Joint DOA Estimation and Dereverberation Based on Multi-channel Linear Prediction Filtering and Azimuth Sparsity

Wenmeng Xiong, Changchun Bao, Jing Zhou, Maoshen Jia, José Picheral

Joint DOA Estimation and Dereverberation Based on Multi-channel Linear Prediction Filtering and Azimuth Sparsity

Wenmeng XIONG, Changchun BAO, *Senior Member, IEEE*, Jing ZHOU, Maoshen JIA, *Senior Member, IEEE*, and
José PICHERAL

***Abstract*—Source localization in reverberant environments has been a prominent research topic in the past two decades. In this paper, instead of the commonly employed time-frequency (TF) bin based methods which rely on empirically selected threshold values, we leverage the microphone array signal model comprising an early reverberant component and a late reverberant component, to propose a novel method for the source localization problem in reverberant environments. Our proposed criterion involves the joint removal of the late reverberant component using the multi-channel linear prediction (MCLP) filter, while estimating the directions of arrival (DOAs) of the actual sources using the early component signals. By applying the azimuth sparsity constraint, the true DOA can be estimated with high resolution and free from the interference of the early reflections. To solve the proposed criterion, DOAs, source signals, and MCLP filter coefficients are estimated by alternative iterations. Additionally, we present a source localization criterion specifically designed for the single source scenario as a special case of the multiple sources scenario. Finally, a source number estimation method and a postprocessing procedure are discussed for searching the global solutions to our proposed criteria. Evaluations with both simulated and realistic data demonstrate the advantages of our proposed methods over the baseline methods.**

***Index Terms*—DOA estimation, MCLP, reverberation, sparsity, PALM**

I. Introduction

Multiple sources localization problem is of great importance for numerous applications such as speech enhancement, source separation, and acoustic imaging. The conventional source localization methods, including the steered response power with phase transform (SRP-PHAT)/beamforming [1], the multiple signal classification (MUSIC) [2], and the maximum likelihood [3], have been widely investigated and proven efficient in real applications. However, the performance of these methods degrades severely in environments with high noise and reverberation.

To address the challenge of reverberation, many methods have been developed based on the multi-channel linear prediction (MCLP) filter in time domain or in short time Fourier transform (STFT) domain. The STFT-based approach is adequate for wideband signals due to its parallel implementation at each frequency bin [4][5]. Early signals can be obtained by removing the delayed microphone signal components of earlier frame indices filtered with the MCLP filter, where the coefficients of filters are to be estimated on-line [6][7][8] or off-line [4][5]. The early signals can be obtained for one single microphone or all the microphones in the array [4]. In [5], the generalized weight prediction error (GWPE) method is proposed, modelling the early signals as time-dependent complex Gaussian and solving a cost function based on the correlation between the early signals iteratively. Various efforts have been made to improve the performance of the MCLP filter. Instead of Gaussian distribution assumption, a sharper distribution such as Lapacian [9] or a distribution with a heavier tail such as Student's t-dsitribution [10] have been found more appropriate for describing the statical characteristics of the information bearing signals including speech and music signals. The order of the prediction filter is assumed to be known *a-priori*. However, a too large order leads to over-estimate the reverberation and distort the desired signals, while a too small order brings more residual reverberation. To solve this problem, the MCLP filter coefficients are modelled with a prior Gaussian distribution in [10], and a statistical model for late reverberation is proposed to avoid the over-estimation of the undesired late reverberant component in [11]. As demonstrated by[12], the MCLP method and the beamforming methods are more suitable for predominant late-reverberant environments and predominant noise environments, respectively. Combinations of beamforming and MCLP filter have been introduced for dereverberation in noisy environments [13][14][15]. To achieve high-efficiency dereverberation in noisy environments, the MCLP can either work alone [16] or combine with a spatial filter [17][18][19] in the Kronecker product array framework, or combine with the generalized sidelobe cancellation filter in the Kalman filter framework [20]. In [21], a deep neural network is trained to predict the direct sound from noisy-reverberant speech, rather than exploiting the linear filter structure. To the best of our knowledge, few work on the MCLP filter has been proposed for source localization in reverberant environments [22].

The commonly used source localization methods in reverberant environments are based on the W-disjoint orthogonal (WDO) assumption [23] that at most one source is dominant

This work was supported by the National Natural Science Foundation of China (Grant No.62101013, No.61831019, No.61971015, and No.62001012).

Wenmeng XIONG, Changchun BAO (corresponding author), Jing ZHOU and Maoshen JIA are with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: wenmeng.xiong@bjut.edu.cn; chchbao@bjut.edu.cn; jiamaoshen@bjut.edu.cn)

José PICHERAL is with Laboratoire des signaux et systèmes (Signal and System Laboratory), Université Paris-Saclay, CNRS, CentraleSupélec, 91190, Gif-sur-Yvette, France (email: Jose.Picheral@centralesupelec.fr)

with respect to other sources at one time-frequency (TF) bin. Extensive research has focused on selecting such bins to enhance the robustness of the source localization methods against reverberations. The single source regions (SSRs) and low reverberant single source (LRSS) regions are selected by coherent tests in [24][25] [26]. Single source confidence measure is proposed in [27]. The direct path dominance (DPD) tests are based on the ratio between the largest and second largest eigenvalues [28] or the dominant eigenvector [29] of the covariance matrix of the microphone signals, or the sound filed directivity [27]. Compared with the planar and spherical microphone arrays, acoustic vector sensors (AVS) for source localization [26][30] require less space and computation, as the bins are selected by comparing the phases of the real and imaginary parts of the microphone signals without eigenvalue decomposition[28]. Further enhancement have been achieved through outlier removal [31], modelling outliers as an extra Gaussian cluster [32], generelizing single source dominance TF bin to single source dominance window [33], and deconvolution of the global pseudo-spectrum by iteratively source detection [34] for the sources with small separations. All these TF bin selection based methods require empirically chosen threshold values for different experiment configurations. A too large threshold leads to miss detection of the sources, while a too small threshold brings more outliers.

In our paper, in order to avoid the need for empirical threshold selection, we leverage the signal model composed of the early reverberant component and the late reverberant component. Specifically, the algorithm is designed for multiple sources scenario, where the late reverberant component is removed with the MCLP filter and the directions of arrival (DOAs) are estimated jointly with the early component signals. To account for the assumptions that the sources are unoverlapped, finite in power, and their direction angles occupy only a fairly small part of the whole angular section of interest, sparsity in the l_1 norm and orthonormalization constraints are imposed on the DOA estimation matrix. By applying the azimuth sparsity constraint, the true DOA can be estimated without the interference of the early reflections. Besides, considering the sparsity of the TF spectrum of information bearing signals, the l_1 norm constraints are imposed both on the source signals and the dereverberated signals. To solve the proposed criterion, the MCLP filter, the DOAs, and the source signals are estimated alternately. Notably, proximal alternating linearizations are introduced for estimating the coupling parameters with non-smooth constraints [35], and the orthogonal Procruste problem framework [36] is employed to deal with the orthonormalization constraint. In the following, we propose a DOA estimation algorithm for the single source scenario, where the orthonormalization constraint on the DOA estimation matrix is replaced by the constraint on the norm of the DOA estimation vector. This algorithm can be considered a special case of the multiple sources localization problem. At last, the source number estimation method and a post-processing procedure for searching the global solution to the proposed algorithm are discussed.

The paper is organized as follows: in section II, the microphone array signals model in reverberant environments and the MCLP filter are briefly reviewed. To clarify the motivation of our work, the disadvantages of the cascade of the MCLP filter and the conventional MUSIC are also outlined. In section III and IV, the localization algorithms are proposed for multiple and single source scenarios, respectively. In section V, pre and post-processing procedures for better performance are discussed. Evaluations on both simulated and real-world data illustrate the advantage of our proposed algorithm in section VI. Finally, conclusions are given in section VII.

II. Preliminaries

A. Signal model

Considering that Q far-field wideband acoustic sources $s_{\mathfrak{t}_q}, q = 1, ..., Q$ impinge on M microphones in a reverberant room, the time domain signals received by the microphone array at time t can be given as:

$$\mathbf{x}_{\mathfrak{t}}(t) = \sum_{q=1}^{Q} \left(\mathbf{h}_{\mathfrak{t}_q} * s_{\mathfrak{t}_q}\right)(t) + \mathbf{b}_{\mathfrak{t}}(t), \tag{1}$$

where $\mathbf{h}_{\mathfrak{t}_q}(t) = \left[h_{\mathfrak{t}_{q,1}}, ..., h_{\mathfrak{t}_{q,M}}\right]^T \in \mathbb{C}^{M\times 1}$ denotes the room impulse response vector from source q to the microphone array at time index t, $*$ denotes the convolution operator, $\mathbf{b}_{\mathfrak{t}}(t)$ denotes an additive noise at the microphone array at time t with variance σ_b^2. In reverberant environments, the signals received at the microphone array in the TF domain are approximately formulated as a convolution of the STFT of the room impulse response $\mathbf{h}_q(n,\omega)$ of order K and the STFT of the signals $s_q(n,\omega)$ along the time frame axis for each frequency bin [37], since the length of the room impulse response is generally much larger than that of the time frame of the STFT. The microphone array signals in the TF domain can be given as:

$$\mathbf{x}(n,\omega) = \sum_{q=1}^{Q}\sum_{k'=0}^{K} \mathbf{h}_q(k',\omega)s_q(n-k',\omega) + \mathbf{b}(n,\omega) \tag{2}$$

where $n = 1, ..., N$ and ω denote the time frame and frequency bin indices, respectively. $s_q(n,\omega)$ and $\mathbf{b}(n,\omega)$ indicate the STFT coefficients of the q^{th} source and the additive noise at the microphone array, respectively.

The microphone array signals $\mathbf{x}(n,\omega)$ can be reformulated as the sum of the early and late reverberant components and the noise:

$$\mathbf{x}(n,\omega) = \mathbf{x}_e(n,\omega) + \mathbf{x}_l(n,\omega) + \mathbf{b}(n,\omega), \tag{3}$$

where $\mathbf{x}_e(n,\omega) = \sum_{q=1}^{Q}\sum_{k'=0}^{k_e} \mathbf{h}_q(k',\omega)s_q(n-k',\omega)$ is defined as the early reverberant component which contains mainly the direct-path and few early reflections of the room impulse response, and $\mathbf{x}_l(n,\omega) = \sum_{q=1}^{Q}\sum_{k'=k_e}^{K} \mathbf{h}_q(k',\omega)s_q(n-k',\omega)$ is defined as the late reverberant component which contains all other reflections of the room impulse response in the TF domain. The value of k_e varies from 0 to 4 according to experimental configurations. In our work, we choose $k_e = 0$ in order to minimize the influence

of the early reflections on the DOA estimation performance. The details of $\mathbf{x}_e(n,\omega)$ can be given as:

$$\mathbf{x}_e(n,\omega) = \sum_{q=1}^{Q}\left(\mathbf{a}(\theta_q,\omega) + \sum_{q'=1}^{Q'}\gamma_{q,q'}\mathbf{a}(\theta_{q,q'},\omega)\right)s_q(n,\omega), \tag{4}$$

where $\mathbf{a}(\theta_q,\omega)$ is called the steering vector of a point source from direction θ_q such that $\mathbf{a}(\theta_q,\omega) = \left[1, e^{-j\omega\Delta t_1(\theta_q)}, ..., e^{-j\omega\Delta t_{M-1}(\theta_q)}\right] \in \mathbb{C}^{M\times 1}$, when the first microphone is taken as the reference, $\Delta t_m(\theta_q)$ is the time delay from the m^{th} microphone to the reference microphone. The term $\mathbf{a}(\theta_q,\omega)$ in (4) denotes the direct-path signals from the q^{th} sources, and $\mathbf{a}(\theta_{q,q'},\omega)$ in (4) denote the q' early reflection from the q^{th} source. The q'^{th} early reflection of the q^{th} source can be considered as sources correlated with s_q but from DOA $\theta_{q,q'}$ which is different from the original DOA θ_q with an attenuation factor $\gamma_{q,q'}$ due to the absorption and time delay during the reflections. In our work, we conduct the experiments with speech signals, in which case the length of the time frame is 20 to 30 ms according to the short time stationary property, the early reverberant component Q' is generally no more than 1.

B. The MCLP filter and the GWPE method

The generalized weighted prediction error (GWPE) method [5] reduces the late reverberation component correlated to the past time frames with a MCLP filter matrix for each current time frame of the microphone array signals in the TF domain. Defining $\mathbf{G}(\omega) = [\mathbf{g}_1, ..., \mathbf{g}_m, ..., \mathbf{g}_M] \in \mathbb{C}^{K_l M\times M}$ as the MCLP filter matrix for frequency bin ω, where K_l gives the prediction order, $\mathbf{g}_m$ are the MCLP filters, the dereverberated speech components can be estimated as:

$$\hat{\mathbf{y}}(n,\omega) = \mathbf{x}(n,\omega) - \mathbf{G}^H(\omega)\tilde{\mathbf{x}}(n,\omega), \tag{5}$$

where $\tilde{\mathbf{x}}(n,\omega) = \left[\mathbf{x}^T(n-\Delta-1,\omega), ..., \mathbf{x}^T(n-\Delta-K_l,\omega)\right]^T$, Δ denotes the delay tap index of the beginning frame of the late reverberation component. Assuming that the estimated speech component $\hat{\mathbf{y}}(n,\omega)$ follows a complex multivariate Gaussian distribution such that $\hat{\mathbf{y}}(n,\omega) \sim \mathcal{N}_{\mathbb{C}}(0,\mathbf{\Lambda}_n)$, $n = 1, ..., N$, where $\mathbf{\Lambda}_n$ denotes the time varying covariance matrix of $\hat{\mathbf{y}}(n,\omega)$, the log-likelihood function can be given as:

$$\mathcal{L}(\mathbf{\Theta}_\omega) = \sum_{n=1}^{N} -\log\left(\det(\mathbf{\Lambda}_n)\right) - \hat{\mathbf{y}}^H(n,\omega)\mathbf{\Lambda}_n^{-1}\hat{\mathbf{y}}(n,\omega) + \text{const}, \tag{6}$$

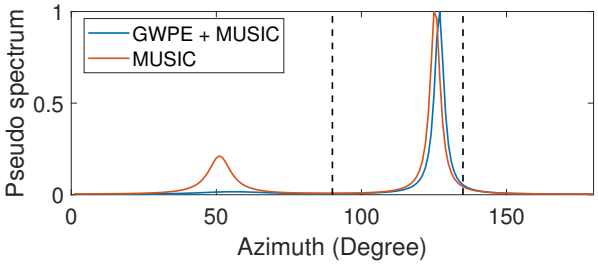
where $\mathbf{\Theta}_\omega = \left\{\mathbf{G}^H(\omega), \mathbf{\Lambda}_n\right\}$ is the set of unknown parameters to be estimated. (6) can be maximized by an alternating two-step scheme: In the first step, (6) is maximized with respect to $\mathbf{G}^H(\omega)$ while $\mathbf{\Lambda}_n$ is kept fixed:

$$\hat{\mathbf{G}}(\omega) = \arg\max_{\mathbf{G}(\omega)} \sum_{n=1}^{N} -\hat{\mathbf{y}}^H(n,\omega)\mathbf{\Lambda}_n^{-1}\hat{\mathbf{y}}(n,\omega), \tag{7}$$

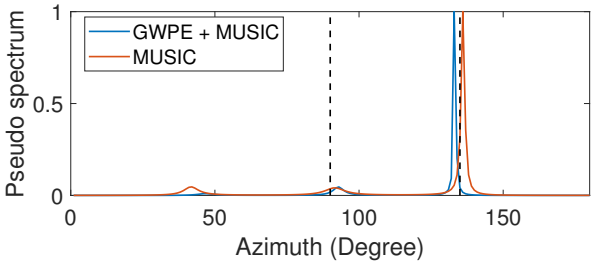
and in the second step, (6) is maximized with respect to $\mathbf{\Lambda}_n$ while $\mathbf{G}^H(\omega)$ is kept fixed:

$$\hat{\mathbf{\Lambda}}_n = \arg\max_{\mathbf{\Lambda}_n} \sum_{n=1}^{N} -\log\left(\det(\mathbf{\Lambda}_n)\right) - \hat{\mathbf{y}}^H(n,\omega)\mathbf{\Lambda}_n^{-1}\hat{\mathbf{y}}(n,\omega). \tag{8}$$

The algorithm can converge in a few iterations.



(a) $\hat{Q} = 2$



(b) $\hat{Q} = 3$

Fig. 1. Pseudo-spectrum of MUSIC with original and dereverberated microphone array signals (sources from $\theta_1 = 90°$ and $\theta_2 = 135°$).

C. Combined GWPE and conventional source localization method

One approach we can try for the source localization problem in reverberant environments is to combine a conventional source localization problem with the GWPE method. In this approach, the source localization method is applied to the output $\hat{\mathbf{y}}(n,\omega)$ of the GWPE method. In our experiments, the sources are located with the famous MUSIC algorithm [38]. With the number of source Q assumed to be known, the DOAs are estimated by searching the corresponding steering vectors which are the most orthogonal to the noise subspace such that:

$$\hat{\theta}_q = \arg\max_{\theta} \frac{1}{\|\mathbf{a}^H(\theta,\omega)\mathbf{V}\|_2^2}, \tag{9}$$

where the matrix $\mathbf{V}$ is the basis of the noise subspace of the covariance matrix of the microphone array signals, $\|\cdot\|_2$ is the l_2 norm operator.

Figure 1 shows the MUSIC pseudo-spectrum with the dereverberated speech components $\hat{\mathbf{y}}(n,\omega)$ (legend: "GWPE+MUSIC") and with the original microphone array signals $\mathbf{x}(n,\omega)$ (legend:"MUSIC"), respectively. Room impulse responses were generated using [39] for an $8\text{m}\times 8\text{m}\times 3\text{m}$ room with a linear microphone array composed of $M = 8$ microphones spaced by 5cm. T_{60} is set to 600 ms and the additive noise is not considered in this simulation. Two incoherent speech sources come from the DOA $90°$ and $135°$,

respectively. The assumed source number for MUSIC is set to $\hat{Q} = 2$ in Fig. 1(a) and $\hat{Q} = 3$ in Fig. 1(b). It can be observed that the conventional MUSIC identifies one peak around $50°$ both in Fig. 1(a) and 1(b) due to reflections. However, by combining the GWPE and MUSIC, this false peak is substantially eliminated and only a small residual peak remains around $50°$, since in (4) the early reverberant component contains also a few early reflections. The source from the DOA $90°$ is omitted in the pseudo-spectra in Fig. 1(a) for both methods, while successfully located with a minor bias in Fig. 1(b) where $\hat{Q}$ is larger than the true source number. The simulation results highlight that the simple combination of the dereverberation method and the DOA estimation method is insufficient for the source localization problem in the reverberant environments.

III. The Proposed Algorithm in Multiple Sources Scenario

In section III and IV, the joint dereverberation and DOA estimation algorithm are investigated for both multiple sources and single source scenarios. For the sake of simplicity, all the algorithms are studied in the same sub-frequency bin in the STFT domain, ω is omitted in the parameters in the following of this paper.

A. The cost function

Assuming that the whole angular sector of interest is discretized by L grid point such that $\phi_k = \phi_1 + (k-1)\delta, k = 1, ..., L$, where δ is the angular separation between two points of the grid. A matrix $\boldsymbol{\alpha} \in \mathbb{C}^{L\times Q}$ can be defined such that the non-zero elements in the q^{th} column of $\boldsymbol{\alpha}$ corresponds to the DOAs of the direct signals and early reflections of the q^{th} source. We also define $\mathbf{H} = [\mathbf{a}(\phi_1), ..., \mathbf{a}(\phi_k), ..., \mathbf{a}(\phi_L)]$ as the matrix composed of the steering vectors corresponding to all the ϕ_k. Therefore, according to (4), a grid model for the early reverberant component can be given as:

$$\mathbf{x}_e(n) \approx \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n), \tag{10}$$

where the elements in the vector $\tilde{\mathbf{s}}(n) = [\tilde{s}_1(n), ..., \tilde{s}_Q(n)]^T$ are linearly proportional to the actual source signals $\mathbf{s}(n) = [s_1(n), ..., s_Q(n)]^T \in \mathbb{C}^{Q\times 1}$ in the TF domain.

The proposed DOA estimation criterion involves an optimization problem with three constraints:

1) The data fitting term: According to (3), we aim to minimize the error between the actual received microphone array signal $\mathbf{x}(n)$ and the signal modelled as the sum of the early and late reverberant components. The early components are given by (10) and the late components are given by $\mathbf{G}^H\tilde{\mathbf{x}}(n)$ obtained in section II-B. The data fitting term can be given as:

$$\mathcal{C}_1 = \sum_{n=1}^{N} \|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)\|_2^2, \tag{11}$$

It is worth noting that (11) is also an approximation of the additive noise $\mathbf{b}(n, \omega)$ in (3). Since the power of the noise is independent with the time index, we omit the covariance matrix of the noise, which is different from (7).

2) Super-Gaussian constraint for the early reverberant components and source signals: Opposed to the Gaussian distribution assumption, previous works [9][10][40][41] have illustrated that the super-Gaussian distribution model is more appropriate for information bearing signals such as speech signals. The optimization problem for the super-Gaussian distributed signals generally involves the minimization of l_p norm of the targets with $0 < p < 2$. Here, we introduce the l_1 norm of the dereverberated speech components $\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n)$ and the Q original sources $\tilde{\mathbf{s}}(n)$:

$$\mathcal{C}_2 = \lambda_{\mathbf{z}_1} \sum_{n=1}^{N} \|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n)\|_1 + \lambda_{\tilde{\mathbf{s}}} \sum_{n=1}^{N} \|\tilde{\mathbf{s}}(n)\|_1, \tag{12}$$

where $\lambda_{\mathbf{z}1}$ and $\lambda_{\tilde{\mathbf{s}}}$ are positive sparsity penalization parameters, $\|\cdot\|_1$ denotes the l_1 norm operator. By minimizing (12), we can exploit the super Gaussian distribution property as well as the sparsity of their STFT spectra of the signals.

3) Azimuth sparsity constraint: The l_1 norm of $\boldsymbol{\alpha}$ is minimized by exploiting the spatial sparsity property of the DOAs of the sources. Indeed, our simulations in section VI have shown that if the sparsity weighting parameter is large enough, the elements corresponding to the early reflections diminish, leaving only those corresponding to the direct signals. The estimated DOAs are obtained from the indices of the non-zero elements in each column of $\boldsymbol{\alpha}$. Besides, we insert the constraint $\boldsymbol{\alpha}^H\boldsymbol{\alpha} = \mathbf{I}$ under the assumption that the sources are not spatially overlapped and have finite power:

$$\mathcal{C}_3 = \|\boldsymbol{\alpha}\|_1, \quad \text{s.t.} \quad \boldsymbol{\alpha}^H\boldsymbol{\alpha} = \mathbf{I}. \tag{13}$$

It is important to note that the normalization constraint on the diagonal of $\boldsymbol{\alpha}^H\boldsymbol{\alpha}$ aims to place the proposed criterion in the framework of the orthogonal Procrustes problem [36] in order to theoretically guarantee convergence. For satisfying the constraint on $\boldsymbol{\alpha}$, the ratio between $\tilde{s}_q(n)$ and $s_q(n)$ varies with the power of the q^{th} source.

Combining $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ in (11)-(13), the parameters for the source localization problem in reverberant environments can be estimated by:

$$\begin{aligned}\left\{\hat{\mathbf{G}}, \hat{\boldsymbol{\alpha}}, \hat{\tilde{\mathbf{s}}}(n)\right\} = &\underset{\mathbf{G},\boldsymbol{\alpha},\tilde{\mathbf{s}}(n)}{\arg\min} \sum_{n=1}^{N} \|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)\|_2^2 \\ &+ \lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1 + \lambda_{\mathbf{z}_1} \sum_{n=1}^{N} \|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n)\|_1 + \lambda_{\tilde{\mathbf{s}}} \sum_{n=1}^{N} \|\tilde{\mathbf{s}}(n)\|_1, \\ &\text{s.t.} \quad \boldsymbol{\alpha}^H\boldsymbol{\alpha} = \mathbf{I},\end{aligned} \tag{14}$$

where $\lambda_{\boldsymbol{\alpha}}$ is the sparsity penalization parameter.

B. Solutions for the cost function

Equation (14) can be solved via Gauss-Seidel iteration scheme [42], the solution converges under the condition that the minimum is uniquely attained in each step [43]. Since the cost function consists of coupling second-order terms and non-smooth constraints, the linearized Gauss-Seidel method, namely the Proximal Alternating Linearized Minimization (PALM) method [35] is utilised to alternately estimate $\mathbf{G}$, $\boldsymbol{\alpha}$, and $\tilde{\mathbf{s}}(n)$ until convergence is reached. All the three parameters

are resolved by sub-iterations rather than simple closed-form expressions. For simplicity, only the indices in sub-iterations are kept and those in the top level iteration are omitted. The alternative iterations are as follows:

1) Solve $\mathbf{G}$:

In the branch for estimating $\mathbf{G}$, $\boldsymbol{\alpha}$ and $\tilde{\mathbf{s}}(n)$ are assumed constant, and equation (14) is simplified as:

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} \sum_{n=1}^{N} \|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)\|_2^2 + \lambda_{\mathbf{z}_1}\sum_{n=1}^{N}\|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n)\|_1. \tag{15}$$

By introducing an extra variable $\mathbf{z}_1(n)$, (15) can be reformulated as:

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} \sum_{n=1}^{N} \|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)\|_2^2 + \lambda_{\mathbf{z}_1}\sum_{n=1}^{N}\|\mathbf{z}_1(n)\|_1, \text{ s.t. } \mathbf{z}_1(n) = \mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n). \tag{16}$$

The augmented Lagrangian of (16) can be given as:

$$\mathcal{L}(\mathbf{G}, \mathbf{z}_1(n), \boldsymbol{\eta}_{\mathbf{G}}(n)) = \sum_{n=1}^{N}\Big(\|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)\|_2^2 + \lambda_{\mathbf{z}_1}\|\mathbf{z}_1(n)\|_1 + \mathcal{R}e\left\{\boldsymbol{\eta}_{\mathbf{G}}^H(n)(\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{z}_1(n))\right\} + \frac{1}{2\rho_{\mathbf{G}}}\|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{z}_1(n)\|_2^2\Big), \tag{17}$$

where $\mathcal{R}e\{\cdot\}$ is the real part operator, $\rho_{\mathbf{G}}$ is the penalization parameter of the convex term.

The problem (17) can be solved via some iterative steps, in the $(l+1)^{th}$ iteration, the closed-form solutions for $\mathbf{G}$ and $\mathbf{z}_1$ are given as:

$$\mathbf{G}^{(l+1)} = \left(\sum_{n=1}^{N}\left(1 + \frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^H(n)\right)^{-1} \cdot \sum_{n=1}^{N}\left(\left(1 + \frac{1}{2\rho_{\mathbf{G}}}\right)\tilde{\mathbf{x}}(n)\mathbf{x}^H(n) + \frac{1}{2}\tilde{\mathbf{x}}(n)\boldsymbol{\eta}_{\mathbf{G}}^{(l)H}(n) - \tilde{\mathbf{x}}(n)\tilde{\mathbf{s}}^H(n)\boldsymbol{\alpha}^H\mathbf{H}^H - \frac{1}{2\rho_{\mathbf{G}}}\tilde{\mathbf{x}}(n)\mathbf{z}_1^{(l)H}(n)\right), \tag{18}$$

and:

$$\mathbf{z}_1^{(l+1)}(n) = \mathcal{S}_{\lambda_{\mathbf{z}_1}/\mu_{\mathbf{z}_1}}\left(\mathbf{z}^{(l)} - \frac{1}{\mu_{\mathbf{z}_1}}\nabla_{\mathbf{z}}V(\mathbf{G}^{(l+1)}, \mathbf{z}_1^{(l)}, \boldsymbol{\eta}_{\mathbf{G}}^{(l)}, n)\right), \tag{19}$$

where $\cdot^{-1}$ is the inverse operator of a matrix, and $\mathcal{S}_{\lambda_{\mathbf{z}_1}/\mu_{\mathbf{z}_1}}(\mathbf{v})$ is the soft thresholding operator of the vector $\mathbf{v}$ such that:

$$\mathcal{S}_{\lambda_{\mathbf{z}_1}/\mu_{\mathbf{z}_1}}(\mathbf{v}) = \begin{cases} \mathbf{v} - \lambda_{\mathbf{z}_1}/\mu_{\mathbf{z}_1}, & \text{if } \mathbf{v} \geq \lambda_{\mathbf{z}_1}/\mu_{\mathbf{z}_1}, \\ \mathbf{v} + \lambda_{\mathbf{z}_1}/\mu_{\mathbf{z}_1}, & \text{if } \mathbf{v} \leq -\lambda_{\mathbf{z}_1}/\mu_{\mathbf{z}_1}, \\ 0, & \text{otherwise,} \end{cases} \tag{20}$$

where the comparisons and subtractions are applied element-wise, and the comparisons are with the absolute values of the elements in $\mathbf{v}$.

The details for obtaining (18) and (19) can be found in Appendix A.

2) Solve $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{s}}(n)$:

In the branch for estimating $\boldsymbol{\alpha}$ and $\tilde{\mathbf{s}}(n)$, $\mathbf{G}$ is assumed constant. The cost function is constructed by introducing $\hat{\mathbf{y}}(n)$ given by (5) into (14):

$$\left\{\hat{\boldsymbol{\alpha}}, \hat{\tilde{\mathbf{s}}}(n)\right\} = \arg\min_{\boldsymbol{\alpha}, \tilde{\mathbf{s}}(n)} \sum_{n=1}^{N}\|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)\|_2^2 + \lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1 + \lambda_{\tilde{\mathbf{s}}}\sum_{n=1}^{N}\|\tilde{\mathbf{s}}(n)\|_1 + \iota(\boldsymbol{\alpha}), \tag{21}$$

where $\iota(\boldsymbol{\alpha})$ is an indicator function such that:

$$\iota(\boldsymbol{\alpha}) = \begin{cases} 0, & \text{if } \boldsymbol{\alpha}^H\boldsymbol{\alpha} = \mathbf{I}, \\ +\infty, & \text{otherwise.} \end{cases} \tag{22}$$

The problem (21) consists of a coupling function $\mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)$ as well as non-smooth functions $\|\boldsymbol{\alpha}\|_1$ and $\|\tilde{\mathbf{s}}(n)\|_1$, which can also be solved by the linearized Gauss-Seidel iteration scheme. Proximal terms are introduced to assure convexity, as the problem includes a non-convex constraint in $\iota(\boldsymbol{\alpha})$. In the $(l+1)^{th}$ iteration, the parameters can be estimated as:

$$\tilde{\mathbf{s}}^{(l+1)}(n) = \arg\min_{\tilde{\mathbf{s}}(n)} \sum_{n=1}^{N}\Big(\|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}^{(l)}\tilde{\mathbf{s}}(n)\|_2^2 + \lambda_{\tilde{\mathbf{s}}}\|\tilde{\mathbf{s}}(n)\|_1 + \frac{d_{\mathbf{s}}}{2}\|\tilde{\mathbf{s}}(n) - \tilde{\mathbf{s}}^{(l)}(n)\|_2^2\Big), \tag{23a}$$

$$\boldsymbol{\alpha}^{(l+1)} = \arg\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N}\|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}^{(l+1)}(n)\|_2^2 + \lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1 + \iota(\boldsymbol{\alpha}) + \frac{d_{\boldsymbol{\alpha}}}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(l)}\|_2^2, \tag{23b}$$

where $d_{\mathbf{s}}$ and $d_{\boldsymbol{\alpha}}$ are penalization parameters of convex terms.

Similarly to (40b) in appendix A, defining that $B(\tilde{\mathbf{s}}^{(l)}(n), \boldsymbol{\alpha}^{(l)}) = \|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}^{(l)}\tilde{\mathbf{s}}^{(l)}(n)\|_2^2$, and the gradient of $B(\tilde{\mathbf{s}}^{(l)}(n), \boldsymbol{\alpha}^{(l)})$ in terms of $\tilde{\mathbf{s}}^{\star(l)}(n)$ is $\nabla_{\tilde{\mathbf{s}}}B(\tilde{\mathbf{s}}^{(l)}(n), \boldsymbol{\alpha}^{(l)}) = -\boldsymbol{\alpha}^{(l)H}\mathbf{H}^H\hat{\mathbf{y}} + \boldsymbol{\alpha}^{(l)H}\mathbf{H}^H\mathbf{H}\boldsymbol{\alpha}^{(l)}\tilde{\mathbf{s}}^{(l)}(n)$, the closed-form solution of (23a) can be given as:

$$\tilde{\mathbf{s}}^{(l+1)}(n) = \mathcal{S}_{\lambda_{\tilde{\mathbf{s}}}/d_{\tilde{\mathbf{s}}}}\left(\tilde{\mathbf{s}}^{(l)} - \frac{1}{d_{\tilde{\mathbf{s}}}}\nabla_{\tilde{\mathbf{s}}}B(\tilde{\mathbf{s}}^{(l)}(n), \boldsymbol{\alpha}^{(l)})\right). \tag{24}$$

To solve (23b), we introduce an extra variable $\mathbf{z}_2 \in \mathbb{C}^{180Q\times 1}$ to solve the non-smooth functions separately:

$$\boldsymbol{\alpha}^{(l+1)} = \arg\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N}\|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}^{(l+1)}(n)\|_2^2 + \lambda_{\boldsymbol{\alpha}}\|\mathbf{z}_2\|_1 + \iota(\boldsymbol{\alpha}) + \frac{d_{\boldsymbol{\alpha}}}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(l)}\|_2^2, \quad \text{s.t. } \mathbf{z}_2 = vec(\boldsymbol{\alpha}), \tag{25}$$

where $vec(\cdot)$ is the vectorization operator of a matrix. The Lagrangian of (25) can be given as:

$$\mathcal{L}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^{(l)}, \mathbf{z}_2, \boldsymbol{\eta}_{\boldsymbol{\alpha}}) = \sum_{n=1}^{N}\|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{\mathbf{s}}^{(l+1)}(n)\|_2^2 + \iota(\boldsymbol{\alpha}) + \lambda_{\boldsymbol{\alpha}}\|\mathbf{z}_2\|_1 + \frac{d_{\boldsymbol{\alpha}}}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(l)}\|_2^2 + \mathcal{R}e\left\{\boldsymbol{\eta}_{\boldsymbol{\alpha}}^H(vec(\boldsymbol{\alpha}) - \mathbf{z}_2)\right\} + \frac{1}{2\rho_{\boldsymbol{\alpha}}}\|\boldsymbol{\alpha} - \mathbf{z}_2\|_2^2, \tag{26}$$

where $\rho_{\boldsymbol{\alpha}}$ is the penalization parameter of the convex term, $\boldsymbol{\eta}_{\boldsymbol{\alpha}}$ is a vector of size $LQ \times 1$.

The problem (26) can be solved by a few iteration steps, in the $(j+1)^{th}$ iteration step, the closed-form solution for $\boldsymbol{\alpha}$ can be derived in the orthogonal Procrustes problem framework as:

$$\boldsymbol{\alpha}^{(l,j+1)} = \mathbf{U}_{L_Q}\mathbf{U}_R, \tag{27}$$

and $\mathbf{z}_2$ is given as:

$$\mathbf{z}_2^{(j+1)} = \mathcal{S}_{\lambda_{\boldsymbol{\alpha}}/d_{\mathbf{z}_2}}\left(\mathbf{z}_2^{(j)} - \frac{1}{\lambda_{\boldsymbol{\alpha}}}\nabla_{\mathbf{z}_2}C(\mathbf{z}_2^{(j)})\right), \tag{28}$$

where $\mathbf{U}_{L_Q}$ is the matrix composed of the first Q columns of $\mathbf{U}_L$, $\mathbf{U}_L$ and $\mathbf{U}_R$ are orthonormal matrix defined in (44), $C(\mathbf{z}_2^{(j)}) = \mathcal{R}e\left\{\boldsymbol{\eta}_{\boldsymbol{\alpha}}^H(vec(\boldsymbol{\alpha}) - \mathbf{z}_2^{(j)})\right\} + \frac{1}{2\rho_{\boldsymbol{\alpha}}}\|\boldsymbol{\alpha} - \mathbf{z}_2^{(j)}\|_2^2$, the gradient of $C(\mathbf{z}_2^{(j)})$ in terms of $\mathbf{z}_2^{\star(j)}(n)$ is $\nabla_{\mathbf{z}_2}C(\mathbf{z}_2^{(j)}) = -\frac{1}{2}\boldsymbol{\eta}_{\boldsymbol{\alpha}} + \frac{1}{2\rho_{\boldsymbol{\alpha}}}(\mathbf{z}_2^{(j)} - \boldsymbol{\alpha})$, $d_{\mathbf{z}_2}$ is the coefficient of the proximal term.

The details for obtaining (27) can be found in Appendix B.

The algorithm for solving (14) is summarized in Algorithm 1. In addition, the time and space complexity of each step in one iteration of the proposed algorithm is given in table I, where the big O notation $\mathcal{O}(\cdot)$ is the asymptotic upper bound of the complexity of an algorithm. The real-time applicability and the hardware requirements of the proposed algorithm can be evaluated by the time complexity and the space complexity, respectively. A few iterations even without full convergence are sufficient for a rough estimation of the DOA. However, to achieve a precise DOA estimation with high resolution that equals the angular separation between two adjacent grid points in the candidate DOA sector (namely 1° in our experiments), about 200 iterations of the outer loop and a few iterations of the inner loop are needed, which requires more computational time than the conventional methods of DOA estimation and dereverberation. Thus, there is a trade-off between the resolution and the computational cost of the DOA estimation.

Algorithm 1 Localization algorithm for multiple sources scenario.

1: Initialization: $l = j = 0$, $\lambda_{\mathbf{z}_1}$, $\lambda_{\tilde{\mathbf{s}}}$, $\lambda_{\boldsymbol{\alpha}}$, $\mu_{\mathbf{z}_1}$, $d_{\tilde{\mathbf{s}}}$, $d_{\boldsymbol{\alpha}}$, $d_{\mathbf{z}_2}$, $\rho_{\mathbf{G}}$, $\rho_{\boldsymbol{\alpha}}$;
2: **repeat**
3: % line 4 to line 10 for $\mathbf{G}$ estimation
4: $l = 0$;
5: **repeat**
6: Calculate $\mathbf{G}^{(l+1)}$ with (18) for solving (40a);
7: Calculate $\mathbf{z}_1^{(l+1)}(n)$ with (19) for solving (40b);
8: Calculate $\boldsymbol{\eta}_{\mathbf{G}}^{(l+1)}(n)$ with (40c);
9: $l = l + 1$;
10: **until** Convergence or up to 10 iterations;
11: % line 12 to line 26 for $\tilde{\mathbf{s}}(n)$ and $\boldsymbol{\alpha}$ estimation
12: $l = 0$;
13: **repeat**
14: % line 15 for $\tilde{\mathbf{s}}(n)$ estimation
15: Calculate $\tilde{\mathbf{s}}^{(l+1)}(n)$ with (24) for solving (23a);
16: % line 17 to line 24 for $\boldsymbol{\alpha}$ estimation in (23b)
17: $j = 0$;
18: **repeat**
19: Calculate $\boldsymbol{\alpha}^{(l,j+1)}$ with (27) for solving (42a);
20: Calculate $\mathbf{z}_2^{(j+1)}$ with (28) for solving (42b);
21: Calculate $\boldsymbol{\eta}^{(j+1)}$ with (42c);
22: $j = j + 1$;
23: **until** Convergence or up to 10 iterations;
24: $\boldsymbol{\alpha}^{(l+1)} = \boldsymbol{\alpha}^{(l,j)}$
25: $l = l + 1$;
26: **until** Convergence or up to 10 iterations;
27: **until** Convergence or up to 10 iterations;

IV. The Proposed Algorithm in Single Source Scenario

The single source scenario can be considered a special case of the multiple sources scenario, where the matrix $\boldsymbol{\alpha}$, the vector $\tilde{\mathbf{s}}(n)$ and $\mathbf{s}(n)$ shrink to a vector of dimension $L \times 1$, the scalar $\tilde{s}(n)$ and $s(n)$, respectively. Since there is no overlap problem of multiple sources in this scenario, the constraint on the orthonormality of $\boldsymbol{\alpha}$ can be replaced by a constraint on the norm of $\boldsymbol{\alpha}$ and only the sparsity penalization of $\boldsymbol{\alpha}$ is kept. The DOA estimation criterion can thus be given as:

$$\begin{aligned}\left\{\hat{\mathbf{G}}, \hat{\boldsymbol{\alpha}}, \hat{\tilde{s}}(n)\right\} = &\arg\min_{\mathbf{G},\boldsymbol{\alpha},\tilde{s}(n)} \sum_{n=1}^{N} \|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{s}(n)\|_2^2 \\ &+ \lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1 + \varrho(\boldsymbol{\alpha}) + \lambda_{\mathbf{z}_1}\sum_{n=1}^{N}\|\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n)\|_1 \\ &+ \lambda_{\tilde{\mathbf{s}}}\sum_{n=1}^{N}|\tilde{s}(n)|,\end{aligned} \tag{29}$$

where:

$$\varrho(\boldsymbol{\alpha}) = \begin{cases} 0, & \text{if } \boldsymbol{\alpha}^H\boldsymbol{\alpha} = 1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{30}$$

The framework for solving (29) is based on the integrated PALM method, which is similar to the multiple sources scenario. The main difference lies in estimation of $\boldsymbol{\alpha}$. Instead of (23b), $\boldsymbol{\alpha}$ can be estimated by:

$$\begin{aligned}\boldsymbol{\alpha}^{(l+1)} = &\arg\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N} \|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}\tilde{s}^{(l+1)}(n)\|_2^2 + \lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1 \\ &+ \varrho(\boldsymbol{\alpha}) + \frac{d_{\boldsymbol{\alpha}}}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(l)}\|_2^2.\end{aligned} \tag{31}$$

No extra augmented Lagrangian is needed here for solving (31). That is to say, the steps between line 17 and line 24 in Algorithm 1 are replaced by the following two-step scheme for estimating $\boldsymbol{\alpha}^{(l+1)}$:

$$\begin{aligned}\boldsymbol{\alpha}_m = &\mathcal{S}_{\lambda_{\boldsymbol{\alpha}}/d_{\boldsymbol{\alpha}}}\left(\boldsymbol{\alpha}^{(l)} - \frac{1}{\lambda_{\boldsymbol{\alpha}}}\sum_{n=1}^{N}(-\mathbf{H}^H\hat{\mathbf{y}}(n)\tilde{s}^{\star(l+1)}(n)\right. \\ &\left.+\mathbf{H}^H\mathbf{H}\boldsymbol{\alpha}\|\tilde{s}^{(l+1)}(n)\|^2)\right),\end{aligned} \tag{32}$$

where the soft thresholding is used for the l_1 norm penalization followed by a projection operator for the unitary norm penalization:

$$\boldsymbol{\alpha}^{(l+1)} = \mathcal{P}_c(\boldsymbol{\alpha}_m), \tag{33}$$

TABLE I: Complexity analyses of Algorithm 1.

parameters	$\tilde{\mathbf{x}}(n) \in \mathbb{C}^{K_l M \times 1}, \tilde{\mathbf{s}}(n) \in \mathbb{C}^{Q \times 1}, \boldsymbol{\alpha} \in \mathbb{C}^{L \times Q}, \mathbf{H} \in \mathbb{C}^{M \times L}, \hat{\mathbf{y}}(n) \in \mathbb{C}^{M \times 1}$						
equation	(18)	(19)	(40c)	(24)	(27)	(28)	(42c)
time	$\mathcal{O}(K_l^2 M^2 N) + \mathcal{O}(K_l^3 M^3)$ $+\mathcal{O}(QLMN) + \mathcal{O}(QK_lMN)$	$\mathcal{O}(K_l M^2 N)$	$\mathcal{O}(K_l M^2 N)$	$\mathcal{O}(QLN)$ $+\mathcal{O}(MLN)$	$\mathcal{O}(QMN) + \mathcal{O}(Q^2 N)$ $+\mathcal{O}(Q^2 L) + \mathcal{O}(QML)$	$\mathcal{O}(1)$	$\mathcal{O}(QL)$
space	$\mathcal{O}(K_l M) + \mathcal{O}(QL)$ $+\mathcal{O}(ML)$	$\mathcal{O}(K_l M^2)$	$\mathcal{O}(K_l M^2)$	$\mathcal{O}(QL)$ $+\mathcal{O}(ML)$	$\mathcal{O}(QL) + \mathcal{O}(ML)$ $+\mathcal{O}(Q^2)$	$\mathcal{O}(QL)$	$\mathcal{O}(QL)$

where:

$$\mathcal{P}_c(\mathbf{v}) = \begin{cases} \mathbf{v}, & \text{if } \mathbf{v}^H\mathbf{v} = \delta_{\mathbf{v}}, \\ \mathbf{v}/\|\mathbf{v}\|_2, & \text{otherwise}, \end{cases} \tag{34}$$

where $\delta_{\mathbf{v}}$ denotes the norm constraints on $\mathbf{v}$.

V. Pre- and post-processing

A. Source number estimation

In (14) the source number is assumed *a-priori* to be known for determining the size of $\boldsymbol{\alpha}$. The K-means algorithm is commonly used for jointly estimating the source number and the DOAs [44][45]. Here, to improve the efficiency, we employ the eigenvalues of the covariance matrix of the microphone array signals to directly estimate the source number. Numerical experiments have shown that although the reverberations may cause perturbations in the eigenvalue calculations, the eigenvalues for the early and late reflections are generally significantly smaller than those for the actual sources [34]. Hence, assuming that the eigenvalues are in descending order such that $\lambda_1 \geq \lambda_2 ... \geq \lambda_M$, the estimated source number $\hat{Q}$ can be obtained using the following criterion:

$$\frac{\lambda_1}{\lambda_{\hat{Q}}} > \varepsilon > \frac{\lambda_1}{\lambda_{\hat{Q}+1}}, \tag{35}$$

where ε is empirically selected and set to 6 when the noise is neglected.

B. Post-processing of DOA estimation

A post-processing step can be added at the output of the DOA estimation algorithm in section III and IV. As the proposed criteria involve non-convex optimization due to $\boldsymbol{\alpha}\tilde{\mathbf{s}}(n)$ in (14) as well as $\boldsymbol{\alpha}\tilde{s}(n)$ in (29), it is crucial to choose well-suited initial value of the parameters, especially of $\boldsymbol{\alpha}$, to avoid local solutions. In our work, we initialize $\boldsymbol{\alpha}$ with J different values $\boldsymbol{\alpha}_{0_1}, ..., \boldsymbol{\alpha}_{0_j}, ..., \boldsymbol{\alpha}_{0_J}$, each vector $\boldsymbol{\alpha}_{0_j}$ is composed of contiguous 0 and 1 such that:

$$\boldsymbol{\alpha}_{0_j} = [0, ..., 0, 1, ..., 1, 0, ..., 0], \tag{36}$$

where the region of component 1 covers the candidate DOAs of the sources. The borders of the components 1 of the J vectors are uniformly distributed around the peaks of the pseudo-spectrum of the conventional beamforming for an *a-priori* rough DOA estimation. At last, we select $\boldsymbol{\alpha}_{0_j}$ for which the final solution yields the minimum value of the cost functions (14) and (29) as the good initial value.

VI. Evaluation

In this section, evaluations are performed on simulated data and realistic data. In all the experiments, our proposed methods are compared with the baseline methods including the conventional beamforming [1], conventional MUSIC [2] in (9) and the DPD based method [29]. Here, to select the direct-path dominant regions, the local covariance matrix $\hat{\mathbf{R}}_{\mathbf{x}}(n,\omega)$ is calculated by averaging the microphone array measurements over D_n time frames and D_ω frequency bins [29] such that $\hat{\mathbf{R}}_{\mathbf{x}}(n,\omega) = \frac{1}{D_n D_\omega}\sum_{d_n=1}^{D_n}\sum_{d_\omega=1}^{D_\omega}\mathbf{x}(n-d_n,\omega-d_\omega)\mathbf{x}^H(n-d_n,\omega-d_\omega)$, D_n and D_ω are generally from 2 to 10 and chosen empirically. Three tests reported in [29] are used:

$$\mathcal{A}_1 = \left\{(n,\omega) : \frac{\sigma_1(n,\omega)}{\sigma_2(n,\omega)} > \mathcal{T}_{th_1}\right\}, \tag{37}$$

$$\mathcal{A}_2 = \left\{(n,\omega) : \frac{\max_\theta |\mathbf{a}^H(\theta,\omega)\mathbf{x}(n,\omega)|^2}{\mathbf{a}^H(n,\omega)\mathbf{a}(n,\omega)} > \mathcal{T}_{th_2}\right\}, \tag{38}$$

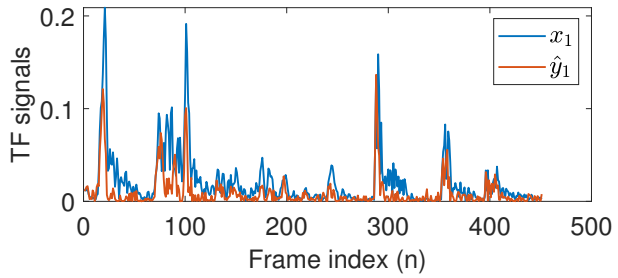
and

$$\mathcal{A}_3 = \left\{(n,\omega) : \max_\theta \frac{1}{\|\mathbf{a}^H(\theta,\omega)\mathbf{P}^\perp_{\mathbf{u}_1(n,\omega)}\|^2} > \mathcal{T}_{th_3}\right\}, \tag{39}$$

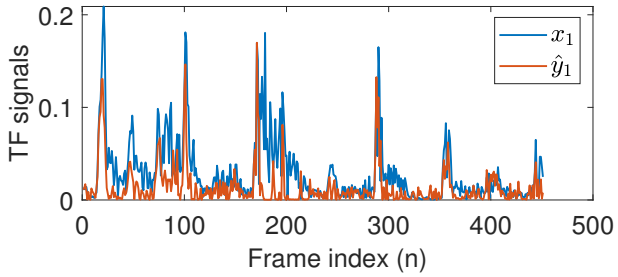
where $\mathcal{A}_r(\cdot), r = 1, 2, 3$ means the set of the TF bins that pass these tests and are considered valid for the DOA estimation, $\sigma_1(n,\omega)$ and $\sigma_2(n,\omega)$ are the largest and the second largest eigenvalues of $\hat{\mathbf{R}}_{\mathbf{x}}(n,\omega)$, $\mathbf{u}_1(n,\omega)$ is the dominant eigenvector of $\hat{\mathbf{R}}_{\mathbf{x}}(n,\omega)$. $\mathbf{P}^\perp_{\mathbf{u}_1(n,\omega)}$ gives the subspace of $\hat{\mathbf{R}}_{\mathbf{x}}(n,\omega)$ which is orthogonal to $\mathbf{u}_1(n,\omega)$. The MUSIC algorithm is applied to each selected single source dominant TF bin or region for estimating the DOA of the corresponding dominant source, and then all the potential DOAs are plotted in a histogram and the peaks are considered as the estimated DOAs of the actual sources. As the thresholds $\mathcal{T}_{th_1} \sim \mathcal{T}_{th_3}$ require empirical selection depending on the experimental configurations, the performance of DOA estimation is evaluated with all the three tests and the best one is chosen as that of the DPD based method.

A. Evaluations with simulated data

For the simulation setup, the RIR is generated using the image method [39] in a room with size of $8\,\mathrm{m} \times 8\,\mathrm{m} \times 2.6\,\mathrm{m}$. The microphone array composed of 8 microphones is centered at $(3, 0.5, 1)\,\mathrm{m}$ and the inter-element space is $0.03\,\mathrm{m}$ along the x axis. Three point sources are placed at $(3, 5.5, 1)\,\mathrm{m}$ for s_1, $(6, 3.5, 1)\mathrm{m}$ for s_2 and $(0.4, 2.9, 1)\mathrm{m}$ for s_3, the corresponding DOAs are $90°$, $135°$ and $45°$, respectively. Source signals are female speeches from the database Librispeech [46] sampled at 16 kHz. All the evaluations are processed at the frequency



(a) One source scenario



(b) Two sources scenario

Fig. 2. Magnitude of STFT signals of microphone array.

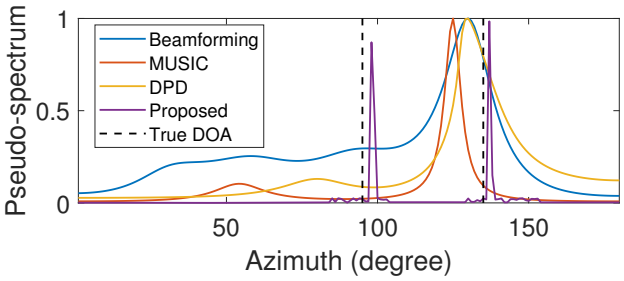
bin of 875 Hz using the STFT with half overlapping Hann windows of 16 ms. Noise is neglected in most cases, except in Fig. 7, where white Gaussian noise is added to the microphones with different signal to noise ratios (SNRs) and 100 Monte Carlo simulations are conducted for each configuration.

In Fig. 2, the dereverberation performance of the proposed algorithm is investigated in single-source and two-source scenarios. The magnitude of the TF signals at 875 Hz are plotted against the frame index ranging from 650 to 1100, since the power of signals dominates in this region. The legend x_1 represents the signals received at the first microphone in (2), and $\hat{y}_1$ stands for the first element in the dereverberated signal vector $\mathbf{x}(n) - \mathbf{G}^H\tilde{\mathbf{x}}(n)$ in (14). It can be observed that the indices of the sharp peaks of x_1 and $\hat{y}_1$ generally coincident. However, compared to x_1, the tails of the sharp peaks of $\hat{y}_1$ are cut off in both signal-source and two-source cases. The differences between x_1 and $\hat{y}_1$ illustrate an approximation of the dereverberation performance of the proposed algorithms.

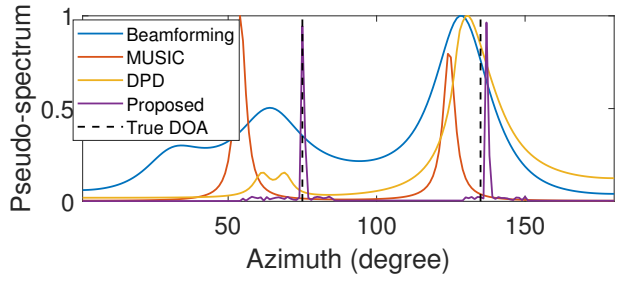
TABLE II: DOA estimation results in Fig. 3.

src separation	80°		60°		40°	
true DOA	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
	55°	135°	75°	135°	95°	135°
beamforming	68°	131°	64°	130°	97°	130°
MUSIC	57°	123°	61°	124°	NAN	125°
DPD	69.5°	132.3°	65°	132°	80°	132.1°
proposed	62°	134°	75°	138°	98°	137°

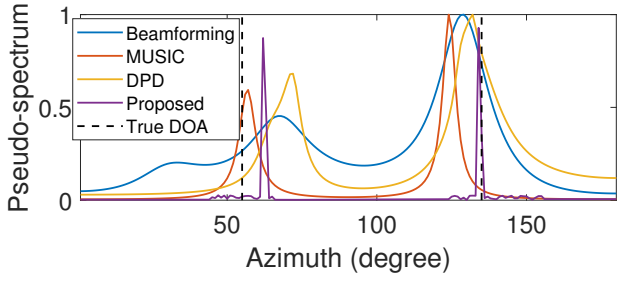
In Fig. 3 the normalized pseudo-spectra of the four DOA estimation methods are compared under different source separation ($40°, 60°$ and $80°$) in two source scenarios, where $T_{60} = 600\text{ms}$. The angular separation between two adjacent points in the x-axis of the pseudo-spectra is $1°$. For the proposed algorithm, we plot $\underline{\boldsymbol{\alpha}} = \sum_{j=1}^{Q} |\boldsymbol{\alpha}_{ij}|$ as the pseudo-spectrum. Fig. 3(a) clearly shows that when the two sources are too close, the three baseline methods only identify one



(a) $\theta_1 = 95°$, $\theta_2 = 135°$



(b) $\theta_1 = 75°$, $\theta_2 = 135°$



(c) $\theta_1 = 55°$, $\theta_2 = 135°$

Fig. 3. Pseudo-spectrum with different source separations, $T_{60} = 600$ ms.

peak and fail to separate the two sources. In contrast, the proposed method successfully separates the two sources due to the sparsity constraint imposed on $\boldsymbol{\alpha}$. Fig. 3(b) demonstrates that the proposed method achieves more accurate peaks for DOA estimation compared to other baseline methods although all the methods succeed in separating the two sources. In Fig. 3(c), the proposed method exhibits the best estimation result for the second source with DOA $\theta_2 = 135°$. However, for the first source with DOA $\theta_1 = 55°$, the peak of $\underline{\boldsymbol{\alpha}}$ deviates more from the true DOA compared to MUSIC. This phenomenon can be attributed to the non-convexity term in the algorithm (14) which leads to local solutions of $\boldsymbol{\alpha}$. To verify this, we use a criterion with the same framework of (14), but with a fixed *a-priori* value of $\boldsymbol{\alpha}$ that is good, and estimate only $\mathbf{G}$ and $\tilde{\mathbf{s}}(n)$. Compared to the results of (14), the results of this new criterion yields a smaller value of (11) which is one part of the cost function. In addition, as stated in section III, when the sparsity weight is sufficiently large, the proposed method eliminates early reflections and only estimates the DOAs of the actual sources. The pseudo-spectra intuitively illustrate the resolution of the DOA estimation algorithms. The main lobes of the proposed method are thinner than those of the baseline methods, which implies its higher resolution. Table II presents the numerical results of DOA estimation in Fig.

3, where "NAN" means the algorithm fails in detecting th source, namely there is no obvious peak near the true DOA From table II it can be seen that the proposed method achieve higher DOA estimation accuracy than the baseline methods.

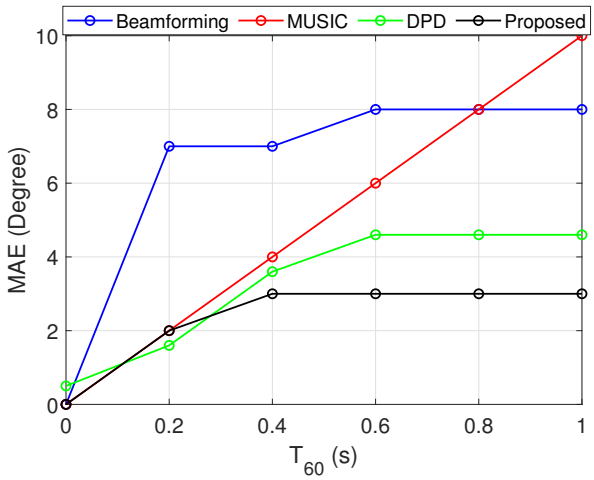


Fig. 4. MAE vs. T_{60}, $\theta = 135°$ in one source case.

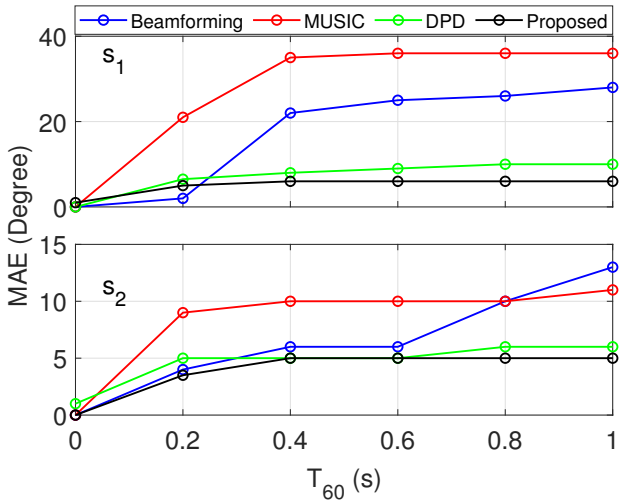


Fig. 5. MAE vs. T_{60}, $\theta_1 = 90°$ and $\theta_2 = 135°$ in two sources case.

Fig. 4 to Fig. 6 illustrate the mean absolute errors (MAEs) between the estimated DOAs and the true DOAs for one source, two sources, and three sources, respectively. Several observations can be made from these figures. Firstly, it is evident that the proposed method outperforms the three baseline methods with the smallest MAEs, following behind is the DPD method; secondly, the MAE increases with the source number and the value of T_{60}. In particular, the sub-figure for s_1 in Fig. 5 shows that the MAE for beamforming and MUSIC exceeds $20°$ degree and approaches $40°$ when T_{60} exceeds 200ms, which indicates that spurious peaks are detected, rendering the DOA estimation results invalid. Similarly, in the sub figures for s_1 and s_2 in Fig. 6 beamforming and MUSIC fail to provide any estimations when T_{60} exceeds 200ms. In contrast, the DPD based method successfully estimate DOAs for all values of T_{60} by utilizing the mask of the single source dominant TF

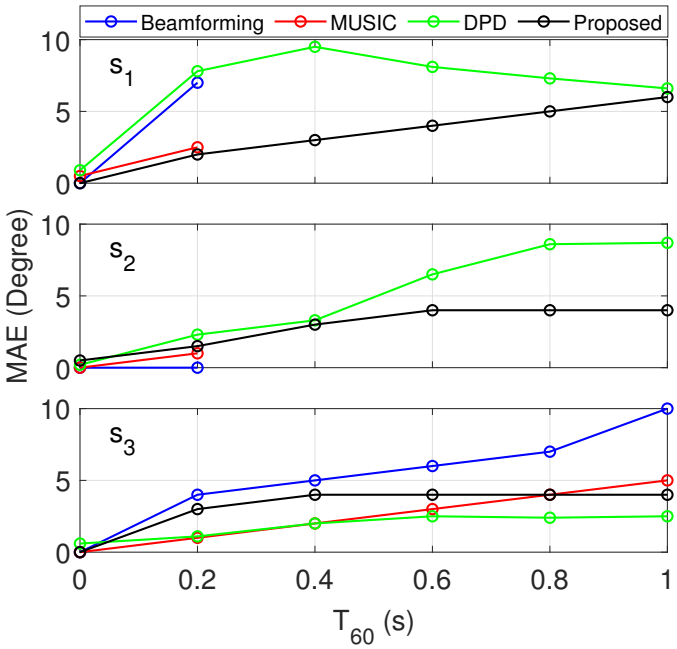


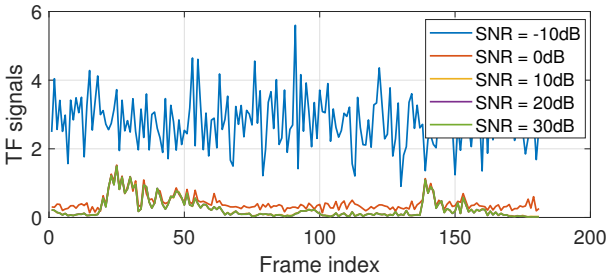
Fig. 6. MAE vs. T_{60}, $\theta_1 = 45°$, $\theta_2 = 90°$, and $\theta_3 = 135°$ in three sources case.

point. The proposed method achieves superior performance in most cases. However, the simulation results also reveal the limitations of the proposed method: firstly, as the MCLP filter cannot eliminate completely the late reverberation, the performance of the proposed algorithm deteriorates as T_{60} increases; secondly, in the sub-figure for s_3 in Fig. 6, a local solution is observed where the performance of the proposed method is slightly inferior to the DPD based methods.

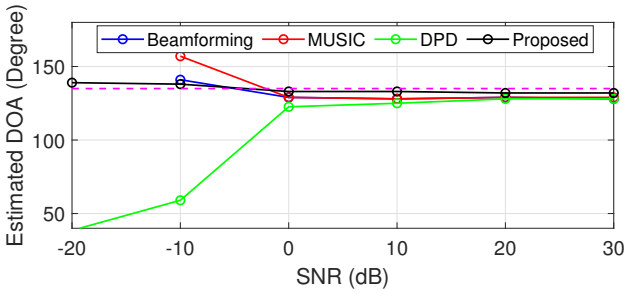
In Fig. 7 the mean value of the estimated DOAs are presented as a function of SNR, with the true DOAs indicated by dashed pink lines. Generally, the three baseline methods fail to give a DOA estimation when the SNR is -20 dB. In Fig. 7(a) it can be observed that the influence of the noise decreases significantly when the SNR exceeds 10 dB, which can be supported by Fig. 7(b) - Fig. 7(d), where the estimated DOAs from all the methods exhibit stability with a SNR larger than 10 dB. In this analysis, instead of the MAEs, the estimated and the true DOA lines are plotted separately, since the crossing points of these lines in Fig. 7(b) and Fig. 7(c) illustrate that the noises at the microphones and the room reverberations have opposite effects on the performance the DOA estimation methods. This intriguing phenomenon motivates further theoretical investigations in future studies.

B. Evaluations with real-world data

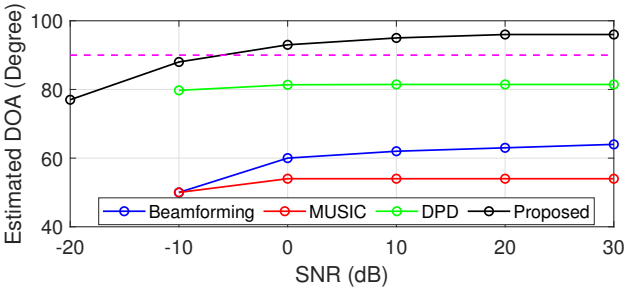
We utilise the single- and multichannel audio recordings database (SMARD) [47] from Aalborg University for the real-world data experiments. Different types of signals are recorded in a $7.34\text{ m} \times 8.09\text{ m} \times 2.87\text{ m}$ box-shaped listening room with a reverberation time of approximately 0.15 seconds. Here male speech samples with the OmniPower 4296 loudspeaker and from the DOAs $90°$ and $126°$ are selected as sources. A



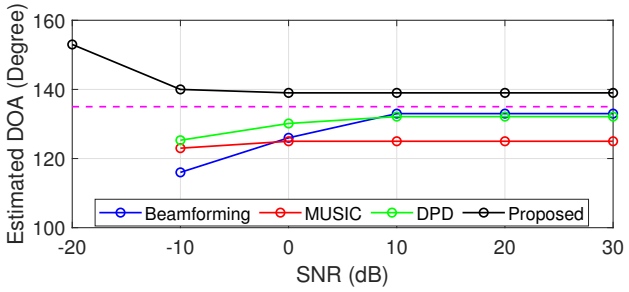
(a) Temporal frequential signals



(b) $\theta = 135°$ in one source case



(c) $\theta_1 = 90°$ in two sources case



(d) $\theta_2 = 135°$ in two sources case

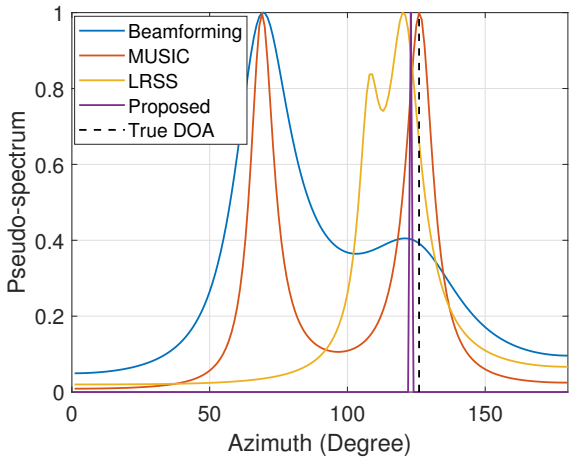
Fig. 7. Mean value of the estimated DOAs vs. SNR, $T_{60} = 600$ ms.

linear microphone array of 7 microphones of Brüel & Kjær Type 2270 is chosen here, centered at $(1.0, 0.559, 1.325)$ m and the inter-element space is 0.05 m along the x-axis. The sampling rate is 48 kHz.

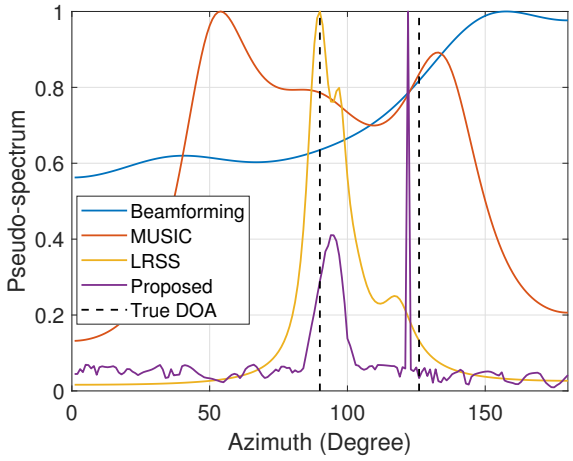
TABLE III: DOA estimation results in Fig. 8.

src number	1	2	
true DOA	θ	θ_1	θ_2
	126°	90°	126°
beamforming	121°	NAN	NAN
MUSIC	126°	NAN	133°
DPD	120°	90°	117°
proposed	123°	94°	122°

In Fig. 8 the performance of the four methods with real



(a) $\theta = 126°$ in one source case



(b) $\theta_1 = 90°$ and $\theta_2 = 126°$ in two sources case

Fig. 8. Pseudo-spectrum obtained by real world data.

data is illustrated in both one source and two sources cases. In one source case (Fig. 8(a)) beamforming and MUSIC find one good peak for the true source and one spurious peak near 70° due to the early reflections. For beamforming the spurious peak is dominant rather than the good peak in the pseudo-spectrum. The DPD-based method provides a more accurate DOA estimation yet a spurious peak also appears due to outliers in the TF bins selecting procedure. The proposed method outperforms the three baseline methods for two reasons. On one hand, spurious peaks are avoided in the pseudo-spectra, since the early reflections are eliminated by the sparsity constraint of $\boldsymbol{\alpha}$. On the other hand, the DOA estimation accuracy is improved, since the late reverberant component is reduced by the MCLP filter matrix $\mathbf{G}$. In two sources case (Fig. 8(b)), beamforming fails to estimate neither of the two sources, MUSIC can give a rough estimate for the source with DOA 126° but fails to find the source with DOA 90°, the DPD-based method gives an accurate estimation for the source wih DOA 90° but fails for the source with DOA 126°, exhibiting several spurious peaks due to the outliers as in Fig. 8(a). The proposed method shows a significant advantage

over the three baseline methods for the estimation of both the two sources. Similarly to Fig. 3, the main lobes in the pseudo-spectra illustrate the resolution of the algorithms, a thinner main lobe implies a higher resolution of the DOA estimation method. Table III present the numerical results of the DOA estimation for the four methods, where "NAN" means the algorithm fails in detecting the source, namely there is no obvious peak near the true DOA. It can be seen that, although the baseline methods may have higher accuracy in some cases, they are prone to fail in detecting the sources or produce false peaks. The proposed method, on the other hand, can consistently estimate the DOA of the sources accurately without spurious peak. All these experiment results illustrate the robustness of the proposed method.

VII. Conclusion

This paper presents a novel approach for joint DOA estimation and dereverberation in reverberant environments. The proposed method effectively removes the late reverberant component using the MCLP filter, allowing for accurate DOA estimation using the early component signals in conjunction. The early reflections are neglected and only the DOAs of the actual sources are estimated when the sparsity weight is large enough. The DOAs, the source signals, and the MCLP filter coefficients are estimated by the linearized alternative iterations, namely PALM. In particular, the orthogonal Procrustes problem framework is introduced to satisfy the orthonormalization constraint on the DOA estimation matrix. Evaluations on both simulated an real-world data demonstrate that compared to the baselines methods, our proposed methods tend to be more accurate and more robust against the reverberations and the noises in case of both multiple sources and single source scenarios.

Appendix A

In the $(l+1)^{th}$ iteration for solving (17), the parameters can be determined by:

$$\mathbf{G}^{(l+1)} = \arg\min_{\mathbf{G}} \mathcal{L}(\mathbf{G}^{(l)}, \mathbf{z}_1^{(l)}, \boldsymbol{\eta}_{\mathbf{G}}^{(l)}(n)), \tag{40a}$$

$$\mathbf{z}_1^{(l+1)}(n) = \arg\min_{\mathbf{z}_1} \mathcal{L}(\mathbf{G}^{(l+1)}, \mathbf{z}_1^{(l)}, \boldsymbol{\eta}_{\mathbf{G}}^{(l)}(n)), \tag{40b}$$

$$\boldsymbol{\eta}_{\mathbf{G}}^{(l+1)}(n) = \boldsymbol{\eta}_{\mathbf{G}}^{(l)}(n) + \gamma(\mathbf{x}(n) - \mathbf{G}^{(l+1)^H}\tilde{\mathbf{x}}(n) - \mathbf{z}_1^{(l+1)}(n)). \tag{40c}$$

For (40a), a closed-form expression of $\mathbf{G}^{(l+1)}$ can be obtained by searching a result such that the gradient of $\mathcal{L}(\mathbf{G}^{(l)}, \mathbf{z}_1^{(l)}, \boldsymbol{\eta}_{\mathbf{G}}^{(l)})$ with respect to $\mathbf{G}^{\star(l+1)}$ equals to 0, where $\cdot^\star$ is the conjugate operator, which results in (18).

For solving (40b), noting that $V(\mathbf{G}^{(l+1)}, \mathbf{z}_1^{(l)}(n), \boldsymbol{\eta}_{\mathbf{G}}^{(l)}(n)) = \mathcal{R}e\left\{\boldsymbol{\eta}_{\mathbf{G}}^{(l)H}(n)(\mathbf{x}(n) - \mathbf{G}^{(l+1)H}\tilde{\mathbf{x}}(n) - \mathbf{z}_1^{(l)}(n))\right\} + \frac{1}{2\rho_{\mathbf{G}}}\|\mathbf{x}(n) - \mathbf{G}^{(l+1)H}\tilde{\mathbf{x}}(n) - \mathbf{z}_1^{(l)}(n)\|^2$ and the gradient of V in $\mathbf{z}_1^{\star(l)}$ is $\nabla_{\mathbf{z}} V(\mathbf{G}^{(l+1)}, \mathbf{z}_1^{(l)}(n), \boldsymbol{\eta}_{\mathbf{G}}^{(l)}(n)) = -\frac{1}{2}\boldsymbol{\eta}_{\mathbf{G}}(n) + \frac{1}{2\rho_{\mathbf{G}}}(\mathbf{z}_1^{(l)}(n) + \mathbf{G}^{(l+1)H}\tilde{\mathbf{x}}(n) - \mathbf{x}(n))$, introducing the proximal terms and ignoring the constant terms, (40b) can be derived as:

$$\mathbf{z}_1^{(l+1)}(n) = \arg\min_{\mathbf{z}_1} < \nabla_{\mathbf{z}} V(\mathbf{G}^{(l+1)}, \mathbf{z}_1^{(l)}(n), \boldsymbol{\eta}_{\mathbf{G}}^{(l)}(n)), \mathbf{z}_1 - \mathbf{z}_1^{(l)}(n) > + \frac{\mu_{\mathbf{z}_1}}{2}\|\mathbf{z}_1 - \mathbf{z}_1^{(l)}(n)\|_2^2 + \lambda_{\mathbf{z}_1}\|\mathbf{z}_1\|_1, \tag{41}$$

where $< \cdot, \cdot >$ is the inner product operator.

With the soft thresholding operator, a closed-from expression for (41) can be given as in (19).

Appendix B

In the $(j+1)^{th}$ iteration for solving (26), the parameters can be determined by:

$$\boldsymbol{\alpha}^{(l,j+1)} = \arg\min_{\boldsymbol{\alpha}} \mathcal{L}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^{(l,j)}, \mathbf{z}_2^{(j)}, \boldsymbol{\eta}_{\boldsymbol{\alpha}}^{(j)}), \tag{42a}$$

$$\mathbf{z}_2^{(j+1)} = \arg\min_{\boldsymbol{\alpha}} \mathcal{L}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^{(l,j+1)}, \mathbf{z}_2^{(j)}, \boldsymbol{\eta}_{\boldsymbol{\alpha}}^{(j)}), \tag{42b}$$

$$\boldsymbol{\eta}_{\boldsymbol{\alpha}}^{(j+1)} = \boldsymbol{\eta}_{\boldsymbol{\alpha}}^{(j)} + \gamma(\boldsymbol{\alpha}^{(l,j+1)} - \mathbf{z}_2^{(j+1)}). \tag{42c}$$

Similarly to (24) and (40b), using the linearized method, (42a) can be reformulated as:

$$\boldsymbol{\alpha}^{(l,j+1)} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(l,j)} + \frac{1}{d_{\boldsymbol{\alpha}}}\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}^{(l,j)})\|_2^2 + \iota(\boldsymbol{\alpha}), \tag{43}$$

where $F(\boldsymbol{\alpha}^{(l,j)}) = \sum_{n=1}^{N}\|\hat{\mathbf{y}}(n) - \mathbf{H}\boldsymbol{\alpha}^{(l,j)}\tilde{\mathbf{s}}^{(l+1)}(n)\|_2^2 + \frac{1}{2\rho_{\boldsymbol{\alpha}}}\|vec(\boldsymbol{\alpha}^{(l,j)}) - \mathbf{z}_2\|_2^2 + \mathcal{R}e\left\{\boldsymbol{\eta}_{\boldsymbol{\alpha}}^H(vec(\boldsymbol{\alpha}^{(l,j)}) - \mathbf{z}_2)\right\}$, and the gradient of $F(\boldsymbol{\alpha}^{(l,j)})$ in terms of $\boldsymbol{\alpha}^{\star(l,j)}$ is $\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}^{(l,j)}) = \sum_{n=1}^{N}(-\mathbf{H}^H\hat{\mathbf{y}}(n)\tilde{\mathbf{s}}^H(n) + \mathbf{H}^H\mathbf{H}\boldsymbol{\alpha}^{(l,j)}\tilde{\mathbf{s}}(n)\tilde{\mathbf{s}}^H(n)) + \frac{\mathcal{M}}{2}\{\boldsymbol{\eta}_{\boldsymbol{\alpha}}\} + \frac{1}{\rho_{\boldsymbol{\alpha}}}(\boldsymbol{\alpha}^{(l,j)} - \mathcal{M}\{\mathbf{z}_2\})$, where $\mathcal{M}\{\cdot\}$ is the operator which reshapes the vector $\boldsymbol{\eta}_{\boldsymbol{\alpha}}$ and $\mathbf{z}_2$ into a matrix of the same size with $\boldsymbol{\alpha}$. The (43) can be considered the Frobenius-norm minimization problem with orthonormalization constraint which is also called the orthogonal Procrustes problem [48][36], and can be solved by means of the singular value decomposition (SVD). Defining matrix $\mathbf{A}$ such that $\mathbf{A} = \boldsymbol{\alpha}^{(l,j)} - \frac{1}{d_{\boldsymbol{\alpha}}}\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}^{(l,j)})$, the SVD of $\mathbf{A}$ can be given as:

$$\mathbf{A} = \mathbf{U}_L\boldsymbol{\Sigma}\mathbf{U}_R, \tag{44}$$

where $\mathbf{U}_L \in \mathbb{C}^{L\times L}$ and $\mathbf{U}_R \in \mathbb{C}^{Q\times Q}$ are orthonormal matrices, $\boldsymbol{\Sigma} \in \mathbb{C}^{L\times Q}$ consists of a diagonal matrix of singular values of dimension $Q \times Q$, and a zero-value matrix of dimension $(L-Q)\times Q$. Thus, $\boldsymbol{\alpha}^{(l,j+1)}$ can be given as in (27).

References

[1] H. Do and H. F. Silverman, "Srp-phat methods of locating simultaneous multiple talkers using a frame of microphone array data," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 125–128.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[3] B. A. Johnson, Y. I. Abramovich, and X. Mestre, "Music, g-music, and maximum-likelihood performance breakdown," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3944–3958, 2008.

[4] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, 2009.

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

[6] S. Braun and E. A. P. 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.

[7] ——, "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.

[8] R. Ikeshita, K. Kinoshita, N. Kamo, and T. Nakatani, "Online speech dereverberation using mixture of multichannel linear prediction models," *IEEE Signal Processing Letters*, vol. 28, pp. 1580–1584, 2021.

[9] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5172–5176.

[10] S. R. Chetupalli and T. V. Sreenivas, "Late reverberation cancellation using bayesian estimation of multi-channel linear predictors and student's t-source prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1007–1018, 2019.

[11] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 101–105, 2017.

[12] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 544–558, 2019.

[13] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined weighted prediction error and minimum variance distortionless response for dereverberation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 446–450.

[14] S. Hashemgeloogerdi and S. Braun, "Joint beamforming and reverberation cancellation using a constrained kalman filter with multichannel linear prediction," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 481–485.

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

[16] G. Huang, J. Benesty, I. Cohen, and J. Chen, "Kronecker product multichannel linear filtering for adaptive weighted prediction error-based speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1277–1289, 2022.

[17] W. Yang, G. Huang, J. Chen, J. Benesty, I. Cohen, and W. Kellermann, "Robust dereverberation with kronecker product based multichannel linear prediction," *IEEE Signal Processing Letters*, vol. 28, pp. 101–105, 2021.

[18] W. Yang, G. Huang, W. Zhang, J. Chen, and J. Benesty, "Dereverberation with differential microphone arrays and the weighted-prediction-error method," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 376–380.

[19] W. Yang, G. Huang, A. Brendel, J. Chen, J. Benesty, W. Kellermann, and I. Cohen, "A bilinear framework for adaptive speech dereverberation combining beamforming and linear prediction," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[20] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 740–754, 2020.

[21] Z.-Q. Wang, G. Wichern, and J. L. Roux, "Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3476–3490, 2021.

[22] W. Xue, Y. Tong, G. Ding, C. Zhang, T. Ma, X. He, and B. Zhou, "Direct-path signal cross-correlation estimation for sound source localization in reverberation." in *INTERSPEECH*, 2019, pp. 2693–2697.

[23] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[24] S. Mohan, M. E. Lockwood, M. L. Kramer, and L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, p. 2136–2147, 2008.

[25] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d localization of multiple sound sources with intensity vector estimates in single source zones," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1556–1560.

[26] K. Wu, V. G. Reju, and A. W. H. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 444–448.

[27] B. Rafaely and K. Alhaiany, "Speaker localization using direct path dominance test based on sound field directivity," *Signal Processing*, vol. 143, pp. 42–47, 2018.

[28] B. Rafaely and D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 6120–6124.

[29] L. Madmoni and B. Rafaely, "Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 131–142, 2019.

[30] K. Wu, V. G. Reju, and A. W. H. Khong, "Multisource doa estimation in a reverberant environment using a single acoustic vector sensor," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1848–1859, 2018.

[31] J. Geng, S. Wang, Q. Liu, and X. Lou, "Multi-level time-frequency bins selection for direction of arrival estimation using a single acoustic vector sensor," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1048–1060, 2022.

[32] M. Jia, Y. Wu, C. Bao, and C. Ritz, "Multi-source doa estimation in reverberant environments by jointing detection and modeling of time-frequency points," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 379–392, 2021.

[33] D. Ying, R. Zhou, J. Li, and Y. Yan, "Window-dominant signal subspace methods for multiple short-term speech source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 731–744, 2017.

[34] B. Yang, H. Liu, C. Pang, and X. Li, "Multiple sound source counting and localization based on tf-wise spatial spectrum clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1241–1255, 2019.

[35] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[36] R. Everson, "Orthogonal, but not orthonormal, procrustes problems," Laboratory for Applied Mathematics, City Univ. New York, New York, NY, USA, and Mount Sinai Medical School, New York, NY, USA, Tech. Rep., 1998.

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

[38] P. Stoica and A. Nehorai, "Music, maximum likelihood, and cramer-rao bound," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 5, pp. 720–741, 1989.

[39] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep. 2.4, 2006. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

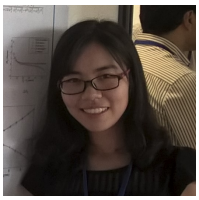
[40] X. Jiang, W.-J. Zeng, A. Yasotharan, H. C. So, and T. Kirubarajan, "Minimum dispersion beamforming for non-gaussian signals," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1879–1893, 2014.

[41] X. Wenmeng, B. Changchun, J. Maoshen, and J. Picheral, "Speech enhancement with robust beamforming for spatially overlapped and distributed sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2778–2790, 2022.

[42] J. M. Ortega and W. C. Rheinboldt, "Iterative solution of nonlinear equations in several variables /," *Academic Pr*, 1970.

[43] D. P. Bertsekas and J. N. Tsitsiklis, "Parallel and distributed computation: Numerical methods," *Prentice-Hall, New Jersey,*, 1989.

[44] Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Decoupled multiple speaker direction-of-arrival estimator under reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3120–3133, 2022.
[45] A. K. Jain, "Data clustering: 50 years beyond k-means," *Springer, Berlin, Heidelberg*, 2008.
[46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
[47] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (smard)," in *Int. Workshop Acoustic Signal Enhancement*, Sep. 2014. [Online]. Available: http://www.smard.es.aau.dk/
[48] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early psd estimation and recursive retf update in reverberant environments by means of the orthogonal procrustes problem," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 755–769, 2020.



XIONG Wenmeng received the B.E. degree in 2011 from Huazhong University of Science and Technology, China. She received the Ph.D. degree in Signal and Image processing in 2017 from CentraleSupelec, France. From 2017 to 2019, She was an algorithm engineer with Ericsson(China), Beijing, China. She is now an assistant professor with the Faculty of Information Technology at the Beijing University of Technology. Her current research interests lie in array signal processing and audio signal processing, including source localization, speech enhancement, dereverberation, and array geometry optimization.

BAO Changchun (IEEE Senior Member, CIE Fellow) is currently a Full Professor with the Faculty of Information Technology, Beijing University of Technology, Beijing, China. His research interests include speech and audio signal processing, speech coding, speech enhancement, speech transcoding, audio coding, audio enhancement, bandwidth extending for speech and audio signals, and 3D audio signal processing. He is the author or co-Author of more than 330 papers in journals and conferences and holds ten patents. He was ever an Associate Editor for the Journal of Communications, and currently the Editor of the Signal Processing and the Editor of the Journal of Data Acquisition & Processing. Prof. Bao is a Board Member of the Chinese Institute of Electronics (CIE), the Vice President of CIE Signal Processing Society Board of Governors, an Associate Editor and the Leader Guest Editor of EURASIP Journal on Audio, Speech, and Music Processing (JASMP), Member of International Speech Communication Association, the Chair of the APSIPASLA TC from 2015 to 2016, and the Chair of National Conference on Man-Machine Speech Communication-Standing Committee.

ZHOU Jing received the B.S. degree in electronic information engineering from Jiangxi University of Science and Technology of China in 2017 and the M.S. degree in electronic and communication engineering from Jiangxi University of Science and Technology of China in 2020. From 2020, he started to pursue doctoral degree in electronic science and technology from Beijing University of Technology. His research interests are in the areas of speech enhancement, microphone array signal processing and machine learning.

JIA Maoshen (M'13-SM'15) received the B.E. degree in Electronic Information Engineering in 2005 from Hebei University, China. He received the Ph.D. degree in Electronic Science and Technology in 2010 from Beijing University of Technology, China. Recently, he is a professor with the Faculty of Information Technology at the Beijing University of Technology. His current research interests include multichannel audio signal processing, audio coding and array signal processing.

José PICHERAL graduated from Supélec, Paris, France, and from the Politecnico di Milano, Milan, Italy, in 1999. He received is PhD from the Université Paris Sud in 2003 for his study about " High resolution methods for joint angle-delay estimation – application to UMTS and geophysics " and the HDR in 2017 for his work about " Parametric methods and inverse approach for coherently distributed sources localization ". He is currently associate Professor at CentraleSupelec and member of the Laboratory of Signals and Systems (L2S).