



Convolutional Weighted Multichannel Wiener Filter Front-end for Distant Automatic Speech Recognition in Reverberant Multispeaker Scenarios

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

The performance of automatic speech recognition (ASR) systems strongly deteriorates when the desired speech signal is contaminated with room reverberation and when the speech of interfering speakers overlaps. To achieve acceptable word error rates (WER) by distant ASR in multispeaker reverberant scenarios, source separation and dereverberation can be performed as front-end processing. An existing optimum filter suitable for this task is the recently proposed weighted power minimization distortionless response convolutional beamformer (WPD). In this paper, we introduce a novel speech enhancement front-end for improving the accuracy of back-end ASR in scenarios with multiple reverberant overlapping speakers. The convolutional weighted multichannel Wiener filter (CW-MWF) is optimum for the joint separation and dereverberation task, and it is derived from the convolutional weighted minimum mean square error (CW-MMSE) optimization criterion, presented recently by the current authors. The WER results of performed experiments indicate superior performance of the CW-MWF in real and simulated rooms, irrespective of the method used for filter parameter estimation and the DNN model used for back-end ASR.

Index Terms: speech enhancement, source separation, dereverberation, automatic speech recognition, optimum filters

1. Introduction

Modern ASR systems achieve tremendous speech recognition accuracy from nearly anechoic recordings of a single speaker or when a close-talk microphone is used in low reverberant conditions [1, 2]. However, in scenarios with multiple speakers with strong speech overlap and when speech is recorded from distance in reverberant rooms, the recognition rates drop drastically, effectively disabling to recognize speech at an acceptable level. A popular approach to tackle this problem is to apply speech enhancement front-end processing to separate and dereverberate the speech of the desired speaker before it enters the back-end ASR module [3]. As multichannel source separation front-end for ASR, the most often used filter is the minimum variance distortionless beamformer (MVDR) [4, 5], while a weighted prediction error (WPE) method [6, 7], and its variants [8, 9], have been successfully applied for dereverberation. Improvements in ASR performance with such front-ends have been widely reported, e.g. at the CHiME-3/4/5 Challenges [10, 11, 12], and the REVERB Challenge [13]. These techniques have also been applied in commercial devices [14, 15].

Despite extensive research, it remains a challenge to reduce the impact of both reverberation and interfering speech simultaneously in an optimum way, i.e. such that strong simultaneous suppression of interfering and reverberant signal components is not compromised by introducing distortions that hinder automatic speech recognition. In [16], a single weighted power

minimization distortionless response convolutional beamformer (WPD) has been proposed, which optimally unifies the WPE and MVDR filtration, outperforming the cascade use of WPE and MVDR as well as many existing approaches in distant ASR [17, 18]. In our very recent work [19], we have proposed a convolutional weighted filter that is optimum for joint separation and dereverberation in a minimum mean square error (MMSE) sense, which has been shown to outperform the WPD and many existing front-ends in terms of objective and perceptual speech enhancement metrics.

In this paper, we derive a convolutional weighted multichannel Wiener filter (CW-MWF) by solving the CW-MMSE optimization problem proposed in [19] and show great improvement in ASR performance achieved by the proposed filter when applied as front-end processing for ASR in reverberant multi-speaker scenarios. The CW-MWF is more convenient and straightforward to compute since, in contrast to CW-MMSE filter [19], it does not require the estimation of *a posteriori* signal-to-interference-and-reverberation-ratio (SIRR). We evaluate the proposed front-end filtering against the popular WPD, recent CW-MMSE, as well as WPE, MVDR, MWF, and cascades thereof in terms of WERs of three state-of-the-art ASR systems [20]. Processed microphone recordings contain speech of two speakers with very high speech overlap, taken in a real and simulated room with reverberation times ranging from 0.3 to 1.5s. A tremendous gain in WER results over other investigated front-ends is reported, irrespective of the front-end parameter estimation method, the reverberation level, and the underlying DNN model of the ASR system.

2. Front-end processing for jointly optimum separation and dereverberation

2.1. Problem formulation

We assume that an I -channel microphone mixture that captures the reverberant signals of J speakers can be modeled within a single time-frequency bin of the short-time Fourier transform (STFT) domain (frequency indices are omitted for brevity) as

$$\mathbf{x}_n = \sum_{j=1}^J \left(\mathbf{v}^{(j)} S_n^{(j)} + \sum_{\tau=b}^{L_a+b-1} \mathbf{a}_\tau^{(j)} S_{n-\tau}^{(j)} \right), \quad (1)$$

where $\mathbf{v}^{(j)} = [1, v_2^{(j)}/v_1^{(j)}, \dots, v_I^{(j)}/v_1^{(j)}] \in \mathbb{C}^I$ is a vector with the relative acoustic transfer functions between the j -th source and the microphones, n is a time frame index, $S_n^{(j)} \in \mathbb{C}$ denotes the desired clean speech signal of the j -th source in the STFT domain as captured by the reference (e.g. first) microphone, and $\mathbf{a}_\tau^{(j)} \in \mathbb{C}^I$ is a convolutional transfer function for $\tau = b, b+1, \dots, L_a+b-1$, 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.

The aim is to restore clean speech signal $S_n^{(j)}$ from the multichannel microphone mixture \mathbf{x}_n via spatial filtration

$$\widehat{S}_n^{(j)} = (\bar{\mathbf{w}}_n^{(j)})^H \bar{\mathbf{x}}_n, \quad (2)$$

where $\bar{\mathbf{x}}_n = [\mathbf{x}_n^T, \mathbf{x}_{n-b}^T, \mathbf{x}_{n-b-1}^T, \dots, \mathbf{x}_{n-L-b+2}^T]^T$ is the extended vector of the microphone mixture and $\bar{\mathbf{w}}_n^{(j)} = [(\mathbf{w}_{0,n}^{(j)})^T, (\mathbf{w}_{n-b}^{(j)})^T, (\mathbf{w}_{n-b-1}^{(j)})^T, \dots, (\mathbf{w}_{n-L-b+2}^{(j)})^T]^T$ is the unified convolutional weighted filter which performs jointly optimum separation and dereverberation, and can be factorized as

$$\mathbf{w}_n^{(j)} = -\mathbf{W}_n \mathbf{w}_{0,n}^{(j)} \quad (3)$$

with $\mathbf{w}_{0,n}^{(j)} = [w_{1,n}^{(j)}, w_{2,n}^{(j)}, \dots, w_{I,n}^{(j)}]^T \in \mathbb{C}^I$ denoting a standard vector of spatial filter coefficients and $\mathbf{W}_n \in \mathbb{C}^{I \times I}$ is the prediction matrix of the WPE algorithm for the filter length L .

2.2. State-of-the-art spatial filtering front-ends

In the following, we discuss four existing optimum spatial filters suitable for dereverberation (WPE algorithm [7]), separation (MVDR [4] and MWF [21] filters), and joint separation and dereverberation (WPD filter [16]). Under the assumption that the desired speech signal (early component of the desired speaker) is modeled by a zero-mean complex Gaussian distribution with time-varying source variance $\phi_{s,n}^{(j)}$, the matrix of optimum WPE prediction coefficients $\bar{\mathbf{W}} = [\mathbf{W}_b, \mathbf{W}_{b+1}, \dots, \mathbf{W}_{L+b}]$ can be computed by minimizing the following cost function [7]

$$\bar{\mathbf{W}} = \arg \min_{\bar{\mathbf{W}}} \sum_n \frac{|\mathbf{x}_n - \sum_{\tau=b}^{L+b} \mathbf{W}_\tau \mathbf{x}_{n-\tau}|^2}{\phi_{s,n}^{(j)}}. \quad (4)$$

A well-known conventional spatial filter is the MVDR beamformer [4], which is derived by minimizing the variance at filter output under the distortionless constraint, i.e. by solving

$$\mathbf{w}_n^{(j)} = \arg \min_{\mathbf{w}_n^{(j)}} \sum_n |(\mathbf{w}_n^{(j)})^H \mathbf{x}_n|^2 \text{ s.t. } (\mathbf{w}_n^{(j)})^H \mathbf{v}^{(j)} = 1. \quad (5)$$

Another well-known filter is the multichannel Wiener filter (MWF) [21] which minimizes the mean square error (MMSE) between the processed output signal and the desired source signal, which can be formulated as the following cost function

$$\mathbf{w}_n^{(j)} = \arg \min_{\mathbf{w}_n^{(j)}} |S_n^{(j)} - (\mathbf{w}_n^{(j)})^H \mathbf{x}_n|^2. \quad (6)$$

Finally, the WPD convolutional beamformer [16] is derived by minimizing the weighted power of the processed signal under the directionless constraint, i.e. by solving

$$\bar{\mathbf{w}}_n^{(j)} = \arg \min_{\bar{\mathbf{w}}_n^{(j)}} \sum_n \frac{|(\bar{\mathbf{w}}_n^{(j)})^H \bar{\mathbf{x}}_n|^2}{\phi_{s,n}^{(j)}} \text{ s.t. } (\bar{\mathbf{w}}_n^{(j)})^H \bar{\mathbf{v}}^{(j)} = 1, \quad (7)$$

where $\bar{\mathbf{v}}^{(j)} = [(\mathbf{v}^{(j)})^T, \mathbf{0}_{I(L-1)}^T]^T \in \mathbb{C}^{IL}$ with $\mathbf{0}_{I(L-1)}$ defined as a vector of zeros of length $I(L-1)$.

2.3. Convolutional weighted multichannel Wiener filter

As robust front-end processing for ASR in reverberant multi-speaker scenarios, we propose to apply a convolutional filter that enables jointly optimum source separation and dereverberation in the MMSE sense. Assuming that each speech signal is modeled by a zero-mean complex Gaussian distribution with

time-varying source variance $\phi_{s,n}^{(j)}$, we formulate an optimization criterion for the convolutional weighted minimum mean square error filter, which is defined as [19]

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

Cost function (8) is similar to (6), however, it is adjusted for the joint task by introducing normalization and the extended vectors $\bar{\mathbf{w}}_n^{(j)}$ and $\bar{\mathbf{x}}_n$. Weighting with the time-varying source variance $\phi_{s,n}^{(j)}$ should allow for a stronger attenuation of undesired components, which has been shown in [18] to be beneficial for ASR using the WPD front-end. The closed-form solution is derived by setting the derivative of (8) to zero. Assuming no correlation between the sources and between the desired early component and the undesired late reverberation [6], we can apply the Sherman-Morrison-Woodbury formula [22] to obtain

$$\bar{\mathbf{w}}_n^{(j)} = \frac{(R_u^{(j)})^{-1} \bar{\mathbf{v}}^{(j)}}{1 + (\bar{\mathbf{v}}^{(j)})^H (R_u^{(j)})^{-1} \bar{\mathbf{v}}^{(j)}}, \quad (9)$$

with the weighted covariance matrix of undesired components

$$R_u^{(j)} = \mathbf{E} \left\{ \frac{\bar{\mathbf{x}}_{u,n}^{(j)} (\bar{\mathbf{x}}_{u,n}^{(j)})^H}{\phi_{s,n}^{(j)}} \right\} \quad (10)$$

and $\bar{\mathbf{x}}_{u,n}^{(j)} \in \mathbb{C}^{IL}$ containing all undesired components of other speakers and the reverberant component of the desired speaker. Expressing the variance of the undesired component as [22]

$$\phi_{u,n}^{(j)} = \left[(\bar{\mathbf{v}}^{(j)})^H \left(\mathbf{E} \left\{ \bar{\mathbf{x}}_{u,n}^{(j)} (\bar{\mathbf{x}}_{u,n}^{(j)})^H \right\} \right)^{-1} \bar{\mathbf{v}}^{(j)} \right]^{-1}, \quad (11)$$

we can transform (9) into the convolutional beamformer \mathbf{h}_{CWBF} followed by a post-filter H_{CWPF} , which yields

$$\bar{\mathbf{w}}_n^{(j)} = \underbrace{\frac{\phi_{s,n}^{(j)}}{\phi_{u,n}^{(j)} + \phi_{s,n}^{(j)}}}_{H_{\text{CWPF}}} \underbrace{\frac{(R_u^{(j)})^{-1} \bar{\mathbf{v}}^{(j)}}{(\bar{\mathbf{v}}^{(j)})^H (R_u^{(j)})^{-1} \bar{\mathbf{v}}^{(j)}}}_{\mathbf{h}_{\text{CWBF}}}. \quad (12)$$

We will refer to (12) as the convolutional weighted multichannel Wiener filter (CW-MWF). The proposed filter constitutes a simplification of the CW-MMSE filter from [19] since, in contrast to the CW-MMSE, the presented post-filter does not require the estimation of a *posteriori* SIRR. Consequently, the overall computational load of (12) is comparable to that of the convolutional beamformer, such as WPD. As shown in Sec. 4, this simplification yields sufficiently accurate enhancement for the ASR task.

2.4. Estimation of filter parameters

The presented CW-MWF requires calculation of the undesired covariance matrix $R_u^{(j)}$, the steering vector $\bar{\mathbf{v}}^{(j)}$, and the post-filter H_{CWPF} . In principle, knowing $\phi_{s,n}^{(j)}$ one could directly infer the variance of undesired components as $\phi_{u,n}^{(j)} = \phi_{x,n}^{(j)} - \phi_{s,n}^{(j)}$. However, this could lead to speech degradation for inaccurate source variance estimates. To avoid such problems, we propose to estimate the variance of undesired components as

$$\phi_{u,n}^{(j)} = \sum_{j'}^J \phi_{s,n}^{(j' \neq j)}. \quad (13)$$

Having inferred $\phi_{s,n}^{(j)}$ and $\phi_{u,n}^{(j)}$, one can conveniently estimate H_{CWPF} , $\bar{\mathbf{v}}^{(j)}$, and $R_u^{(j)}$ (the reader is referred to [19] for details on computing the latter two quantities).

3. Datasets and ASR models

The performance of the proposed front-end processing has been benchmarked against state-of-the-art optimum filters using three recent ASR models from the SpeechBrain toolkit [20]. The first ASR model (denoted as M1) is the *asr-transformer-transformerlm-librispeech*¹, whose structure consists of the transformer-based encoder, followed by the transformer-based decoder with joint CTC and attention beamsearch, coupled with a transformer-based language model. The other two ASR models are non-transformer-based systems, and they comprise the convolutional recurrent encoders (CRDNN [23]) and CTC/Attention decoders. In particular, the second model *asr-crdnn-transformerlm*² (M2) is the pre-trained CRDNN with the transformer-based language model, while the third model *asr-crdnn-rnnlm*³ (M3) uses the RNN-based language model instead. All ASR models were trained using the full 960 h Librispeech train dataset with the default settings of these systems. Specifically, the decoder outputs were processed by in-build language models coupled with beamsearch with default weights. Evaluations were performed using the word error rate (WER) metric. In the first three experiments, the ASR system that achieved the lowest WER on the unprocessed Librispeech *test-clean* part, i.e. model M1, was used.

For testing, two datasets were created with the recordings of two simultaneously active speakers made using a 2-microphone array located in a reverberant room. The reverberant microphone mixtures were synthesized by convolving clean speech signals with the respective room impulse responses (RIRs) and adding them together at the microphones. As clean signals, speech utterances from the Librispeech *test-clean* part [24] were used, such that each recording contained the signals of two different speakers of a similar length. In the first dataset, the RIRs were simulated using the image-source method [25] for the source-array distance of around 2 m and the array with an inter-microphone spacing of 0.05 m located near the center of the room of size $10 \times 10 \times 4$ m with the reverberation time (RT60) ranging from 300 to 1500 ms. The second dataset is based on real RIRs from the MIRD database [26] measured for the randomly selected 13 source positions located on a semi-circle with a 2 m source-array distance and the reverberation time of 610 ms. For each RT60, the presented WER results were obtained based on 2620 transcriptions of both speakers from 1310 reverberant mixtures. Signals were sampled at 16kHz and processed with 512 point STFT with 50% overlap. The filter length L was set to 10 for RIRs with $RT60 \approx 600$ ms and then increased (decreased) by 2 for every increase (decrease) of 150ms in RT60.

4. Experiments and result discussion

4.1. Informed scenario

In the first experiment, the performance of the proposed CW-MWF front-end was compared with other state-of-the-art front-ends, namely the sole MVDR [4], MWF [21], WPD [17], and a cascade of WPE [7] and MVDR or MWF (denoted respectively as WPE&MVDR and WPE&MWF) on the dataset created using simulated RIRs. The reference upper and lower bounds on WER for the applied ASR system (model M1) were found in

¹<https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech>

²<https://huggingface.co/speechbrain/asr-crdnn-transformerlm-librispeech>

³<https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech>

an experiment run on the unprocessed dataset with a mixture of overlapping reverberant source signals (Overlapped srcs) and on the dataset with single sources (Single src) from each mixture. Furthermore, an *informed* approach is used for inferring filter parameters of all compared front-ends, such that $\mathbf{v}^{(j)}$ corresponded to the relative acoustic transfer functions of early RIR parts, while $\phi_{s,n}^{(j)}$ was computed from the non-reverberant signal of the j -th speaker.

The upper two rows of Table 1 present the WER results without any front-end enhancement. In the case of a single speaker (Single src), WER increases alongside the rising reverberation time values from 2.57% for non-reverberant speech (direct) up to 47.4% for $RT60 = 1.5$ s, which shows a detrimental impact of reverberation on ASR accuracy. On the other hand, for overlapped speech (Overlapped srcs), WER always reaches around 100%, regardless of the level of reverberation, which indicates the difficulty to recognize speech of highly overlapping speakers and motivates the application of source separation and dereverberation front-end. As can be observed from the results obtained by processing the signals of overlapped sources using six studied front-ends, all methods improve WER results, and more improvement is achieved for high RT60. In general, front-ends based on filters designed using the MMSE optimization criterion, such as MWF, WPE&MWF, and the proposed CW-MWF, achieve even a few times better WER results than methods based on extensions of MVDR, including MVDR alone, WPE&MVDR, and WPD. Such an improvement in WER is a result of a significantly stronger attenuation of undesired components when the MMSE criterion is used. Adding WPE processing before MVDR or MWF filtration (denoted as WPE&MVDR and WPE&MWF) increases the robustness of such a cascade towards room reverberation. However, notably, better WER results can be achieved when separation and dereverberation are performed simultaneously in a jointly optimum manner. Jointly optimum processing offered by the existing WPD and the proposed CW-MWF front-ends consistently outperforms their non-optimum cascade counterparts (WPE&MVDR and WPE&MWF). All in all, by far the best performance is achieved by the proposed method, which yields the lowest WER amongst all studied front-ends, across all reverberation levels. For the highest considered reverberation of 1.5 s, the gain offered by CW-MWF reaches over 27% absolute WER improvement over WPD and almost 4% improvement over the non-optimum cascade of WPE and MWF.

4.2. Blind scenario

In experiments to follow, all filter parameters were estimated blindly from the microphone mixtures. In particular, we selected to estimate source variances using the sub-source-based Expectation-Maximization algorithm with Multiplicative Updates and Localization Prior (SSEM-MU-LP) [27], which was reported to provide reliable estimates in multispeaker reverberant conditions. In essence, any other similarly robust estimation method could alternatively be used, such as [28, 29, 30, 31, 32].

The results of the second experiment with blindly estimated parameters in simulated rooms (the first dataset) for the selected four most interesting front-ends are depicted in Fig. 1. As can be observed, the WER results of the compared WPE&MVDR, WPE&MWF, WPD, and the proposed CW-MWF are worse than those presented in Table 1 in terms of absolute values, which is due to imperfect parameter estimation. However, the general tendency and the relative differences between the compared front-ends are similar, and the CW-MWF again consis-

Table 1: WER [in %] results of the first experiment for an informed scenario with simulated RIRs. Note that WER above 100% is obtained when the total number of injections, substitutions, and deletions is larger than a number of words in an original transcript.

RT60 [ms]	direct	300	450	600	750	900	1050	1200	1350	1500
Single src	2.57	2.96	3.80	5.35	8.48	14.46	20.21	28.23	36.34	47.4
Overlapped srcs	100.36	100.75	101.82	102.60	103.77	103.72	103.74	102.8	101.82	100.80
MVDR	3.67	42.48	56.92	66.98	75.21	81.88	85.66	88.91	91.28	93.21
MWF	2.68	3.55	4.94	7.15	10.72	15.30	19.09	23.55	28.41	34.08
WPE&MVDR	3.36	11.22	18.17	25.5	33.1	42.71	47.65	53.69	59.52	65.65
WPE&MWF	2.73	3.44	4.33	5.72	7.46	9.81	11.55	13.74	15.82	18.36
WPD	2.70	4.74	8.05	11.94	17.56	24.2	27.87	32.60	37.20	42.48
Proposed	2.65	3.29	4.06	5.16	6.69	8.46	9.65	11.32	12.82	14.61

tently provides by far the best performance. The constant gain in WER presented on a logarithmic scale across various reverberation levels indicates an increasing benefit of using MMSE-based filtration compared with convolutional beamformers.

The third and fourth experiments were performed with real RIRs from the MIRD database (the second dataset). In the third experiment, we compare the performance of the WPD and CW-MWF front-ends for three different parameter estimation methods. The initial steering vectors $\mathbf{v}^{(j)}$ were obtained using the steered response power phase transform (SRP-PHAT) technique [33], while the source variances were estimated from the microphone mixtures using either (i) the MVDR beamformer, (ii) the WPE algorithm followed MVDR beamforming, or (iii) the aforementioned SSEM-MU-LP algorithm [27].

As can be clearly seen from the results of the third experiment depicted in Fig. 2, the overall performance highly depends on the accuracy achieved by the estimation method. Good performance was achieved by the SSEM-MU-LP algorithm, while for the relatively straightforward yet less robust estimation techniques, namely MVDR and WPE&MVDR, the overall system struggles to effectively recognize speech with WER exceeding 47% even at a moderate reverberation level of 610 ms. Nonetheless, the proposed front-end always leads to lower WERs than WPD, regardless of the parameter estimation method.

The fourth experiment compares the performance of the proposed CW-MWF with the popular WPD as front-end processing to all three analyzed ASR systems, denoted as ASR models M1, M2, and M3. In addition, we benchmarked the proposed filter against its predecessor – the recently proposed CW-MMSE filter [19]. The results of the experiment with real RIRs and parameters estimated using SSEM-MU-LP algorithm are presented in Table 2. The proposed front-end consistently outperforms WPD irrespective of the ASR system used for speech recognition. Furthermore, it also performs slightly better than CW-MMSE, and we hypothesize that this behavior may be caused by a flawed estimation of *a posteriori* SIRR. This confirms that the proposed filter is not only simpler implementation-wise but also more robust in practical deployment.

5. Conclusions

This paper shows the clear benefits of using the proposed convolutional weighted multichannel Wiener filter for distant automatic speech recognition in multispeaker reverberant scenarios. The derived filter, which is jointly optimum for source separation and dereverberation, significantly outperforms state-of-the-art front-ends in real and simulated rooms at various reverberation levels, for all tested ASR systems used as backend.

6. Acknowledgements

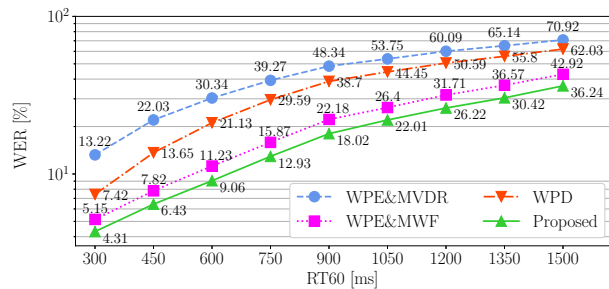


Figure 1: WER [in %] results of the second experiment for a blind scenario (estimated parameters) with simulated RIRs.

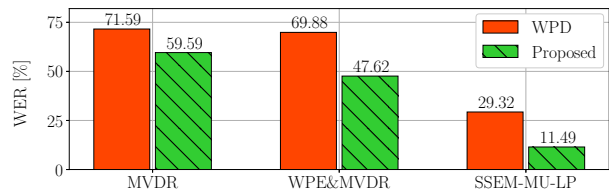


Figure 2: WER [in %] results of the third experiment for a blind scenario with real RIRs. The proposed and WPD front-ends are compared for three filter parameter estimation methods.

Table 2: WER [in %] results for the fourth experiment for a blind scenario with real RIRs. The proposed, WPD, and a recent CW-MMSE front-ends are compared for three ASR systems based on ASR models M1, M2, and M3 described in Sec. 3.

Filter	M1	M2	M3
WPD	29.32	42.95	25.41
CW-MMSE [19]	12.11	23.85	15.92
Proposed	11.49	21.96	13.95

7. References

- [1] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, “A comparative study on transformer vs RNN in speech applications,” in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2019, pp. 449–456.
- [2] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, “State-of-the-art speech recognition with sequence-to-sequence models,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2018, pp. 4774–4778.
- [3] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, “Far-field automatic speech recognition,” *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.
- [4] H. Cox, “Resolving power and sensitivity to mismatch of optimum array processors,” *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, 1973.
- [5] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.
- [6] 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.
- [7] 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.
- [8] 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.
- [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] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.
- [11] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, “The 4th CHiME speech separation and recognition challenge,” *URL: http://spandh.dcs.shef.ac.uk/chime_challenge/* (last accessed on 1 August, 2018), 2016.
- [12] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth ‘CHiME’ speech separation and recognition challenge: dataset, task and baselines,” *arXiv preprint arXiv:1803.10609*, 2018.
- [13] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.
- [14] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafraan, H. Sak, G. Pundak, K. K. Chin *et al.*, “Acoustic modeling for Google Home,” in *Proc. Interspeech*, 2017, pp. 399–403.
- [15] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, “Speech processing for digital home assistants: Combining signal processing with deep-learning techniques,” *IEEE Signal process. magazine*, vol. 36, no. 6, pp. 111–124, 2019.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] M. Fraś, M. Witkowski, and K. Kowalczyk, “Convolutional weighted minimum mean square error filter for joint source separation and dereverberation,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2022, pp. 286–290.
- [20] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, “SpeechBrain: A general-purpose speech toolkit,” *arXiv preprint arXiv:2106.04624*, 2021.
- [21] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.
- [22] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,” in *Microphone arrays*. Springer, 2001, pp. 39–60.
- [23] G. Keren and B. Schuller, “Convolutional RNN: an enhanced model for extracting features from sequential data,” in *Proc. IEEE 2016 Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3412–3419.
- [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)*. IEEE, 2015, pp. 5206–5210.
- [25] 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.
- [26] 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.
- [27] M. Fraś, M. Witkowski, and K. Kowalczyk, “Combating reverberation in NTF-based speech separation using a sub-source weighted multichannel Wiener filter and linear prediction,” *Proc. Interspeech 2021*, pp. 3895–3899, 2021.
- [28] 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.
- [29] 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.
- [30] T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Delcroix, and S. Araki, “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.
- [31] 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.
- [32] 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.
- [33] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, RI, 2000.