applied sciences

MDPI

Article

Dual-Mic Speech Enhancement Based on TF-GSC with Leakage Suppression and Signal Recovery

Hansol Kim and Jong Won Shin *

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdan-gwagiro, Buk-gu, Gwangju 61005, Korea; hansol@gist.ac.kr
* Correspondence: jwshin@gist.ac.kr

Abstract: The transfer function-generalized sidelobe canceller (TF-GSC) is one of the most popular structures for the adaptive beamformer used in multi-channel speech enhancement. Although the TF-GSC has shown decent performance, a certain amount of steering error is inevitable, which causes leakage of speech components through the blocking matrix (BM) and distortion in the fixed beamformer (FBF) output. In this paper, we propose to suppress the leaked signal in the output of the BM and restore the desired signal in the FBF output of the TF-GSC. To reduce the risk of attenuating speech in the adaptive noise canceller (ANC), the speech component in the output of the BM is suppressed by applying a gain function similar to the square-root Wiener filter, assuming that a certain portion of the desired speech should be leaked into the BM output. Additionally, we propose to restore the attenuated desired signal in the FBF output by adding some of the microphone signal components back, depending on how microphone signals are related to the FBF and BM outputs. The experimental results showed that the proposed TF-GSC outperformed conventional TF-GSC in terms of the perceptual evaluation of speech quality (PESQ) scores under various noise conditions and the direction of arrivals for the desired and interfering sources.

Keywords: dual-mic speech enhancement; transfer function-generalized sidelobe canceller; steering error; leakage suppression; signal recovery

Citation: Kim, H.; Shin, J.W. Dual-Mic Speech Enhancement Based on TF-GSC with Leakage Suppression and Signal Recovery. *Appl. Sci.* **2021**, *11*, 2816. https://doi.org/10.3390/app11062816

Academic Editor: José A. González-López

Received: 26 February 2021
Accepted: 20 March 2021
Published: 22 March 2021

1. Introduction

Multi-channel speech enhancement is an essential part for many applications, such as mobile phones, smart TVs, and smart speakers [1,2]. Compared to a single-channel approach, multi-channel speech enhancement can utilize not only spectral information, but also spatial information using multiple microphones, which makes it possible to achieve better performance. Popular examples of spatial filters include multi-channel Wiener filters [1,3–6], adaptive beamformers [2,7,8], and nonlinear filters [9–16].

One of the useful implementations of adaptive beamformers [17] is the generalized sidelobe canceller (GSC) [17], which consists of a fixed beamformer (FBF) that tries to pass only the signals from the desired direction, a blocking matrix (BM) that attempts to filter out the signals from the desired direction, and an adaptive noise canceller (ANC) that removes the signals related to the BM output from the FBF output. The transfer function (TF)-GSC (TF-GSC) [18–27] is one of the most popular algorithms for GSC, which steers a beam to a desired source based on the estimated acoustic TF [2] ratio vector and achieves high noise reduction performance while maintaining low speech distortion [28]. However, the steering error [29] arising from an inaccurate TF ratio estimation may introduce signal leakage in the BM output and speech distortion in the FBF output, which degrades the performance of speech enhancement, especially in the presence of diffuse or nonstationary noise.

One of the typical remedies to this problem is to equip a single-channel speech enhancement module as a post-processor [9–16]. The spectral gain based on the optimally modified log spectral amplitude (OM-LSA) estimator [30] is applied to the output of the

TF-GSC in [12–14], where the transient beam-to-reference ratio (TBRR) [31] is utilized for hypothesis testing [32] to determine if the input contains desired speech, transient interference, or stationary noises, which in turn affects the estimation of a priori signal absence probability, a priori signal-to-noise ratio (SNR), noise power spectra, and spectral gain function. In [15,16], the results of the hypothesis testing were further utilized in the parameter updates for the TF-GSC. Nevertheless, a certain amount of the steering error is practically inevitable, resulting in signal leakage in the BM output and signal attenuation in the FBF output, which limits the performance of speech enhancement using TF-GSC.

In this paper, we propose to modify the outputs of the FBF and the BM by utilizing both of the outputs and the microphone signals. Specifically, we applied a gain function similar to the square-root Wiener filter to the BM output to suppress the desired speech signal leaking into it. Moreover, the attenuated signal in the FBF output was recovered to an extent by adding back a certain amount of appropriate microphone signals. The experimental results show that the proposed TF-GSC achieved better performance than the conventional TF-GSC for various noise scenarios. Moreover, the proposed method can provide consistent performance for various directions of arrivals (DoAs) of the desired and interfering sources in contrast to the conventional TF-GSC.

The remainder of this paper is organized as follows. Section 2 summarizes the conventional TF-GSC and the postfiltering. The proposed TF-GSC with leakage suppression and signal recovery is introduced in Section 3. Section 4 outlines the experimental results. Finally, a conclusion is provided in Section 5.

2. Summary of TF-GSC and Postfiltering

Let $x_p(m)$, $n_p(m)$, and $s(m)$ denote the input microphone signal and interfering signal at the p^{th} microphone and the desired speech at the source for time m, respectively. With additive noise assumption, $x_p(m)$ is represented as:

$$x_p(m) = a_p(m) * s(m) + n_p(m), \quad p = 1, 2 \tag{1}$$

where $a_p(m)$ is the acoustic impulse response from the desired source to the p^{th} microphone and $*$ denotes the convolution operation. The short-time Fourier transform (STFT) coefficients for signals $X_p(n,k)$, $N_p(n,k)$, and $S(n,k)$ for frame n and frequency k are related as:

$$\mathbf{X}(n,k) = \mathbf{A}(n,k)S(n,k) + \mathbf{N}(n,k) \tag{2}$$

in which:

$$\begin{array}{l} \mathbf{X}(n,k) = [X_1(n,k) \quad X_2(n,k)]^T \\ \mathbf{A}(n,k) = [A_1(n,k) \quad A_2(n,k)]^T \\ \mathbf{N}(n,k) = [N_1(n,k) \quad N_2(n,k)]^T \end{array} \tag{3}$$

where $A_p(n,k)$ is the acoustic TF from the desired source to the p^{th} microphone for frame n and frequency k. A block diagram of the TF-GSC is shown in Figure 1 in black, which mainly consists of three blocks [2,7,8,18]. The FBF conserves signals from a desired direction while rejecting other signals as much as possible. The output of the FBF $\mathbf{W}(n,k)$, $Y_{FBF}(n,k)$ is given by:

$$Y_{FBF}(n,k) = \mathbf{W}(n,k)^H \mathbf{X}(n,k), \tag{4}$$

where $\mathbf{W}(n,k)$ is constrained to satisfy $\mathbf{W}(n,k)^H \mathbf{A}(n,k) = \mathcal{F}^*(n,k)$, in which a prespecified filter $\mathcal{F}^*(n,k)$ is usually assumed to be a simple delay [9]. However, since the actual TFs are very difficult to obtain, the TF ratios are estimated instead and the optimal $\mathbf{W}(n,k)$ is represented as [18]:

$$\mathbf{W}(n,k) = \frac{\mathbf{H}(n,k)}{\left\|\mathbf{H}(n,k)\right\|^2} \tag{5}$$

in which **H** is the acoustic TF ratio vector:

$$\mathbf{H}(n,k) = \begin{bmatrix} H_1(n,k) & H_2(n,k) \end{bmatrix}^T = \begin{bmatrix} 1 & \frac{A_2(n,k)}{A_1(n,k)} \end{bmatrix}^T. \tag{6}$$

In the other branch, the BM $\mathbf{B}(n,k)$ tries to block the signals from a desired direction while passing through all other signals, resulting in the noise reference Y_U:

$$Y_U(n,k) = \mathbf{B}(n,k)^H \mathbf{X}(n,k) \tag{7}$$

in which:

$$\mathbf{B}(n,k) = \begin{bmatrix} -H_2(n,k)^* \\ 1 \end{bmatrix}. \tag{8}$$

The third block, the ANC, tries to remove the components related to $Y_U(n,k)$ from the FBF output $Y_{FBF}(n,k)$ using the normalized least-mean-square (NLMS) algorithm:

$$Y(n,k) = Y_{FBF}(n,k) - G(n,k)^H Y_U(n,k) \tag{9}$$

where $G(n,k)$ is the ANC filter updated by:

$$G(n+1,k) = G(n,k) + \mu \frac{Y_U(n,k) Y_{FBF}(n,k)^H}{P_{est}(n,k)} \tag{10}$$

in which μ is a step size and the noise reference power $P_{est}(n,k)$ is updated as:

$$P_{est}(n,k) = \rho_U P_{est}(n-1,k) + (1-\rho_U) Y_U(n,k) Y_U(n,k)^H \tag{11}$$

where ρ_U is a smoothing factor.

In order to effectively operate FBF and BM, the estimation of accurate TF ratios is essential. In [18], the TF ratios were estimated using least squares, assuming that the TF ratios are slowly changing in time compared to the desired signal and the background noise signals are stationary. The TF ratio is updated when the desired signal is present in the last L frames as:

$$H_2(n,k) = \frac{\left\langle \hat{\Phi}_{X_1X_1}(n,k) \hat{\Phi}_{X_2X_1}(n,k) \right\rangle - \left\langle \hat{\Phi}_{X_1X_1}(n,k) \right\rangle \left\langle \hat{\Phi}_{X_2X_1}(n,k) \right\rangle}{\left\langle \hat{\Phi}^2_{X_1X_1}(n,k) \right\rangle - \left\langle \hat{\Phi}_{X_1X_1}(n,k) \right\rangle^2} \tag{12}$$

in which:

$$\left\langle \Phi(n,k) \right\rangle \triangleq \frac{1}{L} \sum_{l=1}^{L} \Phi(n-l+1,k) \tag{13}$$

where $\hat{\Phi}_{X_pX_1}(n,k)$ $(p = 1,2)$ is the estimate of the cross-power spectral density between X_p and X_1 given by:

$$\hat{\Phi}_{X_pX_1}(n,k) = \rho_P \hat{\Phi}_{X_pX_1}(n-1,k) + (1-\rho_P) X_p(n,k) X_1(n,k)^H \tag{14}$$

where ρ_P is a smoothing factor.

However, the TF ratio estimation suffers from diffuse or nonstationary noises, resulting in steering error, which causes leakage of speech components through the BM and speech distortion in the FBF output. One way to mitigate this problem is to apply a postfilter based on the OM-LSA estimator [12–16]. The first step of this postfiltering is the hypothesis test, which determines if the output of the TF-GSC contains desired speech, transient interference, or stationary noises using the TBRR [31]. The result of the hypothesis test is utilized to determine a priori signal absence probability and, finally, the spectral gain

function. Additionally, the hypothesis test result from the postfilter is further used to control the update of the parameters of the TF-GSC in [15,16] as follows:

$$H_2(n,k) = \begin{cases} \dfrac{\left\langle \hat{\Phi}_{X_1X_1}(n,k)\hat{\Phi}_{X_2X_1}(n,k) \right\rangle - \left\langle \hat{\Phi}_{X_1X_1}(n,k) \right\rangle \left\langle \hat{\Phi}_{X_2X_1}(n,k) \right\rangle}{\left\langle \hat{\Phi}^2_{X_1X_1}(n,k) \right\rangle - \left\langle \hat{\Phi}_{X_1X_1}(n,k) \right\rangle^2}, & if\,|\mathcal{L}| \geq N_d \\ H_2(n,k), & otherwise \end{cases} \quad (15)$$

in which:

$$\langle \Phi(n,k) \rangle \triangleq \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \Phi(l,k) \quad (16)$$

where $\mathcal{L}$ is a set of frame indices that contain the desired signal component in the analysis interval from the $n - L + 1^{th}$ frame to the n^{th} frame, $|\mathcal{L}|$ is the size of $\mathcal{L}$, and N_d is the threshold for the TF ratio update. It is noted that the value of L should be carefully chosen to provide accurate estimates of ensemble averages with a large enough number of frames, while the quasi-stationary assumption for the acoustic TF is not violated by too big of an L [15,16,18]. It has the effect of relaxing the assumption on the stationarity of the TF ratio, which reduces the steering error to an extent. Even with the improved TF ratio estimation, however, a certain amount of the steering error is practically inevitable, leaving room for improvement in the performance.

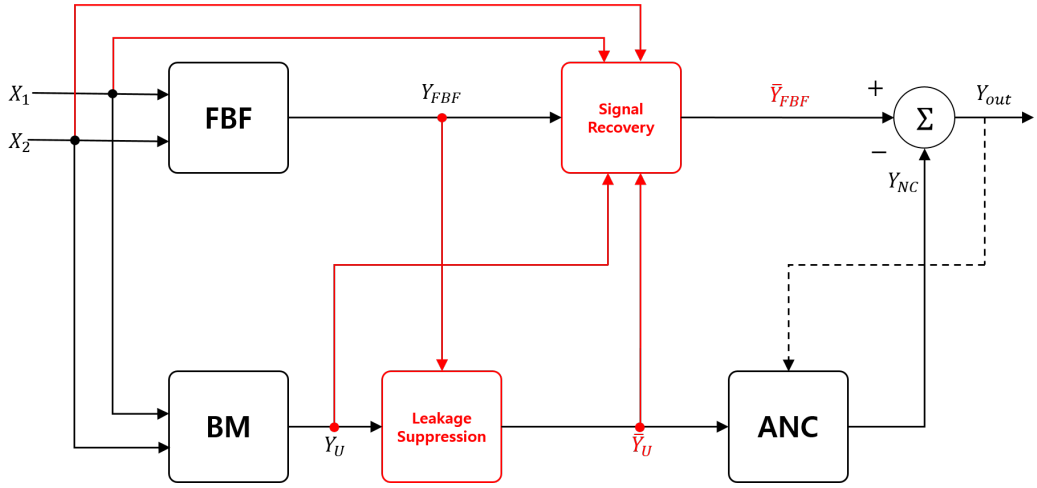


Figure 1. A block diagram of the conventional transfer function-generalized sidelobe canceller (TF-GSC) (black) and the proposed one with leakage suppression and signal recovery (black+red). FBF, fixed beamformer; BM, blocking matrix; ANC, adaptive noise canceller.

3. Proposed TF-GSC with Leakage Suppression and Signal Recovery

To alleviate the problems caused by steering error, this paper proposes a leakage suppression module to suppress the leaked desired signal in the output of the BM and the signal recovery module to restore the attenuated desired signal in the FBF output, which is shown in Figure 1 in red. The postfiltering and the feedback from it are also adopted for both the conventional and the proposed TF-GSC, which is omitted from the figure.

3.1. Leakage Suppression

Although the FBF and BM do not operate perfectly, the ratio of the FBF output Y_{FBF} and the BM output Y_U bears information on the input SNR. The leakage of the desired signal to the BM output would be more severe when the input SNR increases, and vice versa. Therefore, the leakage suppression module applies a spectral gain $G_U(n,k)$ to the BM output $Y_U(n,k)$ to produce the modified noise reference $\bar{Y}_U(n,k)$:

$$\bar{Y}_U(n,k) = G_U(n,k) Y_U(n,k) \quad (17)$$

in which $G_U(n,k)$ has a form similar to the square-root Wiener filter:

$$G_U(n,k) = \sqrt{\frac{|\ Y_U(n,k)\ |^2}{\alpha\ |\ Y_{FBF}(n,k)\ |^2 + |\ Y_U(n,k)\ |^2}} \tag{18}$$

where $\alpha < 1$ is a tuning parameter representing the attenuation of the desired signal component in the BM output. The leakage suppression would attenuate the desired signal component from $Y_U(n,k)$ and thus reduce the signal cancellation in the ANC.

3.2. Signal Recovery

Once the desired signal is attenuated by the FBF with steering error, the following modules such as the ANC and the postfilter cannot restore it, while they can suppress the residual interference in the FBF output. In this regard, the proposed signal recovery module adds a certain amount of the appropriate microphone signal to the FBF output. To determine which microphone signal should be utilized to recover the attenuated desired signal, we evaluated which microphone signal is closer to Y_{FBF} and Y_U, respectively. Specifically, we assessed the cosine similarities between X_1, X_2 and Y_{FBF}, Y_U given by:

$$\begin{aligned} S_{\mathbf{X}_p\mathbf{Y}_{FBF}}(n) &= \lambda S_{\mathbf{X}_p\mathbf{Y}_{FBF}}(n-1) + (1-\lambda)\frac{\mathrm{Re}(\mathbf{X}_p(n) \bullet \mathbf{Y}_{FBF}(n))}{\|\mathbf{X}_p(n)\|\|\mathbf{Y}_{FBF}(n)\|} \\ S_{\mathbf{X}_p\mathbf{Y}_U}(n) &= \lambda S_{\mathbf{X}_p\mathbf{Y}_U}(n-1) + (1-\lambda)\frac{\mathrm{Re}(\mathbf{X}_p(n) \bullet \mathbf{Y}_U(n))}{\|\mathbf{X}_p(n)\|\|\mathbf{Y}_U(n)\|} \end{aligned} \tag{19}$$

in which:

$$\begin{aligned} &\mathbf{X}_p(n) = [X_p(n,1),\ldots,X_p(n,K)]^T, p = 1,2 \\ &\mathbf{Y}_{FBF}(n) = [Y_{FBF}(n,1),\ldots,Y_{FBF}(n,K)]^T \\ &\mathbf{Y}_U(n) = [Y_U(n,1),\ldots,Y_U(n,K)]^T \end{aligned} \tag{20}$$

where λ is a smoothing parameter, $\bullet$ is an inner product operation, K is the number of frequency bins, and $\frac{\mathrm{Re}(\mathbf{X}_p(n)\bullet\mathbf{Y}_{FBF}(n))}{\|\mathbf{X}_p(n)\|\|\mathbf{Y}_{FBF}(n)\|}$ and $\frac{\mathrm{Re}(\mathbf{X}_p(n)\bullet\mathbf{Y}_U(n))}{\|\mathbf{X}_p(n)\|\|\mathbf{Y}_U(n)\|}$ are the cosine values of the Euclidean angles between $\mathbf{X}_p(n)$ and $\mathbf{Y}_{FBF}(n)$ and $\mathbf{Y}_U(n)$, respectively [33,34]. Using these similarities, the two similarity differences $SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}(n)$ and $SD_{\mathbf{X}_{21}\mathbf{Y}_U}(n)$ can be computed:

$$\begin{aligned} &SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}(n) = S_{\mathbf{X}_1\mathbf{Y}_{FBF}}(n) - S_{\mathbf{X}_2\mathbf{Y}_{FBF}}(n), \\ &SD_{\mathbf{X}_{21}\mathbf{Y}_U}(n) = S_{\mathbf{X}_2\mathbf{Y}_U}(n) - S_{\mathbf{X}_1\mathbf{Y}_U}(n). \end{aligned} \tag{21}$$

where $SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}(n)$ is a measure of how much X_1 is more similar to Y_{FBF} compared to X_2, and $SD_{\mathbf{X}_{21}\mathbf{Y}_U}(n)$ represents how X_2 is closer to Y_U compared to X_1. Thus, if both $SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}(n)$ and $SD_{\mathbf{X}_{21}\mathbf{Y}_U}(n)$ are positive (or negative), X_1 (or X_2) contains more of the desired signal. If the absolute value of $SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}$ is small, however, the desired source may be located at the broadside, and thus the average of two microphone signals would provide a better reference of the desired signal. When the signs of $SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}$ and $SD_{\mathbf{X}_{21}\mathbf{Y}_U}$ differ, we anticipate that the signal restoration may not be reliable, and the FBF output is not modified. The selected signal for the signal recovery is summarized by:

$$\mathbf{X}_{selected}(n) = \begin{cases} \mathbf{X}_1(n), & if\ \eta < SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}(n)\ and\ SD_{\mathbf{X}_{21}\mathbf{Y}_U}(n) \geq 0 \\ \dfrac{\mathbf{X}_1(n)+\mathbf{X}_2(n)}{2}, & if\ 0 < SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}(n) \times sgn(SD_{\mathbf{X}_{21}\mathbf{Y}_U}(n)) \leq \eta \\ \mathbf{X}_2(n), & if\ SD_{\mathbf{X}_{12}\mathbf{Y}_{FBF}}(n) \leq -\eta\ and\ SD_{\mathbf{X}_{21}\mathbf{Y}_U}(n) \leq 0 \\ \mathbf{0}, & else \end{cases} \tag{22}$$

which is also described in Figure 2. Using the selected microphone signal, the FBF output $Y_{FBF}(n,k)$ is modified as:

$$\bar{Y}_{FBF}(n,k) = Y_{FBF}(n,k) + G_X(n,k)X_{selected}(n,k) \tag{23}$$

in which $G_X(n,k)$ determines how much of the microphone signal is added to restore the attenuated desired signal. Since adding the microphone signal in low SNR environments may be harmful, $G_X(n,k)$ is designed to have higher values in higher SNRs:

$$G_X(n,k) = \sqrt{\frac{|\, Y_{FBF}(n,k)\,|^2}{\beta\,|\,\bar{Y}_U(n,k)\,|^2 + |\, Y_{FBF}(n,k)\,|^2}} \tag{24}$$

where $\beta > 1$ is a tuning parameter.

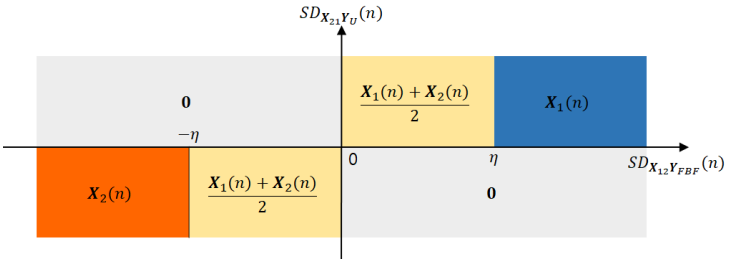


Figure 2. Selected microphone signal to be used for signal recovery depending on the similarity differences between the microphone signals and the outputs of the FBF and BM.

4. Experimental Results

To evaluate the performance of the proposed method, the acoustic environments depicted in Figure 3 were simulated using the image method [35]. The dimension of the room was [6.7 m, 6.1 m, 2.9 m]. Two microphones were located at [3 m, 3 m, 1.5 m] and [3.14 m, 3 m, 1.5 m], respectively, 14 cm away from each other, which is typical for modern smartphones. The reverberation times (RT60s) were 300 ms and 500 ms. The distance between the desired source and the microphone array was 0.4 m, while that for the interfering source was 0.8 m. The DoA for the desired source was $-90°$, $-60°$, $-30°$, and $0°$ when $0°$ indicates the broadside direction, while the interfering source was located at $[-90°, +90°]$ with a $30°$ interval.

Twenty utterances spoken by 13 male and 7 female speakers were selected randomly from the TIMIT database [36] as the desired speech signals. Babble, F16, and Factory1 noises from the NOISEX-92 database [37] and restaurant and street noises from the AURORA 2 database [38] were used as diffuse noise, which was constructed using the arbitrary noise field generator [39]. Five competing talker utterances were selected randomly from the TIMIT database as directional noises. The SNRs for diffuse or directional noise were 0, 5, 10, 15, and 20 dB. The sampling rate was 8 kHz and the 256 point STFT with a Tukey window was used with the frame shift of 160 samples.

The empirically determined values of the parameters for the conventional TF-GSC with postfiltering [16] and the proposed one with two additional modules are summarized in Table 1. The ANC filter G and the noise reference power for ANC update P_{est} were converged before evaluating the performance. The parameters not included in the table were set to be the same as in [16], which produced the highest average performance. Perceptual evaluation of speech quality (PESQ) scores [40] were carried out to evaluate the performance.

Figures 4 and 5 show the average PESQ scores and the 95% confidence intervals (CIs) for the input microphone signal, the output of the conventional TF-GSC with postfiltering, and the proposed scheme in diffuse noise environment according to the SNR, RT60, and the azimuth of the desired source averaged for five noise types. The conventional algorithm improved the PESQ scores in all conditions, but the improvement in performance reduced as the desired source moved from the broadside direction ($0°$) to the end-fire direction ($-90°$). The proposed leakage suppression and signal recovery were considered to be effective in maintaining the performance for all DoAs of the desired source. The improvement in performance over the conventional TF-GSC was statistically significant when the desired source was located near the end-fire direction ($-90°$ and $-60°$) and the SNR

was higher than 0 dB, but was insignificant when the desired source was located at the broadside (0°). This may be because the steering error was smaller when the desired source was located in the broadside direction, and the signal recovery module essentially has no impact for the broadside as it just adds the scaled FBF output $(\mathbf{X}_1(n) + \mathbf{X}_2(n))/2$ to the FBF output. A similar tendency was observed for both RT60s. The PESQ score differences between the proposed and baseline beamformers for each noise type are shown in Figure 6. The proposed beamformer improved the PESQ scores in both stationary and nonstationary noise environments. The signal recovery module may work more effectively when the desired source is located in the end-fire direction by recovering the attenuated desired signal using the closer microphone signal.

Table 1. Parameters used for the baseline and proposed beamformers.

ANC update	$\mu = 0.03$	$\rho_U = 0.97$	
TF ratio identification	$L = 20$	$\rho_P = 0.99$	$N_d = 4$
Leakage suppression	$\alpha = 0.0001$		
Signal recovery	$\lambda = 0.9$	$\eta = 0.1$	$\beta = 10$

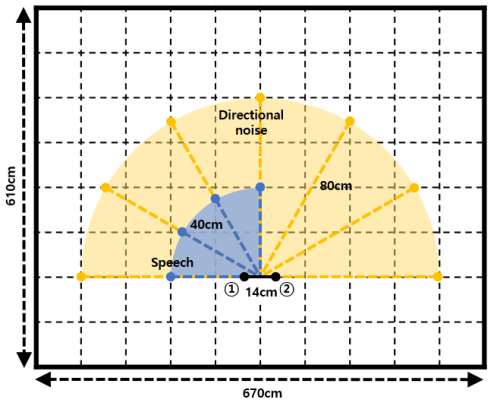


Figure 3. Room configurations with the locations of the microphones and the desired and interfering sources.

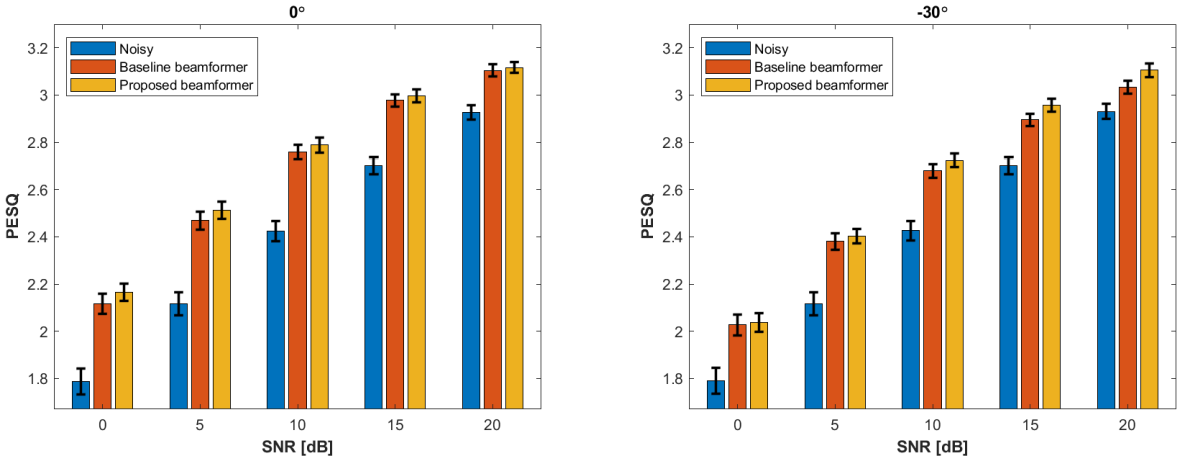


Figure 4. *Cont.*

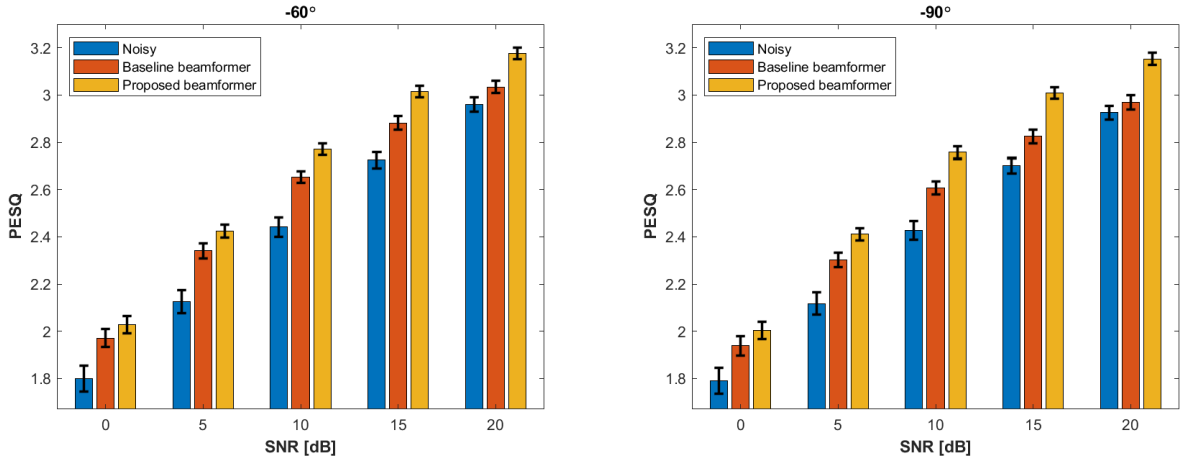


Figure 4. Average perceptual evaluation of speech quality (PESQ) scores and 95% confidence intervals for the noisy input and the outputs of the baseline and proposed beamformers depending on the azimuth of the desired source in diffuse noise environments with an RT60 of 300 ms.

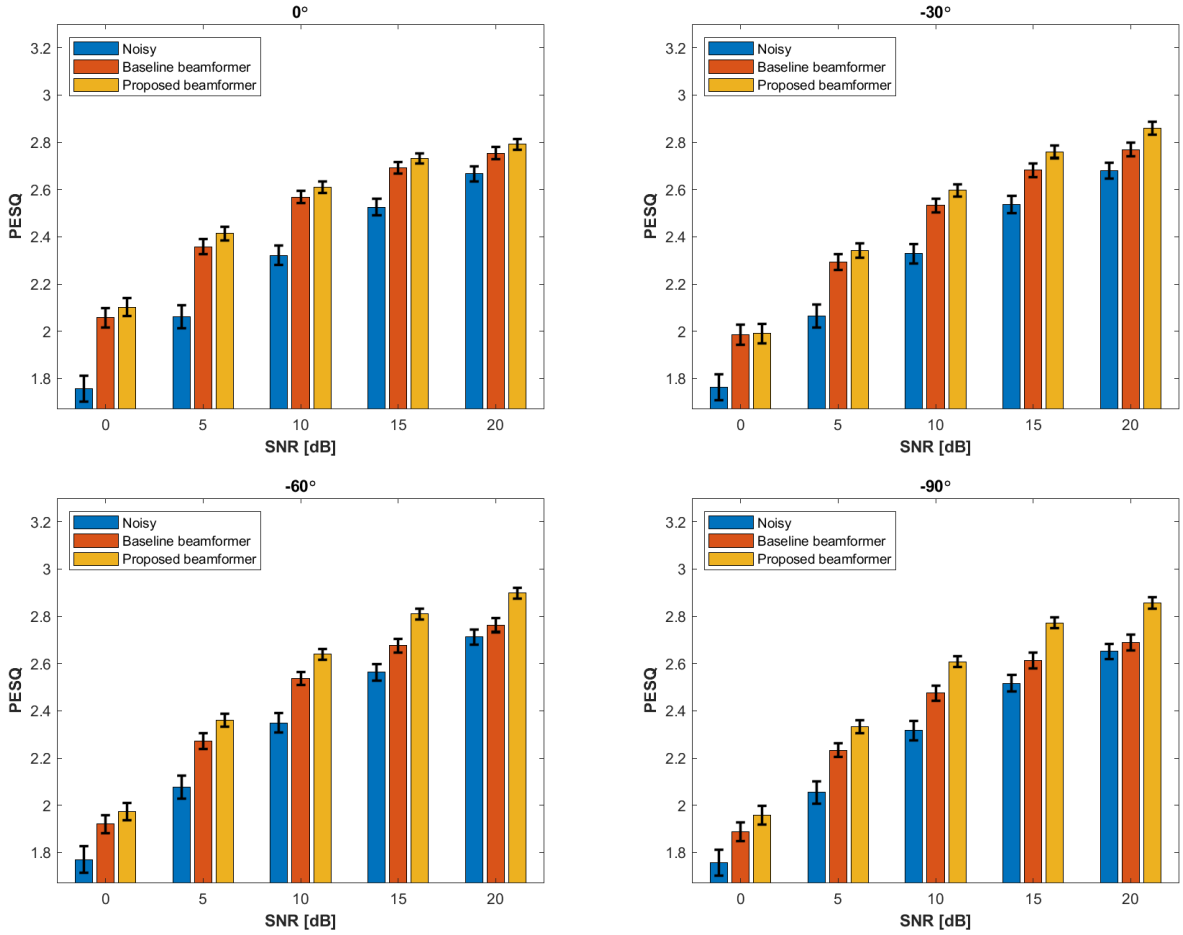
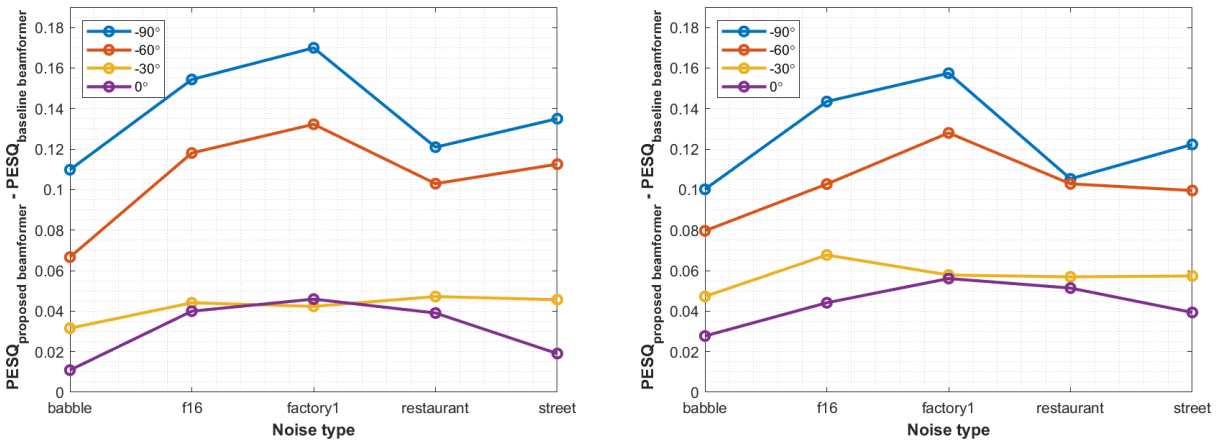


Figure 5. Average PESQ scores and 95% confidence intervals for noisy input and the outputs of the baseline and the proposed beamformers depending on the azimuth of the desired source in diffuse noise environments with the reverberation time (RT60) of 500 ms.



(**a**) (**b**)

Figure 6. Difference of the average PESQ scores for the proposed and baseline beamformers in diffuse noise environments for five noise types with the RT60s of (**a**) 300 ms and (**b**) 500 ms.

Another set of experiments were carried out on the directional interference. We conducted experiments on the speech enhancement in the presence of the competing talker. Figures 7 and 8 show the average PESQ scores and CIs for the conventional and proposed TF-GSC depending on the locations of the desired and interfering sources for two difference RT60s. As in the case of the diffuse noises, the proposed method showed higher PESQ scores than the conventional TF-GSC, achieving almost the same performance for all DoAs of the desired source in contrast to the baseline TF-GSC. The improvement in performance was significant except when the desired source was located at the broadside, as in the diffuse noise case. The PESQ scores were improved even when the desired and interfering sources were located in the same direction, although the improvement was smaller. This may be because the interfering source was located farther than the desired source, as is often the case with practical scenarios. Therefore, Y_U might contain more reverberant components from the competing talker and Y_{FBF} might include more desired signal, which enable the leakage suppression module to operate properly.

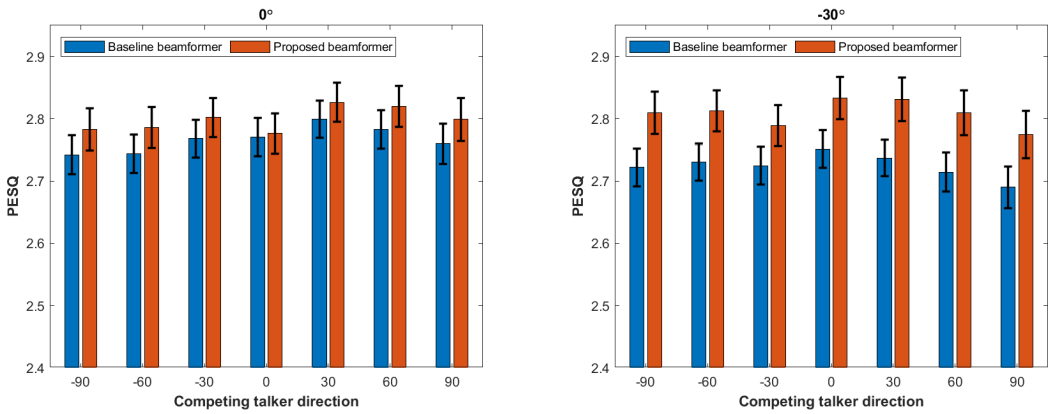


Figure 7. *Cont.*

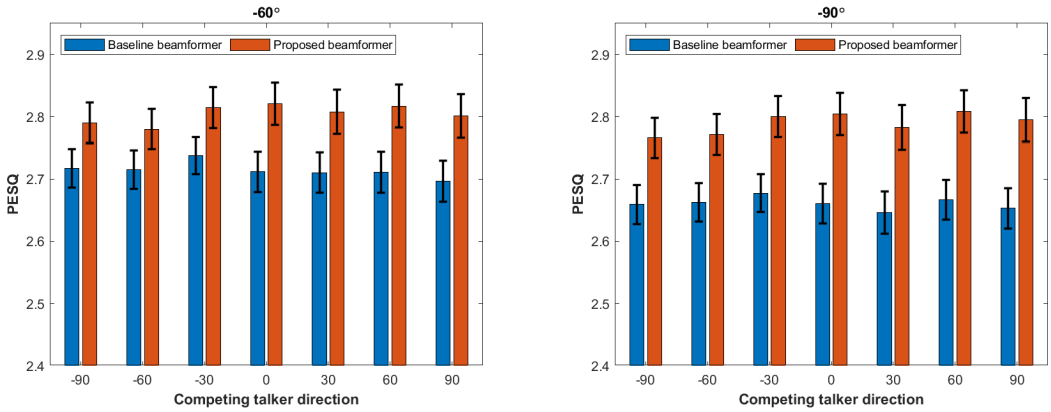


Figure 7. Average PESQ scores and 95% confidence intervals for the outputs of the baseline and proposed beamformers depending on the azimuths of the desired source and a competing talker with an RT60 of 300 ms.

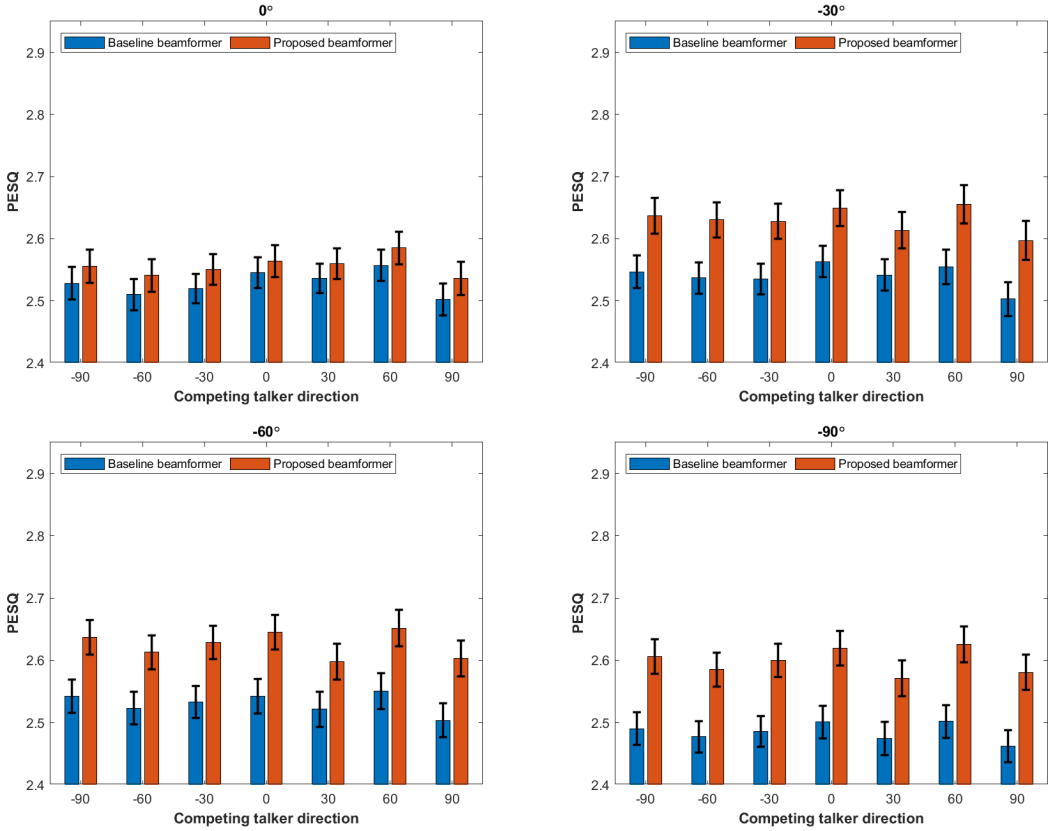


Figure 8. Average PESQ scores and 95% confidence intervals for the outputs of the baseline and proposed beamformers depending on the azimuths of the desired source and a competing talker with an RT60 of 500 ms.

5. Conclusions

In this paper, we introduced two additional modules to mitigate the effect of the steering error of the TF-GSC with postfiltering. The leakage suppression module suppresses the leaked desired signal in the output of the BM by applying a spectral gain similar to the square-root Wiener filter. On the contrary, the signal recovery module restores the attenuated desired signal in the FBF output by adding a certain amount of the appropriate microphone signal, which is chosen by examining the cosine similarities between the microphone signals and the outputs of the FBF and BM. The experimental results showed that the two proposed modules improved the performance of the conventional TF-GSC

with postfiltering, both in diffuse noise environments and in the presence of a competing talker, achieving almost the same PESQ scores for all of the DoAs of the desired signal.

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

Funding: This research was funded by the Ministry of Culture, Sports, and Tourism (MCST) and the Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research and Development Program (project number: R2019080018) and Samsung System LSI (SLSI-201801GE002S).

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

References

1. Benesty, J.; Chen, J.; Huang, Y. *Microphone Array Signal Processing*; Springer-Verlag: Berlin/Heidelberg, Germany, 2008.
2. 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]
3. Chen, J.; Benesty, J.; Huang, Y.; Doclo, S. New insights into the noise reduction wiener filter. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1218–1234. [CrossRef]
4. Bogaert, T.V.; Doclo, S.; Wouters, J.; Moonen, M. Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. *J. Acoust. Soc. Am.* **2009**, *125*, 360–371. [CrossRef] [PubMed]
5. 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]
6. Jin, Y.G.; Shin, J.W.; Kim, N.S. Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement. *JASA Express Lett.* **2017**, *141*, EL228–EL233. [CrossRef] [PubMed]
7. Doclo, S.; Gannot, S.; Moonen, M.; Spriet, A.; Haykin, S.; Liu, K.R. Acoustic beamforming for hearing aid applications. In *Handbook on Array Processing and Sensor Networks*; Haykin, S., Liu, K., Eds.; Wiley: Hoboken, NJ, USA, 2010; pp. 269–302.
8. Doclo, S.; Kellermann, W.; Makino, S.; Nordholm, S.E. Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Process. Mag.* **2015**, *32*, 18–30. [CrossRef]
9. Benesty, J.; Sondhi, M.M.; Huang, Y. *Springer Handbook of Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2007.
10. Dashtbozorg, B.; Abutalebi, H.R. Joint Noise Reduction and Dereverberation of Speech Using Hybrid TF-GSC and Adaptive MMSE Estimator. In Proceedings of the 10th Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009; pp. 1355–1358.
11. Zhang, M.; Wu, S.; Guo, W.; Ji, J. A microphone array dereverberation algorithm based on TF-GSC and postfiltering. In Proceedings of the 2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Nara, Japan, 1–3 June 2016.
12. Cohen, I.; Berdugo, B. Microphone array post-filtering for non-stationary noise suppression. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 13–17 May 2002.
13. 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]
14. Cohen, I. Multichannel post-filtering in nonstationary noise environments. *IEEE Trans. Signal Process.* **2004**, *52*, 1149–1160. [CrossRef]
15. Cohen, I.; Gannot, S.; Berdugo, B. Real-Time TF-GSC in Nonstationary Noise Environments. Available online: https://israelcohen.com/wp-content/uploads/2018/05/iwaenc03.pdf (accessed on 22 March 2021).
16. Cohen, I.; Gannot, S.; Berdugo, B. An integrated real-time beamforming and postfiltering system for nonstationary noise environments. *EURASIP J. Adv. Signal Process.* **2003**, *2003*, 1064–1073. [CrossRef]
17. Van Trees, H.L. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
18. 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]
19. Reuven, G.; Gannot, S.; Cohen, I. Multichannel acoustic echo cancellation and noise reduction in reverberant environments using the transfer-function GSC. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007.
20. Reuven, G.; Gannot, S.; Cohen, I. Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller. *Speech Commun.* **2007**, *49*, 623–635. [CrossRef]
21. Reuven, G.; Gannot, S.; Cohen, I. Dual-source transfer-function generalized sidelobe canceller. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 711–727. [CrossRef]

22. Kim, K.; Baran, R.H.; Ko, H. Extension of two-channel transfer function based generalized sidelobe canceller for dealing with both background and point-source noiser. *Speech Commun.* **2009**, *51*, 521–533. [CrossRef]
23. Talmon, R.; Cohen, I.; Gannot, S. Convolutive transfer function generalized sidelobe canceler. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1420–1434. [CrossRef]
24. Talmon, R.; Cohen, I.; Gannot, S. Multichannel speech enhancement using convolutive transfer function approximation in reverberant environments. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 3885–3888.
25. Lee, I.; Yoon, J.; Lee, Y.; Ko, H. Reinforced blocking matrix with cross channel projection for speech enhancement. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 957–960.
26. Nogueira, W.; Lopez, M.; Rode, T.; Doclo, S.; Buechner, A. Individualizing a monaural beamformer for cochlear implant users. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5738–5742.
27. Barnov, A.; Cohen, A.; Agmon, M.; Bracha, V.B.; Markovich-Golan, S.; Gannot, S. A dynamic TF-GSC beamformer for distributed arrays with dual-resolution speech-presence-probability estimators. In Proceedings of the 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), Eilat, Israel, 16–18 November 2016.
28. Ephraim, Y.; Van Trees, H.L. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 251–266. [CrossRef]
29. Wax, M.; Anu, Y. Performance analysis of the minimum variance beamformer in the presence of steering vector errors. *IEEE Trans. Signal Process.* **1996**, *44*, 938–947. [CrossRef]
30. Cohen, I.; Berdugo, B. Speech enhancement for non-stationary noise environments. *Signal Process.* **2001**, *81*, 2403–2418. [CrossRef]
31. Cohen, I.; Berdugo, B. Multichannel signal detection based on the transient beam-to-reference ratio. *IEEE Signal Process. Lett.* **2003**, *10*, 259–262. [CrossRef]
32. Kay, S.M. *Fundamentals of Statistical Signal Processing*; Prentice Hall PTRs: Englewood Cliffs, NJ, USA, 1993.
33. Scharnhorst, K. Angles in Complex Vector Spaces. *Acta Appl. Math.* **2001**, *69*, 95–103. [CrossRef]
34. Mandolesi, A.L. Grassmann angles between real or complex subspaces. *arXiv* **2019**, arXiv:1910.00147.
35. Habets, E.A. Room Impulse Response Generator. 2010. Available online: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator (accessed on 22 March 2021).
36. Garofolo, J.S. *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*; National Institute of Standards and Technology (NIST): Gaithersburgh, MD, USA, 1988; Volume 107, p. 16.
37. Varga, A.; Steeneken, H.J. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
38. Hirsch, H.G.; Pearce, D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In Proceedings of the ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), Paris, France, 18–20 September 2000.
39. Habets, E.A.; Cohen, I.; Gannot, S. Generating nonstationary multisensor signals under a spatial coherence constraint. *J. Acoust. Soc. Am.* **2008**, *124*, 2911–2917. [CrossRef] [PubMed]
40. ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. 2008. Available online: https://www.itu.int/rec/T-REC-P.862-200102-I/en (accessed on 22 March 2021).