

# Multichannel Online Dereverberation based on Spectral Magnitude Inverse Filtering

Xiaofei Li, Laurent Girin, Sharon Gannot and Radu Horaud

**Abstract**—This paper addresses the problem of multichannel online dereverberation. The proposed method is carried out in the short-time Fourier transform (STFT) domain, and for each frequency band independently. In the STFT domain, the time-domain room impulse response is approximately represented by the convolutive transfer function (CTF). The multichannel CTFs are adaptively identified based on the cross-relation method, and using the recursive least square criterion. Instead of the complex-valued CTF convolution model, we use a nonnegative convolution model between the STFT magnitude of the source signal and the CTF magnitude, which is just a coarse approximation of the former model, but is shown to be more robust against the CTF perturbations. Based on this nonnegative model, we propose an online STFT magnitude inverse filtering method. The inverse filters of the CTF magnitude are formulated based on the multiple-input/output inverse theorem (MINT), and adaptively estimated based on the gradient descent criterion. Finally, the inverse filtering is applied to the STFT magnitude of the microphone signals, obtaining an estimate of the STFT magnitude of the source signal. Experiments regarding both speech enhancement and automatic speech recognition are conducted, which demonstrate that the proposed method can effectively suppress reverberation, even for the difficult case of a moving speaker.

## I. INTRODUCTION

This work addresses the problem of multichannel online dereverberation of speech signals, emitted either by a static or a moving speaker. The objective of dereverberation is to improve speech intelligibility for human listening or automatic speech recognition (ASR). The output of a dereverberation system may include some early reflections, since they deteriorate neither the speech quality nor speech intelligibility [1].

Multichannel dereverberation includes the following different techniques. Spectral enhancement techniques [2], [3], [4], which are performed in the short-time Fourier transform (STFT) domain, remove late reverberation by spectral subtraction. To iteratively estimate the room filters and the speech source signal, other techniques minimize a cost function between the microphone signal(s) and a generative model thereof (or equivalently maximize an objective function). The generative model here mainly indicates the convolutive model between the room filters and the source signal. Sometimes the model that the source signal is assumed to be generated by a random process is also used. These techniques are also usually applied in the STFT domain, where the time-domain

RIR is represented by a subband convolutive transfer function (CTF). An expectation-maximization (EM) algorithm is used in [5] to maximize the likelihood of the microphone signals. The idea is extended to joint dereverberation and source separation in [6]. In [7], [8], [9], a nonnegative convolution approximation is assumed, namely the STFT magnitude of the microphone signal is approximated by the convolution between the STFT magnitude of the source signal and the CTF magnitude. Based on this nonnegative model, tensor factorization [7], iterative auxiliary functions [8] and iterative multiplicative update [9] are used to minimize the fit cost between the STFT magnitude of the microphone signal and its nonnegative generative model. Inverse filtering techniques aim at inverting the room convolution process and recovering the source signal. Depending on the way inverse filters are estimated, inverse filtering techniques can be classified into two groups:

- Linear prediction based techniques model the convolution with the RIR as an auto-regressive (AR) process. This AR process can be carried out either in the time domain or in the STFT domain. In the linear-predictive multi-input equalization (LIME) algorithm [10], the speech source signal is estimated as the multichannel linear prediction residual, which however is excessively whitened. The whitening effect is then compensated by estimating the average speech characteristics. To avoid such whitening effect, a prediction delay is used in the delayed linear prediction techniques [11], [12]. These techniques only model late reverberation into the AR process and leave early reflections of the speech signal in the prediction residual. To account for the time-varying characteristics of speech, the statistical model-based approach [12] iteratively estimates the time-varying speech variance and normalizes the linear prediction with this speech variance. This variance-normalized delayed linear prediction method is also called weighted prediction error (WPE);
- Techniques based on system identification first blindly identify the room filters. Then, the corresponding inverse filters are estimated and applied on the microphone signals to recover the source signal. The cross-relation method [13] is a widely-used system identification method. Inverse filter estimation techniques include the multiple-input/output inverse theorem (MINT) method [14] and some of its variants, such as channel shortening [15] and partial MINT [16]. In [17], [18], the cross-relation method was applied in the STFT domain for CTF estimation. Several variants of subband MINT were proposed based on filter banks [19], [20] or CTF model [21], [22].

X. Li and R. Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France.

L. Girin is with GIPSA-lab and with Univ. Grenoble Alpes, Saint-Martin d'Hères, France.

Sharon Gannot is with Bar Ilan University, Faculty of Engineering, Israel. This work was supported by the ERC Advanced Grant VHIA #340113.

For dynamic scenarios with moving speakers or speech turns among speakers, an online dereverberation method is required. Based on the CTF model, an online likelihood maximization method was proposed in [23], [24] using a Kalman filter and an EM algorithm. An online extension of LIME was proposed in [25] using several different adaptive estimation criteria, such as normalized least mean squares (LMS), steepest descent, conjugate gradient and recursive least square (RLS). RLS-based adaptive WPE (AWPE) [26], [27], [28], [29] has become a popular online dereverberation method. For example, it is now used in the GoogleHome smart loudspeaker [30]. In AWPE, the anechoic speech variance is estimated using a spectral subtraction method in [27], and is simply approximated by the microphone speech variance in [26], [28], [29]. In [31], [32], a probabilistic model and a Kalman filter were used to implement the delayed linear prediction method, which can be seen as a generalization of the RLS-based AWPE. A class of adaptive cross-relation methods were proposed in [33] for online system identification, with the adaptive estimation criteria of normalized LMS and multichannel Newton method. Adaptive multichannel equalization methods were proposed in [34], [35] based on time-domain MINT and gradient descent update. These methods reduce the computational complexity of the original MINT, however they were only used for offline multichannel equalization in static scenarios.

In our previous work [18], a blind dereverberation method was proposed in batch mode for static scenarios, which consists of a blind CTF identification algorithm and a sparse source recovery algorithm. The CTF identification algorithm is based on the cross-relation method. For source recovery, instead of the complex-valued CTF convolution model, we used its nonnegative convolution approximation [7], [8], [9], since the latter was shown to be less sensitive to the CTF perturbations than the former. More precisely, the STFT magnitude of the source signal is recovered by solving a basis pursuit problem that minimizes the  $\ell_1$ -norm of the STFT magnitude of the source signal while constraining the fit cost, between the STFT magnitude of the microphone signals and the nonnegative convolution model, to be below a tolerance.

In the present work, we propose an online dereverberation method. First, we extend the batch formulation of CTF identification in [18] to an adaptive method based on an RLS-like recursive update. The RLS-like method has a better convergence rate than the normalized LMS method used in [33], which is crucial for its application in dynamic scenarios. For source recovery, also based on the nonnegative convolution model, we propose an online STFT magnitude inverse filtering method: the inverse filters of the CTF magnitudes are estimated and applied to the STFT magnitude of the microphone signals to obtain an estimate of the STFT magnitude of the source signal. The inverse filters estimation is based on the MINT theorem [14]. Due to the use of the nonnegative CTF convolution model, the proposed magnitude MINT is different from the conventional MINT methods, such as [15], [16], [21], mainly in aspect to that multichannel fusion and target response. Following the spirit of normalized LMS, we propose to adaptively update the inverse filters based on a gradient descent method.

In summary, the proposed method consists of two novelties 1) an online RLS-like CTF identification technique, and ii) an online STFT-magnitude inverse filtering technique. To the best of our knowledge this is the first time such procedures are proposed for online speech dereverberation. Experimental comparison with AWPE shows that the proposed method performs better for the moving speaker case, mainly due to the use of the less sensitive magnitude convolution model.

The remainder of this paper is organized as follows. The adaptive CTF identification is presented in Section II. The STFT magnitude inverse filtering method is presented in Section III. Experiments with two datasets are presented in Section IV. Section V concludes the work.

## II. ONLINE CTF IDENTIFICATION

We consider a system with  $I$  channels and one speech source. In the time domain, the  $i$ -th microphone signal  $x_i(n)$  is

$$x_i(n) = s(n) \star a_i(n) + e_i(n), \quad i = 1, \dots, I \quad (1)$$

where  $n$  is the time index,  $\star$  denotes convolution,  $s(n)$  is the speech source signal, and  $a_i(n)$  is the RIR from the speech source to the  $i$ -th microphone. The additive noise term  $e_i(n)$  will be discarded in the following, since we do not consider noise in this work. In the STFT domain, based on the CTF approximation, we have

$$x_{i,p,k} \approx s_{p,k} \star a_{i,p,k}, \quad i = 1, \dots, I \quad (2)$$

where  $x_{i,p,k}$  and  $s_{p,k}$  are the STFT coefficients of the corresponding signals, and the CTF  $a_{i,p,k}$  is the subband representation of the RIR  $a_i(n)$ .  $p = 1, \dots, P$  denotes the STFT frame index and  $k = 0, \dots, N-1$  denotes the frequency index,  $P$  is the number of signal frames in a given processed speech sequence, and  $N$  is the STFT frame (window) length. The convolution is executed along the frame index  $p$ . The length of the CTF, denoted as  $Q$ , is assumed to be identical for all frequency bins and is approximately equal to the length of the corresponding RIR divided by  $L$ , where  $L$  denotes the STFT frame step.

### A. Batch CTF Identification

In [18], we proposed a CTF identification method in batch mode. It is based on the following cross-relation between channels [13]:

$$x_{i,p,k} \star a_{j,p,k} = s_{p,k} \star a_{i,p,k} \star a_{j,p,k} = x_{j,p,k} \star a_{i,p,k}. \quad (3)$$

However, this equation cannot be directly used. The reason is that, for the oversampling case (i.e.  $L < N$ ), there is a common region with magnitude close to zero in the frequency response of the CTFs for all channels, caused by the non-flat frequency response of the STFT window. This common zero frequency region is problematic for the cross-relation method. It can be alleviated by using critical sampling (i.e.  $L = N$ ), which however leads to a severe frequency aliasing of the

signals. To achieve a good trade-off, it was proposed in [18] that the signal STFT coefficients are oversampled to avoid frequency aliasing, but the multichannel CTF coefficients are forced to be critically sampled to avoid the common zero problem. More precisely, the Hamming window<sup>1</sup> is used, and we set  $L = N/4$  and  $L_f = N$ , where  $L_f$  denotes the frame step of CTF. Since the channel identification algorithm presented in this section and the inverse filtering algorithm presented in the next section are both applied frequency-wise, hereafter the frequency index  $k$  will be omitted for clarity of presentation.

The critically sampled CTF is defined in vector form as  $\tilde{\mathbf{a}}_i = [a_{i,0}, a_{i,4}, \dots, a_{i,4(\tilde{Q}-1)}]^\top$ , where  $^\top$  denotes matrix/vector transpose and  $\tilde{Q} = \lceil Q/4 \rceil$  ( $\lceil \cdot \rceil$  denotes ceiling function). From the oversampled STFT coefficients of microphone signals, we define the convolution vector as  $\tilde{\mathbf{x}}_{i,p} = [x_{i,p}, x_{i,p-4}, \dots, x_{i,p-4(\tilde{Q}-1)}]^\top$ ,  $p = 1, \dots, P$ . Note that, when  $p < 4(\tilde{Q} - 1) + 1$ , the vector  $\tilde{\mathbf{x}}_{i,p}$  is constructed by padding zeros. Then, the cross-relation can be recast as

$$\tilde{\mathbf{x}}_{i,p}^\top \tilde{\mathbf{a}}_j = \tilde{\mathbf{x}}_{j,p}^\top \tilde{\mathbf{a}}_i. \quad (4)$$

This convolution formulation can be interpreted as that 3/4 of the original oversampled CTF coefficients are forced to be zero. This cross-relation is defined for each microphone pair. To present the cross-relation equation in terms of the CTF of all channels, i.e.

$$\tilde{\mathbf{a}} = [\tilde{\mathbf{a}}_1^\top, \tilde{\mathbf{a}}_2^\top, \dots, \tilde{\mathbf{a}}_I^\top]^\top, \quad (5)$$

we define:

$$\tilde{\mathbf{x}}_{ij,p} = \underbrace{[0, \dots, 0]^\top}_{(i-1)\tilde{Q}}, \underbrace{[\tilde{\mathbf{x}}_{j,p}^\top, 0, \dots, 0]^\top}_{(j-i-1)\tilde{Q}}, \underbrace{[-\tilde{\mathbf{x}}_{i,p}^\top, 0, \dots, 0]^\top}_{(I-j)\tilde{Q}}, \quad j > i. \quad (6)$$

Then the cross-relation can be written as:

$$\tilde{\mathbf{x}}_{ij,p}^\top \tilde{\mathbf{a}} = 0. \quad (7)$$

There is a total of  $M = I(I-1)/2$  distinct microphone pairs, indexed by  $(i, j)$  with  $j > i$ . For notational convenience, let  $m = 1, \dots, M$  denote the microphone-pair index. Then let the subscript  $ij$  be replaced with  $m$ . For the static speaker case, the CTF  $\tilde{\mathbf{a}}$  is time invariant, and can be estimated by solving the following constrained least square problem in batch mode

$$\min \sum_{p=1}^P \sum_{m=1}^M |\tilde{\mathbf{x}}_{m,p}^\top \tilde{\mathbf{a}}|^2 \quad \text{s.t.} \quad \mathbf{g}^\top \tilde{\mathbf{a}} = 1, \quad (8)$$

where  $|\cdot|$  denotes the (entry-wise) absolute value, and  $\mathbf{g}$  is a constant vector

$$\mathbf{g} = [1, \underbrace{0, \dots, 0}_{\tilde{Q}-1}, 1, \underbrace{0, \dots, 0}_{\tilde{Q}-1}, \dots, 1, \underbrace{0, \dots, 0}_{\tilde{Q}-1}]^\top. \quad (9)$$

Here we constrain the sum of the first entries of the  $I$  CTFs to 1, i.e.  $\sum_{i=1}^I a_0^i = 1$ . As discussed in [18], in contrast to the

eigendecomposition method proposed in [13], this constrained least square method is robust against noise interference. The solution to (8) is

$$\check{\mathbf{a}} = \frac{\mathbf{R}^{-1} \mathbf{g}}{\mathbf{g}^\top \mathbf{R}^{-1} \mathbf{g}}, \quad (10)$$

where  $\mathbf{R}$  is the covariance matrix of the microphone signals, i.e.  $\mathbf{R} = \sum_{p=1}^P \sum_{m=1}^M \tilde{\mathbf{x}}_{m,p}^* \tilde{\mathbf{x}}_{m,p}^\top$ .

### B. Recursive CTF Identification

In dynamic scenarios, the CTF vector  $\tilde{\mathbf{a}}$  is time-varying, which thus is rewritten as  $\tilde{\mathbf{a}}^{(p)}$  to specify the frame-dependency. Note that we need to distinguish the superscript  $(p)$ , which represents the time index with respect to the online update, from the subscript  $p$ , which represents the frame index of the signals and filters. At frame  $p$ ,  $\tilde{\mathbf{a}}^{(p)}$  can be calculated by (10) using the microphone signals at frame  $p$  and recent frames. However, this requires needs a large amount of computations of inverse matrix, which is computationally expensive. In this work, we adopt the RLS-like algorithm for recursive CTF identification. At the current frame  $p$ , RLS aims to solve the minimization problem

$$\min \sum_{p'=1}^p \lambda^{p-p'} \left( \sum_{m=1}^M |\tilde{\mathbf{x}}_{m,p'}^\top \tilde{\mathbf{a}}^{(p)}|^2 \right) \quad \text{s.t.} \quad \mathbf{g}^\top \tilde{\mathbf{a}} = 1. \quad (11)$$

The forgetting factor  $\lambda^{p-p'}$  with  $\lambda \in (0, 1]$  gives exponentially decaying weight to older frames. This time-weighted minimization problem can be solved using (10) with  $\mathbf{R}$  replaced by a frame-dependent sample covariance matrix  $\mathbf{R}^{(p)} = \sum_{p'=1}^p \lambda^{p-p'} (\sum_{m=1}^M \tilde{\mathbf{x}}_{m,p'}^* \tilde{\mathbf{x}}_{m,p'}^\top)$ , namely

$$\check{\mathbf{a}}^{(p)} = \frac{(\mathbf{R}^{(p)})^{-1} \mathbf{g}}{\mathbf{g}^\top (\mathbf{R}^{(p)})^{-1} \mathbf{g}}. \quad (12)$$

The sample covariance matrix  $\mathbf{R}^{(p)}$  can be recursively updated as

$$\mathbf{R}^{(p)} = \lambda \mathbf{R}^{(p-1)} + \sum_{m=1}^M \tilde{\mathbf{x}}_{m,p}^* \tilde{\mathbf{x}}_{m,p}^\top. \quad (13)$$

The covariance matrix is updated in  $M$  steps, where each step modifies the covariance matrix by adding a rank-one matrix  $\tilde{\mathbf{x}}_{m,p}^* \tilde{\mathbf{x}}_{m,p}^\top$ ,  $m = 1, \dots, M$ . To avoid the explicit inverse matrix computation, instead of  $\mathbf{R}^{(p)}$  itself, we recursively estimate its inverse  $(\mathbf{R}^{(p)})^{-1}$  based on the Sherman-Morrison formula (14). This procedure is summarized in Algorithm 1, where the Sherman-Morrison formula is applied in each of  $M$  loops. As an initialization, we set  $(\mathbf{R}^{(0)})^{-1}$  to  $1,000\mathbf{I}$ , where  $\mathbf{I}$  denotes identity matrix.

As a least square problem, the number of frames used to estimate  $\tilde{\mathbf{a}}^{(p)}$  should be proportional to the length of the critically sampled CTF, i.e.  $\tilde{Q}$ , and is thus denoted with  $\tilde{P} = \rho \tilde{Q}$ . On the one hand, a large  $\tilde{P}$  is required to ensure the estimation accuracy. On the other hand,  $\tilde{P}$  should be set as small as possible to reduce the dependency of the estimation on the past frames, namely to reduce the latency of the estimation, which is especially important for the moving speaker case. Similar

<sup>1</sup>Other commonly used windows, such as Hanning and *sine* windows, are also applicable.

---

**Algorithm 1** Recursive estimation of  $(\mathbf{R}^{(p)})^{-1}$  at frame  $p$ 


---

Input:  $\tilde{\mathbf{x}}_{m,p}$ ,  $m = 1, \dots, M$ ;  $(\mathbf{R}^{(p-1)})^{-1}$   
 Initialization:  $\mathbf{P} \leftarrow \lambda^{-1}(\mathbf{R}^{(p-1)})^{-1}$   
**for** each microphone pair  $m = 1$  to  $M$  **do**

$$\mathbf{P} \leftarrow \mathbf{P} - (\mathbf{P}\tilde{\mathbf{x}}_{m,p}^*\tilde{\mathbf{x}}_{m,p}^\top\mathbf{P})/(1 + \tilde{\mathbf{x}}_{m,p}^\top\mathbf{P}\tilde{\mathbf{x}}_{m,p}^*) \quad (14)$$

**end for**

Output:  $(\mathbf{R}^{(p)})^{-1} \leftarrow \mathbf{P}$

---

to the RIR samples, the critically sampled CTF coefficients can be assumed to be temporally uncorrelated. However, the microphone signals STFT coefficients are highly correlated due to the temporal correlation of time-domain speech samples and to the oversampling of signals STFT coefficients (i.e. large overlapping of STFT frames). Empirically, we set  $\rho = 4$  to compensate the signal oversampling effect, which is a good tradeoff between the accuracy and latency of CTF estimation. To approximately have a memory of  $\tilde{P}$  frames, we can set  $\lambda = \frac{\tilde{P}-1}{\tilde{P}+1}$ .

### III. ADAPTIVE STFT MAGNITUDE INVERSE FILTERING

In [18], it was found that the estimated complex-valued CTF  $\hat{\mathbf{a}}$  is not accurate enough for effective inverse filtering, due to the influence of noise interference and the frequency aliasing caused by critical sampling. To reduce the sensitivity of the inverse filtering procedure to the CTF perturbations, instead of the complex-valued CTF convolution (2), its magnitude approximation was used, i.e.

$$|x_{i,p}| \approx |s_p| \star |a_{i,p}|, \quad i = 1, \dots, I. \quad (15)$$

This magnitude convolution model is widely used in the context of dereverberation, e.g. [7], [8], [9]. In [21], [22], we proposed a MINT method based on the complex-valued CTF convolution for multisource separation and dereverberation. In the present work, we adapt this MINT method to the magnitude domain, and develop its adaptive version for online dereverberation.

#### A. Adaptive MINT in the Magnitude Domain

The CTF estimate of each channel can be extracted from  $\check{\mathbf{a}}^{(p)}$ , denoted by  $\check{\mathbf{a}}_i^{(p)}$ ,  $i = 1, \dots, I$ . Let  $\bar{\mathbf{a}}_i^{(p)} = |\check{\mathbf{a}}_i^{(p)}|$  denote the CTF magnitude vector, and  $\bar{a}_{i,0}^{(p)}, \dots, \bar{a}_{i,\tilde{Q}-1}^{(p)}$  its elements. Define the inverse filters of  $\bar{\mathbf{a}}_i^{(p)}$  in vector form as  $\mathbf{h}_i^{(p)} \in \mathbb{R}^{\tilde{O} \times 1}$ ,  $i = 1, \dots, I$ , where  $\tilde{O}$  is the length of the inverse filters. Note that both  $\bar{\mathbf{a}}_i^{(p)}$  and  $\mathbf{h}_i^{(p)}$  are critically sampled. To apply the magnitude inverse filtering using  $\mathbf{h}_i^{(p)}$ , we construct the STFT magnitude vector of microphone signals as  $\bar{\mathbf{x}}_{i,p} = [|x_{i,p}|, |x_{i,p-4}|, \dots, |x_{i,p-4(\tilde{O}-1)}|]^\top$ . The output of the multichannel inverse filtering, namely

$$\bar{s}_p = \sum_{i=1}^I \mathbf{h}_i^{(p)\top} \bar{\mathbf{x}}_{i,p} \quad (16)$$

should target the STFT magnitude of the source signal, i.e.  $|s_p|$ .

To this aim, the multichannel equalization, i.e. MINT, should target an impulse function, namely

$$\sum_{i=1}^I \bar{\mathbf{A}}_i^{(p)} \mathbf{h}_i^{(p)} = \mathbf{d}, \quad (17)$$

where the impulse function  $\mathbf{d}$  is defined by  $\mathbf{d} = [1, 0, \dots, 0]^\top \in \mathbb{R}^{(\tilde{Q}+\tilde{O}-1) \times 1}$ , and the convolution matrix  $\bar{\mathbf{A}}_i^{(p)}$  is defined by

$$\bar{\mathbf{A}}_i^{(p)} = \begin{bmatrix} \bar{a}_{i,0}^{(p)} & 0 & \cdots & 0 \\ \bar{a}_{i,1}^{(p)} & \bar{a}_{i,0}^{(p)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \bar{a}_{i,\tilde{Q}-1}^{(p)} & \vdots & \ddots & 0 \\ 0 & \bar{a}_{i,\tilde{Q}-1}^{(p)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \bar{a}_{i,\tilde{Q}-1}^{(p)} \end{bmatrix} \in \mathbb{R}_{\geq 0}^{(\tilde{Q}+\tilde{O}-1) \times \tilde{O}}. \quad (18)$$

In a more compact form, we can write

$$\bar{\mathbf{A}}^{(p)} \mathbf{h}^{(p)} = \mathbf{d}, \quad (19)$$

where  $\bar{\mathbf{A}}^{(p)} = [\bar{\mathbf{A}}_1^{(p)}, \dots, \bar{\mathbf{A}}_I^{(p)}] \in \mathbb{R}_{\geq 0}^{(\tilde{Q}+\tilde{O}-1) \times I\tilde{O}}$  and  $\mathbf{h}^{(p)} = [\mathbf{h}_1^{(p)\top}, \dots, \mathbf{h}_I^{(p)\top}]^\top \in \mathbb{R}^{I\tilde{O} \times 1}$ . The inverse filter estimation amounts to solving problem (19), or equivalently minimizing the squared error

$$J^{(p)} = \|\bar{\mathbf{A}}^{(p)} \mathbf{h}^{(p)} - \mathbf{d}\|^2, \quad (20)$$

where  $\|\cdot\|$  denotes  $\ell_2$ -norm. The size of  $\bar{\mathbf{A}}^{(p)}$  can be adjusted by tuning the length of the inverse filter, i.e.  $\tilde{O}$ . If  $\bar{\mathbf{A}}^{(p)}$  is square or wide, i.e.  $(\tilde{Q} + \tilde{O} - 1) \leq I\tilde{O}$  and thus  $\tilde{O} \geq \frac{\tilde{Q}-1}{I-1}$ , (19) has an exact solution and (20) has zero error, which means an exact inverse filtering can be achieved. Otherwise, (19) is a least square problem, and only an approximate inverse filtering can be achieved. In practice, the configuration with  $\tilde{O} = \lceil \frac{\tilde{Q}-1}{I-1} \rceil$ , i.e.  $\bar{\mathbf{A}}^{(p)}$  is square or slightly wide, achieves the optimal performance, see, e.g., [16], [21], [36].

The minimization problem (20) has a closed-form solution. However, this needs the computation of an inverse matrix for each frame and frequency, which is computationally expensive. In this work, we propose to adaptively estimate  $\mathbf{h}^{(p)}$  following the principle of normalized LMS. For a summary of normalized LMS design and analysis, please refer to Chapter 10.4 of [37]. The proposed LMS-like adaptive estimation method in the following is based on a stationary filtering system, but can be directly used for the nonstationary case due to its natural adaptive characteristic. In a stationary system, the filter to be estimated, i.e. the inverse filter  $\mathbf{h}$  in the present work, is assumed to be time-invariant. Note that with the superscript  $^{(p)}$  removed,  $\mathbf{h}$  denotes the stationary filter. The instantaneous filtering process (19) and squared error (20) are a random instance of the stationary system. The aim of LMS

is to adaptively minimize the mean squared error  $\mathbb{E}[J]$ , where  $\mathbb{E}[\cdot]$  denotes expectation. Note that  $J$  and  $\bar{\mathbf{A}}$  presented later denote the corresponding random variables. At frame  $p$ , the adaptive updation uses the gradient of the instantaneous error  $J^{(p)}$  at the previous estimation point  $\mathbf{h}^{(p-1)}$ , i.e.

$$\Delta J^{(p)}|_{\mathbf{h}^{(p-1)}} = 2\bar{\mathbf{A}}^{(p)\top}(\bar{\mathbf{A}}^{(p)}\mathbf{h}^{(p-1)} - \mathbf{d}). \quad (21)$$

An estimate of  $\mathbf{h}^{(p)}$  based on the gradient descent updation is

$$\mathbf{h}^{(p)} = \mathbf{h}^{(p-1)} - \frac{\mu}{\text{Tr}(\bar{\mathbf{A}}^{(p)\top}\bar{\mathbf{A}}^{(p)})} \Delta J^{(p)}|_{\mathbf{h}^{(p-1)}}, \quad (22)$$

where  $\text{Tr}(\cdot)$  denotes the matrix trace, and  $\frac{\mu}{\text{Tr}(\bar{\mathbf{A}}^{(p)\top}\bar{\mathbf{A}}^{(p)})}$  is the *step-size* for gradient descent. The normalization term  $\frac{1}{\text{Tr}(\bar{\mathbf{A}}^{(p)\top}\bar{\mathbf{A}}^{(p)})}$  is set to make the gradient descent updation converge to an optimal solution, namely to ensure the update stability. It is proven in [37] that, to guarantee the stability, the *step-size* should be set to be lower than  $\frac{1}{\text{Tr}(\mathbb{E}[\bar{\mathbf{A}}^\top\bar{\mathbf{A}}])}$ . Following the principle of normalized LMS, we replace the expectation  $\mathbb{E}[\bar{\mathbf{A}}^\top\bar{\mathbf{A}}]$  with the instantaneous matrix  $\bar{\mathbf{A}}^{(p)\top}\bar{\mathbf{A}}^{(p)}$ . The matrix trace can be computed as  $\text{Tr}(\bar{\mathbf{A}}^{(p)\top}\bar{\mathbf{A}}^{(p)}) = \tilde{Q} \sum_{i=1}^I \bar{\mathbf{a}}_i^{(p)\top} \bar{\mathbf{a}}_i^{(p)}$ . The constant step factor  $\mu$  ( $0 < \mu \leq 1$ ) should be empirically set to achieve a good tradeoff between convergence rate (and tracking ability for the dynamic scenarios with time-varying CTFs) and update stability.

### B. Pairwise Processing

To facilitate the presentation and understanding, the MINT problem (19) is formulated for the general multiple-channel case, namely the multiple channels are simultaneously equalized to a single-channel target function  $\mathbf{d}$ . However, experiments show that this formulation only performs well for the two-channel case. The length of the critically sampled CTF, i.e.  $\tilde{Q}$ , is relative small. As will be shown in the experiments section,  $\tilde{Q}$  is related to both the STFT setting and the reverberation time, and is set to 4 in this work. As mentioned above, the length of the inverse filters should be optimally set to  $\tilde{O} = \lceil \frac{\tilde{Q}-1}{I} \rceil$ . For the two-channel case, the length of the inverse filters is close to the length of CTFs, i.e.  $\tilde{O} = \tilde{Q} - 1$ . Note that we actually set  $\tilde{O} = \tilde{Q}$  in the experiments. The STFT magnitude of the microphone signals for the current frame includes the information of the past  $\tilde{Q} - 1$  frames due to the CTF convolution. Therefore, by intuition, the setting with  $\tilde{O} = \tilde{Q}$  is reasonable that the magnitude inverse filtering at the current frame explores the past  $\tilde{Q} - 1$  frames, more precisely subtracts the reflection information of the past  $\tilde{Q} - 1$  frames from the current frame. When the number of channels is larger than two, the length of the inverse filters will be very small. The magnitude convolution (15) is just a rough approximation, and thus the magnitude inverse filtering (16) is also a rough approximation of the complex-valued inverse filtering. Even though the magnitude MINT (19) can be exactly solved with very short inverse filters, it seems not able to reasonably approximate the complex-valued MINT, and thus the magnitude inverse filtering cannot efficiently suppress reverberation. By setting  $\tilde{O}$  to  $\tilde{Q}$ , as is done for the two-channel case, the dereverberation performance can be

largely improved relative to the very small  $\tilde{O}$  case. With fixed  $\tilde{O} = \tilde{Q}$ , the dereverberation performance is improved along with the increase of the number of channels. However, the improvement is not always significant. The possible reason is that, even if the accuracy of the magnitude MINT (19) can be largely improved by increasing the number of channels, the dereverberation performance cannot be accordingly improved since it is limited by the inaccuracy of the approximate magnitude inverse filtering.

To fully employ all the channels in a proper way, we propose to perform the following pairwise magnitude inverse filtering method, which is shown by experiments to systematically outperform the above mentioned multichannel method with fixed  $\tilde{O} = \tilde{Q}$ . First, the adaptive MINT (and inverse filtering) presented in Section III-A is separately applied for each microphone pair. Then the estimates of the source magnitude obtained by all the  $M$  microphone pairs are averaged as a new source magnitude estimate, which is still denoted with  $\bar{s}_p$  for brevity. The source magnitude estimate provided by each microphone pair is assumed to be independent from the other ones, thence the average of them is hopefully suffering from lower interferences and distortions than each of them.

The above STFT magnitude inverse filtering does not automatically guarantee the non-negativity of  $\bar{s}_p$ , which is infeasible solution for the STFT magnitude of the source signal. The negative values generally appear for the microphone signal frames with a magnitude that is considerably smaller than the magnitude in the preceding frames. Indeed, in that case, applying negative inverse filter coefficients to the preceding frames produces a negative magnitude estimate. Such negative frames are normally following a high-energy speech region, but themselves include very low source energy or purely reverberations. To overcome this, one way is to add the non-negativity constraint of the inverse filtering output to (20), which however leads to a larger complexity for both algorithm design and computation. Instead, we constrain the lower limit of the STFT magnitude of source signal according to the (averaged) STFT magnitude of microphone signals. Formally, the final estimate of the STFT magnitude of source signal is

$$\hat{s}_p = \max(\bar{s}_p, G_{\min} \frac{1}{I} \sum_{i=1}^I |x_{i,p}|), \quad (23)$$

where  $G_{\min}$  is a constant lower limit gain factor. This type of lower limit is widely used in the single-channel speech enhancement methods, e.g. in [38], mainly to keep the noise naturalness. The proposed pairwise processing-based STFT magnitude inverse filtering method at one frame is summarized in Algorithm 2, which is recursively executed frame by frame. As an initialization, we set  $\mathbf{h}^{(0)}$  to a vector with all entries being zero.

Finally, the STFT phase of one of the microphone signals, e.g. the first microphone is used in this work, is taken as the phase of the estimated STFT coefficient of source signal, then we have  $\hat{s}_p = \hat{s}_p e^{j \arg[x_p^1]}$ , where  $\arg[\cdot]$  is the phase of complex number. The time-domain source signal  $\hat{s}(n)$  is obtained by applying the inverse STFT. The MINT formulation (19) implies

---

**Algorithm 2** Adaptive STFT magnitude inverse filtering at frame  $p$ 


---

Input:  $\check{\mathbf{a}}^{(p)}$  computed by (12).  
**for** each microphone pair  $m = 1$  to  $M$  **do**  
  Input:  $\mathbf{h}^{(p-1)}$ .  
  1 Construct  $\bar{\mathbf{A}}^{(p)}$  using (18),  
  2 Compute gradient using (21),  
  3 Update inverse filter using (22),  
  4 Inverse filtering using (16), obtain STFT magnitude estimate with microphone pair  $m$ ,  
**end for**  
5 Average the estimates of all microphone pairs,  
6 Apply lower limit using (23).  
Output: final estimate  $\check{s}_p$ .

---

that the proposed inverse filtering method aims at recovering the signal corresponding to the first CTF frame, which not only includes the direct-path impulse response, but also the early reflections within the period of one STFT frame. As a result, according to the STFT window size, the estimated source signal  $\hat{s}(n)$  includes both direct-path source signal and early reflections within  $N/f$  seconds following the direct-path propagation, where  $f$  is the signal sampling rate.

### C. Difference from Conventional MINT Methods

Due to the use of i) the magnitude convolution model, ii) the critically sampled CTFs and inverse filters, and iii) the adaptive update of the inverse filters, the present adaptive MINT method is largely different from the complex-valued CTF MINT [21], [22] and the time-domain MINT, such as [15], [16], [36], [39], [40]. Besides the pairwise processing scheme, the two main differences are the following.

1) *Desired Response of MINT*: In many time-domain methods, to improve the robustness of MINT to microphone noise and filter perturbations, the target function (desired response) is designed to have multiple non-zero taps. This can be done either by explicitly filling the target function with multiple non-zero taps, such as the partial MINT in [16], or by relaxing the constraint for some taps, such as the relaxed multichannel least-squares in [36]. This way, the desired response with multiple non-zero taps includes both the direct-path propagation and some early reflections. In the present work, the impulse function  $\mathbf{d}$  is used as the desired response of MINT in the CTF domain, namely only one non-zero tap is sufficient, since one tap of CTF corresponds to a segment of RIR that includes both direct-path propagation and early reflections.

It was shown in [21], [22] that, due to the effect of short time STFT windows, the oversampled CTF of multiple channels have common zeros, which is problematic for MINT. A target function incorporating the information of the STFT windows was proposed to compensate the common zeros. In the present work, the critically sampled CTFs do not suffer from this problem.

A modeling delay is always used in the time-domain MINT and complex-valued CTF MINT methods, i.e., in the target function, a number of zeros are inserted prior to the first non-zero tap. It is shown in [21], [39] that the optimal length of the modeling delay is related to the direct-path tap and the length of the room filters. In the present method, the room filters, i.e. CTFs, are blindly estimated, with the direct-path lying in the first tap. In addition, the CTF length is very small as mentioned above. Therefore, the modeling delay is set to 0, which achieved the best performance in the experiments.

2) *Energy Regularization*: An energy regularization is used in [16], [21], [39] to limit the energy of the inverse filters derived by MINT, since high energy inverse filters will amplify microphone noise and filter perturbations. For example, in the present problem, the optimal solution of MINT (20) could have a very large energy, especially when the matrix  $\bar{\mathbf{A}}^{(p)\top} \bar{\mathbf{A}}^{(p)}$  is ill-conditioned. However, for the proposed method, the inverse filters are adaptively updated based on the previous estimation. The step size is set with the guaranteed update stability. Thence, the energy of the inverse filters will not be boosted once the inverse filters are properly initialized.

## IV. EXPERIMENTS

### A. Experimental Configuration

1) *Dataset*: We evaluate the proposed method using two datasets.

- REVERB challenge dataset [41]. We test the evaluation set of SimData-room3 and RealData. SimData-room3 data was generated by convolving clean WSJCAM0 [42] signals with RIRs measured in a room with the reverberation time  $T_{60} = 0.7$  s, and adding measured stationary ambient noise with SNR (signal-to-noise ratio) of 20 dB. The microphone-to-speaker distance is set to 1 m (*near*) and 2 m (*far*), respectively. RealData was recorded in a noisy room with  $T_{60} = 0.7$  s (different room from *SimData-room3*), where human speakers spoke MC-WSJ-AV [43] utterances at a distance from microphones of 1 m (*near*) and 2.5 m (*far*), respectively. We use the data captured with two-channel (2-ch) or eight-channel (8-ch) circular microphone arrays. In addition to speech enhancement performance, we also test the automatic speech recognition (ASR) performance obtained with the enhanced signals. ASR system provided by [44] is taken as the baseline system, and its Kaldi recipe<sup>2</sup> is used. This speech recognizer uses Mel-frequency cepstral coefficient feature, GMM-HMM (Gaussian mixture model-hidden Markov model) back-end, and trigram language model. In addition, several advanced speech recognition techniques are employed, including linear discriminant analysis, semi-tied covariance matrices, multi-condition training, maximum likelihood linear regression, discriminative training based on maximum mutual information criterion.
- Dynamic dataset [24] was recorded by an eight-channel linear microphone array in a room with  $T_{60} = 0.75$  s.

<sup>2</sup>kaldi-trunk/egs/reverb

TABLE I: SRMR, PESQ and STOI scores for the REVERB challenge dataset.

ch	SRMR						PESQ			STOI		
	SmiData-room3			RealData			SmiData-room3			SmiData-room3		
	near	far	Average	near	far	Average	near	far	Average	near	far	Average
unproc.	2.35	2.29	2.32	2.29	2.20	2.24	1.89	1.55	1.72	0.89	0.71	0.80
AWPE	2-ch	2.61	2.83	2.72	2.98	2.96	2.97	2.30	1.77	2.04	0.78	0.76
	8-ch	2.60	2.89	2.75	3.04	3.01	3.03	2.48	1.90	2.19	0.80	0.79
Prop.	2-ch	2.49	2.64	2.56	2.86	2.79	2.82	2.25	1.75	2.00	0.77	0.73
	8-ch	2.50	2.73	2.61	2.98	2.89	2.93	2.42	1.86	2.14	0.78	0.74

The recording SNR is about 20 dB. The human speakers read an article from the New-York Times. Speakers could be static, or moving slightly, such as standing up, sitting down and head turning, or moving largely such as moving from one point to another. Speakers could be facing or not facing the microphone array. The total length of the dataset is 48 minutes. We split the data into three subsets: i) Static and facing array (Static-FA) with speakers being static (or moving slightly) and facing the microphone array. Note that some slight movements are inevitable even if human speakers are asked to be static; ii) Static and not facing array (Static-NFA), and iii) Moving from one point to another. We use the central two channels (2-ch) or all the eight channels (8-ch).

2) *Parameter Settings*: The following parameter settings are used for both datasets, and all the experimental conditions. The sampling rate is 16 kHz. We set the STFT to use a Hamming window with length of  $N = 768$  (48 ms) and frame step  $L = N/4 = 192$  (16 ms). As a result, the 48 ms early reflections will be preserved in the dereverberated signal. It is shown in [45] that, to achieve a better ASR performance, early reflections should be removed as much as possible when late reverberation is perfectly removed. However, when the remaining late reverberation is not low, ASR performance benefits from preserving more early reflections up to 50 ms. Therefore, as we are aiming at adverse acoustic, such as with intense reverberation/noise or with moving speakers, where late reverberation cannot be perfectly suppressed, we have decided to preserve the early reflection in the first 48 ms. The CTF length  $Q$  (and  $\tilde{Q}$ ) is related to the reverberation time, and is the only prior knowledge that the proposed method requires. It is set to  $Q = 16$  (and  $\tilde{Q} = 4$ ), which covers the major part of the RIRs, and also excludes a heavy tail. According to the CTF length, the forgetting factor  $\lambda$  is set to  $\frac{16-1}{16+1} \approx 0.88$ . The constant step factor  $\mu$  is set to 0.05. The constant lower limit gain factor  $G_{\min}$  is set to correspond to  $-15$  dB. These parameters are set to achieve the best ASR performance for RealData of REVERB challenge dataset, and are directly used for other experimental conditions.

3) *Comparison Method*: We compare the proposed method with the adaptive weighted prediction error (AWPE) method presented in [28]. The STFT uses a Hanning window with a length of 512 (32 ms) and frame step of 128 (8 ms). For the 2-ch and 8-ch cases, the length of the linear prediction filters is set to 14 and 8, respectively. The prediction delay is set to 6 to also involve 48 ms of early reflections in the

dereverberated signal. In RLS, the prediction filter vector to be estimated has the length equalling the length of filters times the number of channels. Some pilot experiments show that, to obtain the optimal performance, the number of frames used to estimate the prediction filter vector should be set to be twice the vector length. Accordingly, the forgetting factor in RLS is set to 0.965 and 0.985 for the 2-ch and 8-ch cases, respectively. The first channel is taken as the target channel. Note that these parameters are also set to achieve the best ASR performance for RealData of REVERB challenge dataset, and are directly used for other experimental conditions.

4) *Performance Metrics*: To evaluate the speech enhancement performance, three measures are used, i) a non-intrusive metric, i.e. normalized speech-to-reverberation modulation energy ratio (SRMR) [46], which mainly measures the amount of reverberation and noise; and two intrusive metrics ii) perceptual evaluation of speech quality (PESQ) [47] evaluates the quality of the enhanced signal in terms of both reverberation reduction and speech distortion; iii) short-time objective intelligibility (STOI) [48] is a metric that highly correlates with speech intelligibility as perceived by humans. To measure PESQ and STOI, the clean source signal is taken as reference signal. For Dynamic dataset, the source signal was recorded by a close-talk microphone. For RealData in REVERB challenge dataset, the clean signal is not available, thus PESQ and STOI are not reported. For all of these three metrics, the higher the better.

For REVERB challenge dataset, the ASR performance is measured by word error rate (WER) in percentage. For the unprocessed signals, the baseline recognizer provided in the Kaldi recipe is used, where the multi-condition training uses the simulated reverberated signals. For AWPE and the proposed method, to fit the desired enhanced signals, we trained a recognizer where the multi-condition training is conducted using the simulated early reverberated signals that are generated by convolving the source signal with the first 48 ms (starting from the direct-path) of the first-channel RIR.

Note that both the proposed and comparison methods do not perform noise reduction, thence the outputs used to calculate the scores may contain some amount of noise.

## B. Results for REVERB Challenge Dataset

In REVERB challenge dataset, each subset involves several hundreds of individual signals, with each signal being one

TABLE II: WER for the REVERB challenge dataset.

ch	SmiData-room3			RealData		
	<i>near</i>	<i>far</i>	Average	<i>near</i>	<i>far</i>	Average
unproc.	11.70	20.40	16.05	32.16	30.18	31.17
AWPE 2-ch	9.14	14.48	11.81	24.31	24.24	24.27
8-ch	8.02	12.08	10.05	22.23	21.37	21.80
Prop. 2-ch	9.95	16.31	13.13	23.70	24.17	23.94
8-ch	8.31	12.35	10.33	21.59	20.43	21.01

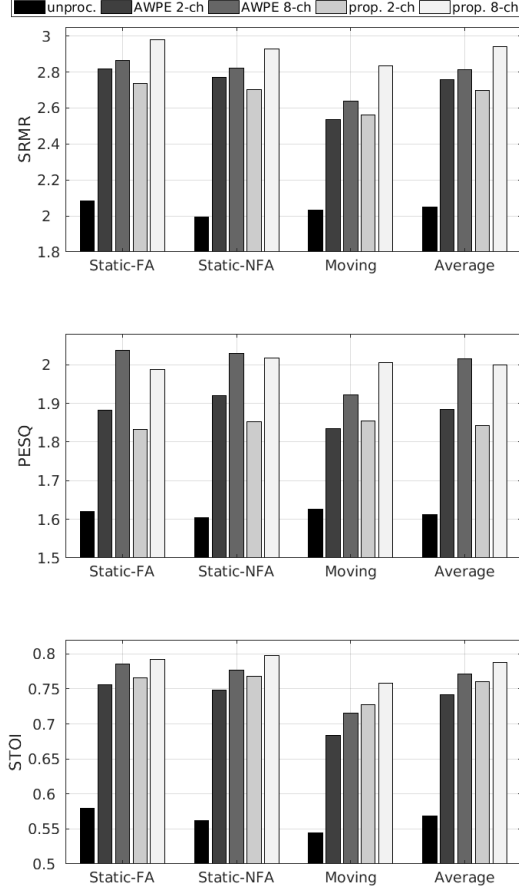


Fig. 1: Dereverberation performance, i.e. SRMR, PESQ and STOI scores (from top to bottom), for Dynamic dataset.

utterance spoken by one static speaker. To simulate a realistic turn-taking scenario, for each subset, all the individual signals are first concatenated as a long signal, which is then processed by the online dereverberation methods, i.e. AWPE and the proposed method. The long enhanced signal is finally separated corresponding to the original individual signals. The performance measures are computed using the individual enhanced signals.

Table I presents the dereverberation results. Both methods noticeably improve the SRMR and PESQ scores compared to the unprocessed signals. For all the conditions, AWPE achieves noticeably better SRMR scores than the proposed method. In terms of PESQ, AWPE still outperforms the proposed method, but the performance gaps between the two methods is not prominent. Relative to the unprocessed signal, the two methods

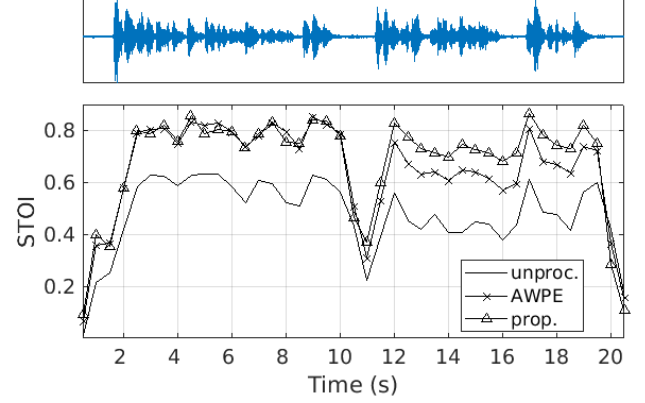


Fig. 2: The short-time STOI scores computed with a 1-s sliding window and 0.5 s sliding step. Eight microphones are used. One speaker was standing at one point within 0-11 s, and started walking to another point from 11 s.

slightly improve the STOI scores for the *far* case, but reduce the STOI scores for the *near* case. This is possibly since the parameters are set based on the RealData data, and in particular the length of the (inverse) filters may be too large for the *near* simulation data. The proposed method is based on the STFT-magnitude convolution and inverse filtering, which is a coarse approximation of the real filtering process. By contrast, AWPE is based on a more accurate complex-valued inverse filtering. As a result, the dereverberated signals obtained with the proposed method are likely to have more late reverberation, extra noise and speech distortions, especially for the 2-ch case. For the 8-ch case, by averaging the source magnitude separately estimated by multiple microphone pairs, the residual late reverberation can be largely removed. Informal listening tests show that the residual late reverberation can be sometimes noticeably perceived for the 2-ch case, while it is not clearly audible for the 8-ch case.

Table II presents the WER. It is seen that the WERs are largely reduced by both methods. For instance, as for RealData, the proposed method achieves 24.5% and 33.7% relative WER improvement with 2-ch and 8-ch, respectively. As for SmiData-room3, AWPE performs noticeably better than the proposed method for the 2-ch case, and slightly better for the 8-ch case. As for RealData, oppositely, the proposed method slightly outperforms AWPE. The possible reasons are three-fold: i) For real recordings, the linear convolution model in the time domain, and thus the complex-valued convolution model in the STFT domain, are less accurate models compared to the case of simulated data, since real recording can deviate from linear convolution. Thence the advantage of complex-valued inverse filtering over STFT magnitude inverse filtering becomes less prominent; ii) Compared to the complex-valued inverse filtering, the STFT magnitude inverse filtering is less sensitive to additive noise, filter perturbations and other unexpected distortions [41]; iii) The remaining late reverberation and extra noise caused by the proposed method may not influence the ASR performance to a large extent.



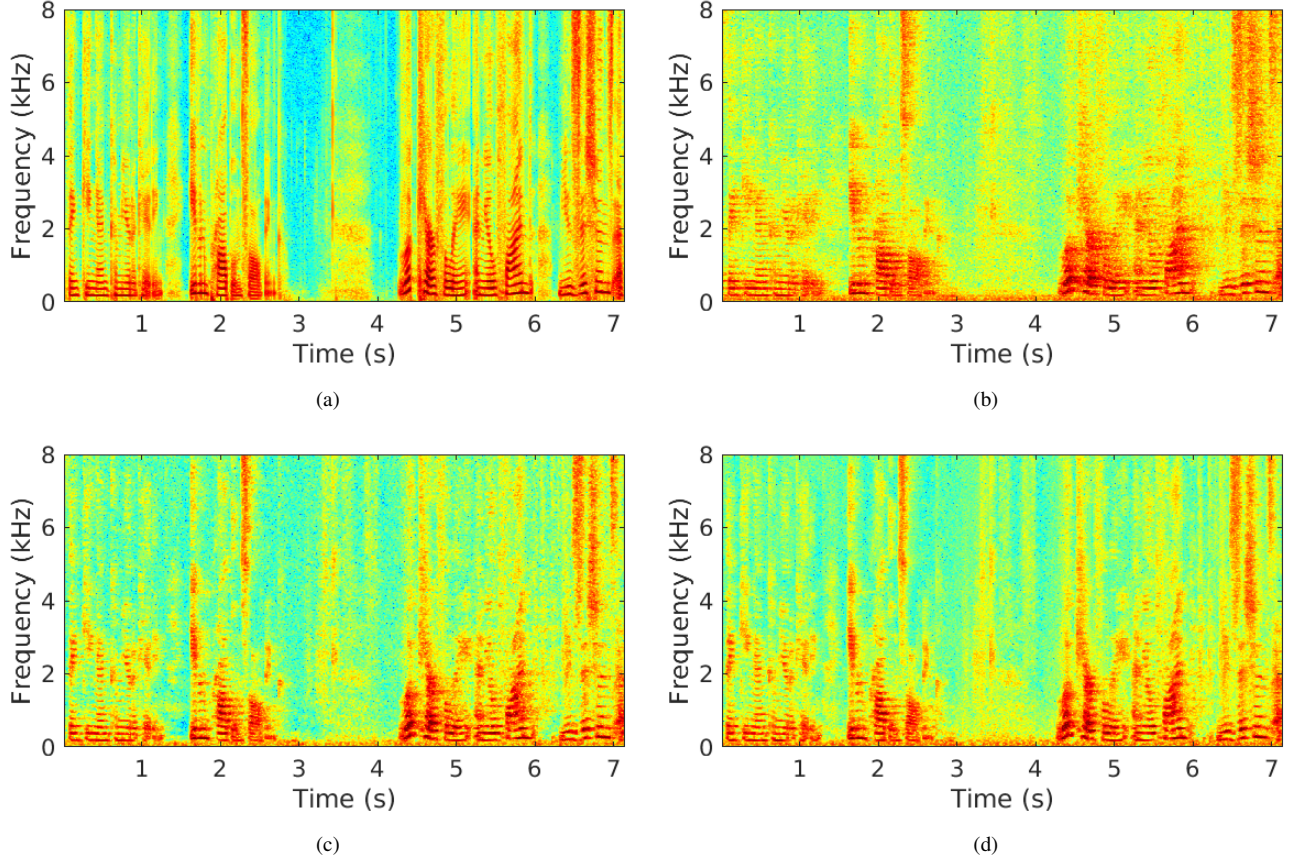


Fig. 3: Spectrogram examples for the Dynamic dataset. (a) close-talk clean signal, (b) microphone signal, (c) 8-ch enhanced signal by AWPE and (d) 8-ch enhanced signal by the proposed method. The speaker was static with in 0-4 s, and started walking from one point to another from 4 s.

### C. Results for Dynamic Dataset

Fig. 1 presents the dereverberation results for the three subsets in Dynamic dataset. For the unprocessed data, all the scores are very low due to the intense reverberation. The Static-NFA set has the lowest SRMR and PESQ scores. When speakers do not face the microphones, the direct-path speech signal received by microphones becomes smaller relative to the reverberation and ambient noise, in other words, the microphone signals are more reverberated and noisy. The Moving case has the lowest STOI scores.

For both methods, the SRMR performance slightly degrades from the Static-FA set to the Static-NFA set, and further noticeably degrade for the Moving set. AWPE achieves larger PESQ scores than the proposed method for the static cases, but has a large performance degradation for the Moving set. By contrast, the proposed method achieves almost the same PESQ scores for the three subsets. The two methods achieve similar STOI scores for the static cases, and lower STOI scores for the Moving set. Overall, compared to AWPE, the proposed method performs noticeably better for the moving speaker case. For the adaptive (inverse) filter estimation, the moving speaker case suffers a larger estimation error compared to the static speaker case, due to the imperfect tracking ability. As already mentioned above, the proposed STFT magnitude inverse fil-

tering method is more robust against filter perturbations than the complex-valued inverse filtering method.

Fig. 2 shows the STOI scores computed with a 1-s sliding window for one audio recording. This result is consistent with Fig. 1 depicting that the two methods have comparable STOI scores when the speaker is static, and the proposed method achieves higher STOI scores when the speaker is moving. When the speaker starts speaking after a silent period, the two methods adapt from background noise to speech, and quickly converge. It is observed from Fig. 2 that the two methods have a similar convergence speed, i.e. less than 1 s. Fig. 3 depicts the spectrograms of the middle part (around the point where the speaker starts moving) of the recording in Fig. 2. It can be seen that reverberation is largely removed by both methods. However, the difference between the two methods and the difference between the static and moving cases cannot be clearly observed from the spectrograms. Informal listening tests show that, the proposed method is not perceived to have more residual reverberation for the moving speaker case compared to the static speaker case. Audio examples for all experiments presented in this paper are available in our website.<sup>3</sup>

<sup>3</sup><https://team.inria.fr/perception/research/ctf-dereverberation>

## V. CONCLUSIONS

In this paper, a blind multichannel online dereverberation method has been proposed. The batch algorithm for multichannel CTF identification proposed in our previous work [22] was extended to an online method based on the RLS criterion. Then, a gradient descent-based adaptive magnitude MINT was proposed to estimate the inverse filters of the identified CTF magnitude. Finally, an estimate of the STFT magnitude of the source signal can be obtained by applying the inverse filtering onto the STFT magnitude of the microphone signals. Experiments were conducted in terms of both speech quality and intelligibility. Compared to the AWPE method, the proposed method achieves comparable ASR performance on the REVERB challenge dataset. Experiments with Dynamic dataset show that the proposed method performs better than AWPE for the moving speaker case due to the robustness of the STFT magnitude-based scheme. Even though the proposed method does not account for noise reduction at all, the dereverberation experiments were performed on data including some additive noise. The experimental results indicate that the dereverberation capability of the proposed method is not significantly deteriorated by the additive noise. However, the noise in the dereverberated signal still has a large influence on both human listening and machine recognition. A noise reduction method that fits well the proposed dereverberation method will be investigated in the future.

## REFERENCES

- [1] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, 2011.
- [2] E. A. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.
- [3] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.
- [4] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1595–1608, 2016.
- [5] O. Schwartz, S. Gannot, E. Habets, *et al.*, "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.
- [6] X. Li, L. Girin, and R. Horaud, "An EM algorithm for audio source separation based on the convolutive transfer function," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.
- [7] S. Mirsamadi and J. H. Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [8] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 276–289, 2016.
- [9] D. Baby and H. Van Hamme, "Joint denoising and dereverberation using exemplar-based sparse representations and decaying norm constraint," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 2024–2035, 2017.
- [10] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, 2007.
- [11] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.
- [12] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.
- [13] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.
- [14] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 2, pp. 145–152, 1988.
- [15] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping-a study," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, pp. V101–V104, 2006.
- [16] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, 2013.
- [17] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.
- [18] X. Li, S. Gannot, L. Girin, and R. Horaud, "Multichannel identification and nonnegative equalization for dereverberation and noise reduction based on convolutive transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1755–1768, 2018.
- [19] S. Weiss, G. W. Rice, and R. W. Stewart, "Multichannel equalization in subbands," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 203–206, 1999.
- [20] N. D. Gaubitch and P. A. Naylor, "Equalization of multichannel acoustic systems in oversampled subbands," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1061–1070, 2009.
- [21] X. Li, S. Gannot, L. Girin, and R. Horaud, "Multisource mint using the convolutive transfer function," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2018.
- [22] X. Li, L. Girin, S. Gannot, and R. Horaud, "Multichannel speech separation and enhancement using the convolutive transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.
- [23] B. Schwartz, S. Gannot, and E. A. Habets, "An online dereverberation algorithm for hearing aids with binaural cues preservation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–5, 2015.
- [24] B. Schwartz, S. Gannot, and E. A. Habets, "Online speech dereverberation using kalman filter and EM algorithm," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 2, pp. 394–406, 2015.
- [25] J.-M. Yang and H.-G. Kang, "Online speech dereverberation algorithm based on adaptive multichannel linear prediction," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 3, pp. 608–619, 2014.
- [26] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3733–3736, 2009.
- [27] T. Yoshioka and T. Nakatani, "Dereverberation for reverberation-robust microphone arrays," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2013.
- [28] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition," in *Proc. Interspeech*, 2017.
- [29] T. Xiang, J. Lu, and K. Chen, "RLS-based adaptive dereverberation tracing abrupt position change of target speaker," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 336–340, 2018.
- [30] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, *et al.*, "Acoustic modeling for google home," *Proc. Interspeech*, pp. 399–403, 2017.
- [31] S. Braun and E. A. Habets, "Online dereverberation for dynamic scenarios using a kalman filter with an autoregressive model," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.
- [32] S. Braun and E. A. Habets, "Linear prediction based online dereverberation and noise reduction using alternating kalman filters," *IEEE/ACM*

- Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1119–1129, 2018.
- [33] Y. Huang and J. Benesty, “A class of frequency-domain adaptive approaches to blind multichannel identification,” *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.
  - [34] W. Zhang, A. W. Khong, and P. A. Naylor, “Adaptive inverse filtering of room acoustics,” in *Asilomar Conference on Signals, Systems and Computers*, pp. 788–792, IEEE, 2008.
  - [35] D. Liu, R. S. Rashobh, A. W. Khong, and M. Yukawa, “A subspace-based adaptive approach for multichannel equalization of room acoustics,” in *Proc. Asia-Pacific Signal and Info. Process. Assoc. Annual Summit and Conf.*, 2011.
  - [36] F. Lim, W. Zhang, E. A. Habets, and P. A. Naylor, “Robust multichannel dereverberation using relaxed multichannel least squares,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, 2014.
  - [37] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*. McGraw-Hill Boston, 2000.
  - [38] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.
  - [39] T. Hikichi, M. Delcroix, and M. Miyoshi, “Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, 2007.
  - [40] A. Mertins, T. Mei, and M. Kallinger, “Room impulse response shortening/reshaping with infinity-and-norm optimization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 249–259, 2010.
  - [41] K. Kinoshita, M. Delcroix, S. Gannot, E. A. 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.
  - [42] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, “Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition,” in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 81–84, 1995.
  - [43] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, “The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 357–362, 2005.
  - [44] F. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J. Geiger, B. Schuller, and G. Rigoll, “The merl/melco/tum system for the reverb challenge using deep recurrent neural network feature enhancement,” in *Proc. REVERB Workshop*, 2014.
  - [45] A. Sehr, E. A. Habets, R. Maas, and W. Kellermann, “Towards a better understanding of the effect of reverberation on speech recognition performance,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2010.
  - [46] J. F. Santos and T. H. Falk, “Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2197–2206, 2014.
  - [47] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, 2001.
  - [48] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.