COMBINED WEIGHTED PREDICTION ERROR AND MINIMUM VARIANCE DISTORTIONLESS RESPONSE FOR DEREVERBERATION

Alejandro Cohen *Georg Stemmer* *Seppo Ingalsuo* *Shmulik Markovich-Golan*

Intel Corporation

Email:{alejandro.cohen,georg.stemmer,seppo.ingalsuo,shmulik.markovich-golan}@intel.com

ABSTRACT

Considering the dereverberation problem using multichannel processing, two main paradigms exist. The first paradigm utilizes the long-term correlation of the reverberant component for reducing it, e.g. Weighted Prediction Error (WPE) [1]. The second paradigm, treats the reverberation as a diffuse noise field, statically independent of the direct speech component, and aims to reduce it using a superdirective beamformer, e.g. [2].

Here we propose to combine the two paradigms in a two-stages algorithm. The first stage comprises of the WPE method, and the second stage comprises of a Minimum Variance Distortionless Response (MVDR) beamformer for treating the residual reverberant component. We conjecture that the coherence of the reverberant component at the output of the WPE is similar to the coherence of the reverberant component at the microphones which should theoretically correspond to a diffuse noise field. By estimating the coherence from the reverberant components, linearly predicted by the WPE, non-ideal factors such as microphone positions errors, non-equalized frequency responses and acoustic shading are accounted for. The advantageous performance of the proposed method is exemplified in an experiment study using simulations.

1. INTRODUCTION

In recent years performance of Automatic Speech Recognition (ASR) methods have improved remarkably thanks to algorithmic and technological advances [3]. Yet, distant speech recognition [4] remains a challenging problem, as conventional ASR methods degrade dramatically when the microphones are far away from the speech source (due to reduced Signal-to-Noise Ratio (SNR) and direct Speech-to-Reverberant Ratio (SRR)). Utilizing spatial-diversity is a key component in distant speech recognition systems [5]. Dereverberation methods for improving speech intelligibility and Word Error Rate (WER) are surveyed in [6]. In a recent challenge [7], the performances of state-of-the-art methods for speech enhancement and ASR are compared in various scenarios. Two different paradigms for the reverberant component are to either treat it is a long-term correlative signal, i.e. reflections are considered as delayed and attenuated replica of the speech source, or treat it as noise comprising a large number of statistically independent sources.

The Weighted Prediction Error (WPE) method which adopts the first paradigm is proposed in [1, 8] and performs very well in the above mentioned challenge. The signals are processed in the Short-Time Fourier Transform (STFT) domain. A criterion combining correlation to previous frames and a Linear Prediction Coefficients (LPC) model for the *dry* speech is optimized in an iterative procedure. A Minimum Variance Distortionless Response (MVDR) beamformer based on [9] is incorporated at the output of a WPE filter in [7] for reducing noise.

In [2], the authors adopt the second paradigm and model the reverberation as a diffuse noise field. A superdirective beamformer [10], implemented in a Generalized Sidelobe Canceler (GSC) structure [11], is applied in the STFT domain. The steering vector towards the desired speaker is defined using the early component of the Impulse Responses (IRs) between the source and the microphone array. The latter responses are estimated using single-channel Wiener filters applied to each of the microphones, and using estimates of the reverberant components [12]. This model is also used in [13], where estimates of the direct speech to diffuse noise ratio, based on spatial coherence between omnidirectional microphones, is used for speech detection.

In this contribution we propose to adopt the two paradigms in a two stage approach. At the first stage we apply the WPE algorithm. We use a Multiple-Input Multiple-Output (MIMO) version of the latter, yielding a dereverberated version for each of the microphones. In practice, the enhanced signals at this stage still comprise of a residual reverberant component, which we aim to reduce at the second stage MVDR beamformer. We conjecture that the spatial properties of the residual reverberation are similar to the spatial properties of the reverberation at the microphones, which ideally follow the diffuse field model. Microphone position errors, non-uniform reverberation field, shading of objects in the room and of the device itself and diverse frequency responses of the microphones affect the accuracy of the theoretical diffuse field. Therefore, we propose to utilize coherence of the reverberant components at microphones, as linearly predicted at the first stage WPE, for modelling the coherence of the residual reverberant component at the output of the first stage and construct the second stage MVDR. Thereby, any miss-modelling is avoided since the coherence is estimated from the signals. Furthermore, since the direct SRR at the output of the WPE is improved compared to the input, we estimate the steering vector of the early speech component after the first stage using the Covariance Whitening (CW) method [14–16].

The paper is structured as follows. In Section 2, the problem is formally described and in Section 3, we provide a brief overview on related methods. Section 4 is dedicated to presenting the proposed method and in Section 5, an experimental study for verifying our assumptions and exemplifying the performance of the proposed

method is described. Finally, we conclude the paper in Section 6.

2. PROBLEM FORMULATION

Let $\underline{s}(t)$ denote a speech signal uttered by the desired speaker, where $\underline{\bullet}$ denotes terms in the time domain and t denotes the discrete time-index with a sampling rate of f_s. The speech signal propagates in a reverberant enclosure and impinges on an array comprising of M microphones. The M-dimensional vector of received microphone signals is:

$$\underline{\mathbf{x}}(t) = \sum_{k=0}^{\infty} \underline{\mathbf{h}}_k \cdot \underline{s}(t-k) + \underline{\mathbf{v}}(t) \tag{1}$$

where $\underline{\mathbf{h}}_k = \begin{bmatrix} h_{k,1} & \dots & h_{k,M} \end{bmatrix}^T$ denotes a vector comprising of the k-th tap coefficients of the multichannel acoustic IRs and $\underline{\mathbf{v}}(t)$ denotes an additive sensors noise with variance σ_v^2, statistically independent of the speech source.

Due to the long IR and the spectral structure of the speech signal, a common practice is to process the microphone signals in the STFT domain. Denote by F the length of analysis and synthesis windows and by D the overlap between consecutive frames. For practical reasons, we assume that F is shorter compared to the length of the IR and adopt the signal model given in [17]. Therefore the problem is formulated in the STFT domain as a convolution along the time-frame axis per frequency bin, while neglecting the cross-band filters. A discussion on the accuracy of this approximation, as well as the more general model is given in [18]. Hence, the received microphone signals are formulated as:

$$\mathbf{x}(n,f) = \mathbf{d}(n,f) + \mathbf{r}(n,f) + \mathbf{v}(n,f) \tag{2}$$

with n and f denote the time-frame and frequency-bin indices. The notations $\mathbf{d}(n,f)$ and $\mathbf{r}(n,f)$ correspond to the *early* and *reverberant* components of the received speech

$$\mathbf{d}(n,f) = \mathbf{h}_0(f) \cdot s(n,f) \tag{3a}$$

$$\mathbf{r}(n,f) = \sum_{\tau=1}^{\infty} \mathbf{h}_\tau(f) \cdot s(n-\tau,f) \tag{3b}$$

where $\mathbf{h}_\tau(f)$ for $\tau = 0, 1, \dots, \infty$ is the Convolutive Transfer Function (CTF), and $\mathbf{v}(n,f)$ corresponds to the sensor noise in the STFT domain. The zeroth tap of the CTF, i.e. $\mathbf{h}_0(f)$, is dominated by the direct arrival and few low-order reflection components of the IR. The rest of the CTF, i.e. $\mathbf{h}_\tau(f);\ \tau = 1, 2, \dots, \infty$, consists of all other high-order reflections of the IRs transformed to the STFT domain. We assume that the early and reverberant components, i.e. $\mathbf{d}(n,f)$ and $\mathbf{r}(n,f)$, are statistically independent.

In the current contribution we consider the problem of dereverberating the received speech, and retrieving the early speech component $\mathbf{d}(n,f)$. Apart for some low-level sensors noise, we assume a quiet environment, and leave the problem of combined dereverberation and noise-reduction for future research.

3. BACKGROUND ON RELATED METHODS

The method that we propose in this paper is based on a combination of two algorithms which stem from different approaches to speech dereverberation, namely, the WPE [1] and the MVDR [2]. In this section we briefly review these algorithms. The WPE, see Section 3.1, treats the reverberation process as a convolutive filter in the STFT domain, and aims at de-correlating the current frame from past frames via linear filtering. The MVDR, see Section 3.2 treats the reverberant component as an interference, and tries to attenuate it spatially, by using a superdirective beamformer.

3.1. Weighted Prediction Error

The WPE algorithm was first introduced in a landmark work [19] for long-term linear prediction in the STFT domain. For a detailed description please refer to [8].

Consider the problem of dereverberating the speech component at the first microphone by using all M microphones. The basic idea is to reduce reverberation by de-correlating past time-frames from the current time-frame and utilizing a time-varying LPC model for the *early* speech component at the first microphone. Denote by $\theta(n) \triangleq \left\{\sigma^2(n), \mathbf{a}(n)\right\}$ the LPC parameters corresponding to the n-th time-frame. The early speech component is modeled in the STFT domain as a complex Gaussian random variable with zero mean, and variance of $\sigma_s^2(n,f) \triangleq \frac{\sigma^2(n)}{|\mathrm{DFT}\{\mathbf{a}(n)\}|^2}$ where $\mathrm{DFT}\{\bullet\}$ denotes the Discrete Fourier Transform (DFT). The enhanced signal, estimating the early speech component at the first microphone, is obtained through the following linear filtering process:

$$y_1(n,f) = x_1(n,f) - \hat{r}_1(n,f) \tag{4}$$

where $\hat{r}_1(n,f) = \sum_{\tau=n_s}^{n_e} \mathbf{p}(\tau,f)^H \mathbf{x}(n-\tau,f)$ is the estimated reverberant component at the first microphone, $\{\mathbf{p}(\tau,f)\}_{\tau=n_s}^{n_e}$ are the linear prediction filters which process past frames in the range of $[n-n_e, n-n_s]$ for enhancing the n-th frame and $(\bullet)^H$ denotes the Hermitian operator. The first microphone signal is modeled in the STFT domain as a complex Gaussian random variable given past microphone frames, the speech model parameters and the linear prediction filters. Denote the set of linear prediction filters by:

$$\mathcal{P} \triangleq [\mathbf{p}(n_s,0), \dots, \mathbf{p}(n_e,0), \dots, \mathbf{p}(n_s,F-1), \dots, \mathbf{p}(n_e,F-1)] \tag{5}$$

and the set of LPC parameters by $\Theta \triangleq \{\theta(0), \dots, \theta(N-1)\}$. The log-likelihood of the observed first microphone given $\mathcal{P}$ and Θ is shown to equal:

$$\mathcal{L}(\Theta,\mathcal{P}) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \log \sigma_s^2(n,f) + \frac{|x_1(n,f) - \sum_{\tau=n_s}^{n_e} \mathbf{p}(\tau,f)^H \mathbf{x}(n-\tau,f)|^2}{\sigma_s^2(n,f)}. \tag{6}$$

An iterative algorithm which alternates between optimizing Θ and $\mathcal{P}$ for maximizing the log-likelihood is proposed. The step for optimizing Θ consists of a standard Yule-Walker solution (linear system solver), and the step for optimizing $\mathcal{P}$ can be interpreted as an extension to the Yule-Walker solution. Only a few iterations are required to provide adequate dereverberation performance. The basic WPE can be extended to the MIMO case and can also be implemented using sub-bands (please refer to [8] for further details). In [7], it is proposed to incorporate an MVDR beamformer at the output of the WPE. However, this beamformer (based on [9]) aims at noise reduction, and not at dereverberation as the proposed method.

3.2. Minimum Variance Distortionless Response

An alternative approach for dereverberation, which is based on the MVDR criterion, is proposed in [2]. The authors treat the reverberant component as a diffuse noise field (see [20]) and design a

beamformer which minimizes the interference while maintaining a distortionless response towards the early speech component at a reference microphone. Similarly to [21], define the Relative Transfer Function (RTF) of the early speech component by:

$$\mathbf{g}_0(f) \triangleq \frac{\mathbf{h}_0(f)}{h_{0,1}(f)} \tag{7}$$

where $h_{0,1}(f)$ denotes the Transfer Function (TF) between the speech source and early speech component at the first microphone.

The dereverberation MVDR beamformer is obtained by:

$$\mathbf{w}(f) = \frac{\mathbf{\Phi}^{-1}(n,f)\mathbf{g}_0(f)}{\mathbf{g}_0^H(f)\mathbf{\Phi}^{-1}(n,f)\mathbf{g}_0(f)} \tag{8}$$

where $\mathbf{\Phi}(n,f) = \sigma_r^2(n,f)\mathbf{\Gamma}(f) + \mathbf{\Phi}_{vv}(f)$ is the covariance matrix of the total interference, comprising both reverberant speech and noise. The term $\sigma_r^2(n,f)$ denotes the spectrum of the reverberant component, the matrix $\mathbf{\Phi}_{vv}(f)$ denotes the noise covariance matrix and $\mathbf{\Gamma}$ denotes the spatial coherence matrix of a diffuse noise field [22, 23]. The theoretical coherence between the received diffuse noise components at the m-th and m'-th microphones is:

$$\gamma_{mm'}(f) \triangleq \text{sinc}\left(\frac{2\pi f \delta_{mm'}}{\nu}\right) \tag{9}$$

where $\text{sinc}(\alpha) \triangleq \frac{\sin\alpha}{\alpha}$, $\delta_{mm'}$ denotes the distance between microphones m and m' and ν denotes the sound velocity. Practically, the measured coherence might vary due to non-ideal conditions.

Given a rough estimate of the exponential decay of the reverberant component (related to the Reverberation Time (RT) of the room), the authors suggest to estimate the spectrum of the reverberant component using spectral subtraction similarly to the single-channel dereverberation method in [24]. The output signals of the the latter single-channel dereverberation procedure, applied to each of the microphones, are also used for estimating the RTF of the early speech component.

4. PROPOSED ALGORITHM

We propose a two-stages algorithm which combines the two approaches for dereverberation which are presented in Section 3.

The first-stage consists of applying the WPE algorithm for constructing *dryer* microphones signals, i.e. use the multichannel inputs to dereverberate each of the microphone signals. A significant level of dereverberation is attained by this stage, however, due to practical considerations, model mismatch and estimation errors, a residual reverberant component at the output of this stage (including late reverberations) is inevitable.

The second stage consists of applying an MVDR for suppressing the latter residual reverberant speech component. A high level block-diagram of the proposed algorithm is depicted in Fig. 1.

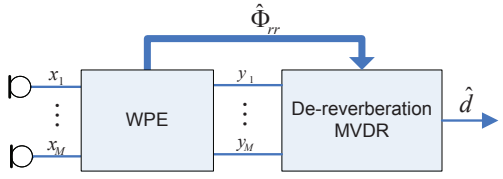

Fig. 1: Combined WPE and MVDR for dereverberation.

In the following we provide a detailed description of the proposed method. The output signals of the first-stage WPE algorithm are given by $\mathbf{y}(n,f) = \mathbf{d}(n,f) + \mathbf{c}(n,f) + \mathbf{u}(n,f)$, with

$$\mathbf{c}(n,f) = \mathbf{r}(n,f) - \sum_{\tau=n_s}^{n_e} \mathbf{p}(\tau,f)^H \left(\mathbf{d}(n-\tau,f) + \mathbf{r}(n-\tau,f)\right) \tag{10}$$

where $\mathbf{c}(n,f)$ and $\mathbf{u}(n,f)$ are the residual reverberant component and noise at the output of the WPE, respectively.

For constructing the MVDR beamformer at the second stage, we require estimates of the RTF of the early speech component $\mathbf{g}_0(f)$ and of the covariance matrix of the interference at the output of the first stage, i.e. $\mathbf{\Phi}_{cc}(f) + \mathbf{\Phi}_{uu}(f)$, where $\mathbf{\Phi}_{cc}(f)$ and $\mathbf{\Phi}_{uu}(f)$ are the covariance matrices of the components $\mathbf{c}(n,f)$ and $\mathbf{u}(n,f)$, respectively. Although, the reverberant component $\mathbf{c}(n,f)$ is non-stationary, we propose to use a time-invariant model for its covariance using long-term averaging.

As presented in Sec. 3, a common model for the spatial properties of the reverberant component is the diffuse noise field, since it comprises of a large number of statistically-independent speech reflections (due to large delays) arriving from all directions. Here, we make a similar argument for the residual reverberant speech at the output of the WPE, i.e. $\mathbf{c}(n,f)$. Assuming that $\mathbf{c}(n,f)$ consists of the speech source filtered by the late reverberant component of the IR, we conjecture that it should also follow the diffuse noise field model. The various components of the IR are depicted in Fig. 2. Furthermore, although theoretically the coherence of a diffuse noise between a pair of microphones can be expressed by Eq. (9), in practice due to estimation errors and model miss-match, the actual coherence may be different (e.g., due to microphone position errors, non-ideal and non-equal microphone frequency-responses, acoustic shading of certain directions by the device itself or by other objects and a non-uniform reverberation field). We propose to alleviate these errors, which might compromise the dereverberation performance, by utilizing the coherence of the reverberant component at the received microphones, as estimated by the WPE, i.e. $\hat{\mathbf{r}}(n,f)$.

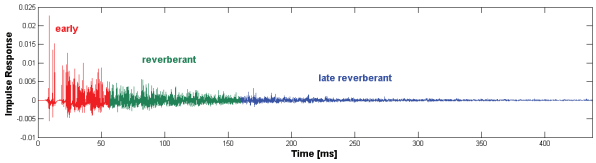

Fig. 2: A synthetic IR decomposed to its different components: early component in red; reverberant component in green; and late reverberant component in blue.

Explicitly, assuming high SNR we approximate that $\mathbf{\Phi}_{rr}(f) + \mathbf{\Phi}_{uu}(f) \approx \mathbf{\Phi}_{rr}(f)$ and estimate $\hat{\mathbf{\Phi}}_{rr}(f)$ using long-term covariance averaging of $\hat{\mathbf{r}}(n,f)$ (generated by the WPE in the first stage).

Next, similarly to the CW method for RTF estimation [14–16], we estimate the RTF of the early component by:

$$\mathbf{g}_0 \triangleq \frac{\hat{\mathbf{\Phi}}_{rr}^{1/2}(f)\mathbf{q}(f)}{e_1^H \hat{\mathbf{\Phi}}_{rr}^{1/2}(f)\mathbf{q}(f)} \tag{11}$$

where the operator $(\bullet)^{1/2}$ denotes the Cholesky decomposition, $\mathbf{q}(f)$ is the principal eigenvector of $\hat{\mathbf{\Phi}}_{rr}^{-1/2}\hat{\mathbf{\Phi}}_{yy}(f)\left(\hat{\mathbf{\Phi}}_{rr}^{-1/2}\right)^H$, $\mathbf{e}_1 \triangleq [1,0,\ldots,0]^H$ is a selection vector and $\hat{\mathbf{\Phi}}_{yy}(f)$ is an estimate for the long-term averaged covariance matrix of $\mathbf{y}(n,f)$.

Finally, the second-stage MVDR, denoted $\mathbf{w}_r(f)$ is computed:

$$\mathbf{w}_r(f) \triangleq \frac{\hat{\mathbf{\Phi}}_{rr}^{-1}(n,f)\hat{\mathbf{g}}_0(f)}{\hat{\mathbf{g}}_0^H(f)\hat{\mathbf{\Phi}}_{rr}^{-1}(n,f)\hat{\mathbf{g}}_0(f)}. \quad (12)$$

5. EXPERIMENTAL RESULTS

We verify our conjecture that the coherence of the residual reverberant component at the output of the first stage WPE algorithm, $\mathbf{c}(f)$, can be modeled as the coherence of a diffuse noise field in Section 5.1. Next, in Section 5.2 we evaluate the performance of the proposed algorithm and compare it to the unprocessed signal, to the output of the first stage WPE, and to a MVDR beamformer (computed according to an ideal coherence matrix $\mathbf{\Gamma}$).

A transcribed 5min *dry* speech recording [25], at a sampling rate of 16kHz, is filtered through simulated IRs [26], generated according to the image model [27]. A circular array with a diameter of 10cm comprising $M = 8$ uniformly spaced microphones is placed at the center of a 7m $\times$ 7m $\times$ 3m simulated room. The received signals are contaminated by an additive sensors noise at an SNR level of 50dB. We evaluate the algorithm in two RTs, 0.4s and 0.6s.

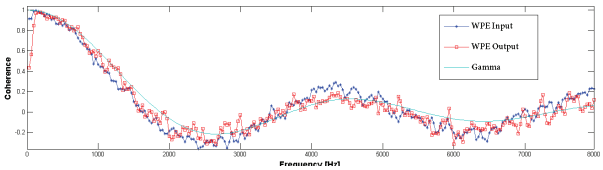

Fig. 3: Coherence of the reverberant components at the microphones, at the WPE output and of a theoretical diffuse field (for microphone spacing of 9.3cm).

5.1. Diffuse model verification

In order to verify the diffuse noise statistical model for the residual reverberant component, we examine the spatial coherence matrix of $\mathbf{c}(n,f)$. The coherence is averaged over 50 different positions of the speech source, uniformly spaced on a 2m circle around the microphone array. The empirical average coherence for each pair of signals, taken from the reverberant components at either the microphones or at the output of the first stage WPE, and the respective theoretical diffuse field coherence are compared. An example for the average coherence between a pair of microphones at a distance of 9.3cm with RT of 0.4s is depicted in Fig. 3. Clearly from this figure, all three coherence measures match closely, namely: 1) of the reverberant component at the microphones, $\mathbf{r}(n,f)$; 2) of the residual reverberant component at the output of the first stage WPE, $\mathbf{c}(n,f)$; and 3) of an ideal diffuse noise field. Similar match is obtained for all tested microphone pairs and RTs. Note that the Room Impulse Responses (RIRs) are obtained using ideal simulation (according to [26]) , and therefore the attained coherence matches an ideal diffuse field.

5.2. Performance evaluation

We evaluate and compare the performance measures of the following signals: 1) the unprocessed reference microphone signal, x_1; 2) the output of an MVDR beamformer; 3) the output of the WPE filter, y_1; 4) the output of the propose method (combining WPE and MVDR), $\hat{d}$. The following performance criteria are tested: Cepstral Distortion (CD), direct SRR [28] and WER. For the CD and SRR criteria, the desired signal is defined as the early speech component, $\mathbf{d}(n)$. We also evaluated the performance of applying multiple output MVDR beamformer followed by the WPE filter. We do not reports of the latter since they are significantly degraded compared to the other methods.

The ASR engine used for the experiments in this paper is a conventional continuous large-vocabulary speech recognizer which has been developed in Intel. The acoustic models are trained using the Kaldi open source toolkit [29] and the language model has been estimated with the MIT language modeling toolkit. Acoustic or language models have not been optimized or tuned for the test data.

The performance is evaluated for speaker to microphone array distances selected from $\{1\text{m}, 2\text{m}, 3\text{m}\}$ and for RT selected from $\{0.4\text{s}, 0.6\text{s}\}$. The results are summarized in Table.1. Evidently, from this summary, the performance in terms of all tested criteria is improved by using the proposed algorithm. Spectrograms at different stages of the proposed method for a source-array distance of 2m and a RT of 0.4s are depicted as an example in Fig. 4.

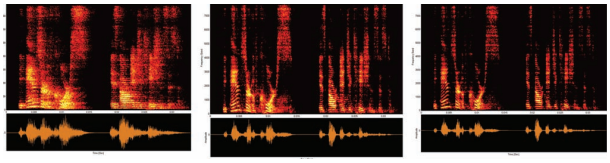

Fig. 4: Example for spectrograms of different stages of the proposed method: microphone (left); output of first stage WPE (center); output of second stage MVDR (right)

Table 1: Performance comparison at unprocessed microphone signal (denoted Mic.), MVDR using ideal diffuse noise field, WPE and the proposed method (denoted Prop.).

RT/	WER [%]			SRR [dB]			CD [dB]		
Stage	1m	2m	3m	1m	2m	3m	1m	2m	3m
RT 0.4s									
Mic.	11.2	21.3	19.8	6.6	2.2	-0.2	3.0	3.8	3.9
MVDR	5.8	10.1	11.8	13.1	6.68	3.8	2.1	2.9	3.4
WPE	4.6	9.6	10.6	18.8	12.0	4.4	1.9	2.7	3.1
Prop.	3.5	7.1	7.6	21.7	14.1	8.0	1.5	2.2	2.5
RT 0.6s									
Mic.	20.3	35.5	37.0	2.9	-1.4	-3.7	3.6	4.2	4.2
MVDR	9.1	17.8	28.4	9.1	1.8	-1.5	3.2	3.7	3.9
WPE	7.1	15.7	13.3	11.9	4.4	0.4	2.3	3.2	3.4
Prop.	5.6	7.1	10.6	14.6	7.6	2.2	1.7	2.4	2.6

6. CONCLUSIONS

The dereverberation problem using a microphone array was considered. A two stage algorithm combining the WPE, at a first stage, and an MVDR beamformer for residual dereverberation at a second stage is proposed. It is verified that the spatial coherence of the reverberant components at the microphones and at the WPE outputs are similar, and is ideally diffuse. The estimated reverberant component at the WPE is used for estimating the coherence of the residual reverberations, which is incorporated in the MVDR. Thereby, the latter beamformer accounts for non-equal microphone frequency responses, acoustic shading and non-uniform reverberation field which may compromise the theoretical diffuse field model. The improved performance of the proposed method is exemplified in simulation.

7. REFERENCES

[1] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. H. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 69–84, 2011.

[2] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.

[3] D. Yu and L. Deng, *Automatic Speech Recognition*. Springer, 2012.

[4] M. Wölfel and J. McDonough, *Distant speech recognition*. Wiley Online Library, 2009.

[5] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[6] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.

[7] K. Kinoshita, M. Delcroix, S. 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.

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

[9] M. Souden, J. Benesty, and S. Affes, "On optimal beamforming for noise reduction and interference rejection," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 109–112.

[10] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer, 2001, pp. 19–38.

[11] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[12] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proceedings of the 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC)*. Citeseer, 2004.

[13] M. Taseska and E. A. Habets, "Mmse-based blind source extraction in diffuse noise fields using a complex coherence-based a priori sap estimator," in *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*. VDE, 2012, pp. 1–4.

[14] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[15] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 60, pp. 233–246, Jan. 2012.

[16] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 544–548.

[17] R. T. I. Cohen and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[18] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[19] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 85–88.

[20] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

[21] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[22] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, no. 1, pp. 43–56, 1988.

[23] E. A. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[24] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1433–1451, 2008.

[25] A. Hoje, "Learning english naturally: a speech recording," http://traffic.libsyn.com/effortlessenglish/Learn_English_Naturally.mp3.

[26] E. Habets, "Room impulse response (RIR) generator," http://home.tiscali.nl/ehabets/rir_generator.html, Jul. 2006.

[27] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[28] P. A. Naylor, N. D. Gaubitch, and E. A. Habets, "Signal-based performance evaluation of dereverberation algorithms," *Journal of Electrical and Computer Engineering*, vol. 2010, p. 1, 2010.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.