CONVOLUTIONAL WEIGHTED MINIMUM MEAN SQUARE ERROR FILTER FOR JOINT SOURCE SEPARATION AND DEREVERBERATION

Mieszko Fraś, Marcin Witkowski, and Konrad Kowalczyk

AGH University of Science and Technology, Institute of Electronics, Kraków, Poland
{fras, witkow, konrad.kowalczyk}@agh.edu.pl

ABSTRACT

Practical scenarios with multiple simultaneously active speakers recorded using one or more microphones in reverberant rooms pose a challenging problem when the extraction of the desired speaker signal is sought for. The majority of techniques found in the literature facilitate either source separation or dereverberation, which can at best be performed as subsequent, cascade processing. Recently, a solution to the joint task has been proposed, which is known as the weighted power minimization distortionless response (WPD) beamformer. In this paper, we derive a convolutional multichannel filter which performs jointly optimum dereverberation and desired source signal extraction. We formulate a single optimization criterion which minimizes the convolutional source-variance weighted mean square error (CW-MMSE), thereby effectively unifying the weighted prediction error (WPE) based dereverberation and MMSE filtering for the desired source extraction from reverberant mixtures of speakers. Experimental results show a significant performance improvement over the compared state-of-the-art methods such as WPD for datasets with simulated and recorded impulse responses.

Index Terms— source separation, dereverberation, minimum mean square error, Wiener filter, weighted prediction error

1. INTRODUCTION

We consider the task of sound source separation in reverberant environments. For source separation, adaptive beamforming techniques can be used, for example, the minimum variance distortionless response (MVDR) beamformer [1, 2] that minimizes the variance of the output signal subject to a distortionless constraint on the signals coming from the desired direction. Another approach often encountered in signal enhancement is the minimum mean square error (MMSE) estimate of the desired signal obtained by the so-called multichannel Wiener filter (MWF) [3]. This method exploits the knowledge about the time-varying variance to more strongly attenuate the undesired signal component. The latter approach is able to surpass the results obtained by MVDR, especially for small microphone arrays. However, its performance strongly depends on the accuracy with which the parameters are estimated and may distort the source signal [4]. For dereverberation, convolutional filtering such as the weighted prediction error (WPE) method or variants thereof have been successfully used in a wide range of applications [5–10]. Recent developments in robust estimation of the source parameters in challenging highly reverberant conditions, such as those in [11–16], are prompting the search for filters that exploit these estimates for more aggressive signal enhancement. However, the disjoint separation and dereverberation strongly affect each other, and the integration of both approaches still remains a challenge.

Recently, a method that unifies the adaptive MVDR beamformer and the WPE-based convolutional filter into a single weighted power minimization distortionless response convolutional beamformer (WPD) has been proposed in [17]. Although WPD outperforms many conventional methods, including the simple cascade of WPE [6] and MVDR filters [18, 19], it struggles to enhance speech signal in time-frequency bins dominated by the desired source or when the undesired signal components impinge at the microphone array from the direction of the desired speaker.

In this paper, we propose a novel convolutional multichannel filter which allows for jointly optimal dereverberation and source signal extraction from a convolutive mixture of multiple sources. We formulate the proposed optimization criterion which minimizes the source-variance weighted mean square error between the unobserved source signal and the signal estimated from the reverberant mixture. We then derive the closed-form solution for jointly optimum WPE-based dereverberation and the multichannel MMSE-based filtering. The proposed filtering is suitable for the joint task of source separation and dereverberation. In experiments performed in real and simulated rooms, we compare the proposed CW-MMSE filter with the WPD beamformer and with a cascade processing of WPE-based dereverberation and MWF-based filtering using standard source separation and dereverberation evaluation measures. The results show great performance improvement offered by the proposed filter over the WPD and disjointly optimized filters.

The rest of the paper is organized as follows. Section 2 describes the signal model. The derivation of the proposed convolutional variance weighted MMSE filter and source parameter estimation are described in Section 3. Section 4 presents experimental results and their discussion, while Section 5 provides concluding remarks.

2. SIGNAL MODEL

Consider a scenario in which J source signals are captured by I microphones in a reverberant environment. The resulting microphone mixture $\boldsymbol{x}_n = [X_{1n}, X_{2n}, \ldots, X_{In}]^{\mathrm{T}} \in \boldsymbol{C}^I$ in the Short-Time Fourier Transform (STFT) domain can be modeled as

$$\boldsymbol{x}_n = \sum_{j=1}^{J} \boldsymbol{y}_n^{(j)}, \tag{1}$$

where vector $\boldsymbol{y}_n^{(j)} = [Y_{1n}^{(j)}, Y_{2n}^{(j)}, \ldots, Y_{In}^{(j)}]^{\mathrm{T}} \in \boldsymbol{C}^I$ contains the spatial image that represents the j-th source as captured by I microphones and n denotes a time frame index. In this paper, the frequency indices are omitted for brevity, assuming that the processing

This research received financial support from the Foundation for Polish Science under grant number First TEAM/2017-3/23 (POIR.04.04.00-00-3FC4/17-00) which is co-financed by the European Union under the European Regional Development Fund and from the National Science Centre under grant number DEC-2017/25/B/ST7/01792.

is performed independently for each frequency bin. The spatial image $\boldsymbol{y}_n^{(j)}$ for each source can be further decomposed into a sum of the early component $\boldsymbol{s}_n^{(j)}$ and late reverberation component $\boldsymbol{r}_n^{(j)}$ as

$$\boldsymbol{y}_n^{(j)} = \boldsymbol{s}_n^{(j)} + \boldsymbol{r}_n^{(j)}, \tag{2}$$

$$\boldsymbol{s}_n^{(j)} = \boldsymbol{v}^{(j)} S_n^{(j)}, \tag{3}$$

$$\boldsymbol{r}_n^{(j)} = \sum_{\tau=b}^{L_a+b-1} \boldsymbol{a}_\tau^{(j)} S_{n-\tau}^{(j)}, \tag{4}$$

where $\boldsymbol{v}^{(j)} = [1, v_2^{(j)}/v_1^{(j)}, \ldots, v_I^{(j)}/v_1^{(j)}] \in \boldsymbol{C}^I$ is a relative acoustic transfer function under the assumption that the duration of the room impulse response (RIR) corresponding to the early component is short enough in comparison with the length of the analysis time window [20], $S_n^{(j)} \in \boldsymbol{C}$ denotes an anechoic source signal as captured by the reference (first) microphone, and $\boldsymbol{a}_\tau^{(j)} \in \boldsymbol{C}^I$ is a convolutional transfer function for $\tau = b, b+1, \ldots, L_a+b-1$, where b is the frame index that divides the signal into the early and late components, and L_a is the length of the convolutional transfer function.‘

3. OPTIMUM MMSE FILTER FOR JOINT SOURCE SEPARATION AND DEREVERBERATION

In this work, the aim is to retrieve the desired early component of each j-th source $\boldsymbol{s}_n^{(j)}$ by reducing its late reverberation $\boldsymbol{r}_n^{(j)}$ and signals of all non-target sources $\boldsymbol{y}_n^{(j')}$ for $j' \neq j$. Without loss of generality, this section describes a method which finds $\widehat{S}_n^{(j)}$ that denotes the estimate of the desired signal at the first, reference microphone.

3.1. Proposed convolutional weighted MMSE filter

This section presents the derivation of the proposed convolutional weighted MMSE (CW-MMSE) filter which retrieves the desired early source signal component under a single unified optimization criterion for joint dereverberation and signal extraction from the source mixture. Let us begin by formulating the MMSE filter that is applied to the multichannel output of the WPE algorithm [6] on I microphone signals, which after [18] can be written as

$$\begin{aligned} \widehat{S}_n^{(j)} &= (\boldsymbol{w}_{0,n}^{(j)})^{\mathrm{H}} \left(\boldsymbol{x}_n - \sum\nolimits_{\tau=b}^{L+b-2} \boldsymbol{W}_\tau^{\mathrm{H}} \boldsymbol{x}_{n-\tau} \right) \\ &= (\boldsymbol{w}_{0,n}^{(j)})^{\mathrm{H}} \boldsymbol{x}_n + \sum\nolimits_{\tau=b}^{L+b-2} (\boldsymbol{w}_\tau^{(j)})^{\mathrm{H}} \boldsymbol{x}_{n-\tau} \\ &= (\bar{\boldsymbol{w}}_n^{(j)})^{\mathrm{H}} \, \bar{\boldsymbol{x}}_n, \end{aligned} \tag{5}$$

where $\boldsymbol{w}_{0,n}^{(j)} = [w_{1,n}^{(j)}, w_{2,n}^{(j)}, \ldots, w_{I,n}^{(j)}]^{\mathrm{T}} \in \boldsymbol{C}^I$ is the vector of filter coefficients, $\boldsymbol{W}_\tau \in \boldsymbol{C}^{I \times I}$ for $\tau = b, b+1, \ldots, b-L-2$ is a prediction matrix of the WPE algorithm, and L denotes the length of the convolutional filter. Next, the MMSE and WPE filters may be unified by forming vector $\boldsymbol{w}_n^{(j)} = -\boldsymbol{W}_n \boldsymbol{w}_{0,n}^{(j)}$ based on which the extended vectors $\bar{\boldsymbol{x}}_n \in \boldsymbol{C}^{IL}$ and $\bar{\boldsymbol{w}}_n^{(j)} \in \boldsymbol{C}^{IL}$ are defined as

$$\bar{\boldsymbol{x}}_n = [\boldsymbol{x}_n^{\mathrm{T}}, \boldsymbol{x}_{n-b}^{\mathrm{T}}, \boldsymbol{x}_{n-b-1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_{n-L-b+2}^{\mathrm{T}}])^{\mathrm{T}}, \tag{6}$$

$$\bar{\boldsymbol{w}}_n^{(j)} = [(\boldsymbol{w}_{0,n}^{(j)})^{\mathrm{T}}, (\boldsymbol{w}_{n-b}^{(j)})^{\mathrm{T}}, (\boldsymbol{w}_{n-b-1}^{(j)})^{\mathrm{T}}, \ldots, (\boldsymbol{w}_{n-L-b+2}^{(j)})^{\mathrm{T}}]^{\mathrm{T}}. \tag{7}$$

Note that until this point, the formulations are in essence similar to those presented in [18] for derivation of the WPD beamformer.

In this work, we aim to derive an optimum convolutional multichannel filter which minimizes the variance-weighted mean square error between the unobserved desired source signal $S_n^{(j)}$ and the estimated signal $\widehat{S}_n^{(j)}$ from the reverberant mixture of multiple speakers. Assuming that $\boldsymbol{y}_n^{(j)}$ is modeled by a zero-mean complex Gaussian distribution with time-varying variance $\phi_{s,n}^{(j)}$, we propose to define the optimization criterion for the convolutional weighted MMSE as

$$\bar{\boldsymbol{w}}_n^{(j)} = \arg\min_{\bar{\boldsymbol{w}}_n^{(j)}} \frac{|S_n^{(j)} - (\bar{\boldsymbol{w}}_n^{(j)})^{\mathrm{H}} \, \bar{\boldsymbol{x}}_n|^2}{\phi_{s,n}^{(j)}} \, . \tag{8}$$

The closed-form solution to the proposed CW-MMSE criterion can be derived by computing the derivative of the cost function given by (8) with the respect to $(\bar{\boldsymbol{w}}_n^{(j)})^{\mathrm{H}}$ and setting it to zero, which yields

$$\mathrm{E}\left\{ \frac{\bar{\boldsymbol{x}}_n \bar{\boldsymbol{x}}_n^{\mathrm{H}} \bar{\boldsymbol{w}}_n^{(j)}}{\phi_{s,n}^{(j)}} - \frac{\bar{\boldsymbol{x}}_n (S_n^{(j)})^*}{\phi_{s,n}^{(j)}} \right\} = 0 \, , \tag{9}$$

where $\mathrm{E}\{\cdot\}$ denotes the statistical expectation operator. Rearranging (9) we can derive the optimum vector with filter coefficients as

$$\bar{\boldsymbol{w}}_n^{(j)} = \left(\frac{\mathrm{E}\{\bar{\boldsymbol{x}}_n \bar{\boldsymbol{x}}_n^{\mathrm{H}}\}}{\phi_{s,n}^{(j)}} \right)^{-1} \frac{\mathrm{E}\{\bar{\boldsymbol{x}}_n (S_n^{(j)})^*\}}{\phi_{s,n}^{(j)}}. \tag{10}$$

Next, under the assumption that $\boldsymbol{y}_n^{(j)}$ among different sources are mutually uncorrelated, as well the desired early and undesired late reverberation components of the desired source, $\boldsymbol{s}_n^{(j)}$ and $\boldsymbol{r}_n^{(j)}$, are also uncorrelated [5], we can express the terms under statistical expectation in (10) as

$$\mathrm{E}\left\{ \bar{\boldsymbol{x}}_n (S_n^{(j)})^* \right\} = \bar{\boldsymbol{v}}^{(j)} \, \phi_{s,n}^{(j)}, \tag{11}$$

$$\mathrm{E}\left\{ \bar{\boldsymbol{x}}_n \bar{\boldsymbol{x}}_n^{\mathrm{H}} \right\} = \bar{\boldsymbol{v}}^{(j)} \phi_{s,n}^{(j)} (\bar{\boldsymbol{v}}^{(j)})^{\mathrm{H}} + \mathrm{E}\left\{ \bar{\boldsymbol{x}}_{u,n}^{(j)} (\bar{\boldsymbol{x}}_{u,n}^{(j)})^{\mathrm{H}} \right\}, \tag{12}$$

where $\bar{\boldsymbol{v}}^{(j)} = [(\boldsymbol{v}^{(j)})^{\mathrm{T}}, \boldsymbol{0}_{I(L-1)}^{\mathrm{T}}]^{\mathrm{T}} \in \boldsymbol{C}^{IL}$ with $\boldsymbol{0}_{I(L-1)}$ defined as a vector of zeros of length $I(L-1)$, the source signal variance is given by $\phi_{s,n}^{(j)} = \mathrm{E}\{S_n^{(j)} (S_n^{(j)})^*\}$, and $\bar{\boldsymbol{x}}_{u,n}^{(j)} \in \boldsymbol{C}^{IL}$ is the extended vector with undesired components of $\bar{\boldsymbol{x}}_n$ which contains the reverberant component of the j-th source and the entire signals of all other sources. Substituting (11) and (12) and applying the well-known Sherman-Morrison formula [21], we can rewrite (10) as

$$\bar{\boldsymbol{w}}_n^{(j)} = \left[(R_u^{(j)})^{-1} - \frac{(R_u^{(j)})^{-1} \bar{\boldsymbol{v}}^{(j)} (\bar{\boldsymbol{v}}^{(j)})^{\mathrm{H}} (R_u^{(j)})^{-1}}{1 + (\bar{\boldsymbol{v}}^{(j)})^{\mathrm{H}} (R_u^{(j)})^{-1} \bar{\boldsymbol{v}}^{(j)}} \right] \bar{\boldsymbol{v}}^{(j)}, \tag{13}$$

where we define the weighted undesired signal covariance matrix as

$$R_u^{(j)} = \mathrm{E}\left\{ \frac{\bar{\boldsymbol{x}}_{u,n}^{(j)} (\bar{\boldsymbol{x}}_{u,n}^{(j)})^{\mathrm{H}}}{\phi_{s,n}^{(j)}} \right\}. \tag{14}$$

We then note that the power spectral density of the entire undesired signal component $\phi_{u,n}^{(j)}$ can be expressed by the following relation:

$$\phi_{u,n}^{(j)} = \left[(\bar{\boldsymbol{v}}^{(j)})^{\mathrm{H}} \left(\mathrm{E}\left\{ \bar{\boldsymbol{x}}_{u,n}^{(j)} (\bar{\boldsymbol{x}}_{u,n}^{(j)})^{\mathrm{H}} \right\} \right)^{-1} \bar{\boldsymbol{v}}^{(j)} \right]^{-1}. \tag{15}$$

Substituting (15) into (13), we reformulate the proposed CW-MMSE filter into the convolutional variance weighted beamformer $\boldsymbol{h}_{\mathrm{CWB}}$ followed by a single-channel post-filter H_{CWPF}. The final closed-form solution to the proposed CW-MMSE optimization is given by

$$\bar{\boldsymbol{w}}_n^{(j)} = \underbrace{\frac{\xi_n^{(j)}}{1 + \xi_n^{(j)}}}_{H_{\mathrm{CWPF}}} \; \underbrace{\frac{(R_u^{(j)})^{-1} \bar{\boldsymbol{v}}^{(j)}}{(\bar{\boldsymbol{v}}^{(j)})^{\mathrm{H}} (R_u^{(j)})^{-1} \bar{\boldsymbol{v}}^{(j)}}}_{\boldsymbol{h}_{\mathrm{CWB}}}, \tag{16}$$

where the weighted convolutive covariance matrix of undesired signal $R_u^{(j)}$ can be conveniently estimated using equation

$$R_u^{(j)} = \sum_{n=0}^{N} \frac{\bar{\boldsymbol{x}}_{u,n}^{(j)} \, (\bar{\boldsymbol{x}}_{u,n}^{(j)})^{\mathrm{H}}}{\phi_{s,n}^{(j)}}, \tag{17}$$

and $\xi_n^{(j)}$ denotes the a posteriori SNR estimated using equation [4]

$$\xi_n^{(j)} = \frac{\boldsymbol{h}_{\mathrm{CWB}}^{\mathrm{H}} \, \bar{\boldsymbol{x}}_{s,n}^{(j)} \, (\bar{\boldsymbol{x}}_{s,n}^{(j)})^{\mathrm{H}} \, \boldsymbol{h}_{\mathrm{CWB}}}{\boldsymbol{h}_{\mathrm{CWB}}^{\mathrm{H}} \, \bar{\boldsymbol{x}}_{u,n}^{(j)} \, (\bar{\boldsymbol{x}}_{u,n}^{(j)})^{\mathrm{H}} \, \boldsymbol{h}_{\mathrm{CWB}}}, \tag{18}$$

where $\bar{\boldsymbol{x}}_s^{(j)}$ is the desired j-th source component of $\bar{\boldsymbol{x}}$.

Compuation of $R_u^{(j)}$ using (17) yields robust weighted convolutive covariance matrix estimation, similarly to the estimation procedure described in [17–19]. A final note can be made on the computational complexity of the proposed CW-MMSE filter compared to the WPD beamformer [17–19]. Since the complexity of computing the post-filter weights is negligible compared with the cost of computing the convolutional weighted beamformer $\boldsymbol{h}_{\mathrm{CWB}}$, the overall computational cost of the proposed CW-MMSE and WPD filters is practically the same.

3.2. Estimation of source parameters

For computation of the proposed CW-MMSE filter given by (16), one needs to build the extended vectors with the desired source components $\bar{\boldsymbol{x}}_{s,n}^{(j)}$, the undesired components $\bar{\boldsymbol{x}}_{u,n}^{(j)}$, and the steering vector $\bar{\boldsymbol{v}}^{(j)}$. All of these signals can be conveniently formed based on only two parameters related to the desired j-th source to be dereverberated and separated from the J-source mixture, namely the steering vector $\boldsymbol{v}^{(j)}$ and the time-varying source variance $\phi_{s,n}^{(j)}$. Based on the estimates of these parameters for the desired source, we can compute the required extended vectors using relations:

$$\boldsymbol{x}_{s,n}^{(j)} = \bar{\boldsymbol{v}}^{(j)} \phi_{s,n}^{(j)}, \tag{19}$$

$$\bar{\boldsymbol{x}}_{s,n}^{(j)} = [(\boldsymbol{x}_{s,n}^{(j)})^{\mathrm{T}}, (\boldsymbol{x}_{s,n-b}^{(j)})^{\mathrm{T}}, (\boldsymbol{x}_{s,n-b-1}^{(j)})^{\mathrm{T}}, \ldots, (\boldsymbol{x}_{s,n-b-L+2}^{(j)})^{\mathrm{T}}]^{\mathrm{T}}, \tag{20}$$

and

$$\bar{\boldsymbol{x}}_u^{(j)} = \bar{\boldsymbol{x}}^{(j)} - \bar{\boldsymbol{x}}_s^{(j)}. \tag{21}$$

3.2.1. Oracle scenario

In the oracle scenario, the desired j-th source power $\phi_{s,n}^{(j)}$ is calculated for each time-frequency bin from the STFT spectrum of the original non-reverberant speech sample convolved with the early RIR component. For the steering vectors $\boldsymbol{v}^{(j)}$, the relative acoustic transfer function corresponding to the early RIR component is used.

3.2.2. Blind scenario

In the blind scenario, the parameters are estimated using the recently proposed source separation algorithm based on nonnegative tensor factorization that is known as sub-source-based Expectation-Maximization algorithm with Multiplicative Updates and Localization Prior SSEM-MU-LP [16]. This algorithm is able to provide reliable estimates of source variances and steering vectors even for a small microphone array which records several sound sources in a highly reverberant environment. This algorithm was chosen arbitrarily and any other method capable of estimating time-varying source variances and their steering vectors could alternatively be used, a selection of suitable techniques are e.g. described in [11–15].

4. EXPERIMENTAL RESULTS

4.1. Experimental setup and evaluation criteria

The evaluation of the proposed CW-MMSE filter was performed using two datasets of two-channel reverberant microphone mixtures synthesized by convolving anechoic speech recordings with the room impulse responses. In the first dataset, the RIRs were generated using the image-source method [22]. We simulated a room of size $10 \times 10 \times 4$ m with the reverberation time (RT) of 300, 600 and 900 ms. The microphone array with inter-microphone spacing of 0.05 m was positioned at random around the room center, while the two speakers were located randomly at a distance of around 2 m from the array. In the second dataset, the RIRs were taken from the MIRD database [23] for the reverberation time of 610 ms. We randomly selected 5 pairs of RIRs from 13 available measurement positions for a constant speaker-microphone distance of 2 m and a 2-element microphone array with inter-microphone spacing of 0.06 m. The Librispeech *test-clean* part [24] was used to select pairs of non-reverberant speech signals with the sample length adjusted to the longer recording by means of zero-padding. In particular, we made sure that different utterances and speakers appear in each pair of speech signals. The microphone mixture signals were then obtained by convolving all pairs of non-reverberant speech with randomly selected pairs of RIRs resulting in total in 1310 reverberant two-channel recordings of two speakers for each of the four considered reverberant scenarios. The signals were sampled at 16 kHz and processed with a 512 point STFT with 50% overlap.

The proposed CW-MMSE method is compared against two other methods. The first reference method is the WPD beamformer described in [18] which has been shown to outperform many conventional methods such as a cascade processing of the WPE and the MPDR beamformer. The second compared method, which we refer to as Cascade, consists of the WPE [6] for dereverberation followed by source separation with the well-known multichannel Wiener filter (MWF) [4] implemented similarly to the proposed method, i.e. as MVDR beamformer with a post-filter. For each performed experiment and method, the same set of parameters $\phi_{s,n}^{(j)}$ and $\bar{\boldsymbol{v}}^{(j)}$, $b = 1$, and $L = 15$ was used for both the oracle and blind scenarios. Only in simulated scenarios with the lowest reverberation time of 300, the length of the convolutional filter was reduced to $L = 7$.

The efficacy of the compared processing was assessed using standard source separation and dereverberation metrics. Specifically, the signal-to-interference ratio (SIR) [25] and the signal-to-distortion ratio (SDR) [25] were used to assess the separation performance, while the frequency-weighted segmental signal-to-noise ratio (FWSNR) [26] and cepstral distance (CD) [26] were used to evaluate the overall enhancement of desired signal by dereverberation and source separation. All presented results show improvements in those measure values between the filtered (output) and unprocessed (input) signals averaged over 1310 recordings and $J = 2$ source signals. As a reference signal, we use the early component of the desired speech as captured by the reference microphone. Irrespective of the metric, the larger improvement value always indicates better performance.

4.2. Results and discussion

The results of experiments performed on the dataset created using the synthesized RIRs in rooms with different reverberation time values of 0.3, 0.6 and 0.9 s are presented in Table 1. As can be clearly observed, the proposed CW-MMSE filter significantly outperforms

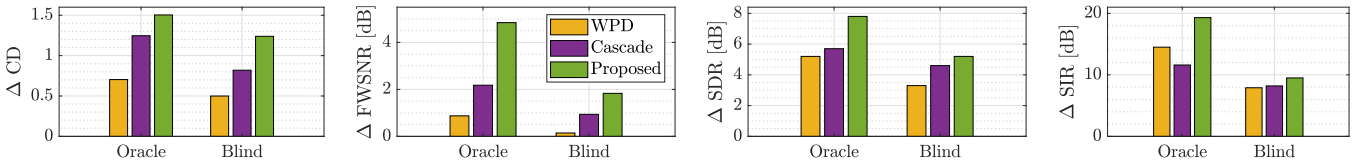

Fig. 1. Improvements (denoted with Δ) in CD, FWSNR, SDR, and SIR metrics for the WPD, Cascade and Proposed filters over the unprocessed (input) microphone signal for genuine RIRs from the MIRD database [23] measured using a 2-element array located 2 m from the source in a room with RT of 0.61 s. All presented values for the oracle and blind scenarios were averaged over 2620 (2x1310) results.

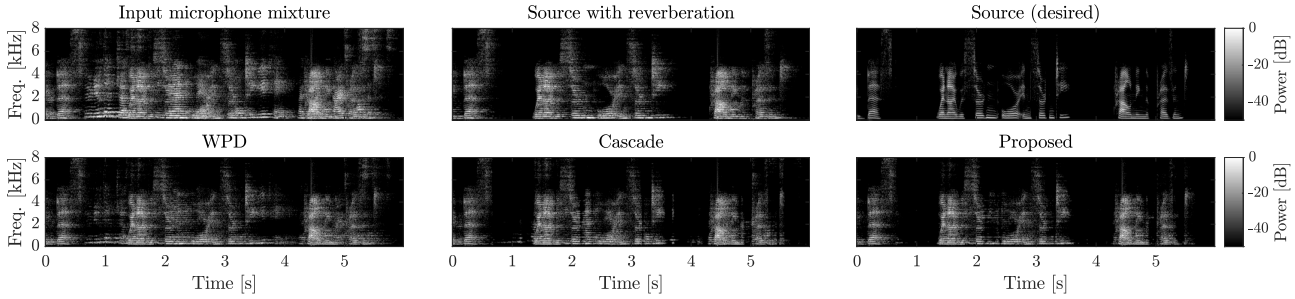

Fig. 2. Examplary spectrograms of the unprocessed (input) microphone signal, the reverberant source signal, the desired source signal and the outputs of the WPD, the Cascade and the Proposed filters for the blind scenario with genuine (measured) RIRs at the RT of 0.61 s.

both compared techniques in terms of all presented evaluation measures in the oracle and blind scenarios, irrespective of the reverberation time values. In particular, the results for the oracle scenario, which can be seen as a performance upper bound, show that the knowledge of the time-varying variance of the desired speech component allows for a much stronger suppression of undesired sources and late reverberation by the proposed filter in comparison with the WPD and the Cascade of WPE and MWF (reaching as much as 50% relative gain in ΔSIR). The results for the blind scenario also confirm that the proposed CW-MMSE method can be successfully used in practical setups when the source parameters need to be estimated blindly from the microphone mixture.

The results of experiments performed with the dataset created using real RIRs are depicted in Fig. 1, and in general they confirm the aforementioned main conclusions. Interestingly, although the Cascade filtering achieves similar performance as WPD in source separation for all oracle scenarios, WPD is shown to be more robust in source separation in practical blind scenarios. On the other hand, Cascade filtering always outperforms WPD in term of dereverberation. By far the best performance in joint dereverberation and source separation is nonetheless always achieved by the proposed CW-MMSE filter. For instance, the improvement over the WPD reaches as much as 300% for ΔFWSNR and 200% for ΔCD. Poor performance of WPD can be attributed to the fact that the use of the time-varying variance only as a weighting factor in covariance matrix computation has a limited capability to enhance the signal in those time-frequency bins that are dominated by the desired source.

Figure 2 shows exemplary spectrograms of the unprocessed (input) microphone mixture, the desired and reverberant source signal, and the signals resulting from all three compared filters. As can be seen, the proposed CW-MMSE filter clearly outperforms the other two filters in suppression of late reverberation of the desired speech and speech of interfering speakers. The benefit of using a post-filter by the Cascade and Proposed filters is exhibited e.g. near 1 and 4 s in which the interfering speaker is more effectively removed than by using the WPD filter.

Table 1. Improvements (Δ) in CD, FWSN [dB], SDR [dB], and SIR [dB] metrics for the WPD, Cascade and Proposed filters over the unprocessed microphone signal for simulated rooms with the RT of 0.3, 0.6 and 0.9 s. All scores were averaged over 2620 results.

Scen.	RT	Method	ΔCD	ΔFWSNR	ΔSDR	ΔSIR
Oracle	0.3	WPD	1.4	4.6	12.8	22.9
		Cascade	2.2	4.6	10.9	18.0
		Proposed	2.4	6.0	14.0	29.5
	0.6	WPD	1.0	2.4	7.3	16.2
		Cascade	1.7	4.9	7.3	16.3
		Proposed	1.7	7.0	9.7	24.1
	0.9	WPD	0.7	1.1	6.0	13.6
		Cascade	1.3	3.5	5.8	15.5
		Proposed	1.4	5.6	7.7	21.0
Blind	0.3	WPD	1.3	3.9	10.6	19.6
		Cascade	0.8	2.6	8.5	12.7
		Proposed	1.8	4.0	10.7	22.4
	0.6	WPD	0.9	1.6	6.8	13.6
		Cascade	1.2	3.1	5.4	10.7
		Proposed	1.5	4.1	8.1	16.9
	0.9	WPD	0.6	0.7	5.2	10.9
		Cascade	0.9	2.3	4.2	9.8
		Proposed	1.2	3.0	5.8	13.2

5. CONCLUSIONS

In this paper we have introduced a novel filter for joint source separation and dereverberation. The derived convolutive weighted MMSE filter has been shown to significantly outperform state-of-the-art signal enhancement techniques for the joint task in terms of all investigated evaluation measures in the performed experiments.

6. REFERENCES

[1] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, 1973.

[2] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[4] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2009.

[5] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[6] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[7] K. Kinoshita *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. on Advances in Signal Process.*, vol. 2016, no. 1, pp. 1–19, 2016.

[8] M. Witkowski, M. Rybicka, and K. Kowalczyk, "Sparse linear prediction-based dereverberation for signal enhancement in distant speaker verification," in *Proc. European Signal Process. Conf. (EUSIPCO)*. IEEE, 2021, pp. 461–465.

[9] M. Witkowski and K. Kowalczyk, "Split Bregman approach to linear prediction based dereverberation with enforced speech sparsity," *IEEE Signal Process. Letters*, vol. 28, pp. 942–946, 2021.

[10] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[11] T. Nakatani *et al.*, "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2020, pp. 6399–6403.

[12] N. Ito, C. Schymura, S. Araki, and T. Nakatani, "Noisy cGMM: Complex Gaussian mixture model with non-sparse noise model for joint source separation and denoising," in *Proc. IEEE Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2018, pp. 1662–1666.

[13] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[14] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2018, pp. 31–35.

[15] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[16] M. Fraś and K. Kowalczyk, "Maximum a Posteriori estimator for convolutive sound source separation with sub-source based NTF model and the localization probabilistic prior on the mixing matrix," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2021, pp. 526–530.

[17] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Letters*, vol. 26, no. 6, pp. 903–907, 2019.

[18] T. Nakatani, K. Kinoshita, R. Ikeshita, H. Sawada, and S. Araki, "Simultaneous denoising, dereverberation, and source separation using a unified convolutional beamformer," in *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust. (WASPAA)*. IEEE, 2019, pp. 224–228.

[19] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2267–2282, 2020.

[20] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[21] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays*. Springer, 2001, pp. 39–60.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[23] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IEEE Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[25] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Int. Conf. on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.

[26] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2007.