Gerkmann, T.; Breithaupt, C.; Martin, R. Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 910–919. [[CrossRef](#)]
60. International Telecommunication Union. *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codec*; International Telecommunication Union: Geneva, Switzerland, 2007.
61. Jensen, J.; Taal, C.H. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2009–2022. [[CrossRef](#)]
62. Le Roux, J.; Wisdom, S.; Erdogan, H.; Hershey, J.R. SDR—half-baked or well done? In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 626–630.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.