Joint acoustic localization and dereverberation through plane wave decomposition and sparse regularization

Niccolò Antonello, Enzo De Sena, *Member, IEEE,* Marc Moonen, *Fellow, IEEE,* Patrick A. Naylor, *Senior Member, IEEE,* and Toon van Waterschoot *Member, IEEE*

***Abstract*—Acoustic source localization and dereverberation are formulated jointly as an inverse problem. The inverse problem consists of the approximation of the sound field measured by a set of microphones. The recorded sound pressure is matched with that of a particular acoustic model based on a collection of plane waves arriving from different directions at the microphone positions. In order to achieve meaningful results, spatial and spatio-spectral sparsity can be promoted in the weight signals controlling the plane waves. The large-scale optimization problem resulting from the inverse problem formulation is solved using a first order optimization algorithm combined with a weighted overlap-add procedure. It is shown that once the weight signals capable of effectively approximating the sound field are obtained, they can be readily used to localize a moving sound source in terms of direction of arrival (DOA) and to perform dereverberation in a highly reverberant environment. Results from simulation experiments and from real measurements show that the proposed algorithm is robust against both localized and diffuse noise exhibiting a noise reduction in the dereverberated signals.**

***Index Terms*—Dereverberation, Source localization, Sparse sensing, Inverse problems**

I. Introduction

While there are many source localization algorithms that work well in free-field acoustic scenarios, source localization in highly reverberant environments is challenging [1], [2]. Reverberant environments are also problematic for speech intelligibility and significant research efforts have been focusing on dereverberation [3]. Dereverberation and source localization are often connected: for example many dereverberation algorithms require the knowledge of the direction of arrival (DOA) of the sound source [4], [5]. Instead, other algorithms rely either on channel equalization which requires estimation of the room impulse responses (RIRs) [6] or on multi-channel linear prediction (MCLP) which requires no a priori knowledge of the acoustics but is non-robust to additive noise [7], [8].

Recently, source localization has been posed as an inverse problem where physical acoustic models are used to reconstruct and localize the sound source in terms of spatial coordinates using microphones scattered in the room [9]–[11]. This is achieved by exploiting compressed sensing (CS) techniques [12] *i.e.*, by including a *sparse regularization* in the inverse problems that exploits the fact that the sound field is generated by sound sources that are *spatially sparse*. Such methods allow precise localization of the source position inside the room but require detailed knowledge of the room geometry and boundary conditions. Alternatively, the plane wave decomposition method (PWDM) has been shown to approximate well any sound field in source-free volumes [13]. This allows sound sources to be localized without knowledge of the room geometry. In [14] a narrowband source is localized in terms of spatial coordinates. This is achieved by splitting a sound field into its direct and reverberant components and by modeling the latter using the PWDM. Once this step is achieved, it is shown that standard localization techniques can be readily applied to the estimated direct sound field component to retrieve the coordinates of the sound sources. However, splitting the sound field into its reverberant and direct sound field components requires a large number of microphones, particularly when these are scattered in a large volume. This number can be reduced when partial knowledge of the room geometry is available [15]. Additionally, in [16] the low-rank nature of the reverberant sound field component is exploited and combined with the spatial sparsity of the direct sound field component. When localization is sought only in terms of DOA compact microphone arrays are typically employed. In this context CS techniques have also been found useful. In [17] the microphone measurements are matched with an over-complete dictionary of steering vectors. The promotion of a sparse solution enables a precise estimate of the DOA of multiple sound sources with an increased resolution. A similar approach is proposed in [18] using the spherical harmonic decomposition method (SHDM), a method that is

N. Antonello is at Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland. M. Moonen, T. van Waterschoot are with KU Leuven, ESAT–STADIUS, Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. T. van Waterschoot is also with KU Leuven, ESAT–ETC, e-Media Research Lab, Andreas Vesaliusstraat 13, 1000 Leuven, Belgium. E. De Sena is with University of Surrey, Institute of Sound Recording, GU2 7XH, Guilford, Surrey, UK. P. A. Naylor is with Imperial College, Electrical & Electronic Engineering, Exhibition Road SW7 2AZ, London, UK.

This research work was carried out at the ESAT Laboratory of KU Leuven, the frame of the FP7-PEOPLE Marie Curie Initial Training Network “Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)”, funded by the European Commission under Grant Agreement no. 316969, KU Leuven Impulsfonds IMP/14/037, KU Leuven Internal Funds VES/16/032, KU Leuven C2-16-00449 “Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking”. The research leading to these results has received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors’ views and the Union is not liable for any use that may be made of the contained information.

closely related to the PWDM. Here the SHDM is used to construct an over-complete dictionary that accounts for the presence of a rigid baffle of a spherical microphone array. Group sparsity has also been proposed to model sound fields leading to spatial, spatio-temporal and spatio-spectral sparsity [19], [20] and has been shown to improve DOA estimation particularly when combined with speech modeling [21]. Similar approaches have been used also for a dereverberation task. For example, channel equalization and beamforming are employed in [22], [23] respectively after estimating DOAs using sparse regularization. More recently, in [24] joint dereverberation and DOA estimation has been achieved through a sparse signal reconstruction task. In particular, a beamformer with enhanced resolution is obtained by exploiting sparse Bayesian learning (SBL) to automatically tune the hyperparameters that control the level of sparsity outperforming affirmed beamforming algorithms such as the minimum variance distortionless response (MVDR) [25].

In this paper, a recently proposed RIR interpolation algorithm [19], is reformulated and modified to be able to perform joint source localization and dereverberation. The proposed algorithm, called acoustic dereverberation and localization through field approximation (ADeLFi), relies on the approximation of the sound field recorded by a set of microphones which is formulated as a regularized inverse problem. This consists of an optimization problem that matches the sound pressure measured by the microphones with the sound pressure predicted by an acoustic model. Here the PWDM is used, which is capable of approximating the sound field in a source-free volume where the microphones are positioned. The PWDM is based on a large collection of plane waves that contribute to the sound field from a particular direction and are associated with a DOA. Plane waves are controlled by signals, named *weight signals*, that are estimated through the optimization problem. It is shown that, employing specific sparsity-inducing regularization, different kinds of sparse priors can be promoted in the weight signals: spatial sparsity and spatio-spectral sparsity. Spatial sparsity promotes only few weight signals to be active, hence limiting the number of directions from which the plane waves can arrive. On the other hand, spatio-spectral sparsity lead to weight signals that have a spectrum composed by few frequency components. Notice that in [19] *spatio-temporal sparsity* is promoted as well by employing a different acoustic model named time-domain equivalent source method (TESM). Being a time-domain method, TESM can easily allow for the promotion spatio-temporal sparsity. While in the context of RIR interpolation spatio-temporal sparsity outperformed spatial and spatio-spectral sparsity, in the context of speech dereverberation this regularization is not effective as the sound field is not generated by a temporally sparse sound source [26].

The resulting optimization problem has a large scale and nonsmooth cost function: this is solved using an accelerated version of the proximal gradient (PG) algorithm [27] and combined with a weighted overlap-add (WOLA) procedure. Once the approximation step is achieved, the weight signals can be used to estimate the DOA of a sound source. It is observed that the weight signal with *strongest energy* is associated with a particular DOA which is more likely to correspond to the DOA of the original sound source. This enables to localize the sound source and as a consequence a dereverberated signal can be readily obtained by selecting the weight signal corresponding to the estimated DOA. Alternatively, one can compute the sound pressure inside the source-free volume with the acoustic model using a small set of weight signals with corresponding DOAs close to the estimated one. In fact, solving the inverse problem accounts for decomposing the sound field into different plane waves with specific directions. This effectively creates a *spatial distribution* of reverberation among the weight signals controlling these plane waves, resulting in the weight signals to effectively be *dereverberated signals*. Additionally, this decomposition will occur also in the presence of a noise field; the contribution of the noise field will also be spatially distributed among the weight signals which therefore exhibit a *noise reduction*. It is shown that the WOLA procedure also enables the DOA estimation and dereverberation of a *moving sound source*.

The formulation of joint DOA estimation and dereverberation through a sound field approximation task allows to propose a new procedure for tuning the level of regularization, which represents the major element of novelty of this paper. In particular, the level of sparsity is not extrapolated from signals statistics as it is commonly pursued in sparsity-based beamforming, like *e.g.*, in [24], but rather by assessing the quality of the approximation through an additional microphone. This is achieved by adopting a modified version of K-fold cross validation (KCV), a procedure often employed in machine learning. The KCV strategy is simplified to be suited for online audio processing. Simulated and real measurement results show that in a sound field generated by a speech source, spatio-spectral and spatial sparsity based regularization have similar performances both in terms of sound field approximation and dereverberation. The proposed algorithm is shown to be robust even when localized and diffuse noise are present in the microphone signals; accurate DOA estimation and noise reduction in the dereverberated signal are achieved. Notice that this paper will not focus on the computational complexity of the proposed algorithm which is rather large and could be effectively reduced by employing parallel computing and fast transformations [28], [29]. Instead, the aim of the paper is to to introduce a novel approach and to compare it qualitatively to state-of-the-art algorithms in a variety of scenarios.

This paper is organized as follows: in Section II the PWDM is described. Section III describes the inverse problem that is used to perform the sound field approximation. In Section IV the ADeLFi algorithm is presented describing the optimization algorithm, showing the WOLA procedure and the regularization tuning strategy. Finally, in Section V the algorithm is validated using simulated and real measurements and in Section VI conclusions are drawn.

Preliminary results have been presented in [26]. The main novelties of this paper are: (i) a novel processing of the weight signals to reduce artifacts in the dereverberated signals, (ii) modifications to the proposed algorithm which allow the possibility of tracking the position of a moving sound source, (iii) a comparison with state-of-the-art dereverberation and DOA estimation algorithms, (iv) the inclusion of more objective perceptual performance measures and (v) the validation of the proposed algorithm using real measurements.

II. Acoustic model

A plane wave is defined as

$$\phi_{l,m}(f) = e^{ik_f \mathbf{n}_l^\intercal \mathbf{x}_m}, \tag{1}$$

and is the homogeneous solution of the Helmholtz equation. Here $\mathbf{x}_m$ is the the m-th microphone position, $\mathbf{n}_l$ is the unit vector indicating the direction of the l-th plane wave, f is the frequency in Hz, k_f is the wave number defined as $k_f = 2\pi f/c = \omega_f/c$, where c is the speed of sound. A sound field in a source-free volume can be represented by a finite weighted sum of plane waves coming from N_w different directions [13]:

$$p(\mathbf{x}, f)|_{\mathbf{x}=\mathbf{x}_m} \approx \sum_{l=0}^{N_w-1} \phi_{l,m}(f) w_l(f) \ \forall \mathbf{x}_m \in \Omega, \tag{2}$$

where the *weight* $w_l(f)$ is a complex scalar that weights the l-th plane wave at the frequency f. Equation (2) describes the plane wave decomposition method (PWDM). This equation can be generalized for N_m discrete positions $\mathbf{x}_m \in \Omega$ and N_f discrete frequencies:

$$\mathbf{P} = \mathsf{D}\mathbf{W}, \tag{3}$$

where $\mathbf{P} \in \mathbb{C}^{N_f \times N_m}$ is a matrix in which the m-th column is the discrete Fourier transform (DFT) of the sound pressure signal $p(\mathbf{x}, n)|_{\mathbf{x}=\mathbf{x}_m}$ at a particular time window (snapshot) and $\mathbf{W} \in \mathbb{C}^{N_f \times N_w}$ is a matrix containing the weights $w_l(f)$. The linear mapping $\mathsf{D} : \mathbb{C}^{N_f \times N_w} \to \mathbb{C}^{N_f \times N_m}$ represents a *dictionary* of plane waves. In this paper D will be constructed such that $N_w > N_m$, leading to an *over-complete* dictionary. Equation (3) should not be confused with a linear matrix equation, *i.e.*, D is indeed a mapping rather than a matrix multiplier.

The dictionary D can be also viewed as a dictionary of *steering vectors*. In particular, the row of $\mathbf{P}$ corresponding to the f-th frequency can be expressed as

$$[p_{f,0}, \dots p_{f,N_m-1}]^\intercal = \mathbf{A}_f [w_{f,0}, \dots w_{f,N_w-1}]^\intercal, \tag{4}$$

where $\mathbf{A}_f \in \mathbb{C}^{N_m \times N_w}$ is a matrix having in its columns steering vectors, commonly referred to as *sensing matrix*. Notice that in the following D will indicate the PWDM. Other acoustic models could be employed as well, such as acoustic models of microphones mounted on spherical rigid baffles [18] or of human heads using head-related transfer function (HRTF) [24].

III. The inverse problem

Consider a single sound source in the far field and a set of N_m microphones. It is assumed that the microphones are far from any scattering object and have the sound source in their line of sight. Additionally, it is assumed that $\mathbf{P}$ can be decomposed as follows:

$$\mathbf{P} = \mathbf{P}_e + \mathbf{P}_d. \tag{5}$$

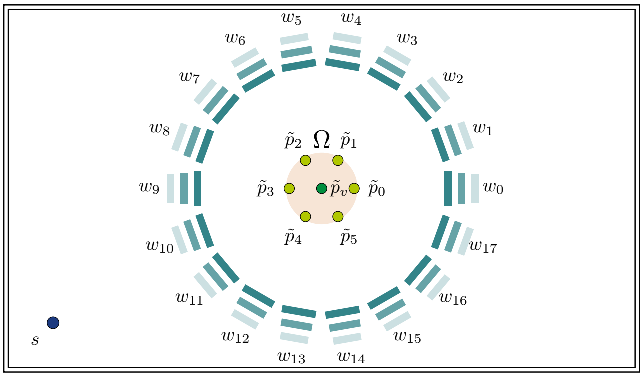

Figure 1. Horizontal cross-section of a room. A sound source in the front left corner generates a reverberant sound field. A spherical microphone array (light green dots) measures the sound field of the room. A set of plane waves (depicted with three lines) represents the acoustic model that is used to match the sound pressure measured by the microphones. The sound field is approximated in the shaded volume Ω. An additional microphone is placed at the center of the array to validate the quality of the approximation.

The first term $\mathbf{P}_e$ represents early reflections which are generated by a relatively small number of plane waves. The early reflections component $\mathbf{P}_e$ also includes the line of sight which is assumed to be the plane wave with the strongest energy. This plane wave is associated with the location of the sound source in terms of DOA. The second component $\mathbf{P}_d$ represents the diffuse sound field and consists of uncorrelated plane waves arriving from a large number of directions. The microphones are positioned at the boundary of the source-free volume $\Omega \in \mathbb{R}^3$ as depicted in Figure 1. The directions of the plane waves can be selected from a spherical lattice centered at the center of the microphone array. In this paper a Fibonacci lattice [30] is used to provide a nearly uniform sampling of the surface of a sphere. Notice that other types of lattices with even more uniform spherical sampling exist [31], [32]. In order to achieve accurate approximation, a large number of plane waves must be used.

The aim is to approximate the sound field inside this volume to jointly dereverberate and localize the sound source. What is sought by the inverse problem is to estimate the weight signals that lead to the optimal sound field approximation. This can be achieved by matching the microphone measurements with the acoustic model described in Section II. An interior Dirichet problem can be formulated as the following optimization problem [33]:

$$\mathbf{W}_r^\star = \underset{\mathbf{W}}{\text{argmin}} \quad q(\mathbf{W}) = \frac{1}{2} \|\mathsf{D}\mathbf{W} - \tilde{\mathbf{P}}_r\|_F^2, \tag{6}$$

where $\| \cdot \|_F$ is the Frobenius norm and the columns of the matrix $\tilde{\mathbf{P}}_r$ contain the microphone measurements, *i.e.*, the N_f-long measured complex sound pressure at the r-th time window.

Problem (6) is heavily *ill-posed*. This will in general lead to *over-fitting*: the measured sound pressure will coincide with the sound pressure of the acoustic model but only at the microphone positions, resulting in a poor sound field approximation in other positions. Additionally, (6) can be *ill-conditioned*: it is possible that some of the elements of

$\mathbf{W}_r^\star$ become unbounded when their corresponding frequency coincides with the eigenvalues of the interior Dirichet problem [33]. This instability is also known as the forbidden frequency problem [34, §8.10.2]. To avoid both instability and over-fitting, a *regularization term* g is added to to q, *i.e.*:

$$\mathbf{W}_r^\star = \underset{\mathbf{W}}{\operatorname{argmin}} \quad q(\mathbf{W}) + \lambda g(\mathbf{W}), \tag{7}$$

where λ is a scalar often referred to as *hyperparameter*. The regularization term g acts as a soft constraint limiting the magnitude of $\mathbf{W}$ (instability) and avoiding $q(\mathbf{W})$ becoming too small (over-fitting). A possible choice for g is the *sum of l_2-norms regularization* corresponding to

$$g(\mathbf{W}) = \sum_{l=0}^{N_w-1} \|\mathbf{w}_l\|_2, \tag{8}$$

where $\mathbf{w}_l$ indicates the l-th column of $\mathbf{W}$. This regularization consist of a convex function, often referred to as $l_{2,1}$ *mixed norm* with the notation $\|\cdot\|_{2,1}$. If a large value of λ is used, *group sparsity* is promoted and only few columns of $\mathbf{W}_r^\star$ will be non-zero. In practice, this would mean only few plane waves being active, meaning that the sum of l_2-norms regularization promotes *spatial sparsity*. Another common regularization is the l_1*-norm regularization* corresponding to

$$g(\mathbf{W}) = \|\text{vec}(\mathbf{W})\|_1 = \sum_{l=0}^{N_w-1} \|\mathbf{w}_l\|_1, \tag{9}$$

which is a convex function that promotes *sparsity* in $\mathbf{W}_r^\star$, *i.e.*, only few elements of the matrix to be non-zero. When a frequency domain acoustic model is used, as in the case of the PWDM, *spatio-spectral sparsity* is promoted. Notice that unlike the sum of l_2-norms, the l_1-norm promotes spatial sparsity but it fails to do it consistently between different frequencies. This is due to the fact that non-zero elements are not constrained to belong to any particular column.

The choice of these sparsity promoting regularization terms is motivated by the fact that $\mathbf{P}_e$, consists of a *sparse* set of plane waves arriving from a limited number of directions. However, (7) aims at reconstructing the whole sound field $\mathbf{P}$, which due to the presence of the diffuse field component $\mathbf{P}_d$ is not a sparse set of plane waves. The parameter λ should be tuned such that both $\mathbf{P}_e$ and $\mathbf{P}_d$ are jointly reconstructed with accuracy while preserving a sufficient level of sparsity to enable joint localization and dereverberation. As it will be described in Section IV-C, an additional microphone will be used to tune λ and find the best balance between sparsity and sound field approximation. There exist other types of regularization that can promote sparsity within group sparsity: these can enable both the presence of few non-zero columns in $\mathbf{W}_r^\star$ and let these columns be sparse vectors [20], [35]. However, these types of regularization are not treated in this paper and left for future work as they may lead to nonconvex problems or to the nontrivial tuning of multiple hyperparameters.

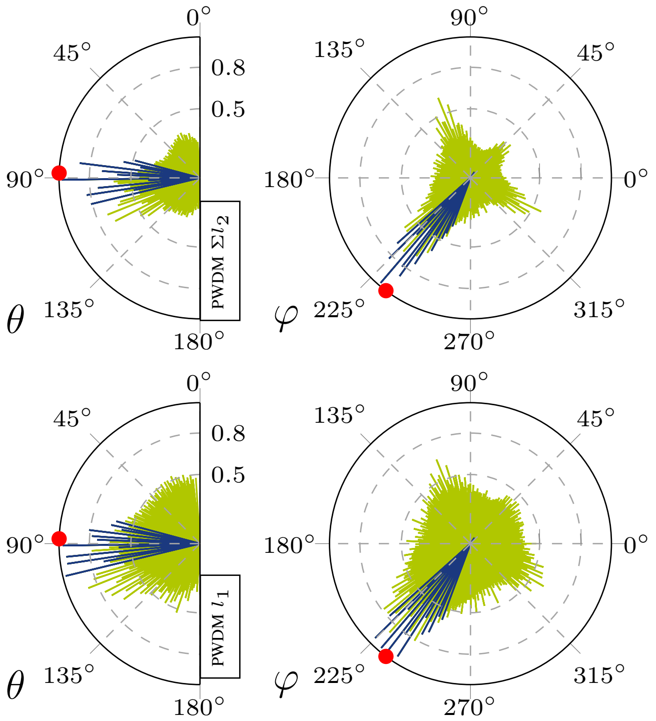

Figure 2. Visualization of the mean of the energy of the weight signals, i.e. $1/N_r \sum_{r=0}^{N_r-1} \|\mathbf{w}_{r,l}^\star\|_2^2$, as a function of the azimuth angle φ and polar angle θ. The red dots represent the true source position. Darker lines represent the N_b neighbor weight signals around the estimated DOA. The weight signals are obtained from simulation results using $N_w = 500$ plane wave directions and $N_m = 16$ microphones with a sensor noise of 40 dB SNR. These results will be presented in detail in Section V-A.

Once a solution is obtained, the DOA of the sound source can be inferred by finding the weight signal with the *strongest energy*:

$$\|\mathbf{w}_{r,b}^\star\|_2^2 = \max\left\{\|\mathbf{w}_{r,0}^\star\|_2^2, \dots, \|\mathbf{w}_{r,N_w-1}^\star\|_2^2\right\}, \tag{10}$$

since the b-th weight signal is associated with a plane wave direction of the Fibonacci lattice and hence to a polar angle $\theta_r^\star$ and azimuth angle $\varphi_r^\star$. Here $\mathbf{w}_{r,l}^\star$ is the l-th column of the solution of (7) of the r-th snapshot $\mathbf{W}_r^\star$. Figure 2 shows the mean of the energy of the weight signals of different snapshots as a function of the azimuth and polar angles for the simulation results presented in Section V-A. A clear maximum is visible towards the direction of the sound source shown by the red dot.

A dereverberated signal is also readily available: the weight signal $\mathbf{w}_{r,b}^\star$ will have less reverberation compared to the microphone signals. That is because $\mathbf{w}_{r,b}^\star$ accounts only for a single plane wave which contributes to the sound field from a specific direction. If a sufficient level of sound field approximation is reached, the effect of reverberation is *spatially distributed* among the plane waves of the acoustic model. Promoting sparsity is fundamental since the prior knowledge given by the sparsity regularization term biases the solution $\mathbf{W}_r^\star$ towards a better sound field approximation of the early reflection component $\mathbf{P}_e$, rather than of the diffuse sound field $\mathbf{P}_d$. This bias enables DOA estimation using (10). Nevertheless, if spatial sparsity is promoted too strongly, reverberation would not be spatially distributed among the plane waves resulting in only few plane waves trying to approximate the entire sound field $\mathbf{P}$. This would result in $\mathbf{w}^\star_{r,b}$ strongly contributing to $\mathbf{P}_d$ and hence containing a level of reverberation close to the one of the unprocessed microphone signals. This condition can be avoided when spatial distribution of reverberation among the plane waves is achieved, that is when λ is properly tuned.

However, since a finite number of directions is used, it is possible that other plane waves with similar directions to the plane wave corresponding to $\mathbf{w}^\star_{r,b}$ contribute significantly to the sound field generated from that particular direction. It is therefore advantageous to employ multiple weight signals to produce the dereverberated signal. This can be achieved by selecting N_b weight signals that are the nearest neighbors of the plane wave directions corresponding to $\mathbf{w}^\star_{r,b}$ in the Fibonacci lattice. These weight signals, together with $\mathbf{w}^\star_{r,b}$, can be used as the columns of $\mathbf{W}^\star_{r,b} \in \mathbb{C}^{N_f \times (N_b+1)}$. The energy of these weight signals is visualized with darker lines in Figure 2. A dereverberated signal can then be obtained by generating a new sound field $\mathbf{P}_{r,b} = \mathsf{D}_{r,b}\mathbf{W}^\star_{r,b}$, using the same acoustic model employed in the inverse problem. Here, $\mathsf{D}_{r,b} : \mathbb{C}^{N_f \times (N_b+1)} \to \mathbb{C}^{N_f \times N_m}$ maps a smaller number of weight signals to $\mathbf{P}_{r,b}$ which now represents a new sound field created by a limited number of plane waves corresponding to the selected directions. Although here $\mathsf{D}_{r,b}$ utilizes the same microphone positions of the measurements, any position inside Ω can be chosen as well. This is achieved by setting different microphone positions during the construction of $\mathsf{D}_{r,b}$. As it will be shown in Section V, the artifacts present in the $\mathbf{P}_{r,b}$ signals are often less audible than those in $\mathbf{w}^\star_{r,b}$, particularly when spatio-spectral sparsity is promoted.

IV. The ADeLFi algorithm

In this section the proposed algorithm is presented, referred to as *acoustic dereverberation and localization through field approximation (ADeLFi)*. The pseudo-code of ADeLFi is given in Algorithm 1 and a detailed explanation is provided in the next subsections.

A. Optimization algorithm

Problem (7) is nonsmooth and can easily become of large scale. A well known algorithm that can address this type of problems is the *proximal gradient (PG) algorithm* which is a first-order optimization algorithm suitable for nonsmooth cost functions and having minimal memory requirements [36], [37]. The PG algorithm generalizes the gradient descent algorithm to a class of nonsmooth problems such as the problem in (7) where q is smooth and g is nonsmooth, as the regularization terms of (8) and (9). The PG algorithm consists of iterating

$$\mathbf{W}^{s+1} = \mathrm{prox}_{\gamma\lambda g}\left(\mathbf{W}^s - \gamma \nabla q(\mathbf{W}^s)\right), \tag{11}$$

starting from an initial guess $\mathbf{W}^0$. Here ∇q is the gradient of q, γ is the step-size, and $\mathrm{prox}_{\gamma g}$ is the *proximal mapping* of the function g [36].

For the regularization terms described in (8) and (9) the proximal mapping consists of a computationally cheap operation. If $g(\mathbf{W}) = \|\mathrm{vec}(\mathbf{W})\|_1$, its proximal mapping reads:

$$\mathrm{prox}_{\gamma\lambda g}(\mathbf{W}) = \mathcal{P}_+(\mathbf{W} - \lambda\gamma) - \mathcal{P}_+(-\mathbf{W} - \lambda\gamma), \tag{12}$$

where $\mathcal{P}_+$ is the element-wise mapping performing $\max\{0, |\cdot|\}$ with $|\cdot|$ indicating the modulus of a complex number. On the other hand, if $g(\mathbf{W}) = \sum_{l=0}^{N_w-1} \|\mathbf{w}_l\|_2$, the proximal mapping becomes:

$$\begin{aligned}&\mathrm{prox}_{\gamma\lambda g}(\mathbf{W}) = \\ &[\mathbf{w}_0 \mathcal{P}_+(1 - \tfrac{\lambda\gamma}{\|\mathbf{w}_0\|_2}) \dots \mathbf{w}_{N_w-1} \mathcal{P}_+(1 - \tfrac{\lambda\gamma}{\|\mathbf{w}_{N_w-1}\|_2})].\end{aligned} \tag{13}$$

In both cases the proximal mapping performs a soft-thresholding of either the elements of $\mathbf{W}$ (l_1-norm) or its columns (sum of l_2-norms) which is a computationally simple operation.

Another fundamental operation needed in the PG algorithm is the evaluation of the gradient of q which is given by

$$\mathbf{J} = \nabla q(\mathbf{W}) = \mathsf{D}^*(\underbrace{\mathsf{D}\mathbf{W} - \tilde{\mathbf{P}}_r}_{\mathbf{R}}), \tag{14}$$

where $\mathbf{J}$ is a matrix with the gradient with respect to $\mathbf{w}_l$ in the l-th column, $\mathbf{R}$ is the residual matrix, *i.e.*, the difference between the sound pressure recorded by the microphones and the sound pressure predicted by the acoustic model, and $\mathsf{D}^* : \mathbb{C}^{N_f \times N_m} \to \mathbb{C}^{N_f \times N_w}$ is the *adjoint mapping* of D. In this context, D is often referred to as the *forward mapping*. The adjoint mapping is a generalization of the transpose of a matrix to linear mappings. In general, when D is a linear mapping between two large finite-dimensional spaces, it is not ideal to use a matrix-vector multiplication based algorithm for its evaluation, as this in fact would require the storage of a very large matrix. Instead, it is possible to compute the mapping using its definition, *i.e.*, by directly applying (2) iteratively or by utilizing fast transformations. The same strategy can be adopted for the adjoint mapping whose definition is similar to the definition of the forward mapping. The adjoint mapping of the PWDM is obtained as the cross-spectrum between $\hat{\phi}_{l,m}$ and $\hat{r}_m$:

$$j_l(f) = \sum_{m=0}^{N_m-1} \phi^*_{l,m}(f) r_m(f), \tag{15}$$

where $\hat{r}_m$ is the residual of the m-th microphone in the frequency domain. In both cases j_l indicates the complex signal appearing in the l-th column of $\mathbf{J}$. The gradient at the iterate $\mathbf{W}^s$ can be efficiently computed together with the evaluation of $q(\mathbf{W}^s)$: this strategy is also known as back-propagation in machine learning [38] or automatic differentiation [39] and leads to *matrix-free optimization* [37], [40].

Finally, an accelerated variant of the PG algorithm is used employing a limited-memory quasi-Newton method [27]. An implementation of the algorithm written in the Julia language is also available online [41].

B. Weighted overlap-add procedure

Solving the optimization problem in (7) using long time windows is not feasible since evaluating the linear mapping

Algorithm 1 Acoustic Dereverberation and Localization through Field Approximation (ADeLFi) method

```
 1: Inputs:
 2: $\tilde{\mathbf{P}}_t \in \mathbb{R}^{N_t \times N_m}$, $\tilde{\mathbf{p}}_{t,v} \in \mathbb{R}^{N_t}$
 3: $g$, $N_w$, $N_\tau$, $N_b$, $N_o$, $\mathbf{u}_a$, $\mathbf{u}_s$, $\beta \in (0,1]$, $\eta$
 4: Outputs:
 5: $\bar{\mathbf{P}}_t \in \mathbb{R}^{N_t \times N_m}$, $\bar{\mathbf{w}}_t \in \mathbb{R}^{N_t}$, $\boldsymbol{\varphi}^\star \in \mathbb{R}^{N_r}$, $\boldsymbol{\theta}^\star \in \mathbb{R}^{N_r}$
 6: Construct the acoustic model
 7: $\mathsf{D} : \mathbb{C}^{N_f \times N_w} \to \mathbb{C}^{N_f \times N_m}$
 8: $\mathsf{D}_v : \mathbb{C}^{N_f \times N_w} \to \mathbb{C}^{N_\tau}$ using $N_w$ directions
 9: Compute candidate angles $\breve{\boldsymbol{\varphi}} \in \mathbb{R}^{N_w}$, $\breve{\boldsymbol{\theta}} \in \mathbb{R}^{N_w}$
10: Set $k = 0$, $r = 0$, $\epsilon_{v,-1} = +\infty$, $\bar{\mathbf{e}} = \mathbf{0} \in \mathbb{R}^{N_w}$,
11: $\bar{\mathbf{P}}_t = \mathbf{0}$, $\bar{\mathbf{w}}_t = \mathbf{0}$, $\boldsymbol{\varphi}^\star = \mathbf{0}$, $\boldsymbol{\theta}^\star = \mathbf{0}$.
12: while $k + N_\tau \leq N_t$ do
13:     Select samples of $r$-th snapshot and weight with $\mathbf{u}_a$
14:     $\tilde{\mathbf{P}}_r \leftarrow \mathsf{F}\mathsf{S}_k\tilde{\mathbf{P}}_t$, $\tilde{\mathbf{p}}_{r,v} \leftarrow \mathsf{F}[p_{v,t}(k) \dots p_{v,t}(k+N_\tau-1)]^\intercal$
15:     where $\mathsf{S}_k : \mathbb{R}^{N_t \times N_m} \to \mathbb{R}^{N_\tau \times N_m}$ selection operator
16:     and $\mathsf{F} : \mathbb{R}^{N_\tau \times N_m} \to \mathbb{C}^{N_f \times N_m}$ real DFT
17:     $\lambda \leftarrow 10^{-6}\lambda_{\text{max}}$
18:     for $z = 0, \dots, N_\lambda - 1$ do
19:         $\mathbf{W}_r^{\star,z} \leftarrow \underset{\mathbf{W}}{\text{argmin}} \frac{1}{2}\|\mathsf{D}\mathbf{W} - \tilde{\mathbf{P}}_r\|_F^2 + \lambda g(\mathbf{W})$
20:         $\epsilon_{v,z} \leftarrow \|\mathsf{D}_v\mathbf{W}_r^{\star,z} - \tilde{\mathbf{p}}_{r,v}\|_2^2 / \|\tilde{\mathbf{p}}_{r,v}\|_2^2$
21:         if $\epsilon_{v,z} > \epsilon_{v,z-1} + \eta$ then break
22:         else
23:             $\bar{\mathbf{W}}_r \leftarrow \mathbf{W}_r^{\star,z}$
24:             increase $\lambda$ logarithmically
25:     $\mathbf{e}_r \leftarrow [\|\bar{\mathbf{w}}_{r,0}\|_2^2, \dots, \|\bar{\mathbf{w}}_{r,N_w-1}\|_2^2]^\intercal$,
26:     $\bar{\mathbf{e}} \leftarrow \mathbf{e}_r + \beta\bar{\mathbf{e}}$
27:     Set $b$ as the index of the maximum element of $\bar{\mathbf{e}}$
28:     Set $\varphi_r^\star \leftarrow \breve{\varphi}_b$, $\theta_r^\star \leftarrow \breve{\theta}_b$
29:     Weight $\mathsf{F}^{-1}\bar{\mathbf{w}}_{r,b}^\intercal$ with $\mathbf{u}_s$ and append to $\bar{\mathbf{w}}_t$
30:     Find $N_b$ neighbors plane waves
31:     Construct $\mathsf{D}_{r,b}$ and $\bar{\mathbf{W}}_{r,b}$
32:     Compute $\bar{\mathbf{P}}_{r,b} = \mathsf{D}_{r,b}\bar{\mathbf{W}}_{r,b}$
33:     Weight $\mathsf{F}^{-1}\bar{\mathbf{P}}_{r,b}$ with $\mathbf{u}_s$ and append to $\bar{\mathbf{P}}_t$
34:     $r \leftarrow r + 1$, $k \leftarrow k + N_\tau - N_o$
```

D and its adjoint becomes too costly and the optimization problem becomes too large. Additionally, it is well known that speech is sparse in the short-time Fourier transform (STFT) domain. Therefore, a weighted overlap-add (WOLA) procedure is used for processing single-snapshots (SSs): the N_t-long microphone time-domain signals appearing in the column of the matrix $\tilde{\mathbf{P}}_t \in \mathbb{R}^{N_t \times N_m}$ are split into N_r frames of N_τ samples each. An analysis window function is applied to the r-th frame which is then converted to a complex signal by applying a real DFT (Line 14). If $N_\tau = 512$ and $N_w = 500$ the optimization problem will then have $N_f N_w = 128.5 \times 10^3$ complex optimization variables, which is manageable. Notice that only $N_f = \lfloor N_\tau/2 \rfloor + 1$ frequencies need to be processed since a real DFT that exploits Hermitian symmetry is used. Analysis and synthesis window functions, here chosen to be both square-rooted Hann windows, are indicated in Algorithm 1 with $\mathbf{u}_a$ and $\mathbf{u}_s$ respectively. The frames are overlapped by N_o samples: here an overlap of 50% is used.

The size of the volume Ω, and hence the microphone array geometry, imposes a constraint on the frame length N_τ. Assuming that a common phase shift is introduced in (2) such that all of the acoustic delays of the plane waves are causal, the following inequality must be satisfied:

$$\frac{cN_\tau}{F_s} > \max\{\|\mathbf{x} - \mathbf{y}\|_2 \mid \mathbf{x}, \mathbf{y} \in \Omega\}. \tag{16}$$

This means that the length of the frame should at least allow for the plane waves to reach all of the microphones. In practice it is better to choose a longer frame in order to minimize the duration of the transient which should correspond only to a short initial part of the frame. If this is the case the effect of the transient will then be effectively canceled by the weighting and averaging operations of the WOLA procedure. The above inequality also suggests that the use of a microphone array scattered in a large volume should be avoided when using the ADeLFi algorithm. In the following, a frame length of $N_\tau = 512$ will be used, corresponding to time window of 64 ms at $F_s = 8$ kHz for all the microphone array configurations used in the simulations and real measurements of Section V. This frame length always significantly exceeds the lower bound (16).

C. Tuning of parameter λ

The parameter λ, scaling the regularization term g in (7), controls the level of regularization and should be tuned properly. One of the most popular tuning strategies is the K-fold cross validation (KCV): this involves solving the optimization problem multiple times with different values of λ and K *folds* (partitions) of the available data. However, this strategy is not ideal for online audio processing: if N_λ candidate values for λ are used, it is then required to solve KN_λ optimization problems per frame.

Therefore, a novel simplification of the KCV strategy is adopted as follows. An additional microphone, referred to as *validation microphone*, is positioned inside the volume Ω to record the time-domain sound pressure $\tilde{\mathbf{p}}_{t,v}$ and validate the quality of the approximation. For each frame, the optimization problem is solved multiple times using different values of λ. In Algorithm 1 the best λ is chosen in each frame inside the for-loop with counter z (Line 18). A low level of regularization is initially used, *i.e.*, λ is first chosen to be $10^{-6}\lambda_{\text{max}}$, where λ_{max} is the value for which $\mathbf{W}_r^\star = \mathbf{0}$ [19]. Hence the first optimization problem utilizes a low level of regularization and is initialized with a null initial guess. Once a solution is obtained, the *validation error* $\epsilon_{v,0}$ is computed (Line 20), namely the normalized mean squared error (NMSE) between the validation microphone frequency-domain signal $\tilde{\mathbf{p}}_{r,v}$ in the r-th frame, weighted as well by the square-rooted Hann window, and the validation microphone sound pressure signal predicted by the acoustic model when regularized by λ. In practice $\epsilon_{v,0}$, gives a measure of the quality of the approximation, which for small values of λ is expected to be poor due to over-fitting. The problem is then solved again after increasing λ logarithmically.

This is also warm-started using the previous solution which helps in reducing the number of iterations of the current optimization problem. This procedure is stopped once the validation error stops decreasing, $\epsilon_{v,z} > \epsilon_{v,z-1} + \eta$, namely when the regularization ceases to be beneficial in terms of the quality of the approximation. Here $\eta = 10^{-4}$ is a small value that prevents the procedure being stopped too early if $\epsilon_{v,z}$ and $\epsilon_{v,z-1}$ are very close. Finally the solution with the best regularization parameter is chosen, that is $\mathbf{W}_r^{\star,z-1}$ which was copied to $\bar{\mathbf{W}}_r$ during the previous iteration. The optimization algorithm solving the problems in Line 19 is stopped whenever the number of iterations exceeds 200 or when the following condition is satisfied $\|\mathrm{vec}(\mathbf{W}^s - \mathbf{W}^{s-1})\|_\infty/\gamma < 10^{-3}$.

D. DOA estimation and dereverberation

The last part of Algorithm 1 (starting from Line 25) consists of estimating the DOA and obtaining a dereverberated signal from the weight signals. As described in Section III, the DOA can be inferred from the weight signal with strongest energy. However, the frame-based WOLA procedure offers the possibility of estimating DOAs in different time windows allowing the localization of a *moving sound source*.

After processing the r-th frame of the microphone signals to obtain $\bar{\mathbf{W}}_r$, the energy of each of its columns can be calculated and stored in a vector $\mathbf{e}_r \in \mathbb{R}^{N_w}$. The index of the maximum element of $\mathbf{e}_r$ will then correspond to the weight signal with strongest energy corresponding to the DOA at the r-th frame. However, it is possible to include the *memory* of the previous DOA estimates by performing a recursive averaging with forgetting factor β in order to give more weight to the latest estimates (Line 26). This will prevent abrupt changes of the DOA estimates. An index b will then be retrieved (Line 27), which can be used to obtain the azimuth and polar angles $\varphi_r^\star$ and $\theta_r^\star$ of the r-th frame DOA (Line 28). These angles are selected out of the candidate angles stored in $\breve{\boldsymbol{\varphi}} = [\breve{\varphi}_0, \ldots, \breve{\varphi}_{N_w-1}]^\intercal$ and $\breve{\boldsymbol{\theta}} = [\breve{\theta}_0, \ldots, \breve{\theta}_{N_w-1}]^\intercal$ which correspond to spherical coordinates with origin at the center of the microphone array, obtained from the Fibonacci lattice (Line 9). This enables *tracking* the DOA of a moving sound source, *i.e.*, creating the vectors $\boldsymbol{\varphi}^\star$ and $\boldsymbol{\theta}^\star$ which contain the estimated azimuth and polar angles of each frame, *i.e.*, $\varphi_r^\star$ and $\theta_r^\star$ for $r = 0, \ldots, N_r - 1$, respectively.

Once the b-th weight signal is chosen, this can be appended to the dereverberated time-domain signal $\bar{\mathbf{w}}_t$ (Line 29). Alternatively, as described in Section III, the weight signals of the N_b plane waves with nearest neighbor directions of the one corresponding to $\bar{\mathbf{w}}_{r,b}$ can also be used to produce dereverberated signals. The neighbor weight signals can be selected together with $\bar{\mathbf{w}}_{r,b}$ to construct $\bar{\mathbf{W}}_{r,b} \in \mathbb{R}^{N_\tau \times (N_b+1)}$ (Line 31). Once the the acoustic model $\mathsf{D}_{r,b} : \mathbb{R}^{N_\tau \times (N_b+1)} \to \mathbb{R}^{N_\tau \times N_m}$ is constructed, the selected weight signals $\bar{\mathbf{W}}_{r,b}$ can be used to generate the sound pressure signals $\bar{\mathbf{P}}_{r,b}$ at the microphone position (or at any other positions inside Ω), corresponding to a new sound field with sound waves arriving only from a limited number of directions. Similarly to what is performed for $\bar{\mathbf{w}}_t$ in Line 29, dereverberated signals can then be obtained by appending $\bar{\mathbf{P}}_{r,b}$ to $\bar{\mathbf{P}}_t$ (Line 33). Once all of the frames are processed, the columns of $\bar{\mathbf{P}}_t$ will consist of time-domain N_t-long dereverberated signals, typically with less pronounced audio artifacts than those in $\bar{\mathbf{w}}_t$.

V. Results

A. Simulation results

In this section, results of simulations using the ADeLFi algorithm are presented. A reverberant shoebox-shaped room with dimensions $[L_x, L_y, L_z] = [7.34, 8.09, 2.87]$ m and reverberation time of $T_{60} = 1$ s is modeled using the randomized image method (RIM) [42]. The sound source is placed in the front left corner of the room ($\mathbf{x}_s = [L_x/8,\ L_x/8,\ 1.6]$ m), and a sampling frequency of $F_s = 8$ kHz is used. An anechoic sound sample of 5.3 s of male speech from [43] is convolved with the RIRs to simulate the microphone signals. The microphones are positioned to form a spherical microphone array with a radius of 10 cm, centered at $\mathbf{x}_c = [4.4,\ 5.7,\ 1.4]$ m, position at which the validation microphone is also set. Here, $N_w = 500$ plane wave directions are used. This number is chosen empirically: a lower number leads to a reduction of the performances and a higher number does not particularly change the results while increasing the computational load of the algorithm. Three different scenarios are compared: (i) sensor noise only, (ii) diffuse babble noise generated using the technique proposed in [44] with a SNR of 10 dB (iii) localized noise from a white source signal placed at $[7/8L_x,\ L_x/8,\ 1.6]$ m with a SNR of 15 dB. Spatially incoherent white noise is added with a SNR of 40 dB to simulate sensor noise in all cases. The validation microphone signal is also corrupted by these types of noise. Here, since only static sound sources are used to simulate the microphone signals, a forgetting factor of $\beta = 1$ is employed.

Figure 3(a) shows the median $\bar{\epsilon}_v$ of the validation errors: almost identical validation errors are achieved using ADeLFi with either sum of l_2-norms (spatial sparsity) or l_1-norm (spatio-spectral sparsity) regularization. The performance is slightly worse in the case of diffuse babble noise and decreases for the case of localized noise.

Remarkably, even if the diffuse babble noise has lower SNR than the localized noise, better performance is achieved in the former case. This is due to the different nature of the noise. The diffuse babble noise is highly spatially correlated at low frequencies where most of its energy lies. ADeLFi seems capable of effectively approximating this diffuse noise field as proven by the low NMSE shown in Figure 3(a). On the other hand, the localized noise is white meaning it generates a full band diffuse sound field in such reverberant environment. At high frequencies the noise is spatially uncorrelated making it more difficult to approximate this sound field due to the lack of spatial correlation.

As the lower plots of Figure 3(b) show, all the ADeLFi configurations achieve good localization even when only 4 microphones are used. Here a minimum angular distance of 4.5° is reached, corresponding to an angular similarity of $\sigma_\alpha = 0.97$, which is due to the finite number of plane wave directions. The localization performance is also compared with well established localization algorithms, namely the MUltiple

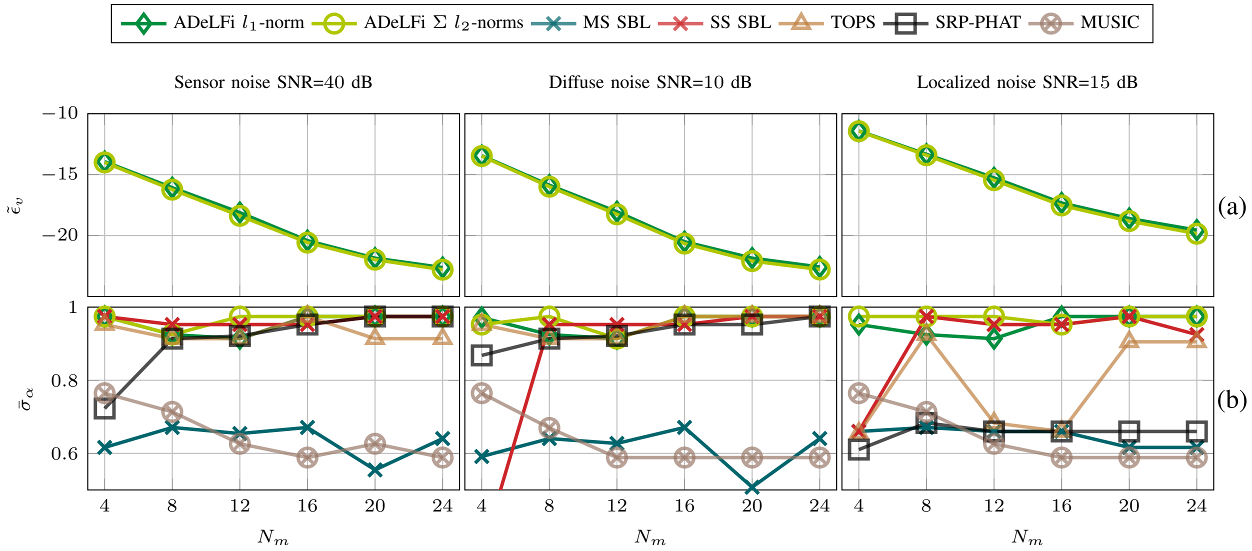

Figure 3. Median of the validation error ϵ_v in dB (a) and of the angular similarity σ_α (b) for different types of acoustic models and types of regularization of the ADeLFi algorithm as a function of the number of microphones (excluding the validation microphone). Notice that for SBL, MUSIC, TOPS and SRP-PHAT the validation microphone is included.

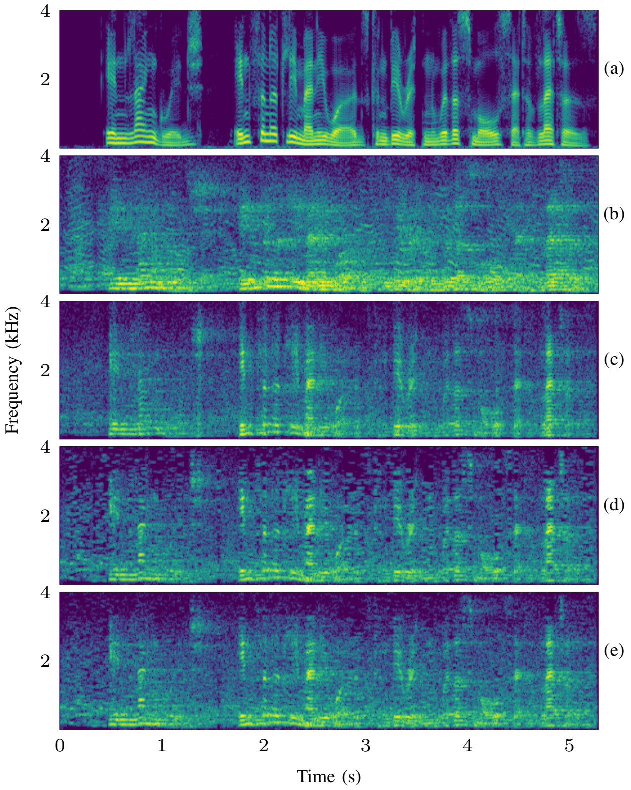

Figure 4. Spectrogram of anechoic speech signal (a), reverberant microphone signal (b), dereverberated signal obtained through ADeLFi with sum of l_2-norms regularization (spatial sparsity) (c), and with l_1-norm regularization (spatio-spectral sparsity) with single component ($\bar{\mathbf{w}}_t$) (d), and multiple components ($N_b + 1 = 12$) (e) using $N_m = 20$ microphones with diffuse babble noise (SNR = 10 dB).

SIgnal Classification (MUSIC) [45], test of orthogonality of projected subspaces (TOPS) [46] and steered response power-phase transform (SRP-PHAT) [47] the implementation of which is found in [48]. For a fair comparison, these algorithms use the same set of candidate directions given by the Fibonacci lattice used in ADeLFi. While MUSIC fails to retrieve the correct DOA due to the highly reverberant environment, TOPS and SRP-PHAT both achieve good localization in the sensor noise and diffuse babble noise scenario, achieving similar performance as ADeLFi. However, in the presence of localized white noise they are outperformed by all of the different configurations of ADeLFi. In fact in this noise scenario, both TOPS and SRP-PHAT often identify either the DOA of the noise source instead of the DOA of the speech source or something in between. Once more the localized white noise scenario is the most difficult one but ADeLFi proves itself robust achieving almost the same performance as in the other scenarios.

Additionally, ADeLFi is compared with a recently proposed algorithm that also performs joint dereverberation and DOA estimation [24]. In particular, the same acoustic model of ADeLFi can be employed in the algorithm of [24] which essentially utilizes a different strategy to tune the sparse regularization based on sparse Bayesian learning (SBL) to promote spatial sparsity. This algorithm, here referred to as SBL, can be employed using either a single-snapshot (SS) or a multi-snapshot (MS) approach. Using the same parameters of the simulation results of [24], the MS approach consists of processing groups of 8 ms time windows with 50% overlap. This results in the estimation of the DOA in longer time windows of 40 ms with 10% overlap. On the other hand, in the SS approach a single DOA is obtained for the whole signal. For the SS configuration, the same choice of time windows of ADeLFi is adopted (64 ms with 50% overlap). In Figure 3 SS SBL reaches similar localization performance to

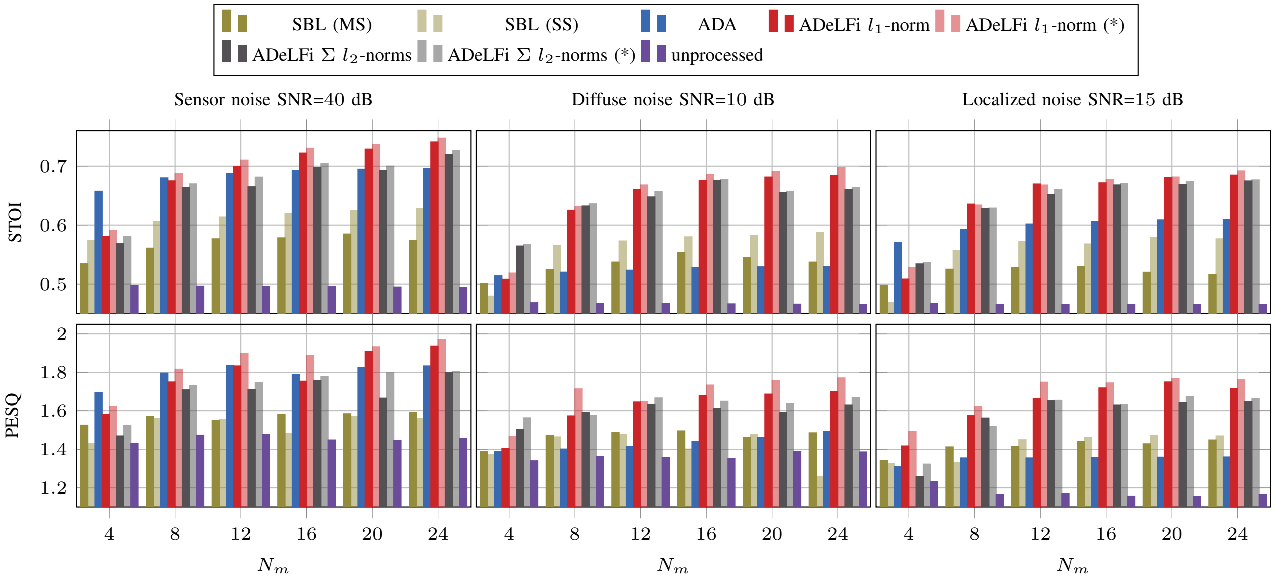

Figure 5. STOI and PESQ scores (RAW) for different types of acoustic models and regularizations of ADeLFi as a function of the number of microphones (excluding the validation microphone). Results with (*) correspond to dereverberated signals generated using multiple weight signals ($N_b + 1 = 12$).

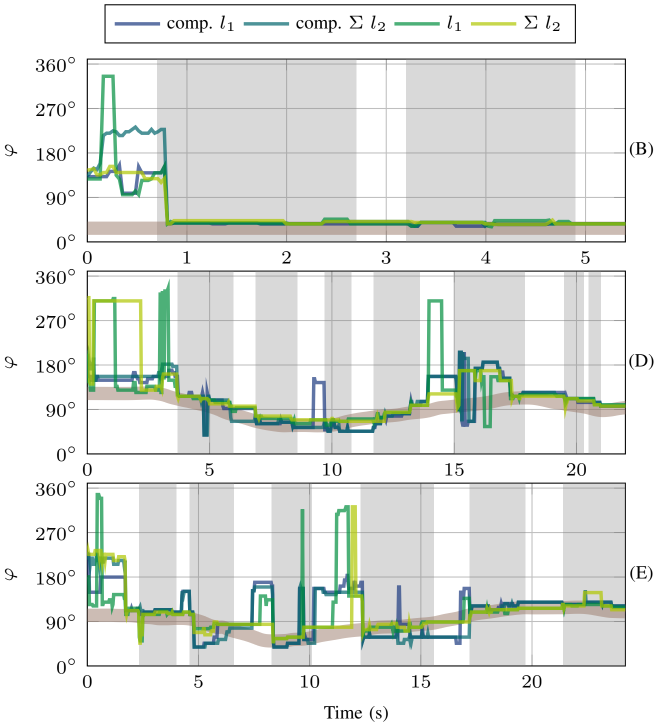

Figure 6. Estimated azimuth angle $\varphi^\star$ using measurements. Thick line indicates the ground truth. Gray areas visualize the time windows with voice activity.

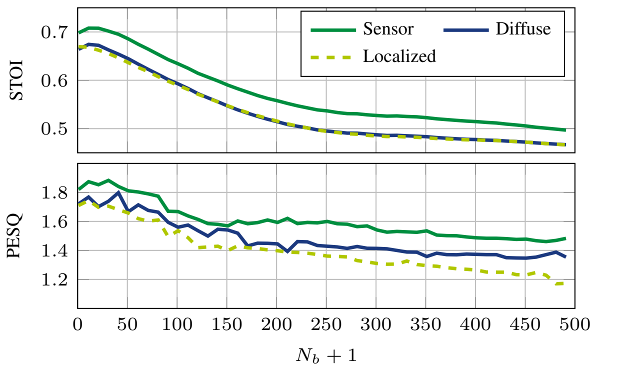

Figure 7. STOI and PESQ scores (RAW) for different choices of the number of weight signal N_b. Results of ADeLFi using l_1-norm regulatization and $N_m = 12$ microphones are shown for different types of noise: sensor noise (40 dB), diffuse noise (10 dB) and localized noise (15 dB).

ADeLFi. Poorer results are only reported for the case of diffuse and localized noise with 4 microphones. On the contrary, the MS SBL seems to fail to reach proper localization. However, it should be pointed out that, unlike in ADeLFi, MS SBL does not implement a recursive averaging of DOA estimates between consecutive frames. Therefore, the poor localization performance of the MS configuration is most likely due to the DOA estimates changing abruptly between time windows without voice activity. The use of a similar strategy such as the one described in Section IV-D or the employment of a voice activity detection (VAD) algorithm would possibly solve this issue.

Figure 4 compares the spectrogram of the dereverberated signals produced by ADeLFi with the original anechoic speech signal and one of the microphone signals. In the latter the presence of the babble noise can be seen as well as the speech components being smeared out by reverberation. Both reverberation and noise are effectively reduced by ADeLFi. Figure 4 (c) and (d) show the spectrogram of the dereverberated signal produced when using spatial sparsity (sum of l_2-norms) and with spatio-spectral sparsity (l_1-norm) respectively. In the latter figure spectral sparsity is particularly visible at higher frequencies. This effect results in audible artifacts, *i.e.*, the presence of musical noise in the dereverberated signal. This is effectively reduced when multiple weight signals are used

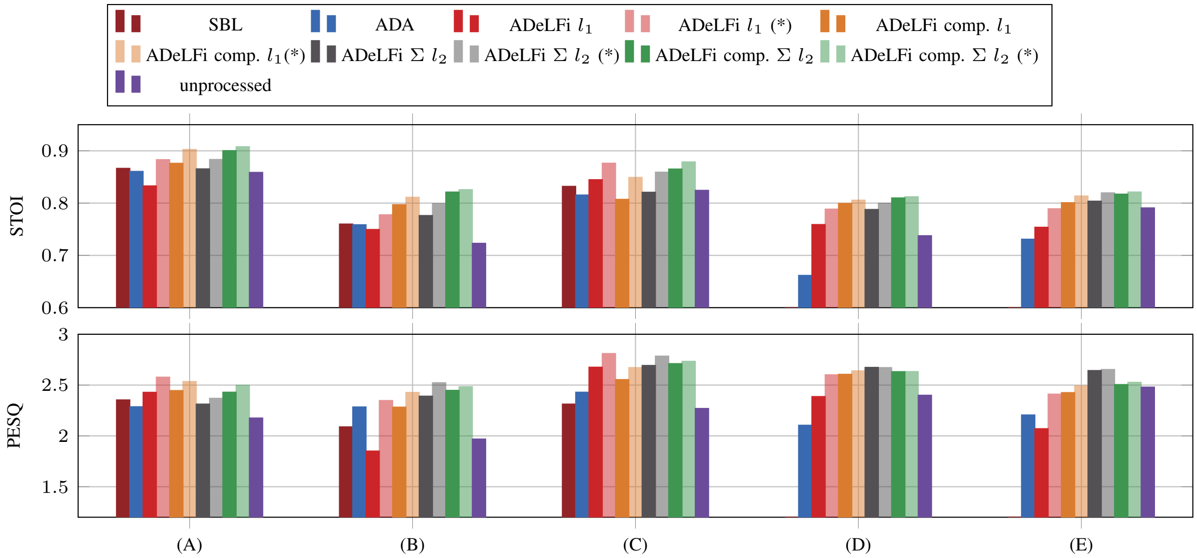

Figure 8. STOI and PESQ scores (RAW) for different types of acoustic models and regularizations of ADeLFi (shown in the legend with bold fonts) using measured data with a single static (a-c) and moving sound source (d-e). Results with (*) correspond to dereverberated signals generated using multiple weight signals ($N_b + 1 = 12$).

as show in Figure 4 (e) visualizing the spectrogram produced at the same microphone position using $N_b + 1 = 12$ weight signals with spatio-spectral sparsity. A clear reduction of the spectral sparsity is seen which leads to a substantial reduction of the musical noise.

Figure 5 shows speech intelligibility and quality scores obtained using the short-time objective intelligibility (STOI) [49] and perceptual evaluation of speech quality (PESQ) [50] respectively. These are in line with the results of Figure 3 with scores increasing with the number of microphones. It can be seen that ADeLFi with spatio-spectral sparsity achieves the best performances in particular when multiple weight signals are used. For all configurations, the results of ADeLFi outperform the ones of SBL achieving a higher level of dereverberation. The algorithms have comparable results only when 4 microphones are used. Notice that the unbiased estimate of SBL was used to produce the dereverberated signals as suggested in [24]. A comparison with sound samples dereverberated using the adaptive sparse MCLP-based speech dereverberation (ADA) [51] algorithm is also reported. This is a state-of-the-art dereverberation algorithm based on MCLP that does not require any prior knowledge on the DOA of the desired source and on the acoustic properties of the room. Here, ADA utilizes a forgetting factor of 1. ADA outperforms ADeLFi in the case of the sensor noise scenario reaching high scores with only 4 microphones. However, the performance of ADA deteriorates significantly in the other scenarios, particularly in the case of diffuse babble noise. On the contrary, ADeLFi is more robust and capable of performing *noise reduction* too: since the algorithm aims at approximating the diffuse noise field as well, it evenly distributes the noise energy among the active plane waves and hence selecting the weight signal with the strongest energy generally leads to a SNR increase. This can be seen in Figure 4 when comparing the spectrogram of the microphone signal (b) with the dereverberated signals produced by ADeLFi (b-e).

Figure 7 shows the scores for STOI and PESQ as function of the number of weight signals N_b. For large values of N_b the scores approach the ones of the unprocessed signals, while for values close to 1 the performance are similar to the one of $\bar{\mathbf{w}}_t$. It can be seen that in many cases an increase of the scores is present in the range $5 \leq N_b + 1 \leq 40$. Here, only the results using the l_1-norm regularization are shown for brevity: similar figures are reached for the sum of l_2-norms regularization. These results justify the choice of $N_b + 1 = 12$ which corresponds to selecting only 2.4% of the plane wave directions.

Informal listening tests indicate that dereverberated signals obtained by adopting either spatial sparsity or spatio-spectral sparsity based regularization are comparable with the latter having less audible distortions. The dereverberation effect increases as more microphone are used. When listening to the dereverberated signals, it is evident that the noise is reduced when compared to the microphone signals. Audio samples can be found in [52].

B. Results using real measurements

In this section the performance of the ADeLFi algorithm is validated using real measurements. The measurements are taken from the LOCATA challenge development database [53]. This database provides different recordings taken in a room with reverberation time of approximately $T_{60} = 0.15$ s. Here 5 different scenarios are tested: three scenarios with a static source (denoted in the figures and tables with (A), (B) and (C) and corresponding to recordings 1, 2 and 3 respectively of Task 1 of the LOCATA database) and two scenarios with a moving sound source (denoted (D) and (E) corresponding to recordings 1 and 3 respectively of Task 3 of the LOCATA database). For the static source scenarios loudspeaker sources (Genelec 1029A & 8020C) were used playing speech signals from the

	(A)		**(B)**		**(C)**		**(D)**		**(E)**	
	$\bar{\sigma}_\alpha$	$\bar{\epsilon}_v$ (dB)	$\bar{\sigma}_\alpha$	$\bar{\epsilon}_v$ (dB)	$\bar{\sigma}_\alpha$	$\bar{\epsilon}_v$ (dB)	$\bar{\sigma}_\alpha$	$\bar{\epsilon}_v$ (dB)	$\bar{\sigma}_\alpha$	$\bar{\epsilon}_v$ (dB)
ADeLFi l_1	0.96	-17.24	0.83	-16.77	0.94	-15.71	0.91	-12.79	0.9	-13.30
ADeLFi l_1 Comp.	0.99	-20.89	0.93	-20.84	0.94	-19.21	0.88	-15.14	0.89	-15.87
ADeLFi Σl_2	0.93	-17.24	0.88	-16.78	0.94	-15.71	0.9	-12.70	0.9	-13.26
ADeLFi Σl_2 Comp.	0.95	-20.84	0.88	-20.82	0.94	-19.10	0.88	-15.10	0.87	-15.83
SBL	0.99	-	0.96	-	0.9	-	-	-	-	-
SBL Comp.	0.96	-	0.96	-	0.85	-	-	-	-	-

Table I

Median of validation error ϵ_v and angular similarity σ_α between the estimated DOA and the ground truth during voice activity time windows for different scenarios of the LOCATA challenge using ADeLFi and SBL with and without compensation of the scattering field of the microphone array rigid baffle.

CSTR VCTK database [54]. The moving sound sources were created by people talking while walking around the room. The ground-truth positions of the speakers were measured using infra-red cameras (type Flex 13) using a tracking system (OptiTrack) with frame rate of 120 Hz [55]. All of the results presented here were obtained using a spherical microphone array of 32 microphones mounted on a rigid sphere with a radius of 4.2 cm (Eigenmike) [56]. Out of these recordings only $N_m = 15$ microphones are used in the ADeLFi algorithm. Two additional microphones are used as validation microphones. It is possible to compensate for the effect of the rigid baffle of the microphone array: the sound field scattered by the sphere can be effectively removed by applying a specific normalization in the spherical harmonic domain (SHD) [57]. This compensation is performed using the MATLAB code of [58], [59]. A sampling frequency of $F_s = 8$ kHz is used. The microphone recordings contain measurement noise and in some cases traffic noise as well coming from outside of the building (particularly in scenarios (B) and (D)). The forgetting factor is empirically chosen to be $\beta = 0.9$ for all the results presented here, including the static source scenarios. Here, $N_w = 500$ plane wave directions are used as well.

Figure 6 shows the estimated azimuth angle $\boldsymbol{\varphi}^\star$ as a function of time. The ground truth is visualized using a thick line. The grey areas in the plots visualize the time windows where voice activity is present. For the static scenarios, only case (B) is shown for brevity. It can be seen that, as soon as voice activity begins, all of the various configurations of ADeLFi succeed in finding a good estimate of the azimuth angle with similar performance. Figure 6 (D-E) shows the case of the moving sound sources where it is seen that the estimated azimuth angle is successfully tracked within the time windows, with few exceptions. Similar results can be observed for the elevation angles and are not reported here for brevity. Instead, Table I summarizes the median angular distances between the estimated DOAs and the ground truth DOAs in the time windows with voice activity. As it can be seen ADeLFi with spatio-spectral sparsity based regularization achieves the most accurate estimates, although it is sometimes almost equaled or surpassed by the spatial sparsity based regularization. Although ADeLFi assumes that no scattering object should in the proximity of the microphones, these results shows that it is still capable of reaching almost equivalent results in terms of dereverberation and localization even when the rigid baffle compensation is not applied to the microphone signals. Concerning the sound field approximation, as it can be seen in Table I, an improvement of around 2 dB in the median of the validation error $\bar{\epsilon}_v$ is seen for all of the scenarios when the rigid baffle compensation is used. This indicates that better sound field approximation is indeed reached when the rigid baffle compensation is employed although this does not substantially increase the DOA estimation and dereverberation performance. In some cases, the median angular similarity $\bar{\sigma}_\alpha$ is slightly lowered but only in case of the moving sources scenarios. This is most likely caused by a different DOA estimation in time windows without voice activity as shown in Figure 6 (D-E) which influences the DOA averaging procedure described in Section IV-D. In Table I, the localization performances of SBL are also reported using a SS approach using the same parameters described in Section V-A. Here, SBL reaches similar performance to ADeLFi. As in the case of ADeLFi, the rigid baffle compensation does not particularly affect the results. The comparison is not carried out for the moving source scenarios, since SBL was not specifically designed for such task.

Finally, Figure 8 shows the STOI and PESQ scores of the dereverberated signals obtained with ADeLFi. The reference signals used to compute these measures are the semi-anechoic speech signals used to drive the loudspeakers for the static sources scenarios while for the moving speaker scenarios the recordings of a microphone near the mouth are used. In most of the cases, ADeLFi improves both the audio quality and the speech intelligibility, with visible improvements when multiple weight signals are used. In most of the cases, the rigid baffle compensation does not lead to a substantial increase of the objective measure scores indicating a particular robustness of ADeLFi against model errors. In all scenarios, the objective measure scores of SBL are only slightly lower than the ones of ADeLFi. Here only the results with rigid baffle compensation are shown for SBL since almost equivalent results are obtained for the unprocessed microphone signals case. The difference between ADeLFi and SBL is less noticeable than in the simulation results of Section V-A, possibly due to the lower amount of reverberation present in the room where the real measurements took place. These results are once more compared with ADA: the scores of the dereverberated signals of ADA are often outperformed by ADeLFi, particularly in the moving source scenarios. Notice that here the forgetting factor of ADA is set to 0.99 and the validation microphones are included in the processing. As for the simulation results of the previous Section, in many cases the best results are achieved using ADeLFi in combination with spatio-spectral sparsity based regularization (l_1-norm), although spatial sparsity based regularization (sum of l_2-norms) often outperforms this, especially for the moving source scenarios. Sound samples can be found in [52].

VI. Conclusions

In this paper a novel algorithm for joint source localization and dereverberation has been proposed. This algorithm relies on approximating the sound field using the measurements of a set of microphones and solving an inverse problem that employs a particular acoustic model. The inverse problem is solved using an accelerated variant of the PG algorithm using optimization and a WOLA procedure in order to obtain the weight signals that control the plane waves which effectively approximate the sound field. The inverse problem is regularized using sparsity-promoting regularization and, depending on the choice of the regularization term, spatial or spatio-spectral sparsity can be promoted in the weight signals. A novel technique for tuning the level of regularization is proposed which is based on comparing the approximated sound field with the sound field measured by an additional microphone. It has been shown that, by finding the weight signal with strongest energy during different time windows, a moving sound source can be localized in terms of DOA. The same weight signal, together with its neighbors can then also be used for a dereverberation task. Simulations have shown that DOA estimation can be achieved using relatively few microphones ($N_m \geq 4$) when a speech source generates the sound field and that spatial and spatio-spectral sparsity based regularizations are comparable in terms of both approximation quality and dereverberation. The proposed algorithm is shown to be robust against different types of noise using both simulated and real measurements. Compared with state-of-the-art algorithms for both DOA estimation and dereverberation, the algorithm shows competitive performance and additionally provides noise reduction in the dereverberated signals. A main drawback of the proposed algorithm is its computational complexity. For example, using $N_m = 12$ microphones and $N_w = 500$ plane waves directions a real-time factor of 381 is reached using a single core on a Intel Core™ i7 2.7 GHz computer. However, many of the numerical operations can be performed in parallel and the use of fast transformations should be investigated. Future research will focus on the reduction of the computational burden and on the extension of the algorithm under more complex scenarios including the localization of multiple sound sources and the use of moving microphones.

Acknowledgements

The authors would like to thank Brian Fitzpatrick for the helpful discussions and the anonymous reviewers for their valuable suggestions.

References

[1] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[2] S. Argentieri, P. Danes, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[3] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer, 2010.

[4] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '07)*, 2007, pp. 901–904.

[5] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, 2015.

[6] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, 2013.

[7] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, 2008.

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

[9] I. Dokmanić and M. Vetterli, "Room helps: Acoustic localization with finite elements," in *Proc. 2012 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '12)*, 2012, pp. 2617–2620.

[10] N. Antonello, T. van Waterschoot, M. Moonen, and P. A. Naylor, "Source localization and signal reconstruction in a reverberant field using the FDTD method," in *Proc. 22nd European Signal Process. Conf. (EUSIPCO '14)*, 2014, pp. 301–305.

[11] S. Kitić, L. Albera, N. Bertin, and R. Gribonval, "Physics-driven inverse problems made tractable with cosparse regularization," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 335–348, 2016.

[12] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[13] A. Moiola, R. Hiptmair, and I. Perugia, "Vekua theory for the Helmholtz operator," *Zeitschrift für angewandte Mathematik und Physik*, vol. 62, no. 5, pp. 779–807, 2011.

[14] G. Chardon, T. Nowakowski, J. De Rosny, and L. Daudet, "A blind dereverberation method for narrowband source localization," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 815–824, 2015.

[15] T. Nowakowski, J. de Rosny, and L. Daudet, "Robust source localization from wavefield separation including prior information," *J. Acoust. Soc. Amer.*, vol. 141, no. 4, pp. 2375–2386, 2017.

[16] S. Koyama and L. Daudet, "Comparison of reverberation models for sparse sound field decomposition," in *Proc. 2015 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '17)*. IEEE, 2017.

[17] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[18] E. Fernandez-Grande and A. Xenaki, "Compressive sensing with a spherical microphone array," *J. Acoust. Soc. Amer.*, vol. 139, no. 2, pp. EL45–EL49, 2016.

[19] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1929–1941, 2017.

[20] N. Murata, S. Koyama, N. Takamune, and H. Saruwatari, "Sparse representation using multidimensional mixed-norm penalty with application to sound field decomposition," *IEEE Trans. Signal Process.*, vol. 66, no. 12, pp. 3327–3338, 2018.

[21] A. Asaei, H. Bourlard, M. J. Taghizadeh, and V. Cevher, "Model-based sparse component analysis for reverberant speech localization," in *Proc. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '14)*, 2014, pp. 1439–1443.

[22] P. K. T. Wu, N. Epain, and C. Jin, "A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques," in *Proc. 2012 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '12)*, 2012, pp. 4053–4056.

[23] N. Epain, T. Noohi, and C. Jin, "Sparse recovery method for dereverberation," in *Proc. REVERB Workshop*, 2014.

[24] A. Xenaki, J. Bünsow Boldt, and M. Græsbøll Christensen, "Sound source localization and speech enhancement with sparse bayesian learning beamforming," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3912–3921, 2018.

[25] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on array processing and sensor networks*, H. Simon and K. J. R. Liu, Eds. Wiley, 2010.

[26] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot, "Joint source localization and dereverberation by sound field interpolation using sparse regularization," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, 2018, pp. 6892–6896.

[27] A. Themelis, L. Stella, and P. Patrinos, "Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone line-search algorithms," *arXiv:1606.06256*, 2016.

[28] A. Duijndam and M. Schonewille, "Nonuniform fast Fourier transform," *Geophysics*, vol. 64, no. 2, pp. 539–551, 1999.

[29] A. Averbuch, R. R. Coifman, D. L. Donoho, M. Elad, and M. Israeli, "Fast and accurate polar Fourier transform," *Applied and Computational Harmonic analysis*, vol. 21, no. 2, pp. 145–167, 2006.

[30] Á. González, "Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices," *Mathematical Geosciences*, vol. 42, no. 1, pp. 49–64, 2010.

[31] R. H. Hardin and N. J. Sloane, "Mclaren's improved snub cube and other new spherical designs in three dimensions," *Discrete & Computational Geometry*, vol. 15, no. 4, pp. 429–441, 1996.

[32] E. B. Saff and A. B. Kuijlaars, "Distributing many points on a sphere," *The mathematical intelligencer*, vol. 19, no. 1, pp. 5–11, 1997.

[33] E. Perrey-Debain, "Plane wave decomposition in the unit disc: Convergence estimates and computational aspects," *Journal of Computational and Applied Mathematics*, vol. 193, no. 1, pp. 140–156, 2006.

[34] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic Press, 1999.

[35] A. Gramfort, D. Strohmeier, J. Haueisen, M. S. Hämäläinen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations," *NeuroImage*, vol. 70, pp. 410–422, 2013.

[36] N. Parikh and S. P. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[37] N. Antonello, L. Stella, P. Patrinos, and T. van Waterschoot, "Proximal gradient algorithms: Applications in signal processing," *arXiv preprint arXiv:1803.01621*, 2018.

[38] S. Theodoridis, *Machine learning: a Bayesian and optimization perspective*. Academic Press, 2015.

[39] A. Griewank and A. Walther, *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

[40] S. Diamond and S. Boyd, "Matrix-free convex optimization modeling," in *Optimization and its Applications in Control and Data Sciences*. Springer, 2016, pp. 221–264.

[41] L. Stella and N. Antonello. (2017) StructuredOptimization.jl. https://github.com/kul-forbes/StructuredOptimization.jl.

[42] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 4, pp. 774–786, 2015.

[43] Bang and Olufsen, "Music for Archimedes," CD B&O 101, 1992.

[44] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[45] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[46] Y.-S. Yoon, L. M. Kaplan, and J. H. McClellan, "TOPS: New DOA estimator for wideband signals," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1977–1989, 2006.

[47] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[48] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, 2018, pp. 351–355.

[49] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[50] ITU-T, "Perceptual evaluation of of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," in *ITU-T Recpmmendation P.862, Int. Telecommun. Union*, 2001.

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

[52] N. Antonello. (2018) ADelFi audio samples. https://nantonel.github.io/adelfi/.

[53] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, 2018, www.locata-challenge.org.

[54] C. Veaux, J. Yamagishi, and K. MacDonald. (2016) CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.

[55] OptiTrack. (2018) Product information about OptiTrack Flex13. http://optitrack.com/products/flex-13/.

[56] mh acoustics. (2013) EM32 eigenmike microphone array release notes (v17.0). http://www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf.

[57] F. Jacobsen, G. Moreno-Pescador, E. Fernandez-Grande, and J. Hald, "Near field acoustic holography with microphones on a rigid sphere (l)," *J. Acoust. Soc. Amer.*, vol. 129, no. 6, pp. 3461–3464, 2011.

[58] A. H. Moore. (2017) sap-sh-doa-estimation. https://github.com/ImperialCollegeLondon/sap-sh-doa-estimation.

[59] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 178–192, 2017.