sensors

MDPI

Article

Postfilter for Dual Channel Speech Enhancement Using Coherence and Statistical Model-Based Noise Estimation

Sein Cheong, Minseung Kim and Jong Won Shin *

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea; seiinjung@gm.gist.ac.kr (S.C.); kms0603@gm.gist.ac.kr (M.K.)
* Correspondence: jwshin@gist.ac.kr

Abstract: A multichannel speech enhancement system usually consists of spatial filters such as adaptive beamformers followed by postfilters, which suppress remaining noise. Accurate estimation of the power spectral density (PSD) of the residual noise is crucial for successful noise reduction in the postfilters. In this paper, we propose a postfilter utilizing proposed *a posteriori* speech presence probability (SPP) and noise PSD estimators, which are based on both the coherence and the statistical models. We model the coherence-based *a posteriori* SPP as a simple function of the magnitude of coherence between two microphone signals and combine it with a single-channel SPP based on statistical models. The coherence-based estimator for the PSD of the noise remaining in the beamformer output in the presence of speech is derived using the pseudo-coherence considering the effect of the beamformers, which is used to construct the coherence-based noise PSD estimator. Then, the final noise PSD estimator is obtained by combining the coherence-based and statistical model-based noise PSD estimators with the proposed SPP. The spectral gain function is also modified, incorporating the proposed SPP. Experimental results demonstrate that the proposed method led to more accurate noise PSD estimation and perceptual evaluation of speech quality scores in various diffuse noise environments, and did not degrade the speech quality under the presence of directional interference, although the proposed method utilizes the coherence information.

Keywords: noise PSD estimation; coherence; dual channel speech enhancement; postfilter; speech presence probability estimation

Citation: Cheong, S.; Kim, M.; Shin, J.W. Postfilter for Dual Channel Speech Enhancement Using Coherence and Statistical Model-Based Noise Estimation. *Sensors* **2024**, *24*, 3979. https://doi.org/10.3390/s24123979

Academic Editor: Alicja Wieczorkowska

Received: 11 May 2024
Revised: 18 June 2024
Accepted: 18 June 2024
Published: 19 June 2024

1. Introduction

Over the past decades, there has been a growing demand for speech enhancement using microphone arrays in speech processing applications such as automatic speech recognition, mobile communications, and hearing aids [1–4]. Multichannel speech enhancement aims to reduce the additive noise and improve the quality of the speech signals obtained by multiple microphones placed in a variety of acoustic environments [5–32]. In many multichannel speech enhancement systems, beamforming algorithms, such as the minimum-variance distortionless-response (MVDR) beamformer [11] and the general transfer function generalized sidelobe canceler (TF-GSC) [12,13], have been employed to extract a desired signal, exploiting spatial information on the location of the sound sources. Although these beamformers successfully reduce the interfering noise without creating too much speech distortion, the amount of noise suppression is not very high in general, and nonstationary interferences and diffuse noises may disrupt some of the beamformers. Therefore, an additional postfilter is usually used to further enhance the output of the beamformer [14–28]. It has been shown that the multichannel Wiener filter (MWF) can be factorized into the MVDR beamformer and a single-channel Wiener postfilter [14,15].

The postfilters used in the literature include the Wiener filter [16–19], short-time spectral amplitude (STSA) estimator [20], and optimally modified log-spectral amplitude (OM-LSA) [21] estimator, for all of which the accurate estimation of the noise power spectral

density (PSD) in the beamformer output is crucial. The noise PSD estimation approaches for the postfilter can be classified into two categories. The methods falling into the first category are essentially single-channel approaches which estimate the noise PSD from the output of the beamformer [22,23] using single-channel noise estimation approaches [33–37]. The advantage of these approaches is that the performance of the postfilter is not severely affected by the steering error of the beamformer. However, the single-channel noise PSD estimation approaches cannot rapidly track the changes in the noise statistics and thus, the noise PSD is underestimated for nonstationary noises, resulting in insufficient noise suppression. The methods in the second category are multichannel approaches which utilize spatial information from the microphone signals or beamformers. In [18], the noise PSD is estimated by a recursive averaging of the power spectrum of the null-beamformed signal. In [21], the noise PSD estimate is obtained by a recursive averaging of the periodogram of the beamformer output, with a smoothing factor dependent on the speech presence probability (SPP). The SPP in [21] is affected by transient beam-to-reference ratio (TBRR), which is the ratio of the transient powers in the beamformer output and the noise reference signals of the TF-GSC obtained by applying the minima controlled recursive averaging (MCRA) to those signals. These methods utilizing the beamformer output signals can effectively deal with moderately nonstationary noises, but the performance deteriorates when the steering error occurs in the beamformer or the noise is highly nonstationary. The leakage of the speech signal into the noise reference leads to the speech attenuation in the postfilter, which is more crucial for the quality of the enhanced speech. On the other hand, the noise PSD estimation based on the microphone signals in [16,17] essentially estimates the PSD of the noise in the microphone signals, which differs from the PSD of the noise in the beamformer output. In [15], the noise PSD at the beamformer output is estimated from the PSD of the microphone signals, room transfer function, and noise coherence matrix, and a two-step approach to estimate the Wiener postfilter is proposed based on the maximum likelihood approach and the Bayesian refinement. While it provides a novel mathematical framework to estimate the noise PSD and obtain the postfilter, it is assumed that the noise coherence matrix is known in advance. There have been several approaches to apply those filters using dual channel noise PSD estimators without applying beamformers [29–32]. Among them, Nelke et al. [29] employ a statistical model-based single-channel noise PSD estimator [36] for low-frequency bins and a coherence-based dual channel noise PSD estimator for high-frequency bins, as the coherences are not discriminative for low frequencies. This method can be applied to the estimation of the noise PSD for the postfilter, but the coherence-based dual channel noise PSD estimator in [29] can only estimate the noise PSD in the microphone signals, which will be higher than the noise PSD in the beamformer output.

In this paper, we propose a postfilter for dual channel speech enhancement combining a statistical model-based single-channel noise estimator and a coherence-based dual channel estimator with a SPP. Specifically, we model the coherence-based *a posteriori* SPP, and combine it with the statistical model-based SPP [36]. We then derive the coherence-based dual channel noise PSD estimator considering the speech presence uncertainty and the difference between the noise PSDs in the microphone signals and the beamformer output. Finally, the spectral gain function of the postfilter is computed by utilizing the noise PSD estimate and *a posteriori* SPP based on both the statistical models and the coherences.

2. System Overview and Review of SPP-Based Noise Estimation

2.1. Problem Formulation and System Overview

Assuming that two microphones capture the desired speech along with the uncorrelated additive noise, the two microphone signals in the STFT domain in a vector form, $\mathbf{Z}(l,k) = [Z_1(l,k), Z_2(l,k)]^T$, with a time index l and a frequency index k, can be written as

$$\begin{aligned} \mathbf{Z}(l,k) &= \mathbf{g}(l,k)S_1(l,k) + \mathbf{V}(l,k) \\ &= \mathbf{S}(l,k) + \mathbf{V}(l,k), \qquad 1 \le l \le L, 1 \le k \le K \end{aligned} \tag{1}$$

where $\mathbf{S}(l,k) = [S_1(l,k), S_2(l,k)]^T$ is the clean speech at the microphones including early reflections, $\mathbf{V}(l,k) = [V_1(l,k), V_2(l,k)]^T$ is the additive noise at the microphones including late reverberations, and $\mathbf{g}(l,k) = [1, g_2(l,k)]^T$ is the relative transfer function (RTF) vector. $\mathbf{S}(l,k)$ and $\mathbf{V}(l,k)$ are assumed to be uncorrelated as in many research works [6,15,19,20,22]. The microphone signals are processed by a filter-and-sum adaptive beamformer $\mathbf{W}^H(l,k) = [W_1^*(l,k), W_2^*(l,k)]$ to produce the beamformer output $Y(l,k) = \mathbf{W}^H(l,k)\mathbf{Z}(l,k)$. $Y(l,k)$ can be considered to be the sum of the filtered speech $X(l,k)$ and the residual noise $N(l,k)$, which are assumed to be mutually uncorrelated as

$$\begin{aligned} Y(l,k) &= \mathbf{W}^H(l,k)\mathbf{Z}(l,k) \\ &= \mathbf{W}^H(l,k)\mathbf{S}(l,k) + \mathbf{W}^H(l,k)\mathbf{V}(l,k) \\ &\triangleq X(l,k) + N(l,k). \end{aligned} \tag{2}$$

The goal of the postfilter is usually to estimate $X(l,k)$ from $Y(l,k)$ with the help of $\mathbf{Z}(l,k)$, although $X(l,k)$ may contain speech distortion to an extent. The output of the postfilter is given as

$$\widehat{X}(l,k) = G(l,k)Y(l,k) \tag{3}$$

in which $G(l,k)$ is the spectral gain of the postfilter. The block diagram of a dual channel speech enhancement system with an adaptive beamformer and a postfilter is illustrated in Figure 1.

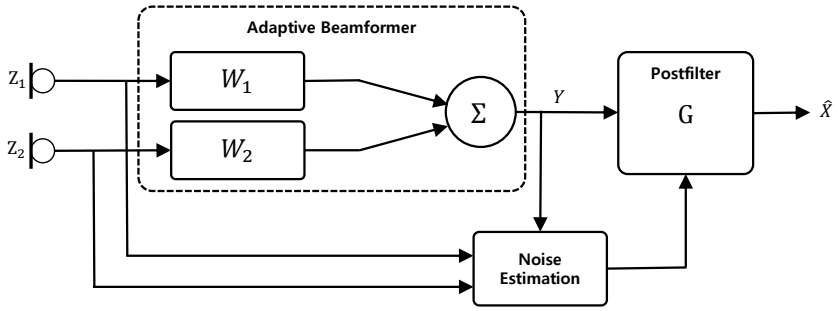


Figure 1. General block diagram of a dual channel speech enhancement system with a beamformer and a postfilter.

2.2. Single-Channel Noise PSD Estimator Based on Speech Presence Probability

In [36], a statistical model-based single-channel noise PSD estimation using a fixed a priori signal-to-noise ratio (SNR) for speech presence is proposed. Under the assumption that the speech and noise STFT coefficients are distributed according to the complex Gaussian distributions with zero means, the likelihood functions for the hypotheses of speech presence H_1 and speech absence H_0 are modeled as

$$f(Y(l,k) \mid H_0) = \frac{1}{\widehat{\lambda}_{n_s}(l,k)\pi} exp\left(-\frac{|Y(l,k)|^2}{\widehat{\lambda}_{n_s}(l,k)}\right) \tag{4}$$

$$f(Y(l,k) \mid H_1) = \frac{1}{\widehat{\lambda}_{n_s}(l,k)(1+\xi_{H_1})\pi} \cdot exp\left(-\frac{|Y(l,k)|^2}{\widehat{\lambda}_{n_s}(l,k)(1+\xi_{H_1})}\right) \tag{5}$$

where ξ_{H_1} indicates the fixed a priori SNR, which represents the SNR "if speech were present" [36], and $\hat{\lambda}_{n_s}(l,k)$ is the estimate of the noise PSD based on statistical modeling.

According to Bayes' rule, *a posteriori* SPP $P(H_1|Y)$, which is a function of $|Y|$ and thus denoted as $p_{|Y|}(l,k)$ in the next section, is obtained as

$$P(H_1||Y|) = p_{|Y|}(l,k) = \left\{1 + (1+\xi_{H_1})exp\left(-\frac{|Y(l,k)|^2}{\widehat{\lambda}_{n_s}(l-1,k)}\frac{\xi_{H_1}}{\xi_{H_1}+1}\right)\right\}^{-1} \tag{6}$$

where the parameter ξ_{H_1} is set to be 15 dB in the experiments, which is obtained by minimizing the total risk of error as in [36], and a priori probability of speech presence $P(H_1)$ is assumed to be 1/2. To allow the adaptation of the noise PSD estimate when the noise PSD is underestimated, $p_{|Y|}(l,k)$ is constrained to be less than 0.99 when smoothed *a posteriori* SPP is higher than 0.99. With *a posteriori* SPP, the noise PSD periodogram in the current frame is estimated as a weighted summation of the noise PSD estimate from the previous frame $\hat{\lambda}_{n_s}(l-1,k)$ and the power of the beamformer output signal $|Y(l,k)|^2$:

$$\widetilde{\lambda}_{n_s}(l,k) = p_{|Y|}(l,k) \cdot \hat{\lambda}_{n_s}(l-1,k) + (1 - p_{|Y|}(l,k)) \cdot |Y(l,k)|^2. \tag{7}$$

Finally, the noise PSD estimate is obtained by recursive smoothing with a smoothing parameter α_{sm} as

$$\hat{\lambda}_{n_s}(l,k) = \alpha_{sm}\hat{\lambda}_{n_s}(l-1,k) + (1-\alpha_{sm})\widetilde{\lambda}_{n_s}(l,k). \tag{8}$$

3. Postfilter for Dual Channel Speech Enhancement Utilizing Noise Estimation Based on Coherence and Statistical Model

We propose a postfilter for dual channel speech enhancement, combining a statistical model-based single-channel noise estimate and a coherence-based dual channel noise estimate. Assuming that the phase of the coherence between two microphone signals, which is the same as the phase difference between them in the short-time Fourier transform (STFT) domain, is already exploited well by the beamformer, we focus on the magnitude of the coherence as spatial information in the proposed postfilter. Firstly, we model the coherence-based *a posteriori* SPP as a simple function of the magnitude of the coherence, and combine it with the SPP based on the statistical modeling of the beamformer in (6). Then, we derive a dual channel noise PSD estimator for speech presence periods based on coherence, and obtain a noise PSD estimate considering speech presence uncertainty utilizing the *a posteriori* SPP. The final noise PSD estimate is constructed by combining the coherence-based estimate and the statistical model-based single-channel estimate utilizing the coherence-based *a posteriori* SPP. The OM-LSA gain function is used as the postfilter, utilizing the combined noise estimates and the combined *a posteriori* SPP. The block diagram of the proposed method is presented in Figure 2.

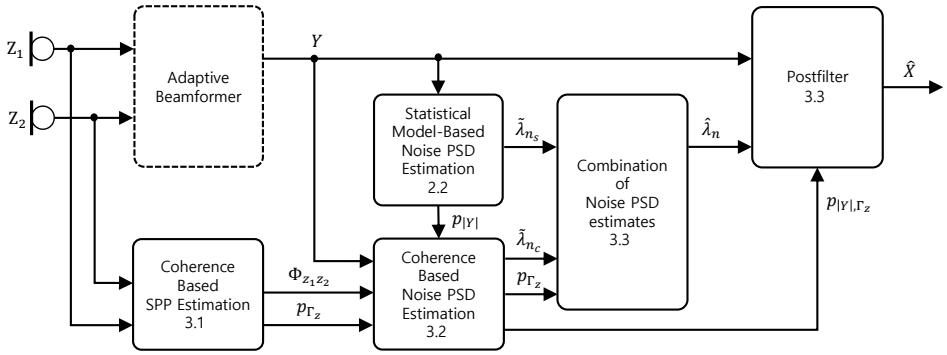


Figure 2. Block diagram of the dual channel speech enhancement system employing the proposed postfilter.

3.1. Modeling of a Posteriori SPP Based on Coherence

One of the spatial properties that may be used to distinguish signals with different spatial characteristics is the coherence. For two microphone signals Z_1 and Z_2, the coherence between them is defined as

$$\Gamma_z(l,k) = \frac{\Phi_{z_1z_2}(l,k)}{\sqrt{\Phi_{z_1z_1}(l,k)\Phi_{z_2z_2}(l,k)}} \tag{9}$$

where $\Phi_{z_1z_2}$ is the cross PSD of Z_1 and Z_2 and $\Phi_{z_1z_1}$ and $\Phi_{z_2z_2}$ are the auto PSDs, which can be estimated by temporal smoothing. While the phase of the coherence is related to the

inter-channel time difference of arrival for a single directional signal, the magnitude of the coherence is related to how many signals from point sources and image sources accounting for reflections are mixed in the corresponding time–frequency bin. For example, any signals from a point source without reverberation show the magnitude of coherence $|\Gamma_z(l,k)|$ to be 1. Another useful example is the spherically isotropic or diffuse noise, for which the coherence function can be derived as [38]

$$\Gamma_z^{diffuse}(l,k) = sinc(\frac{2\pi k f_s d_{mic}}{2Kc}) \tag{10}$$

where $2K$ is the size of the discrete Fourier transform (DFT), f_s is the sampling frequency, d_{mic} is the distance between microphones, and c is the speed of sound. As the interchannel phase differences, which are the phases of the coherences, are already exploited in the beamformers, we focus on the magnitudes of coherences in the postfilter to utilize complementary information on the spatial characteristics. It is also noted that the directional interferences are taken care of in the adaptive beamformers, and therefore, diffuse noises may be the main obstacles that remain in the beamformer output, which can be effectively discriminated from desired speech using the magnitude of coherence, except low frequencies.

In this paper, it is assumed that the target speaker is located closer to a microphone array than other point sources generating directional interferences. As the distance between a sound source and microphones increases, the magnitude of the coherence decreases due to reverberation. In this regard, the magnitude of the coherence would be high if speech is present in that time–frequency bin, and low when only directional interferences or diffuse noises exist. In this paper, we model the coherence-based *a posteriori* SPP as a simple function of the magnitude of the coherence. As the magnitudes of the coherence in individual time–frequency bins may be vulnerable to the local SNR and reverberation and are not discriminative enough in the low-frequency bins, it is beneficial to aggregate the coherences in all frequency bins to determine frame-wise voice activity and apply separate functions to model *a posteriori* SPP depending on the voice activity. Let $\Gamma_z^{(f)}(l) = \frac{1}{K}\sum_{k=1}^{K}|\Gamma_z(l,k)|$ be the frame-wise coherence measure to decide voice activity. The *a posteriori* SPP based on the coherence, $P(H_1|\Gamma_z)$ or $p_{\Gamma_z}(l,k)$, is modeled as

$$P(H_1|\Gamma_z) = p_{\Gamma_z}(l,k) = \begin{cases} \alpha \cdot |\Gamma_z(l,k)| + \alpha_{min} & \text{if } \Gamma_z^{(f)}(l) > \eta \\ \beta \cdot |\Gamma_z(l,k)| & \text{otherwise} \end{cases} \tag{11}$$

where η is the threshold to apply different functions, and α, α_{min} and β are experimentally determined constants between 0 and 1 with $\alpha + \alpha_{min} \leq 1$. For simplicity, the $p_{\Gamma_z}(l,k)$ is designed as a linear combination of $|\Gamma_z(l,k)|$ and 1 or 0 for speech presence or absence, respectively. This coherence-based *a posteriori* SPP is used in the coherence-based noise PSD estimation introduced in Section 3.2 and the combination of the statistical model-based and coherence-based noise PSD estimates explained in Section 3.3.

3.2. Proposed Dual Channel Noise PSD Estimator Based on Coherence

In order to estimate the noise PSD from dual microphone signals, the noise PSDs for high frequencies were derived as a function of the coherences of the speech and noise and the auto- and cross-PSDs of the microphone signals, while the noise PSD in low frequencies were estimated using an SPP computed from the first microphone signal in [29].In this paper, we formulate the PSD of the residual noise in the beamformer output $N(l,k)$ as a function of the pseudo-coherences of speech and noise considering the difference in the noise PSD in the microphone signals and the beamformer output and the speech presence uncertainty.

Assuming that the desired speech and the background noise are uncorrelated, the cross PSD of the microphone signals and the PSD of the beamformer output signal can be described as

$$\Phi_{z_1z_2}(l,k) = \Phi_{s_1s_2}(l,k) + \Phi_{v_1v_2}(l,k) \tag{12}$$

$$\Phi_y(l,k) = \Phi_x(l,k) + \Phi_n(l,k) \tag{13}$$

where $\Phi_{s_1s_2}(l,k)$ and $\Phi_{v_1v_2}(l,k)$ indicate the cross PSD of speech and noise at the first and second microphones, while $\Phi_x(l,k)$ and $\Phi_n(l,k)$ indicate the PSD of speech and noise at the beamformer output, respectively. For speech present regions, we can rewrite the cross PSD of the microphone signals in (12) using the PSDs of the beamformed speech and noise $\Phi_x(l,k)$ and $\Phi_n(l,k)$ in a similar way to [29]

$$\Phi_{z_1z_2}(l,k) = \Lambda_x(l,k)\ \Phi_x(l,k) + \Lambda_n(l,k)\ \Phi_n(l,k) \tag{14}$$

where $\Lambda_x(l,k)$ and $\Lambda_n(l,k)$ are the pseudo-coherence of speech and noise considering the effect of beamformers defined as

$$\Lambda_x(l,k) = \frac{\Phi_{s_1s_2}(l,k)}{\Phi_x(l,k)} \tag{15}$$

$$\Lambda_n(l,k) = \frac{\Phi_{v_1v_2}(l,k)}{\Phi_n(l,k)}, \tag{16}$$

which are not the same as the coherence of the speech and noise at the microphones given by

$$\Gamma_s(l,k) = \frac{\Phi_{s_1s_2}(l,k)}{\sqrt{\Phi_{s_1s_1}(l,k)\Phi_{s_2s_2}(l,k)}} \tag{17}$$

$$\Gamma_v(l,k) = \frac{\Phi_{v_1v_2}(l,k)}{\sqrt{\Phi_{v_1v_1}(l,k)\Phi_{v_2v_2}(l,k)}}. \tag{18}$$

Comparing with the original definition of the coherence, we can see that the denominators, the geometric means of the PSDs for speech and noise in two microphone signals in (17) and (18), are replaced by the PSDs of the speech and noise in the beamformer output in (15) and (16).

From Equations (13) and (14), the coherence-based estimate of $\Phi_n(l,k)$ for speech presence, $\Phi^{coh}_{n|H_1}(l,k)$, can be written as a function of $\Phi_y(l,k)$, $\Phi_{z_1z_2}(l,k)$, $\Lambda_x(l,k)$ and $\Lambda_n(l,k)$:

$$\Phi^{coh}_{n|H_1}(l,k) = \frac{\Phi_{y|H_1}(l,k) - \frac{\Phi_{z_1z_2|H_1}(l,k)}{\Lambda_x(l,k)}}{1 - \frac{\Lambda_n(l,k)}{\Lambda_x(l,k)}}. \tag{19}$$

As for the speech absent regions, the instantaneous value for the power spectrum of the beamformer output provides the most accurate estimate for the noise PSD in the beamformer output. Therefore, the final dual channel noise PSD based on coherences, $\widetilde{\lambda}_{n_c}(l,k)$, is given as a linear combination of $\Phi^{coh}_{n|H_1}(l,k)$ and $|Y(l,k)|^2$ in which the weights are determined by the *a posteriori* SPP considering both the beamformed signal and the coherence for microphone signals, $p_{|Y|,\Gamma_z}(l,k) = P(H_1||Y|,\Gamma_z)$, as follows:

$$\widetilde{\lambda}_{n_c}(l,k) = p_{|Y|,\Gamma_z}(l,k) \cdot \hat{\Phi}^{coh}_{n|H_1}(l,k) + (1 - p_{|Y|,\Gamma_z}(l,k)) \cdot |Y(l,k)|^2. \tag{20}$$

Assuming that both $|Y(l,k)|$ and $\Gamma_z(l,k)$ would indicate speech presence if speech is present in that time–frequency bin, $P(H_1||Y|,\Gamma_z)$ is represented as the product of the *a posteriori* SPP based on the magnitude spectrum of the beamformed signal $P(H_1||Y|)$ in (6), and the *a posteriori* SPP based on the coherence $P(H_1|\Gamma_z)$ in (11), i.e.,

$$p_{|Y|,\Gamma_z}(l,k) = p_{|Y|}(l,k) \cdot p_{\Gamma_z}(l,k) \tag{21}$$

To evaluate (19), the pseudo-coherences of the speech and noise $\Lambda_x(l,k)$ and $\Lambda_n(l,k)$ need to be estimated in addition to $\Phi_{y|H_1}(l,k)$ and $\Phi_{z_1z_2|H_1}(l,k)$, which can be obtained by the temporal smoothing of $|Y(l,k)|^2$ and $Z_1(l,k)Z_2^*(l,k)$ for speech presence periods. The pseudo-coherences can be estimated in a similar way to the coherence estimation in [29] using the SPP in (21). We omit the frame and frequency indices for brevity. The estimate for noise pseudo-coherence, $\hat{\Lambda}_n$, is updated with a smoothing parameter α_Λ during noise-only periods determined by the *a posteriori* SPP $p_{|Y|,\Gamma_z}$ as

$$\hat{\Lambda}_n = \alpha_\Lambda \cdot \hat{\Lambda}_{n,last} + (1-\alpha_\Lambda)\,\hat{\Lambda}_{Y|H_0}, \;\; \text{if } p_{|Y|,\Gamma_z} < p_{th_1} \tag{22}$$

where p_{th_1} is a threshold to update $\hat{\Lambda}_n$, $\hat{\Lambda}_{n,last}$ denotes the estimate of noise pseudo-coherence in the frame it was updated lastly, and $\hat{\Lambda}_{Y|H_0}$ denotes the estimate for the pseudo-coherence of the noisy signal for speech absent regions considering the effect of beamformers defined as

$$\hat{\Lambda}_{Y|H_0} = \frac{\hat{\Phi}_{z_1 z_2|H_0}}{\hat{\Phi}_{y|H_0}}, \tag{23}$$

in which $\Phi_{y|H_0}(l,k)$ and $\Phi_{z_1 z_2|H_0}(l,k)$ are estimated by the temporal smoothing of $|Y|^2$ and $Z_1 Z_2^*$ during noise-only periods in a similar way to (22).

The estimation of Λ_x is not as straightforward as that for Λ_n because the background noises reside also in the speech-active regions. The pseudo-coherence of the noisy signal for speech present regions, $\Lambda_{Y|H_1}$, can be expressed as

$$\begin{aligned}
\Lambda_{Y|H_1} &= \frac{\Phi_{z_1 z_2|H_1}}{\Phi_{y|H_1}} \\
&= \frac{\Phi_{s_1 s_2} + \Phi_{v_1 v_2}}{\Phi_x + \Phi_n} \\
&= \frac{\Phi_{s_1 s_2}}{\Phi_x}\left(\frac{\Phi_x}{\Phi_x + \Phi_n}\right) + \frac{\Phi_{v_1 v_2}}{\Phi_n}\left(\frac{\Phi_n}{\Phi_x + \Phi_n}\right) \\
&= \Lambda_x\left(\frac{\gamma}{1+\gamma}\right) + \Lambda_n \frac{1}{1+\gamma}
\end{aligned} \tag{24}$$

where $\gamma = \frac{\Phi_x}{\Phi_n}$ is the SNR at the beamformer output, which is estimated using the statistical model-based noise PSD estimate in (8) as

$$\hat{\gamma} = \frac{\hat{\Phi}_{y|H_1}}{\hat{\lambda}_{n_s}} - 1 \tag{25}$$

in which $\hat{\Phi}_{y|H_1}$ is obtained by the temporal smoothing of beamformer outputs in speech-active periods, i.e., the time–frequency bins with $p_{|Y|,\Gamma_z} > p_{th_2}$. The left-hand side of (24), $\Lambda_{Y|H_1}$, can also be estimated as $\frac{\hat{\Phi}_{z_1 z_2|H_1}}{\hat{\Phi}_{y|H_1}}$, in which $\hat{\Phi}_{z_1 z_2|H_1}$ is obtained in a similar way to $\hat{\Phi}_{y|H_1}$. Then, the speech pseudo-coherence $\Lambda_x(l,k)$ can be estimated according to (24) using $\hat{\Lambda}_{Y|H_1}$, $\hat{\gamma}$ in (25), and $\hat{\Lambda}_n$ in (22) with additional temporal smoothing in the speech presence periods as

$$\hat{\Lambda}_x = \alpha_\Lambda \cdot \hat{\Lambda}_{x,last} + (1-\alpha_\Lambda)\left[\hat{\Lambda}_{Y|H_1}\frac{\hat{\gamma}+1}{\hat{\gamma}} - \hat{\Lambda}_n \frac{1}{\hat{\gamma}}\right], \;\; \text{if } p_{|Y|,\Gamma_z} > p_{th_2}. \tag{26}$$

3.3. Combining Noise PSD Estimates and Gain Calculation

The proposed dual channel noise PSD estimator based on coherence in (20) shows different characteristics from the single-channel SPP-based noise PSD estimator in (7). Figure 3 shows one example of the noise power spectrum in the beamformer output and the estimates of it for Cafeteria noise at 5 dB SNR. The single-channel SPP-based estimate $\widetilde{\lambda}_{n_s}$ in Figure 3b seems to be stable and reliable, while it cannot track abrupt changes in the noise power spectrum. In contrast, the dual channel coherence-based estimate $\widetilde{\lambda}_{n_c}$ in Figure 3c could deal with rapidly changing noises, but occasionally, a certain portion of the speech power spectrum is included in the noise power spectrum estimate. The speech leakage in the noise power spectrum estimate leads to speech distortion when applying the postfilter, and thus is much more critical for the quality of the enhanced speech than the underestimation of the noise statistics. Therefore, we combine two noise power spectrum estimates $\widetilde{\lambda}_{n_s}$ and $\widetilde{\lambda}_{n_c}$ as

$$\widetilde{\lambda}_n(l,k) = p_{\Gamma_z}(l,k) \cdot \widetilde{\lambda}_{n_s}(l,k) + (1 - p_{\Gamma_z}(l,k)) \cdot \widetilde{\lambda}_{n_c}(l,k), \tag{27}$$

so that it follows $\widetilde{\lambda}_{n_s}$ in the presence of speech signal and becomes $\widetilde{\lambda}_{n_c}$ when it is certain that speech is absent. It is noted that we use p_{Γ_z} instead of $p_{|Y|,\Gamma_z}$ as the weight to react faster to the speech onsets. $\widetilde{\lambda}_n$ is shown in Figure 3d, which looks similar to the true noise power spectrum at the beamformer output $|N(l,k)|^2$ in Figure 3a compared with $\widetilde{\lambda}_{n_s}$ and $\widetilde{\lambda}_{n_c}$.

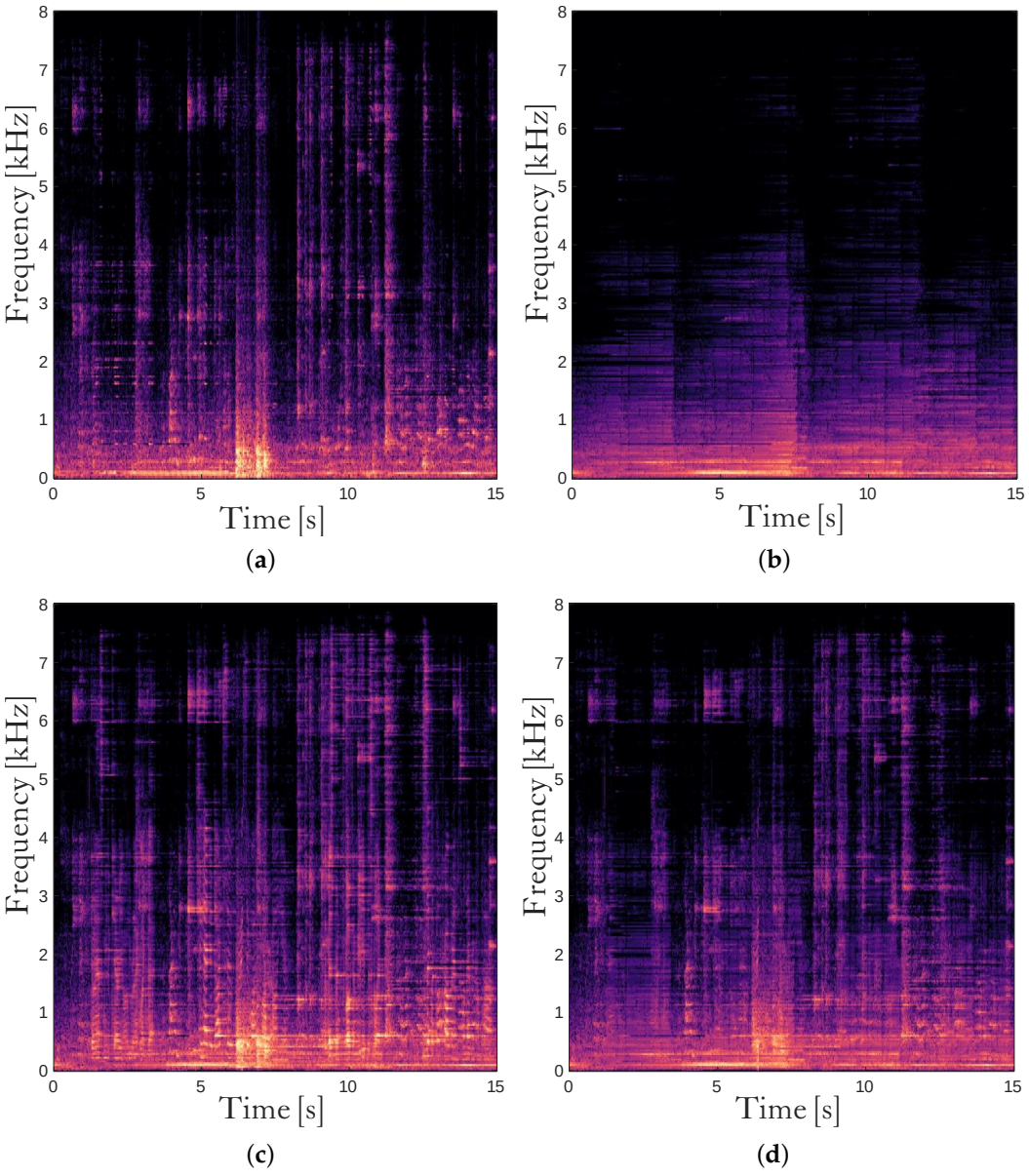


Figure 3. Noise power spectrum and the estimates of it before temporal smoothing for Cafeteria noise at 5 dB SNR in the smaller simulated room. (**a**) True noise power spectrum at the beamformer output, (**b**) the single-channel SPP-based estimate in (7), (**c**) the coherence-based estimate in (20), and (**d**) the combined estimate in (27).

The final estimate of the noise PSD $\hat{\lambda}_n$ is obtained by the temporal smoothing of $\widetilde{\lambda}_n$ with a smoothing parameter α_n as

$$\hat{\lambda}_n(l,k) = \alpha_n \hat{\lambda}_n(l-1,k) + (1-\alpha_n)\widetilde{\lambda}_n(l,k). \tag{28}$$

Using the noise PSD estimate $\hat{\lambda}_n$ in (28) and the *a posteriori* SPP $p_{|Y|,\Gamma_z}$ in (21), the gain function of the postfilter G can be computed as the OM-LSA speech estimator [39]

$$G(l,k) = \left\{\max(\tilde{G}(l,k), G_{min})\right\}^{p_{|Y|,\Gamma_z}(l,k)} \cdot G_{min}^{1-p_{|Y|,\Gamma_z}(l,k)} \tag{29}$$

where G_{min} indicates a minimum value for the gain in speech absent periods and $\tilde{G}(l,k)$ is the spectral gain function of the minimum mean-square error short-time log spectral amplitude (MMSE-LSA) estimator given by [40]

$$\tilde{G}(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)} \exp\left[\frac{1}{2}\int_{v(l,k)}^{\infty} \frac{e^{-t}}{t} dt\right] \tag{30}$$

where $v(l,k) \triangleq \gamma(l,k)\xi(l,k)$, $\xi(l,k) \triangleq E|X(l,k)|^2/\hat{\lambda}_n(l,k)$ indicates the a priori SNR at the beamformer output estimated using a decision-directed (DD) approach [5], and $\gamma(l,k) \triangleq |Y(l,k)|^2/\hat{\lambda}_n(l,k)$ is the *a posteriori* SNR.

4. Experimental and Results

4.1. Experimental Configurations

To demonstrate the performance of the proposed coherence and statistical model-based dual channel noise PSD estimator and the postfilter, we simulated two rooms of dimensions 6.7 m × 6.1 m × 2.9 m and 9 m × 7.5 m × 3.5 m using the image method [41,42]. The reverberation times were RT_{60} = 300 ms and RT_{60} = 500 ms, respectively. The microphones were located at (3 m, 3 m, 1.5 m) and (3.14 m, 3 m, 1.5 m) for both of the rooms, which corresponded to the form factor of the modern smartphones in the landscape orientation. We assumed the "hand-held handsfree" scenario [43] in which the desired speaker was located at the broadside of the microphone array, 0.4 m away from the center of the microphones.

In addition, we also utilized a real-recorded room impulse response (RIR) from the multi-channel impulse response database (MIRD) [44] with the room dimensions of 6 m × 6 m × 2.4 m and the RT_{60} of 360 ms. The desired speaker was assumed to be located 1 m away from the center of two microphones at the broadside direction. The 1 m distance was not a typical one for the "hand-held handsfree" use cases and was not favorable to the proposed method utilizing the coherence information, but we could not find a more suitable real-recorded RIR database. The distance between the two microphones we utilized was 14 cm as in the simulated RIR cases, in accordance with the size of the recent smartphones.

Twelve utterances from the TIMIT database [45] were used as desired speech signals, and the Cafeteria, Crossroad, Kindergarten 1, Pub, Train Station, Callcenter, and Mensa noises from ES 202 396-1 [46] were used to generate diffuse noises using the arbitrary noise field generator [47]. The SNRs for diffuse noises were −5, 0, 5, 10, and 15 dB. The signals were sampled at 16 kHz, and 512-point STFT was applied to the 32 ms of windowed signal with 20 ms frame shift, in which the Tukey window with the cosine fraction of 75% was adopted.

We compared the performance of the proposed postfilter to those of the postfilter using the TBRR-based multichannel noise PSD estimator [21], which is denoted as *TBRR*, and the one adopting the SPP-based single-channel noise estimator introduced in Section 2.2 [36], which is denoted as *Single-SPP*. Although there have been several recent research studies on better spatial filtering [9,10], not much effort has been devoted to improve the postfilters for the spatial filtering recently, except for the deep learning-based approaches. Deep learning-based postfilters using single-channel [24,25] or multichannel information [23,27,28] have been proposed, but these approaches often require high computational complexity and large training datasets.

The general transfer function-generalized sidelobe canceler (TF-GSC) [21] was used as the adaptive beamformer with the same parameter values as in [21]. The OM-LSA speech estimator in (29) and (30) was employed as a postfilter for all three methods, in which different SPP and noise PSD estimators were adopted. The parameter values for the proposed method used for the experiments are summarized in Table 1. The smoothing parameters to compute $\Phi_{y|H_0}$ and $\Phi_{z_1z_2|H_0}$ in (23) were 0.4, while those to obtain $\Phi_{y|H_1}$ and $\Phi_{z_1z_2|H_1}$ that were used to compute $\hat{\Lambda}_{Y|H_1}$ and $\hat{\gamma}$ were 0.7. The parameters for the compared methods were selected to maximize the average PESQ score for the enhanced speech, which were the same as the values in the original papers, except α_λ was 0.8 instead of 0.85 in [21].

As the performance measure for the noise PSD estimation accuracy, we used the segmental logarithmic error (LogErr) defined by

$$\text{LogErr} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\left|10\,log_{10}\left[\frac{\lambda_n(l,k)}{\hat{\lambda}_n(l,k)}\right]\right| \tag{31}$$

where $\lambda_n(l,k)$ indicates the true noise PSD at the beamformer output obtained by processing the noise with the TF-GSC computed for the noisy input and applying temporal smoothing in (28). The logarithmic error can be represented as a summation of the overestimation error $\text{LogErr}_{\text{ov}}$ and the underestimation error $\text{LogErr}_{\text{un}}$ of the noise PSD, which are defined as [36]

$$\text{LogErr}_{\text{ov}} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\left|\min\left(0, 10\,log_{10}\left[\frac{\lambda_n(l,k)}{\hat{\lambda}_n(l,k)}\right]\right)\right| \tag{32}$$

$$\text{LogErr}_{\text{un}} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\max\left(0, 10\,log_{10}\left[\frac{\lambda_n(l,k)}{\hat{\lambda}_n(l,k)}\right]\right). \tag{33}$$

The $\text{LogErr}_{\text{ov}}$ may indicate the degree of speech attenuation caused by the postfilter, while $\text{LogErr}_{\text{un}}$ would be related to the amount of residual noises. As for the speech enhancement performance, the ITU-T Recommendation P.862.2 wideband perceptual evaluation of speech quality (PESQ) [48] scores and the segmental SNR (SSNR) improvement were evaluated. The SSNR is defined as

$$\text{SSNR} = \frac{10}{|\mathbb{L}|}\sum_{l\in\mathbb{L}}\log_{10}\frac{\sum_{n=1}^{N}s^2(lN+n)}{\sum_{n=1}^{N}(\hat{s}(lN+n)-s(lN+n))^2} \tag{34}$$

where $N = 160$, $\mathbb{L}$ is the set of speech active segments and $s(l)$, and $\hat{s}(l)$ indicate the clean speech signal and the estimate of it in the time domain, respectively.

As the proposed method relies on the coherence information and the coherences for the directional interferences would be higher than those for the diffuse noises, the performance for the proposed method may deteriorate in the presence of directional interference. To demonstrate that the proposed system does not show inferior performance to the previous approaches in the presence of both diffuse noise and directional interference, we conducted an additional experiment on the smaller simulated room. Directional interference was constructed by the image method [41,42] using three utterances from the Wall Street Journal (WSJ0) dataset [49] in which the noise source was located at -30° from the broadside direction, 0.8 m away from the center of the microphones. The signal-to-interference ratios (SIRs) for the directional interferences were 0, 5, and 10 dB.

Table 1. Parameters for the proposed statistical model and coherence-based postfilter.

α_{sm}	α	α_{min}	β	η	α_Λ	α_n	p_{th_1}	p_{th_2}	G_{min}
0.8	0.75	0.2	0.5	0.377	0.95	0.8	0.1	0.6	−9 dB

4.2. Experimental Results

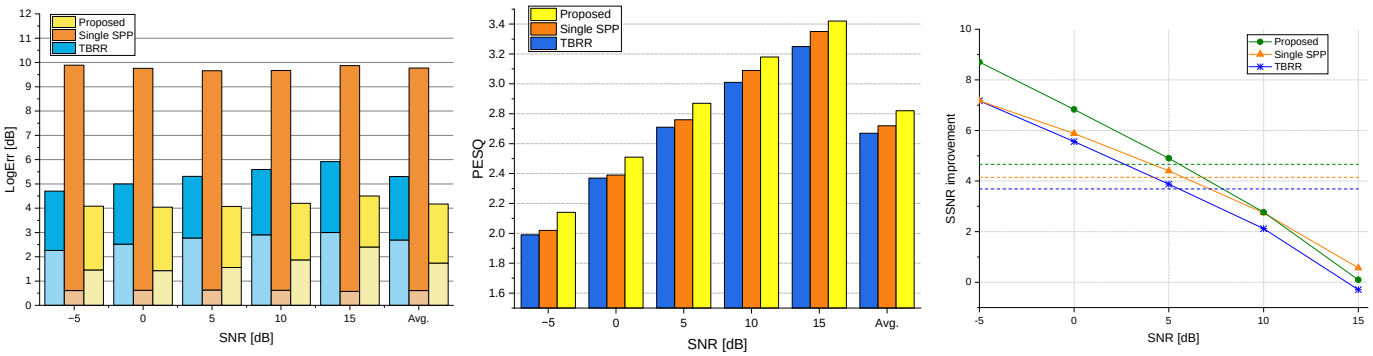
Figures 4–6 show the logarithmic errors, PESQ scores, and SSNR improvements for the proposed and compared methods averaged over seven types of diffuse noise for various SNRs. Figures 4 and 5 are for the simulated rooms with the RT_{60} of 300 ms and 500 ms, respectively, and the distance to the desired speaker of 40 cm. Figure 6 is for the real-recorded RIR with the RT_{60} of 360 ms and the distance to the desired speaker of 1 m. The performance results averaged over all SNR conditions are shown as the rightmost bar graphs or dashed lines. As for the noise PSD estimation accuracy in terms of LogErr, the proposed method exhibited the lowest LogErr for all cases. The *TBRR* showed lower LogErr than the *Single-SPP*, but the $\text{LogErr}_{\text{ov}}$ for the *TBRR* was higher. Compared with the *Single-SPP* [36], the noise underestimation errors for the proposed method were reduced, while the noise overestimation errors were slightly increased. It implies that the proposed

coherence-based noise PSD estimator in (20) could track abrupt change in the noise power spectrum as illustrated in Figure 3c, which resulted in the final noise PSD estimate in (28) close to the true PSD. On the other hand, it can be found that the noise overestimation of the *TBRR* [21] was higher than that of the proposed method, which would be more crucial to the perceptual quality of enhanced speech. Figure 7 shows the logarithmic errors for two highly nonstationary noises, the Kindergarten noise 1 and the Cafeteria noise, and two more stationary noises, the Train Station noise and Crossroad noise, for each SNR averaged over three room conditions. The proposed method marked the lowest LogErr for all noise types and SNRs. The *TBRR* [21] tended to overestimate the noise PSD more in the presence of highly nonstationary noises, although it was originally proposed to tackle relatively nonstationary diffuse noises. The TBRR in the speech active region was occasionally underestimated in the presence of highly nonstationary noises since the transient power in the noise reference of the TF-GSC became high, leading to a low *a posteriori* SPP and overestimation of the noise PSD. The $\text{LogErr}_{\text{un}}$ of the *Single-SPP* increased as the noise became more nonstationary, i.e., from the Crossroad noise in Figure 7d through the Cafeteria noise in Figure 7b to the Kindergarten 1 noise in Figure 7a, as the single-channel noise estimation would regard highly nonstationary noise as speech. It is noted that the Train Station noise in Figure 7c is relatively stationary on average but includes occasional nonstationary events, and thus it did not show a clear tendency among the other three noises. The $\text{LogErr}_{\text{ov}}$ for the proposed method also increased for more nonstationary noises, maybe because the local SNRs in specific time–frequency bins could be very low in highly nonstationary noises for the same input SNR, which led to low *a posteriori* SPP and noise PSD overestimation.

Figure 8 presents the spectrogram of the desired clean speech and the *a posteriori* SPPs estimated by the proposed and competing methods for the concatenation of two utterances in the Cafeteria noise at 5 dB SNR for the smaller simulated room. The *Single-SPP* [36] shows overestimation of the *a posteriori* SPP in many TF bins in Figure 8b, as it cannot easily discriminate the noise onset from the speech onset. The noise onset causes overestimation of the SPP in the speech absent region, which leads to the underestimation of the noise PSD, and the underestimated noise PSD in turn results in the overestimation of the SPP for the upcoming frames. The *TBRR* [21] makes blue horizontal lines in Figure 8c, in which the inaccurate estimation of the acoustic transfer function in the TF-GSC brings about the leakage of the speech to the noise references, which lowers the TBRR together with the transient noises and then results in low *a posteriori* SPP. In contrast, both the coherence-based and statistical model-based noise PSD estimators in the proposed method are not tightly coupled with the performance of the beamformer, and thus occasional failure of the TF-GSC does not affect the postfilter critically. We can also see that the proposed method shows better speech onset detection and a clearer harmonic structure.

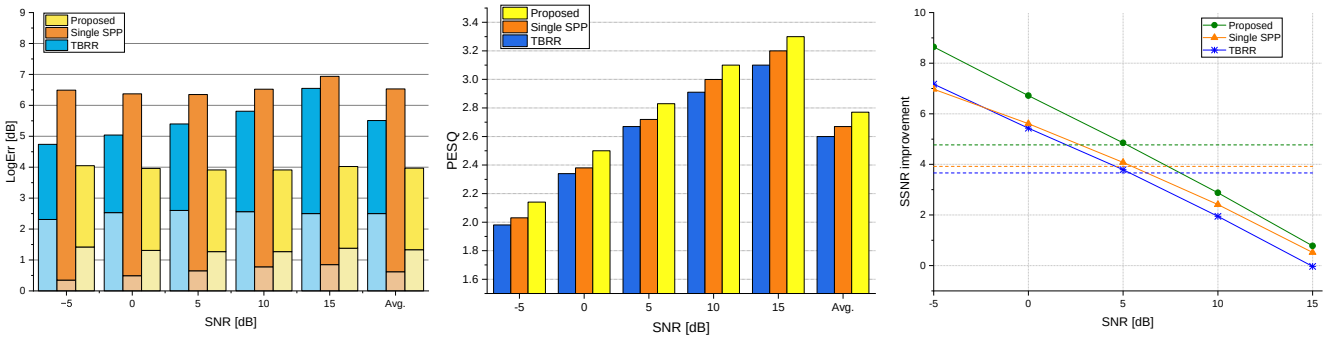
The speech enhancement performances for the proposed and compared postfilters in the presence of diffuse noises are shown in the PESQ scores and the SSNR improvements in Figures 4–6. Since the $\text{LogErr}_{\text{ov}}$ is lower for the *Single-SPP* compared with that for the *TBRR* although the LogErr for the *TBRR* is lower, the average PESQ scores are higher for the *Single-SPP* than those for the *TBRR*. The proposed postfilter with new noise PSD and *a posteriori* SPP estimators result in the highest PESQ scores for all SNRs and all room conditions. The PESQ score for the proposed method averaged over all room and noise conditions is 2.71, 0.09 higher than that for the second best one, *Single-SPP*, with the p-value of 0.014. As the SNR increases, the PESQ score improvement of the proposed method over the *Single-SPP* decreases, whereas the difference between the PESQ scores for the *Single-SPP* and the *TBRR* increases. It may be because the *Single-SPP* tends to underestimate the noise PSD, resulting in less speech attenuation and more residual noise, which is less crucial in the high SNR environments where the noise power is low from the start. In terms of the SSNR improvement, the proposed postfilter demonstrates the best performance for all SNRs and all room conditions except the 15 dB SNR in the smaller simulated room. As the power of the noise in comparison with the speech power is small for the 15 dB

SNR, the power of the residual noise is small with a similar $\mathrm{LogErr}_{\mathrm{un}}$, and thus the SSNR improvement could become high for the *Single-SPP* with low $\mathrm{LogErr}_{\mathrm{ov}}$ and high $\mathrm{LogErr}_{\mathrm{un}}$ as in the PESQ scores. In the same regard, the SSNR improvement for the *Single-SPP* subtracted by that for the *TBRR* increases with the SNR. Although the PESQ scores for the *Single-SPP* are higher than those for the *TBRR* in all conditions, the SSNR improvements for these approaches are comparable, possibly because the speech attenuation caused by the noise PSD overestimation is more critical for the PESQ scores compared with the SSNR improvement. The SSNR improvements for the *TBRR* are higher than those for the *Single-SPP* under the conditions where the $\mathrm{LogErr}_{\mathrm{ov}}$ for the *TBRR* are not high.



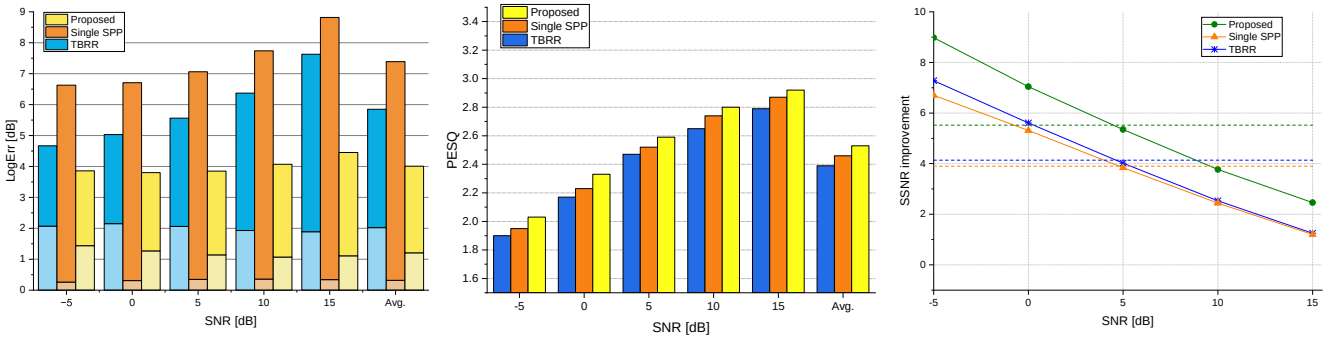
(**a**) LogErr (**b**) PESQ (**c**) SSNR improvement

Figure 4. The logarithmic errors, PESQ scores, and SSNR improvements for the proposed and competing postfilters averaged over all types of diffuse noise in various SNRs for the smaller simulated room with the RT_{60} of 300 ms. The lower and upper bars in the LogErr plot represent the overestimation and underestimation errors, respectively. The average scores are shown as the rightmost bars (**a**,**b**) or dashed lines (**c**).



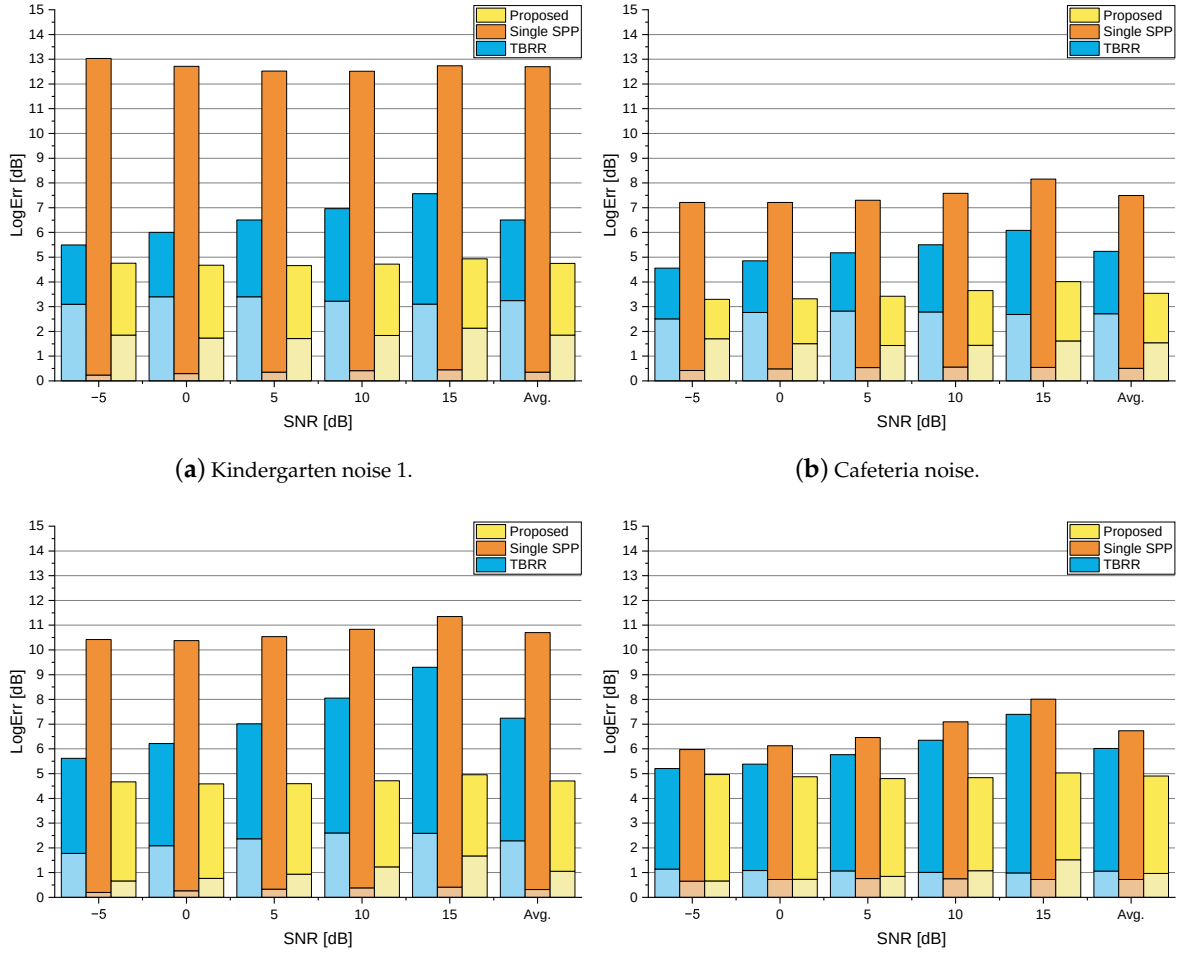
(**a**) LogErr (**b**) PESQ (**c**) SSNR improvement

Figure 5. The logarithmic errors, PESQ scores, and SSNR improvements for the proposed and competing postfilters averaged over all types of diffuse noise in various SNRs for the larger simulated room with the RT_{60} of 500 ms. The lower and upper bars in the LogErr plot represent the overestimation and underestimation errors, respectively. The average scores are shown as the rightmost bars (**a**,**b**) or dashed lines (**c**).



(**a**) LogErr (**b**) PESQ (**c**) SSNR improvement

Figure 6. The logarithmic errors, PESQ scores, and SSNR improvements for the proposed and competing postfilters averaged over all types of diffuse noise in various SNRs for the real-recorded RIR database with the RT_{60} of 360 ms. The lower and upper bars in the LogErr plot represent the overestimation and underestimation errors, respectively. The average scores are shown as the rightmost bars (**a**,**b**) or dashed lines (**c**).



(**a**) Kindergarten noise 1. (**b**) Cafeteria noise.

(**c**) Train Station noise. (**d**) Crossroad noise.

Figure 7. The logarithmic errors of the proposed and competing noise PSD estimators for four types of diffuse noises averaged over three room conditions depending on the input SNR. The lower and upper bars indicate the overestimation and underestimation errors, respectively.

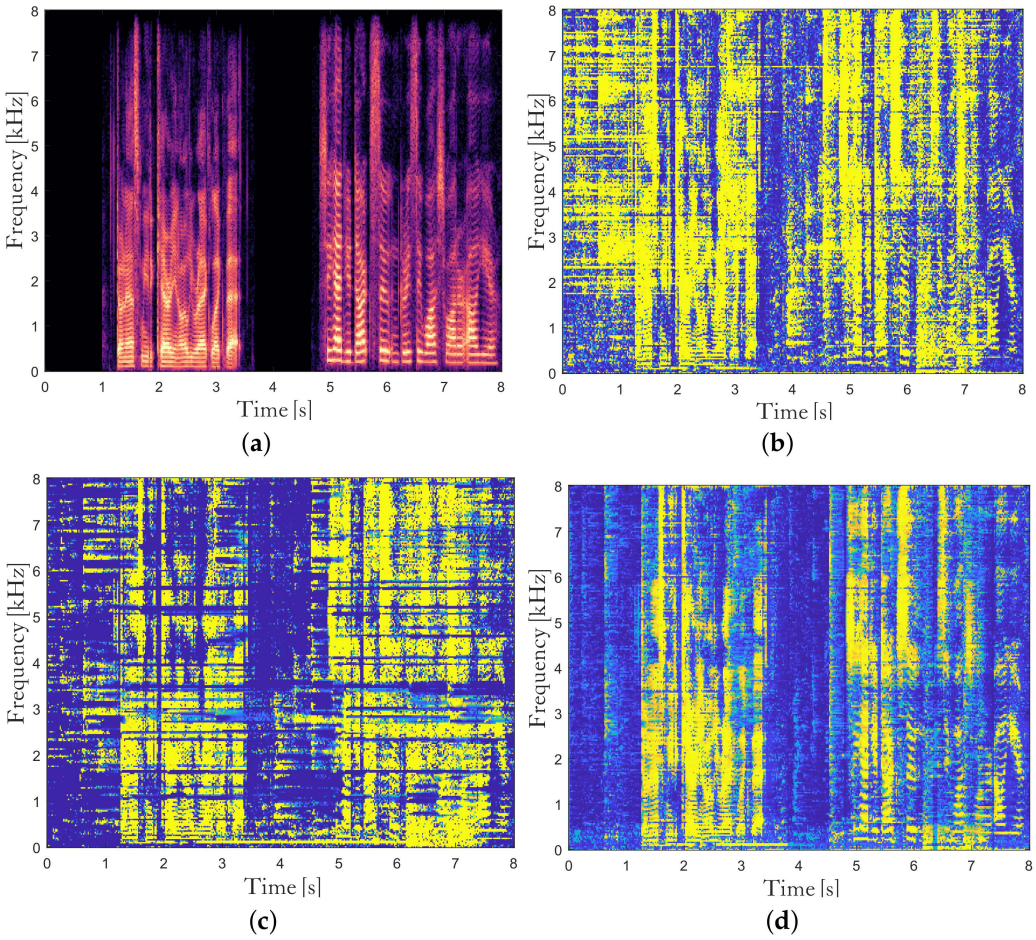
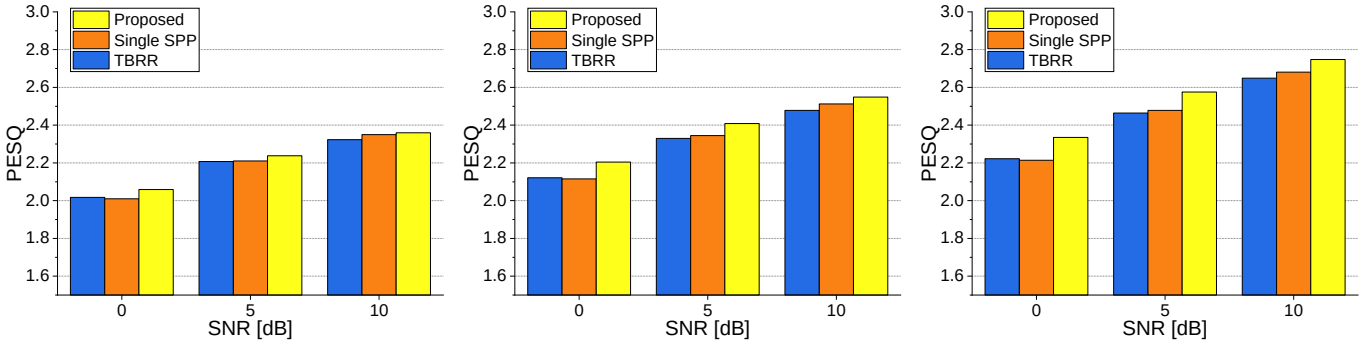


Figure 8. The spectrogram of the desired clean speech (**a**) and the *a posteriori* SPPs estimated by the *Single-SPP* [36] (**b**), the *TBRR* [21] (**c**), and the proposed method (**d**) for two utterances in the Cafeteria noise at 5 dB SNR in the smaller simulated room with the RT_{60} of 300 ms.

As the *a posteriori* SPP based on the coherence in (11) and the coherence-based noise PSD estimator in (20) relies on the difference of the coherences for the desired speech and noise, the performance of the proposed method in the presence of directional interference, which would have a higher magnitude coherence, than the diffuse noises may be questionable. To ensure that the performance of the speech enhancement does not degrade by the adoption of the proposed coherence-based SPP and noise PSD estimators in the presence of directional interferences, we conducted another set of experiments with both diffuse noises and directional interferences in various SNRs and SIRs in the smaller simulated room with the RT_{60} of 300 ms. Figure 9 presents the average PESQ scores for the proposed method and competing noise PSD estimators with various SNRs and SIRs. Although the performance improvement over other approaches is reduced to 0.06 on average, the *p*-value over the *Single-SPP* is 0.000055. We can verify that the proposed method performs slightly better than other approaches even when directional interferences disrupt the coherence-based estimators.



(**a**) SIR 0dB (**b**) SIR 5dB (**c**) SIR 10dB

Figure 9. Average PESQ scores for the proposed and competing noise PSD estimators with both the diffuse noises and directional interferences in various SNRs and SIRs in the smaller simulated room.

5. Conclusions

In this work, we have proposed a postfilter for dual channel speech enhancement utilizing *a posteriori* SPP and noise PSD estimators based on both coherence and statistical models. We have modeled the *a posteriori* SPP as a function of the magnitude of the coherence between dual microphone signals and integrated it with the statistical model-based SPP, which is then utilized for the noise PSD estimation and the OM-LSA gain function. The coherence-based noise PSD estimator is derived considering the difference between the noises in the microphone signals and the beamformer output and the speech presence uncertainty explicitly, and is combined with the SPP-based single-channel noise PSD estimator using the proposed coherence-based SPP. Experimental results show that the proposed method leads to more accurate estimation of the noise PSD and better speech enhancement in terms of the logarithmic error, PESQ scores, and SSNR improvement under the presence of various types of diffuse noise in three room conditions for the "hand-held handsfree" scenario. It is also demonstrated that the proposed method slightly outperforms competing methods in the presence of both diffuse noises and directional interferences even though the proposed approach utilizes the coherence information.

Author Contributions: Conceptualization, S.C., M.K. and J.W.S.; methodology, S.C. and J.W.S.; software, S.C.; validation, S.C., M.K. and J.W.S.; formal analysis, J.W.S.; investigation, S.C. and M.K.; resources, J.W.S.; writing—original draft preparation, S.C.; writing—review and editing, J.W.S.; visualization, S.C.; 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 MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2021-0-01835) 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: Data are contained within the article.

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

References

1. Vary, P.; Martin, R. *Digital Speech Transmission: Enhancement, Coding and Error Concealment*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
2. Benesty, J.; Chen, J.; Huang, Y. *Microphone Array Signal Processing*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 1.
3. Kates, J.M. *Digital Hearing Aids*; Plural Publishing: San Diego, CA, USA, 2008.
4. Rabiner, L. *Fundamentals of Speech Recognition*; PTR Prentice Hall: Hoboken, NJ, USA, 1993.
5. 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] [PubMed]

6. 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]
7. 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]
8. Rascon, C. A corpus-based evaluation of beamforming techniques and phase-based frequency masking. *Sensors* **2021**, *21*, 5005. [CrossRef] [PubMed]
9. Neo, V.W.; Evers, C.; Naylor, P.A. Enhancement of noisy reverberant speech using polynomial matrix eigenvalue decomposition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3255–3266. [CrossRef]
10. Moore, A.H.; Hafezi, S.; Vos, R.R.; Naylor, P.A.; Brookes, M. A compact noise covariance matrix model for MVDR beamforming. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2049–2061. [CrossRef]
11. Van Trees, H.L. *Optimum Array Processing*; John Wiley & Sons, Inc.: New York, NY, USA, 2002.
12. Gannot, S.; Burshtein, D.; Weinstein, E. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. Signal Process.* **2001**, *49*, 1614–1626. [CrossRef]
13. Kim, H.; Shin, J.W. Dual-mic speech enhancement based on TF-GSC with leakage suppression and signal recovery. *Appl. Sci.* **2021**, *11*, 2816. [CrossRef]
14. Simmer, K.U.; Bitzer, J.; Marro, C. Post-Filtering Techniques. In *Microphone Arrays: Signal Processing Techniques and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2001; pp. 39–60.
15. Thüne, P.; Enzner, G. Maximum-likelihood approach with Bayesian refinement for multichannel-Wiener postfiltering. *IEEE Trans. Signal Process.* **2017**, *65*, 3399–3413. [CrossRef]
16. Zelinski, R. A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms. In Proceedings of the ICASSP-88, International Conference on Acoustics, Speech, and Signal Processing, New York, NY, USA, 11–14 April 1988; pp. 2578–2579.
17. McCowan, I.A.; Bourlard, H. Microphone array post-filter based on noise field coherence. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 709–716. [CrossRef]
18. Kumatani, K.; Raj, B.; Singh, R.; McDonough, J. Microphone Array Post-Filter Based on Spatially-Correlated Noise Measurements for Distant Speech Recognition. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
19. Kim, M.; Cheong, S.; Song, H.; Shin, J.W. Improved speech spatial covariance matrix estimation for online multi-microphone speech enhancement. *Sensors* **2022**, *23*, 111. [CrossRef] [PubMed]
20. Lefkimmiatis, S.; Maragos, P. A generalized estimation approach for linear and nonlinear microphone array post-filters. *Speech Commun.* **2007**, *49*, 657–666. [CrossRef]
21. Gannot, S.; Cohen, I. Speech enhancement based on the general transfer function GSC and postfiltering. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 561–571. [CrossRef]
22. Cauchi, B.; Kodrasi, I.; Rehr, R.; Gerlach, S.; Jukic, A.; Gerkmann, T.; Doclo, S.; Goetze, S. Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. *EURASIP J. Adv. Signal Process.* **2015**, *2015*, 61. [CrossRef]
23. Cheng, R.; Bao, C. Speech Enhancement Based on Beamforming and Post-Filtering by Combining Phase Information. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 4496–4500.
24. Zhou, Y.; Chen, Y.; Ma, Y.; Liu, H. A real-time dual-microphone speech enhancement algorithm assisted by bone conduction sensor. *Sensors* **2020**, *20*, 5050. [CrossRef] [PubMed]
25. Šarić, Z.; Subotić, M.; Bilibajkić, R.; Barjaktarović, M.; Stojanović, J. Supervised speech separation combined with adaptive beamforming. *Comput. Speech Lang.* **2022**, *76*, 101409. [CrossRef]
26. Tao, T.; Zheng, H.; Yang, J.; Guo, Z.; Zhang, Y.; Ao, J.; Chen, Y.; Lin, W.; Tan, X. Sound localization and speech enhancement algorithm based on dual-microphone. *Sensors* **2022**, *22*, 715. [CrossRef]
27. Kim, M.; Cheong, S.; Shin, J.W. DNN-based Parameter Estimation for MVDR Beamforming and Post-Filtering. In Proceedings of the INTERSPEECH, Dublin, Ireland, 20–24 August 2023; pp. 3879–3883.
28. Sun, T.; Lei, T.; Zhang, X.; Hu, Y.; Zhu, C.; Lu, J. A Lightweight Hybrid Multi-Channel Speech Extraction System with Directional Voice Activity Detection. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1486–1490.
29. Nelke, C.M.; Beaugeant, C.; Vary, P. Dual Microphone Noise PSD Estimation for Mobile Phones in Hands-Free Position Exploiting the Coherence and Speech Presence Probability. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 7279–7283.
30. Kim, K.; Jeong, S.Y.; Jeong, J.H.; Oh, K.C.; Kim, J. Dual Channel Noise Reduction Method Using Phase Difference-Based Spectral Amplitude Estimation. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 217–220.
31. Jeub, M.; Herglotz, C.; Nelke, C.; Beaugeant, C.; Vary, P. Noise Reduction for Dual-Microphone Mobile Phones Exploiting Power Level Differences. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1693–1696.

32. Jin, W.; Taghizadeh, M.J.; Chen, K.; Xiao, W. Multi-Channel Noise Reduction for Hands-Free Voice Communication on Mobile Phones. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 506–510.
33. Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **2001**, *9*, 504–512. [CrossRef]
34. Cohen, I.; Berdugo, B. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Process. Lett.* **2002**, *9*, 12–15. [CrossRef]
35. Hendriks, R.C.; Heusdens, R.; Jensen, J. MMSE Based Noise PSD Tracking with Low Complexity. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 4266–4269.
36. Gerkmann, T.; Hendriks, R.C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 1383–1393. [CrossRef]
37. Cohen, I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 466–475. [CrossRef]
38. Cron, B.F.; Sherman, C.H. Spatial-correlation functions for various noise models. *J. Acoust. Soc. Am.* **1962**, *34*, 1732–1736. [CrossRef]
39. Cohen, I.; Berdugo, B. Speech enhancement for non-stationary noise environments. *Signal Process.* **2001**, *81*, 2403–2418. [CrossRef]
40. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, *33*, 443–445. [CrossRef]
41. Lehmann, E.A.; Johansson, A.M. Prediction of energy decay in room impulse responses simulated with an image-source model. *J. Acoust. Soc. Am.* **2008**, *124*, 269–277. [CrossRef]
42. Lehmann, E.A.; Johansson, A.M.; Nordholm, S. Reverberation-Time Prediction Method for Room Impulse Responses Simulated with the Image-Source Model. In Proceedings of the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 21–24 October 2007; pp. 159–162. [CrossRef]
43. *ETSI TS 126 132*; Speech and Video Telephony Terminal. European Telecommunications Standards Institute: Sophia Antipolis, France, 2014.
44. Hadad, E.; Heese, F.; Vary, P.; Gannot, S. Multichannel Audio Database in Various Acoustic Environments. In Proceedings of the 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), Juan-les-Pins, France, 8–11 September 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 313–317.
45. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L. *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
46. *ETSI ES 202 396-1*; Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise Part 1: Background Noise Simulation Technique and Background Noise Database. European Telecommunications Standards Institute: Sophia Antipolis, France, 2008.
47. Habets, E.A.P.; Cohen, I.; Gannot, S. Generating nonstationary multisensor signals under a spatial coherence constraint. *J. Acoust. Soc. Am.* **2008**, *124*, 2911–2917. [CrossRef]
48. *P.862.2*; Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codec. International Telecommunication Union: Geneva, Switzerland, 2007.
49. Garofolo, J.; Graff, D.; Paul, D.; Pallett, D. *Csr-i (wsj0) Complete ldc93s6a*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993; Volume 83.