applied sciences

MDPI

Article

Crossband Filtering for Weighted Prediction Error-Based Speech Dereverberation

Tomer Rosenbaum [1,*], Israel Cohen [1] and Emil Winebrand [2]

[1] Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel; icohen@ee.technion.ac.il
[2] Insoundz Ltd., Tel Aviv 6473104, Israel; emil.winebrand@insoundz.com
* Correspondence: tomer11r@campus.technion.ac.il

Abstract: Weighted prediction error (WPE) is a linear prediction-based method extensively used to predict and attenuate the late reverberation component of an observed speech signal. This paper introduces an extended version of the WPE method to enhance the modeling accuracy in the time–frequency domain by incorporating crossband filters. Two approaches to extending the WPE while considering crossband filters are proposed and investigated. The first approach improves the model's accuracy. However, it increases the computational complexity, while the second approach maintains the same computational complexity as the conventional WPE while still achieving improved accuracy and comparable performance to the first approach. To validate the effectiveness of the proposed methods, extensive simulations are conducted. The experimental results demonstrate that both methods outperform the conventional WPE regarding dereverberation performance. These findings highlight the potential of incorporating crossband filters in improving the accuracy and efficacy of the WPE method for dereverberation tasks.

Keywords: crossband filtering; speech dereverberation; speech enhancement; weighted prediction error

Citation: Rosenbaum, T.; Cohen, I.; Winebrand, E. Crossband Filtering for Weighted Prediction Error-Based Speech Dereverberation. *Appl. Sci.* **2023**, *13*, 9537. https://doi.org/10.3390/app13179537

Academic Editor: Lijiang Chen

Received: 25 June 2023
Revised: 19 August 2023
Accepted: 22 August 2023
Published: 23 August 2023

1. Introduction

When a distant microphone captures a speech signal in a room, it is inevitably subjected to adverse acoustic effects, including background noise and reverberation. These effects can harm the quality of the observed speech signal, significantly degrading the performance of crucial applications like automatic speech recognition (ASR). To address this issue, extensive research has been conducted on speech dereverberation. The primary objective of speech dereverberation is to eliminate or reduce late reflections in the observed speech signal. It is well known that while the early reflections are not harmful and, in some cases, even might improve the speech intelligibility [1–3], the late reflections are a significant contributing factor to the degradation in speech quality and intelligibility [4–6]. By mitigating the effects of reverberation, dereverberation techniques aim to restore the clarity and intelligibility of the captured speech, ultimately enhancing the performance of various speech processing applications. As a result, developing efficient and reliable dereverberation methods plays a vital role in advancing the field of speech processing and facilitating the deployment of robust speech-based applications in diverse settings.

Over the years, numerous dereverberation methods have been developed, employing different approaches to address the challenge of reverberation in speech signals [7–10]. One prominent approach is beamforming, which leverages an array of microphones to enhance the desired speech signal while suppressing unwanted background noise and reverberation. Notable beamforming methods include the Minimum Variance Distortionless Response (MVDR) beamformer [11] and its two-stage variant [12]. These methods estimate the optimal weights for the microphone array to enhance the desired speech source while attenuating the reverberant and noise components. Spectral enhancement methods have

also been widely explored in dereverberation research. These methods operate in the frequency domain and aim to enhance the desired speech signal by modifying the spectral characteristics. One example is the spectral subtraction method [13,14]. This technique estimates the noise and reverberation spectra and subtracts them from the observed signal to enhance the speech component. Other spectral enhancement approaches employ advanced signal processing methods such as Wiener filtering and statistical modeling to separate the desired speech from the interfering components [15].

Another category of dereverberation methods focuses on estimating an inverse filter to predict and suppress the late reflections present in the observed speech signal [16–20]. One notable method in this category that has gained significant attention in the field of speech processing is the weighted prediction error (WPE) method [19,20]. WPE is based on linear prediction (LP), utilizing an inverse filter to estimate and suppress the late reflections in the observed speech signal. By exploiting the statistical properties of the reverberation, WPE effectively separates the desired speech component from the reverberant component. The method estimates the optimal filter coefficients by minimizing the prediction error between the observed signal and its predicted version. WPE has proven highly effective in various applications [21,22]. Due to its effectiveness, WPE has received significant attention and has been extensively studied, leading to the proposal of numerous extensions, generalizations, and variants. For instance, in [20], the model formulation, which assumed a single speech source, was generalized to an arbitrary number of sources. Other extensions and generalizations include the employment of deep neural networks [23,24], switching mechanisms [25], and Kronecker product filtering to improve the computational complexity [26].

In most applications, the WPE method is implemented in the short-time Fourier transform (STFT) domain, which provides a suitable framework for the analysis and processing of the observed speech signal. By decomposing the time-domain-observed signal into subbands, WPE operates on these subbands individually in a frequency-band-wise manner. In the time domain, WPE models the observed speech signal as the result of a linear convolution between the clean speech signal and an unknown room impulse response (RIR). However, when transitioning to the STFT domain, the relationship between the observed and clean signals becomes more intricate. In the STFT domain, the observed subbands are influenced not only by their corresponding clean subbands but also by the adjacent subbands and the crossband filters [27,28]. The exact relation between each observed subband and the clean signal results from convolutions between all clean subbands and their corresponding crossband filters. These filters capture the interdependencies and interactions between different frequency bands. The influence and information exchange between adjacent frequency components is taken into account by considering the crossband filters, enabling more comprehensive modeling and processing of the observed speech signal.

However, the conventional WPE method in the STFT domain neglects the influence of crossband filters. It assumes that each observed subband is solely the result of a convolution between the corresponding clean subband and a convolutive transfer function (CTF), often referred to as the "band-to-band filter" [28]. In other words, the CTF approach approximates each observed subband as solely dependent on its corresponding clean subband. This simplification introduces an inherent error in the WPE model when operating in the STFT domain [19]. The approximation error resulting from neglecting the crossband filters can significantly impact the performance of speech processing methods in the STFT domain. Previous studies have explored the effect of this approximation error in the context of system identification methods [28,29]. However, for WPE-based speech dereverberation, a preliminary study was conducted in [30]. Unfortunately, this study had limited scope and did not provide a comprehensive analysis of the effectiveness of the WPE-based approach incorporating crossband filtering in various real-world scenarios.

Given the importance of accurately modeling the crossband filters, further investigation is necessary to understand their impact on the performance of WPE-based dereverberation. A comprehensive analysis of the effectiveness of the WPE approach with crossband

filtering in diverse real-world scenarios is essential to shed light on the potential benefits and limitations of incorporating crossband filters into the dereverberation process. Such research will contribute to advancing the understanding and practical implementation of WPE-based methods, facilitating their optimization and broader application in real-world speech processing scenarios.

In this paper, we present an extension to the conventional WPE method that enhances the accuracy of the model approximation in the STFT domain by incorporating crossband filters. Our approach aims to capture the interdependencies between adjacent subbands and refine the estimation of the late reflections in the observed speech signal. By considering samples from neighboring subbands, we redefine the WPE observation vector and modify the prediction inverse filter to include both crossband and traditional band-to-band components. We explore two versions of the proposed method to investigate their effectiveness. The first version prioritizes accuracy by improving the model approximation, albeit at the expense of increased computational complexity. In contrast, the second version maintains the same computational complexity as the conventional WPE while enhancing the accuracy. Surprisingly, the second version demonstrates competitive performance compared to the first method.

We conduct a series of experiments to validate the performance of our proposed versions. The results confirm that both versions surpass the conventional WPE regarding dereverberation performance. This highlights the significance of incorporating information from neighboring subbands in the STFT domain in improving dereverberation outcomes. Furthermore, our findings suggest that the early samples of the crossband components might offer greater efficacy in mitigating reverberation than the late samples of the band-to-band component. By presenting these experimental results, we provide empirical evidence supporting the effectiveness of our proposed extensions to the WPE method, offering valuable insights into the potential benefits of considering crossband filters for dereverberation tasks in the STFT domain.

The remainder of this paper is organized as follows. Section 2 presents the model and the problem. Section 3 describes the proposed method. Section 4 details the experimental setup and results. Section 5 concludes this work.

2. Model Formulation

2.1. Signal Model and Crossband Filters

Our study considers an arbitrary room with a single speech source. Let $x(n) \in \mathbb{R}$ be the time-domain clean speech signal, and let $x_{f,t} \in \mathbb{C}$ be the STFT representation of $x(n)$, where $f = 0, \ldots, F-1$ and $t = 0, \ldots, T-1$ are the frequency and time bins, respectively. The speech signal is captured by an array of M microphones. In this work, we assume that the background noise is negligible. Hence, the observed signal in the m-th microphone, $y^{(m)}(n)$, is given by

$$y^{(m)}(n) = \sum_i h^{(m)}(i) x(n-i) \tag{1}$$

where n is the discrete-time index, and $h^{(m)}(n)$ is the RIR from the source $x(n)$ to the m-th microphone. Based on the analysis in [28], the relation between the clean signal $x_{f,t}$ and the observed signal $y^{(m)}_{f,t}$ in the STFT domain is given by

$$y^{(m)}_{f,t} = \sum_{f'=0}^{F-1} \sum_l x_{f',t-l} h^{(m)}_{f;f',l} = \sum_{f'=0}^{F-1} x_{f',t} * h^{(m)}_{f;f',t} \tag{2}$$

where $*$ denotes a linear convolution, and the coefficients $h^{(m)}_{f;f',t} \in \mathbb{C}$ are derived from the time-domain RIR $h^{(m)}(n)$ and from the analysis and synthesis filters that are used to transform the signals from the time domain to the STFT domain and vice versa. Given a single frequency bin f, we consider the time sequence $h^{(m)}_{f;f,t}$ as the band-to-band filter, while

the set of time sequences $\left\{h^{(m)}_{f;f',t}\right\}_{f'\neq f}$ are considered as the crossband filters associated with f.

2.2. Problem Formulation

In the conventional STFT-domain dereverberation problem, the CTF approximation is employed, i.e., the contribution of the crossband filters is neglected, and the observed signal $y_{f,t}$ is approximated as

$$y^{(m)}_{f,t} \approx \sum_{l=0}^{L-1} x_{f,t-l} h^{(m)}_{f;f,l} \tag{3}$$

where we assume a length L band-to-band filter (or CTF) $h^{(m)}_{f;f,t}$. In the context of this paper, the goal of the dereverberation is to predict the late reflection component and subtract it from the observed signal, resulting in an enhanced signal $z^{(m)}_{f,t}$ that consists of the direct sound and the early reflections:

$$z^{(m)}_{f,t} = y^{(m)}_{f,t} - \sum_{l=D}^{L-1} x_{f,t-l} h^{(m)}_{f;f,l} \approx \sum_{l=0}^{D-1} x_{f,t-l} h^{(m)}_{f;f,l} \tag{4}$$

where $0 < D \ll L$ is a predefined parameter that enables separation between early and late reflections.

Extension to Crossband Filters

The model described in (4) can be extended by employing the accurate model in (2). The enhanced signal is then given by

$$z^{(m)}_{f,t} = y^{(m)}_{f,t} - \sum_{f'=0}^{F-1} \sum_{l=D}^{L_{f'}-1} x_{f,t-l} h^{(m)}_{f;f',l}, \tag{5}$$

where $L_{f'}$ is the length of the crossband filter corresponding to frequency bin f'. In terms of computational complexity, the accurate model in (5) is expensive since it increases the complexity by a factor of F compared to the CTF approximation in (4). The analysis in [28] shows that in terms of energy, the band-to-band filter is more significant compared to the crossband filters, and the energy of $h_{f;f',t}$ decreases when $|f - f'|$ increases. Based on this observation, and to improve the accuracy of the CTF model in (3) with a relatively small price of increasing model complexity, we consider the contribution of the two nearest crossband filters, i.e., we approximate the observed and the enhanced signals as

$$y^{(m)}_{f,t} \approx \sum_{f'=f-1}^{f+1} \sum_{l=0}^{L_{|f-f'|}-1} x_{f',t-l} h^{(m)}_{f;f',l}, \tag{6}$$

$$z^{(m)}_{f,t} = y^{(m)}_{f,t} - \sum_{f'=f-1}^{f+1} \sum_{l=D}^{L_{|f-f'|}-1} x_{f,t-l} h^{(m)}_{f;f',l}, \tag{7}$$

where $L_0 := L_{\text{bb}}$ and $L_1 := L_{\text{cb}}$ are the lengths of the band-to-band filter and the crossband filter, respectively.

3. Proposed WPE with Crossband Filtering

3.1. Conventional WPE

The conventional WPE for multichannel input predicts the components of the late reflections based on the LP of an inverse filter [19]. To leverage the spatial information to improve the estimation of the late reflections, the inverse filter predicts the late reflections

based on observations from all channels. More specifically, let $y_{f,t}^{(m)}$ be an observed signal in the STFT domain captured by the m-th microphone of an M length microphone array, and let

$\mathbf{g}_f^{(m)} \in \mathbb{C}^{LM}$ be an LM-order prediction filter. The enhanced signal $z_{f,t}^{(m)}$ is achieved as

$$z_{f,t}^{(m)} = y_{f,t}^{(m)} - \mathbf{g}_f^{(m)H} \mathbf{y}_{f,t;L} \in \mathbb{C}, \tag{8}$$

$$\mathbf{y}_{f,t;L} = \left[\mathbf{y}_{f,t;L}^{(1)}, \ldots, \mathbf{y}_{f,t;L}^{(M)}\right]^T \in \mathbb{C}^{LM}, \tag{9}$$

$$\mathbf{y}_{f,t;L}^{(m)} = \left[y_{f,t-D}^{(m)}, \ldots, y_{f,t-D-L+1}^{(m)}\right] \in \mathbb{C}^{L}, \tag{10}$$

where $(\cdot)^T$ and $(\cdot)^H$ represent the transpose and Hermitian transpose, respectively, and $D > 0$ is the predefined prediction delay. Note that given a frequency bin f, based on the definition of $\mathbf{y}_{f,t;L}$ in (9), the enhanced signal is obtained using only information from samples of the observed signal in the band-to-band frequency bin, ignoring samples from the crossband frequency bins. For simplicity, we select an arbitrary value for m and omit the microphone designation from this point onward.

3.1.1. Filter Estimation

The filter $\mathbf{g}_f$ is estimated in a frequency-wise manner based on the maximum likelihood (ML) criterion, assuming that the signal $\mathbf{z}_f := \left\{z_{f,t}\right\}_t$ follows a complex Gaussian distribution with zero mean and time-dependent variances $\boldsymbol{\lambda}_f := \left\{\lambda_{f,t}\right\}_t$. The filter coefficients $\mathbf{g}_f$ and the variances $\boldsymbol{\lambda}_f$ are alternately estimated to minimize the following objective:

$$\mathcal{L}(\mathbf{g}_f, \boldsymbol{\lambda}_f) = \sum_t \left[\frac{|y_{f,t} - \mathbf{g}_f^H \mathbf{y}_{f,t;L}|^2}{\lambda_{f,t}} + \log \lambda_{f,t}\right]. \tag{11}$$

The entire estimation process for the first channel (i.e., $m = 1$) is described in Algorithm 1. The extension to other channels is straightforward.

Algorithm 1 Conventional WPE Filter Estimation for First Channel

Input:

Observed multichannel signal in STFT domain $\left\{y_{f,t}^{(m)}\right\}_{f,t,m}$

Small constant $\epsilon > 0$, filter length L, number of iterations N

for $f = 0, \ldots, F-1$ **do**

1. Initialize $\lambda_{f,t} = \max\left\{\frac{1}{L}\sum_{t'=t-L/2+1}^{t+L/2}\left|y_{f,t}^{(1)}\right|^2, \epsilon\right\}$
2. **for** $n = 1, \ldots, N$ **do**

   Compute:

   $\Phi_f = \sum_t \frac{\mathbf{y}_{f,t;L}\mathbf{y}_{f,t;L}^H}{\lambda_{f,t}^2}$

   $\phi_f^{(1)} = \sum_t \frac{\mathbf{y}_{f,t;L}y_{f,t}^{(1)}}{\lambda_{f,t}^2}$

   Update:

   Filter: $\mathbf{g}_f^{(1)} = \left(\Phi_f^H \Phi_f\right)^{-1} \Phi_f^H \phi_f^{(1)}$

   Enhanced signal: $z_{f,t}^{(1)} = y_{f,t}^{(1)} - \mathbf{g}_f^{(1)H} \mathbf{y}_{f,t;L}$

   Variances: $\lambda_{f,t} = \max\left\{\frac{1}{L}\sum_{t'=t-L/2+1}^{t+L/2}\left|z_{f,t}^{(1)}\right|^2, \epsilon\right\}$
3. **end for**

end for

3.2. WPE with Crossband Filtering

To consider the contribution of the two nearest crossband filters, we define a prediction filter $\tilde{\mathbf{g}}_f \in \mathbb{C}^{\tilde{L}M}$, where $\tilde{L} = L_{\text{bb}} + 2L_{\text{cb}}$, and an observation vector:

$$\tilde{\mathbf{y}}_{f,t;\tilde{L}} = \left[\mathbf{y}_{f,t;L_{\text{bb}}}^T, \mathbf{y}_{f+1,t;L_{\text{cb}}}^T, \mathbf{y}_{f-1,t;L_{\text{cb}}}^T\right]^T \in \mathbb{C}^{\tilde{L}M}, \tag{12}$$

where we consider $\mathbf{y}_{f,t;L_{\text{bb}}}$ and $\mathbf{y}_{f\pm1,t;L_{\text{cb}}}$ as the "band-to-band component" and "crossband components", respectively. The enhanced signal is now obtained as

$$\tilde{z}_{f,t} = y_{f,t} - \tilde{\mathbf{g}}_f^H \tilde{\mathbf{y}}_{f,t;\tilde{L}} \in \mathbb{C}. \tag{13}$$

The definition of $\tilde{\mathbf{y}}_{f,t;\tilde{L}}$ in (12) forces the enhanced signal to take into account samples from the two nearest crossband frequency bins in addition to the samples from the band-to-band frequency bin. The filter coefficients $\tilde{\mathbf{g}}_f$ are estimated according to the method described in Section 3.1.1 and in Algorithm 1. Here, the term $\mathbf{g}_f^H \mathbf{y}_{f,t;L}$ in (11) is substituted with the term $\tilde{\mathbf{g}}_f^H \tilde{\mathbf{y}}_{f,t;\tilde{L}}$.

We propose two versions of the proposed WPE with crossband filtering. First, we fix $L_{\text{bb}} = L$. The parameter L_{cb} controls this setup's tradeoff between model complexity and model accuracy. When $L_{\text{cb}} = 0$, the proposed method is equivalent to the conventional WPE, and the length of $\tilde{\mathbf{g}}_f$ is equal to the length of $\mathbf{g}_f$. When L_{cb} increases, the accuracy of the model approximation increases, but so does the length of $\tilde{\mathbf{g}}_f$, resulting in larger computational complexity. In the second version, we fix $\tilde{L} = L$. Here, L_{bb} decreases when L_{cb} increases, meaning that early samples from crossband components are taken into account instead of late samples from the band-to-band component. This setup reduces the accuracy of the band-to-band model approximation but maintains fixed computational complexity.

4. Experimental Results

4.1. Data and Setup

To validate the performance of the proposed method, we collected a dataset of 10 clean speech signals from the Deep Noise Suppression (DNS) challenge dataset [31]. To emulate realistic acoustic conditions, we generated acoustic channel RIRs using the image model method [32]. The reverberation levels were controlled by adjusting the wall reflection coefficient parameter. The experimental setup consisted of a uniform linear array with four microphones positioned in a room measuring 6 m by 8 m by 3 m. The speaker was located at coordinates (5, 4, 1.7), while the microphones were placed at different positions along the x-axis. More specifically, the microphones were positioned at $(x, 2, 1.6)$, where x was uniformly distributed from 2.936 to 2.999. The reverberation time (T60) of approximately 300 ms was attained by configuring the absorption coefficient of the room's walls. This choice was informed by our observation that lower values of T60 resulted in a relatively inconspicuous reverberation effect, accompanied by an insubstantial enhancement through the proposed method. Furthermore, various room configurations were comprehensively explored, encompassing alterations in the microphone array's spatial arrangement, inter-microphone spacing, and the speaker's position. More specifically, we conducted a comprehensive exploration of microphone spacing, ranging from 1 cm to 4 cm. Additionally, we meticulously examined various configurations involving offsets in both the microphone array and the speaker's position along both the x-axis and y-axis. Remarkably, despite these deliberate modifications, the experimental outcomes exhibited remarkable consistency across configurations. Given this consistency, we present the outcomes from a representative configuration for brevity and clarity. The clean speech signals and RIRs were sampled at 16 kHz. The multichannel observed signals were generated by convolving the RIRs with the single-channel clean speech signals. Spectral analysis was performed using STFTs with a 512-length Blackman window and a shift of 128 samples between frames. By meticulously designing this experimental setup, we aimed to establish a reliable basis for

evaluating the proposed method's performance. Including diverse speech signals, accurate RIR generation, and careful control of reverberation levels contributed to a comprehensive assessment of the method's effectiveness in enhancing speech signals. Additionally, when performing WPE, we set the predefined prediction delay to $D = 3$ to maintain consistency across the experiments. Using Algorithm 1, we have empirically determined that setting the number of iterations to $N = 3$ is adequate to achieve the convergence of the filter coefficients and the variances. Based on the definition of the prediction filters of order L and $\tilde{L}$ in Section 3, we conducted two types of experiments.

1. Length extension (Ext.): In this first version of the proposed method, denoted as the "length extension", we set the parameter L_{bb} equal to L. This design choice ensured that the samples from the crossband components, which introduced additional computational complexity, were included in the analysis. This specific experiment aimed to demonstrate the significance of the information contained in the crossband components in enhancing the dereverberation performance. While it did introduce computational complexity, this experimental approach allowed us to assess the true potential and effectiveness of the proposed method by leveraging the information-rich crossband components.
2. Length preservation (Pres.): In the second version of the proposed method, denoted as "length preservation", we established $\tilde{L}$ to equal L. To maintain comparable computational complexity to the conventional WPE method, we introduced a modification by discarding the two most recent samples from the band-to-band component for every sample utilized in the crossband components. This adjustment allowed us to strike a balance between computational efficiency and the evaluation of the relative importance of early samples from the crossband components and late samples from the band-to-band component. By discarding the two latest samples from the band-to-band component, we aimed to explore the tradeoff between the temporal characteristics of the crossband and band-to-band components. This experimental design enabled us to assess the respective significance of early samples from the crossband components and late samples from the band-to-band component in the dereverberation process.

4.2. Performance Measure

We varied the length of the crossband components L_{cb} to examine how it affected the performance of the proposed method. We examined the performance in terms of three widely used measures for speech dereverberation [21,33]: the frequency-weighted segmental SNR (FWSegSNR) [34,35], the cepstral distance (CD) [36], and the perceptual evaluation of speech quality (PESQ) [37]. Given a clean ground-truth signal in the STFT domain $x_{f,t}$ and the corresponding enhanced signal $\hat{x}_{f,t}$, FWSegSNR was computed as follows:

$$\mathrm{FWSegSNR} = \frac{10}{T}\sum_{t=0}^{T-1}\frac{\sum_{f=0}^{F-1} w_{f,t}\log_{10}\frac{x_{f,t}^2}{\left(x_{f,t}-\hat{x}_{f,t}\right)^2}}{\sum_{f=0}^{F-1} w_{f,t}}, \tag{14}$$

where F and T are the numbers of frequency bands and time frames, respectively, and $w_{f,t}$ is the weight assigned to the f-th frequency at the t-th frame. We set the weights $w_{f,t}$ according to the standard AI weights [38]. The CD measure is defined as

$$\mathrm{CD} = \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sum_{m=0}^{M-1}\left[C_x(m,t) - C_{\tilde{x}}(m,t)\right]^2}, \tag{15}$$

where $C_x(m,t)$ is the cepstral coefficient of the m-th Mel band of $x_{f,t}$ [36]. It is worth noting that a universally accepted suite of objective quality measures has yet to be fully established within the dereverberation landscape [33]. Given this ongoing evolution, our choice of

performance measures aimed to shed light on the relative strengths and limitations of various approaches. For FWSegSNR and PESQ, larger values indicate better dereverberation performance. For CD, smaller values indicate better performance. To highlight the effectiveness of the method, we considered the "gain" concerning the observed signal, i.e., instead of presenting the absolute measures' values, we offer the following measures:

$$\Delta\text{FWSegSNR} = \text{FWSegSNR}/\text{FWSegSNR}_{\text{observed}}, \tag{16}$$

$$\Delta\text{CD} = \text{CD}_{\text{observed}}/\text{CD}, \tag{17}$$

$$\Delta\text{PESQ} = \text{PESQ}/\text{PESQ}_{\text{observed}}, \tag{18}$$

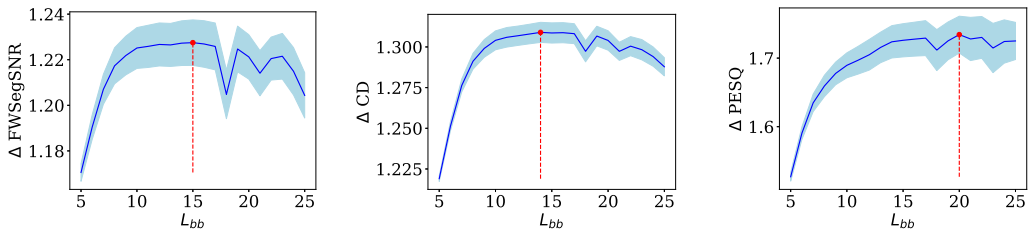
where $(\cdot)_{\text{observed}}$ is the performance measure when considering the observed signal instead of the enhanced signal. Based on this definition, larger values indicate better dereverberation performance across all measures. Values smaller than 1 indicate a degradation in performance. The performance gain was computed individually for each of the 10 speakers. The scores depict the mean improvement across these 10 speakers and the corresponding standard deviation.

4.3. Optimal Band-to-Band Length

To begin our investigation, we performed a series of simulated experiments on the conventional WPE method to identify the optimal filter length L_{bb} within the specific room configuration under consideration. In this set of experiments, we systematically varied the value of L_{bb} in the range of 5 to 25 while keeping the crossband filter length L_{cb} fixed at 0.

The results of these experiments are presented in Figure 1, which showcases the scores obtained for each measure across the range of L_{bb}. Upon close examination, it becomes evident that the optimal filter length varies for different performance measures. Specifically, the FWSegSNR measure attains its peak performance with $L_{\text{bb}} = 15$, while the CD measure achieves its optimal result at $L_{\text{bb}} = 14$. On the other hand, the PESQ measure demonstrates its best performance when L_{bb} is set to 20.

Building upon these findings, we conducted further experiments, focusing exclusively on the optimal values of L_{bb}. Consequently, we set L_{bb} to take on 14, 15, and 20 values, thereby allowing us to thoroughly compare the proposed and conventional WPE methods' performance under these specific settings.



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

Figure 1. Performance of conventional WPE for different filter lengths: (**a**) FWSegSNR—optimal length is 15. (**b**) CD—optimal length is 14. (**c**) PESQ—optimal length is 20.

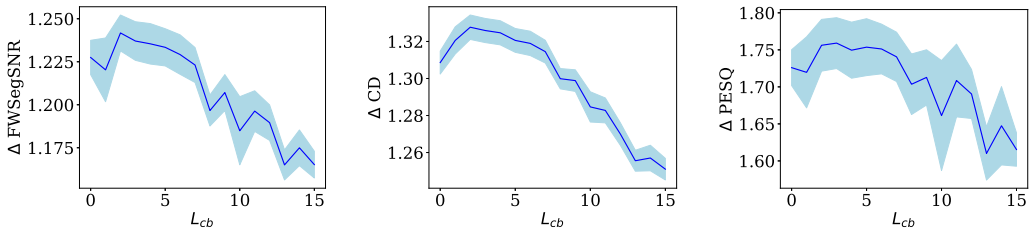
4.4. Crossband Filtering—Length Extension

To thoroughly explore the impact of different crossband filter lengths (L_{cb}) in conjunction with various choices of L_{bb}, we systematically varied L_{cb} within the range of 0 to L_{bb}, while keeping L_{bb} fixed for each specific experiment. Results show that early crossband samples indeed improve the dereverberation performance, and, in all cases (i.e., for each measure and each choice of L_{bb}), the optimal performance is achieved for $L_{\text{bb}} > 0$. To our surprise, introducing late samples from the crossband components leads to a decrease in

performance, even when the length of the band-to-band component remains fixed. This intriguing finding strongly suggests that late samples of the crossband components may have a detrimental effect on the overall performance, even compared to the conventional WPE. Drawing from this observation and the outcomes presented in Figure 1, we propose that both the traditional and proposed methods of WPE attain optimal performance at a specific length choice. Surprisingly, beyond this optimal value, the performance deteriorates, despite the availability of additional information for dereverberation.

4.4.1. Optimal FWSegSNR ($L_{\text{bb}} = 15$)

The observations depicted in Figure 2a,c provide valuable insights into the impact of incorporating the first crossband component on the dereverberation performance, specifically in terms of FWSegSNR and PESQ. Surprisingly, it is evident that introducing the first sample of the crossband component leads to a decrease in performance compared to the conventional WPE method. Conversely, when considering the CD measure (as illustrated in Figure 2b), adding the first sample of the crossband component improves the dereverberation performance. Further analysis reveals that the optimal performance, in terms of FWSegSNR and CD, is achieved when L_{cb} is set to 2, as shown in Table 1. On the other hand, the optimal performance in terms of PESQ is attained when L_{cb} is set to 3. It is worth noting that for $L_{\text{cb}} > 6$, the performance starts to decline, surpassing the level achieved by the conventional WPE method. These intriguing findings shed light on the intricate relationship between different choices of L_{cb} and their impact on the overall dereverberation performance.

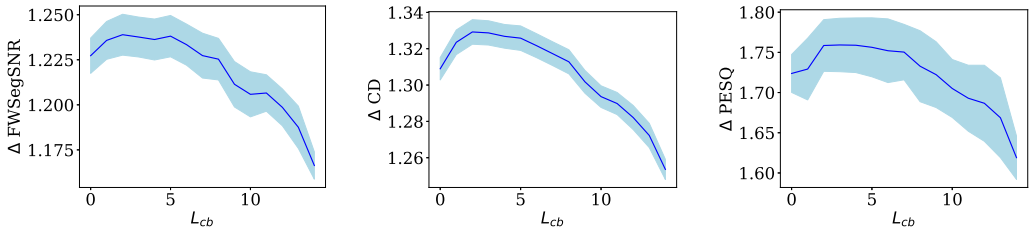


(**a**) (**b**) (**c**)

Figure 2. Performance evaluation (Ext.) for $L_{\text{bb}} = 15$ (optimal FWSegSNR) in terms of (**a**) FWSegSNR, (**b**) CD, and (**c**) PESQ.

4.4.2. Optimal CD ($L_{\text{bb}} = 14$)

The obtained results are demonstrated in Figure 3a–c. Notably, a remarkable improvement in performance is observed from the first sample of the crossband component, as evidenced by the enhancement in all measured metrics. Furthermore, as indicated in Table 1, the optimal performance, both in terms of FWSegSNR and CD, is achieved when L_{cb} is set to 2. Similarly, for optimal performance in terms of PESQ, a value of $L_{\text{cb}} = 3$ is identified. It is worth highlighting that for values of L_{cb} exceeding 7, a noticeable decline in performance is observed compared to the conventional WPE method. This observation further underscores the importance of carefully selecting an appropriate value for L_{cb} to achieve optimal dereverberation results. The presented findings shed light on the effectiveness of integrating the first crossband component and its significant impact in improving the dereverberation performance across various evaluation metrics.

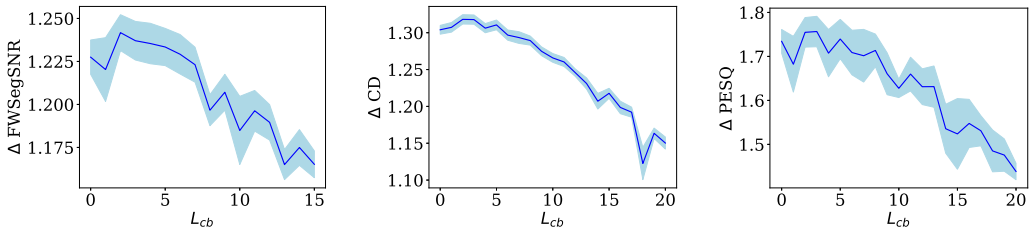


(**a**) (**b**) (**c**)

Figure 3. Performance evaluation (Ext.) for $L_{\text{bb}} = 14$ (optimal CD) in terms of (**a**) FWSegSNR, (**b**) CD, and (**c**) PESQ.

4.4.3. Optimal PESQ ($L_{\text{bb}} = 20$)

The obtained results are depicted in Figure 4a–c, providing insights into the performance characteristics when considering $L_{\text{cb}} = 20$. The observed behavior closely resembles the findings discussed in Section 4.4.1, where the inclusion of the first sample of the crossband component initially leads to a degradation in performance. However, it is noteworthy that a performance improvement becomes evident from the second sample onward. Table 1 reveals that the optimal gain in performance coincides with the choices identified in Section 4.4.1. However, it is worth noting that the optimal gain values are slightly lower for the cases of FWSegSNR and CD. These findings underscore the consistent impact of the crossband component and its potential to enhance the dereverberation performance, albeit with some variation in the optimal gain values across different evaluation metrics.



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

Figure 4. Performance evaluation (Ext.) for $L_{\text{bb}} = 20$ (optimal PESQ) in terms of (**a**) FWSegSNR, (**b**) CD, and (**c**) PESQ.

Table 1. Summary of experimental results (length extension).

Measure	L_{bb}	Optimal Gain	Optimal L_{cb}
FWSegSNR	14	1.23	2
	15	1.24	2
	20	1.22	2
CD	14	1.33	2
	15	1.32	2
	20	1.31	2
PESQ	14	1.76	4
	15	1.76	3
	20	1.75	3

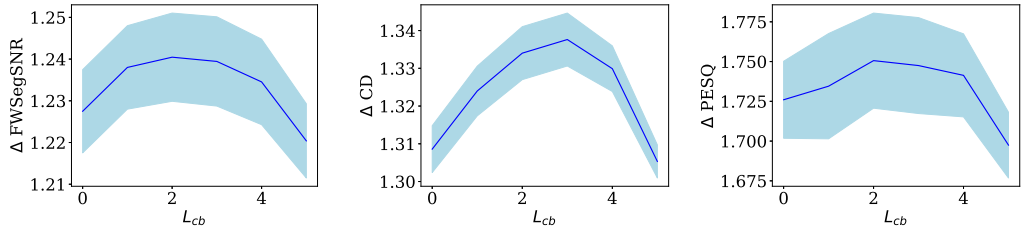
4.5. Crossband Filtering—Length Preservation

In order to evaluate the performance, we set up an experimental configuration where we systematically varied the value of L_{cb} for each chosen L_{bb}. In this setup, we discarded two late samples from the band-to-band component for each increment in the crossband samples in order to maintain fixed computational complexity. Specifically, we explored the range of L_{cb} from 0 to $\lfloor L_{\text{bb}}/3 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function, ensuring that the band-

to-band component retained its significance relative to the crossband components. Notably, introducing early crossband samples led to an overall improvement in dereverberation performance across all measured criteria. This finding underscores the effectiveness of incorporating the information within the crossband components to enhance the quality of the dereverberated speech signals.

4.5.1. Optimal FWSegSNR ($L_{bb} = 15$)

The obtained results are illustrated in Figures 5a–c, providing a comprehensive assessment of the performance. The optimal gains and corresponding values of L_{cb} are summarized in Table 2. Surprisingly, the observed performance is highly competitive with the outcomes presented in Section 4.4.1, despite using fewer data for the dereverberation process. Notably, the introduced method even exhibits an improvement in terms of CD compared to the length extension approach. Optimal performance, in terms of FWSegSNR, is achieved when $L_{cb} = 2$, while, for CD and PESQ, the optimal values are obtained with $L_{cb} = 3$. These findings highlight the efficacy of incorporating early crossband samples, demonstrating their valuable contribution in improving dereverberation outcomes.

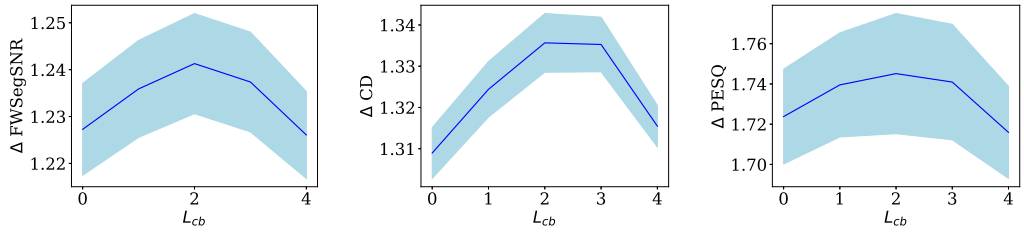


(**a**) (**b**) (**c**)

Figure 5. Performance evaluation (Pres.) for $L_{bb} = 15$ (optimal FWSegSNR) in terms of (**a**) FWSegSNR, (**b**) CD, and (**c**) PESQ.

4.5.2. Optimal CD ($L_{bb} = 14$)

The obtained results are presented in Figures 6a–c and are summarized in Table 2. Interestingly, it is observed that for all measures, the optimal gain is achieved when $L_{cb} = 2$. In this particular setup, the length preservation method outperforms the length extension method in terms of FWSegSNR and CD. However, regarding PESQ, the length extension method provides better and competitive performance. These findings emphasize the significance of considering the specific setup and context when evaluating the performance of different dereverberation methods, as their effectiveness may vary depending on the chosen measures and objectives.



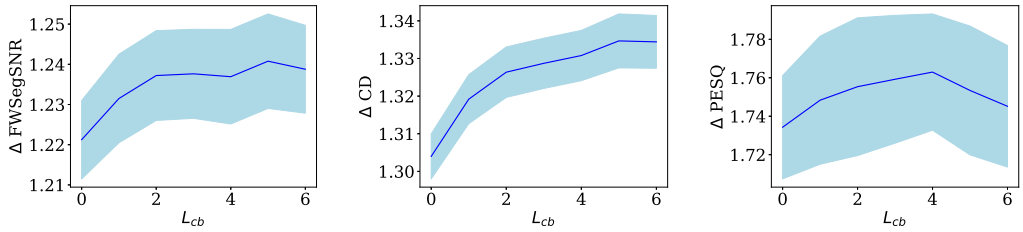
(**a**) (**b**) (**c**)

Figure 6. Performance evaluation (Pres.) for $L_{bb} = 14$ (optimal CD) in terms of (**a**) FWSegSNR, (**b**) CD, and (**c**) PESQ.

4.5.3. Optimal PESQ ($L_{bb} = 20$)

Results are presented in Figures 7a–c and are summarized in Table 2. Notably, in Figure 7a, it is observed that the optimal performance in terms of FWSegSNR is achieved when $L_{cb} = 5$, which corresponds to $L_{bb} = 10$. This finding is further supported by

Figure 1a, which indicates that the performance in terms of FWSegSNR remains relatively stable for L_{bb} in the range of 10 to 16. This suggests that the optimal performance in terms of FWSegSNR, when utilizing the crossband components, is achieved when L_{bb} is in proximity to the optimal value obtained with the conventional WPE. A similar observation can be made for the CD metric, as depicted in Figures 1b and 7b. Furthermore, it is worth mentioning that this experiment yielded the best overall performance in terms of PESQ, as shown in Table 2. These findings highlight the importance of carefully selecting the parameters and considering the specific objectives when evaluating the performance of dereverberation methods.



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

Figure 7. Performance evaluation (Pres.) for $L_{bb} = 20$ (optimal PESQ) in terms of (**a**) FWSegSNR, (**b**) CD, and (**c**) PESQ.

Table 2. Summary of experimental results (length preservation).

Measure	L_{bb}	Optimal Gain	Optimal L_{cb}
FWSegSNR	14	1.24	2
	15	1.24	2
	20	1.23	5
CD	14	1.33	2
	15	1.34	3
	20	1.33	5
PESQ	14	1.74	2
	15	1.75	3
	20	1.77	4

4.6. Discussion

The conducted experiments involving length extension and length preservation methods have provided valuable insights into the performance of the WPE-based dereverberation approach. The results, as summarized in Tables 1 and 2, demonstrate the effectiveness of both methods in improving the performance compared to the conventional WPE while considering different aspects of the evaluation metrics. The length preservation method, which incorporates the early samples of crossband components, has shown competitive performance compared to the length extension method. Remarkably, the length preservation method achieves comparable or superior results across various evaluation metrics, including FWSegSNR, CD, and PESQ. This indicates that by utilizing the crossband components in an optimized manner, the length preservation approach offers an attractive alternative for dereverberation tasks. Notably, the length preservation method achieves these performance gains while maintaining the same computational complexity as the conventional WPE. This is a significant advantage, as it allows for efficient real-time implementation without sacrificing the quality of dereverberation results.

Overall, the findings highlight the importance of considering different approaches and parameters in the WPE-based dereverberation framework. The length preservation method presents a promising avenue for further exploration, offering competitive performance with improved computational efficiency. Further research can investigate the method's

robustness across various real-world scenarios and explore potential optimizations to enhance its effectiveness in different reverberant environments.

5. Conclusions

Our investigation focused on exploring the impact of crossband filters in the STFT domain on WPE-based speech dereverberation. We introduced two extensions to the conventional WPE that specifically accounted for crossband filtering and demonstrated their effectiveness in enhancing the dereverberation performance. Interestingly, the first extension, which increased the model's complexity, naturally improved the performance. However, the second extension maintained the same model complexity as the conventional WPE and exhibited notable performance improvements. This observation suggests that the early samples of the crossband components play a crucial role in dereverberation, surpassing the significance of the late samples from the band-to-band components. Surprisingly, the late samples of the crossband components had an unexpected detrimental effect on the dereverberation performance. To further advance this research area, future investigations can explore the impact of crossband filtering in more complex models, such as scenarios involving speaker switching or time-varying RIRs. Additionally, combining the proposed concept with other extensions of WPE, such as the Kronecker filtering extension, holds promise [26]. The combination of crossband and Kronecker filtering for WPE has the potential to reduce the computational complexity while simultaneously improving the performance, as demonstrated by the recent work on Kronecker filtering for WPE.

Author Contributions: Conceptualization, T.R.; Methodology, T.R., I.C. and E.W.; Software, T.R.; Validation, T.R., I.C. and E.W.; Formal Analysis, T.R.; Investigation, T.R. and I.C; Resources, E.W.; Data Curation, T.R. and E.W.; Writing—Original Draft Preparation, T.R.; Writing—Review and Editing, I.C.; Visualization, T.R.; Supervision, I.C.; Project Administration, I.C. and E.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found in the DNS Challenge dataset—https://github.com/microsoft/DNS-Challenge (accessed on 1 March 2023).

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

References

1. Benesty, J.; Chen, J.; Huang, Y. *Microphone Array Signal Processing*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 1.
2. Brandstein, M.; Ward, D. *Microphone Arrays: Signal Processing Techniques and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2001.
3. Rosenbaum, T.; Cohen, I.; Winebrand, E. Attenuation Of Acoustic Early Reflections In Television Studios Using Pretrained Speech Synthesis Neural Network. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 7422–7426.
4. Boothroyd, A. Room Acoustics and Speech Perception. *Semin. Hear.* **2004**, *25*, 155–166. [CrossRef]
5. Schmid, D.; Enzner, G.; Malik, S.; Kolossa, D.; Martin, R. Variational Bayesian inference for multichannel dereverberation and noise reduction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1320–1335. [CrossRef]
6. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
7. Zheng, C.; Li, X.; Schwarz, A.; Kellermann, W. Statistical analysis and improvement of coherent-to-diffuse power ratio estimators for dereverberation. In Proceedings of the 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, 13–16 September 2016; IEEE: New York, NY, USA, 2016; pp. 1–5.
8. Williamson, D.S.; Wang, D. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1492–1501. [CrossRef]

9. Schwartz, O.; Gannot, S.; Habets, E.A. An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1495–1510. [CrossRef]
10. Inoue, S.; Kameoka, H.; Li, L.; Seki, S.; Makino, S. Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: New York, NY, USA, 2019; pp. 96–100.
11. Habets, E.A.P.; Benesty, J.; Cohen, I.; Gannot, S.; Dmochowski, J. New Insights Into the MVDR Beamformer in Room Acoustics. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 158–170. [CrossRef]
12. Habets, E.A.P.; Benesty, J. A Two-Stage Beamforming Approach for Noise Reduction and Dereverberation. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 945–958. [CrossRef]
13. Habets, E.A. Speech dereverberation using statistical reverberation models. In *Speech Dereverberation*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 57–93.
14. Lebart, K.; Boucher, J.M.; Denbigh, P.N. A new method based on spectral subtraction for speech dereverberation. *Acta Acust. United Acust.* **2001**, *87*, 359–366.
15. Schwartz, O.; Gannot, S.; Habets, E.A. Multi-microphone speech dereverberation and noise reduction using relative early transfer functions. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 240–251. [CrossRef]
16. Miyoshi, M.; Kaneda, Y. Inverse filtering of room acoustics. *IEEE Trans. Acoust. Speech Signal Process.* **1988**, *36*, 145–152. [CrossRef]
17. Jukić, A.; van Waterschoot, T.; Gerkmann, T.; Doclo, S. Multi-channel linear prediction-based speech dereverberation with sparse priors. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1509–1520. [CrossRef]
18. Kinoshita, K.; Delcroix, M.; Nakatani, T.; Miyoshi, M. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 534–545. [CrossRef]
19. Nakatani, T.; Yoshioka, T.; Kinoshita, K.; Miyoshi, M.; Juang, B.H. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 1717–1731. [CrossRef]
20. Yoshioka, T.; Nakatani, T. Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 2707–2720. [CrossRef]
21. Kinoshita, K.; Delcroix, M.; Gannot, S.; Habets, E.A.; Haeb-Umbach, R.; Kellermann, W.; Leutnant, V.; Maas, R.; Nakatani, T.; Raj, B.; et al. A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 7. [CrossRef]
22. Li, B.; Sainath, T.N.; Narayanan, A.; Caroselli, J.; Bacchiani, M.; Misra, A.; Shafran, I.; Sak, H.; Pundak, G.; Chin, K.K.; et al. Acoustic Modeling for Google Home. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 399–403.
23. Kinoshita, K.; Delcroix, M.; Kwon, H.; Mori, T.; Nakatani, T. Neural Network-Based Spectrum Estimation for Online WPE Dereverberation. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 384–388.
24. Ikeshita, R.; Kamo, N.; Nakatani, T. Blind signal dereverberation based on mixture of weighted prediction error models. *IEEE Signal Process. Lett.* **2021**, *28*, 399–403. [CrossRef]
25. Kamo, N.; Ikeshita, R.; Kinoshita, K.; Nakatani, T. Importance of Switch Optimization Criterion in Switching WPE Dereverberation. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 176–180.
26. Huang, G.; Benesty, J.; Cohen, I.; Chen, J. Kronecker product multichannel linear filtering for adaptive weighted prediction error-based speech dereverberation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1277–1289. [CrossRef]
27. Portnoff, M. Time-frequency representation of digital signals and systems based on short-time Fourier analysis. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 55–69. [CrossRef]
28. Avargel, Y.; Cohen, I. System identification in the short-time Fourier transform domain with crossband filtering. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1305–1319. [CrossRef]
29. Avargel, Y.; Cohen, I. On multiplicative transfer function approximation in the short-time Fourier transform domain. *IEEE Signal Process. Lett.* **2007**, *14*, 337–340. [CrossRef]
30. Nakatani, T.; Yoshioka, T.; Kinoshita, K.; Miyoshi, M.; Juang, B.H. Speech dereverberation in short time Fourier transform domain with crossband effect compensation. In Proceedings of the 2008 Hands-Free Speech Communication and Microphone Arrays, Trento, Italy, 6–8 May 2008; IEEE: New York, NY, USA, 2008; pp. 220–223.
31. Dubey, H.; Gopal, V.; Cutler, R.; Matusevych, S.; Braun, S.; Eskimez, E.S.; Thakker, M.; Yoshioka, T.; Gamper, H.; Aichner, R. ICASSP 2022 Deep Noise Suppression Challenge. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022.
32. Allen, J.B.; Berkley, D.A. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950. [CrossRef]
33. Kinoshita, K.; Delcroix, M.; Yoshioka, T.; Nakatani, T.; Habets, E.; Haeb-Umbach, R.; Leutnant, V.; Sehr, A.; Kellermann, W.; Maas, R.; et al. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; IEEE: New York, NY, USA, 2013; pp. 1–4.
34. Ma, J.; Hu, Y.; Loizou, P.C. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.* **2009**, *125*, 3387–3405. [CrossRef] [PubMed]

35. Liu, Z.; Ma, H.T.; Chen, F. A new data-driven band-weighting function for predicting the intelligibility of noise-suppressed speech. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; IEEE: New York, NY, USA, 2017; pp. 492–496.
36. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of the IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Victoria, BC, Canada, 19–21 May 1993; Volume 1, pp. 125–128. [CrossRef]
37. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; Proceedings (Cat. No. 01CH37221); IEEE: New York, NY, USA, 2001; Volume 2, pp. 749–752.
38. *ANSI/ASA S3.5-1997*; Methods for Calculation of the Speech Intelligibility Index. American National Standard: Washington, DC, USA, 1997.