Indoor Sound Source Localization based on Sparse Bayesian Learning and Compressed Data

Zonglong Bai[1,2], Jinwei Sun[1], Jesper Rindom Jensen[2] and Mads Græsbøll Christensen[2]
[1]School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin, China
[2]Audio Analysis Lab, CREATE, Aalborg University, Aalborg, Denmark

***Abstract*—In this paper, the problems of indoor sound source localization using a wireless acoustic sensor network are addressed and a new sparse Bayesian learning based algorithm is proposed. Using time delays for the direct paths from candidate source locations to microphone nodes, the proposed algorithm estimates the most likely source location. To reduce the amount of data that must be exchanged between microphone nodes, a Gaussian measurement matrix is multiplied on to each channel and the proposed method operates directly on the compressed data. This is achieved by exploiting sparsity in both the frequency and space domains. The performance is analysed in numerical simulations, where the performance as a function of the reverberation times in investigated, and the results show that the proposed algorithm is robust to reverberation.**

***Index Terms*—Sound Source Localization, Sparse Bayesian Learning, Array Signal Processing, Reverberation Environment.**

I. Introduction

Sound source localization using microphone arrays is one of the key technologies for many applications such as teleconferencing [1], robot audition [2] and hearing aids [3]. Many algorithms have been derived for tackling the sound source localization problem such as time delay estimation based methods [4], beam-forming methods [5], subspace methods and statistical methods [6]. However, most of the methods suffer a heavy performance loss in the reverberate environments which limits their practical applications.

Indoor source localization is particularly challenging due to reverberation, which causes the sound waves to be reflected by the surrounding walls and mixed with the direct sound. The sound propagation can be treated as several image sources in a free field that are correlated with each other and the sound source [7]. Recently, sparse Bayesian learning (SBL) based sound source localization has attracted widespread attention because it achieves high-resolution performance and typically outperforms conventional methods when localizing correlated and non-stationary sources [8]–[11]. The SBL framework estimates hyper-parameters by maximizing the posterior distribution of candidate sources amplitude. The posterior distribution can be derived from the likelihood function and the prior distribution. Assuming the likelihood function and prior information follows independent and identically distributed (i.i.d) complex Gaussian distributions, the posterior distribution is also a complex Gaussian distribution. Maximizing the likelihood, the hyper-parameters can be determined, which are the amplitudes of candidate sources. The SBL methods have been reported to perform well with correlated sources [10], and methods based on this methodology are thus good candidates for indoor sound source localization, as considered in this paper.

This work is supported by China Scholarship Council, grant ID: 201806120176.

The method proposed builds on SBL and compressed sensing theory, and the two main contributions of the paper are the following. First, a SBL based sound source localization model is proposed using the time delay of the direct-path component of the sound propagation. This model does not make any assumptions on the array geometry and is thus applicable in wireless acoustic sensor networks. Second, a compressed sensing and SBL based sound source localization algorithm is proposed based on the model, which efficiently reduce the amount of data that has to be transmitted within the network.

The remainder of the paper is organised as follows. In Section 2, a sparse signal model for indoor sound source localization is proposed and used for solving the localization problem with SBL. In Section 3, the temporal domain array data is compressed and a SBL based sound source localization algorithm is proposed that operates on the compressed data. The experimental results are provided in Section 4, and we conclude and elaborate on future work in Section 5.

II. Background

In this section, we first construct an array manifold matrix using the direct path components from the candidate sources to each of the microphone nodes. Then, we show how to estimate the sound source location using SBL beamforming.

A. Sparse Model for Indoor Sound Source Localization

We consider the localization of a single source in two dimensions, i.e., we assume that the sound source and all microphone nodes are located in the same horizontal plane described by Cartesian coordinates. The coordinates of the sound source are $\mathbf{s} = (x_0, y_0)$ and the coordinates of the m'th microphone in a wireless acoustic sensor network are $\mathbf{r}_m = (x_m, y_m)$ for $m = 1, 2, \ldots, M$, which are here assumed known. That is, the time delay from the sound source to microphone node $m = 1, 2, \ldots, M$ is

$$\tau_m = \frac{|\mathbf{r}_m - \mathbf{s}|}{c}, \tag{1}$$

where c is the sound velocity in air. The steering vector is $\boldsymbol{a}_f = [e^{j2\pi f\tau_1}, e^{j2\pi f\tau_2}, \cdots, e^{j2\pi f\tau_M}]^{\mathrm{T}}$, where f is the frequency and $(\cdot)^{\mathrm{T}}$ denotes the matrix/vector transpose.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

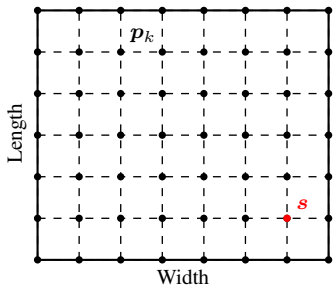

Fig. 1. Illustration of the candidate sound source positions in the target plane.

In the sparse representation framework, we define K candidate source positions $(p_k, k = 1, 2, \cdots, K)$ in the target plane as depicted in Figure 1. The steering vector for the kth candidate source is $\boldsymbol{a}_{kf} = [e^{j2\pi f \tau_{k1}}, e^{j2\pi f \tau_{k2}}, \cdots, e^{j2\pi f \tau_{kM}}]^{\mathrm{T}}$, where $\tau_{km} = \frac{|\mathbf{r}_m - \mathbf{p}_k|}{c}$ is the time delay from the kth candidate source to the m'th microphone node. Considering all K candidate sources, the array manifold matrix, or dictionary, can be constructed as $\mathbf{A}_f = [\boldsymbol{a}_{1f}, \cdots, \boldsymbol{a}_{kf}, \cdots, \boldsymbol{a}_{Kf}]$.

The model for the microphone array data is thus given by

$$\boldsymbol{X}_f = \mathbf{A}_f \boldsymbol{S}_f + \boldsymbol{N}_f, \ f = 1, 2, \cdots, F, \tag{2}$$

where $\boldsymbol{X}_f$ is the array data at the f'th frequency bin, $\boldsymbol{S}_f$ is a sparse vector with non-zero elements corresponding to true source complex amplitude at the truth bearing while other elements are zero. The background noise is denoted $\boldsymbol{N}_f$, which is the noise at the f'th frequency bin, and it is assumed to be white Gaussian noise and uncorrelated with $\boldsymbol{S}_f$.

The model in (2) represents the sound source localization problem as a sparse representation one, hence, we can solve it using SBL as described in the next subsection.

B. Sound Source Localization based on SBL

Solving the problem in (2) using SBL involves determining the posterior distribution of $\boldsymbol{S}_f$ from the prior distribution of $\boldsymbol{S}_f$ and the likelihood. We process by assuming a circular, symmetric white Gaussian to the observed noise. So the likelihood function is

$$p(\boldsymbol{X}_f | \boldsymbol{S}_f; \sigma_f^2) = \mathcal{CN}\left(\boldsymbol{X}_f | \mathbf{A}_f \boldsymbol{S}_f, \sigma_f^2 \mathbf{I}\right), \tag{3}$$

For the f'th frequency bin, the complex Gaussian prior of the source is

$$p(\boldsymbol{S}_f; \boldsymbol{\xi}_f) = \mathcal{CN}(\boldsymbol{S}_f | 0, \boldsymbol{\Sigma}_{Sf}), \tag{4}$$

where, $\boldsymbol{\Sigma}_{Sf} = \mathrm{diag}(\boldsymbol{\xi}_f)$, $\boldsymbol{\xi}_f = [\xi_{1f}, \cdots, \xi_{Kf}]$, the hyper-parameter ξ_{kf} controls the amplitude of the kth candidate source at the f'th frequency bin, and $\mathrm{diag}(\boldsymbol{x})$ is the diagonal operator with the entries of the vector $\boldsymbol{x}$ on the diagonal.and the posterior distribution of $\boldsymbol{S}_f$ is

$$\begin{aligned} p(\boldsymbol{S}_f | \boldsymbol{X}_f; \boldsymbol{\xi}_f, \sigma_f^2) &\propto p\left(\boldsymbol{X}_f | \boldsymbol{S}_f; \sigma_f^2\right) p\left(\boldsymbol{S}_f; \boldsymbol{\xi}_f\right) \\ &= \mathcal{CN}(\boldsymbol{S}_f | \boldsymbol{\Lambda}_f, \boldsymbol{\Sigma}_{xf}), \end{aligned} \tag{5}$$

where,

$$\begin{aligned} \boldsymbol{\Lambda}_f &= \boldsymbol{\Sigma}_{Sf} \mathbf{A}_f^{\mathrm{H}} \boldsymbol{\Sigma}_{xf}^{-1} \boldsymbol{X}_f \\ \boldsymbol{\Sigma}_f &= \boldsymbol{\Sigma}_{Sf} - \boldsymbol{\Sigma}_{Sf} \mathbf{A}_f{}^{\mathrm{H}} \boldsymbol{\Sigma}_{xf}^{-1} \mathbf{A}_f \boldsymbol{\Sigma}_{Sf}, \end{aligned} \tag{6}$$

and

$$\boldsymbol{\Sigma}_{xf} = \mathrm{E}\left[\boldsymbol{X}_f \boldsymbol{X}_f^{\mathrm{H}}\right] = \mathbf{A}_f \boldsymbol{\Sigma}_{Sf} \mathbf{A}_f{}^{\mathrm{H}} + \sigma_f^2 \mathbf{I} \tag{7}$$

is the data covariance matrix, the superscript $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose.

The hyper-parameters $\boldsymbol{\xi}_f$ and σ_f^2 are estimated using the evidence:

$$\begin{aligned} \hat{\boldsymbol{\xi}}_f &= \arg\max_{\boldsymbol{\xi}_f \geqslant 0} \log p(\boldsymbol{X}_f; \boldsymbol{\xi}_f, \sigma_f^2) \\ &= \arg\min_{\boldsymbol{\xi}_f \geqslant 0} \left\{ \log\det(\boldsymbol{\Sigma}_{xf}) + \boldsymbol{X}_f^{\mathrm{H}} \boldsymbol{\Sigma}_{xf}^{-1} \boldsymbol{X}_f \right\}, \end{aligned} \tag{8}$$

where $\det(\cdot)$ denotes taking the determinant of a matrix.

The objective function of (8) is non-convex. However, according to [9] and [10], $\hat{\boldsymbol{\xi}}_f$ can be approximately solved using a fixed point update method

$$\hat{\xi}_{kf}^i = \hat{\xi}_{kf}^{i-1} \frac{\mathbf{a}_{kf}^{\mathrm{H}} \boldsymbol{\Sigma}_{xf}^{-1} \boldsymbol{\Upsilon}_{xf} \boldsymbol{\Sigma}_{xf} \boldsymbol{a}_{kf}}{\boldsymbol{a}_{kf}^{\mathrm{H}} \boldsymbol{\Sigma}_{xf}^{-1} \boldsymbol{a}_{kf}}, \tag{9}$$

where $\hat{\xi}_{kf}^i$ is the estimate of the k'th parameter of $\boldsymbol{\xi}_f$ at the i'th iteration, and $\boldsymbol{\Upsilon}_{xf} = \boldsymbol{X}_f \boldsymbol{X}_f^{\mathrm{H}}$ is the array data cross-spectral matrix. Then, the hyper-parameter σ_f^2, which is the noise variance at frequency f, can be estimated using the maximum likelihood procedure as

$$\hat{\sigma}_f^2 = \frac{1}{M - \bar{K}} \mathrm{Tr}\left[\left(\mathbf{I}_{\mathrm{M}} - \mathbf{A}_{\mathcal{N}} \mathbf{A}_{\mathcal{N}}^{+}\right) \boldsymbol{\Upsilon}_{xf}\right], \tag{10}$$

where $\bar{K}$ is the predefined number of sources, $\boldsymbol{A}_{\mathcal{N}}$ is the subset of $\boldsymbol{A}_f$ and $\mathcal{N}$ is the set of the $\bar{K}$ largest parameters in $\boldsymbol{\xi}_f^i$, $(\cdot)^+$ denote the pseudo inverse operator, and $\mathrm{Tr}[\cdot]$ is the trace operator.

If we assume that $\forall \boldsymbol{S}_f, \ f = 1, 2, \cdots, F$ have a common sparsity structure, the hyper-parameters of each frequency bin $\hat{\boldsymbol{\xi}}_f$ can be combined:

$$\hat{\boldsymbol{\xi}}_{1:F} = \frac{1}{F} \sum_{f=1}^{F} \hat{\boldsymbol{\xi}}_f, \tag{11}$$

where $\hat{\boldsymbol{\xi}}_{1:F}$ denotes the power of the sources at each grid point. Moreover, the update functions become

$$\hat{\xi}_{k,1:F}^i = \hat{\xi}_{k,1:F}^{i-1} \frac{\sum_{f=1}^{F} \boldsymbol{a}_{kf}^{\mathrm{H}} \boldsymbol{\Sigma}_{xf}^{-1} \boldsymbol{\Upsilon}_{xf} \boldsymbol{\Sigma}_{xf} \boldsymbol{a}_{kf}}{\sum_{f=1}^{F} \boldsymbol{a}_{kf}^{\mathrm{H}} \boldsymbol{\Sigma}_{xf}^{-1} \boldsymbol{a}_{kf}} \tag{12}$$

$$\hat{\sigma}^2 = \frac{1}{M - \bar{K}} \mathrm{Tr}\left[\left(\mathbf{I}_{\mathrm{M}} - \mathbf{A}_{\mathcal{N}} \mathbf{A}_{\mathcal{N}}^{+}\right) \boldsymbol{\Upsilon}_{\boldsymbol{x}}\right], \tag{13}$$

where $\hat{\xi}_{k,1:F}^i$ is the k'th parameter of $\hat{\boldsymbol{\xi}}_{1:F}$ at the i'th iteration, $\mathcal{N}$ is the set of the $\bar{K}$ largest parameters in $\hat{\boldsymbol{\xi}}_{1:F}^i$, and $\boldsymbol{\Upsilon}_{\boldsymbol{x}} = \sum_{f=1}^{F} \boldsymbol{X}_f \boldsymbol{X}_f^{\mathrm{H}}$.

III. Proposed Sound Source Localization Methods

In this section, we reduce the volume of array data by adding a measurement matrix to each channel and propose methods for sound source localization using compressed data that utilize both spatial and spectral sparsity.

A. Sound Source Localization Model for Compressed Data

Compressed sensing methods can be used to reduce the data size since most sound sources are sparse in some domains such as the frequency domain [12]. After multiplying the predefined sparse sensing matrix to raw data of each channel, we have

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{\Phi}_1\boldsymbol{\Psi} & & \\ & \ddots & \\ & & \boldsymbol{\Phi}_M\boldsymbol{\Psi} \end{bmatrix} \left(\bar{\boldsymbol{X}} + \boldsymbol{V}\right), \tag{14}$$

where $\boldsymbol{y} = [\boldsymbol{\Phi}_1\bar{\boldsymbol{x}}_1^{\mathrm{T}}, \cdots, \boldsymbol{\Phi}_M\bar{\boldsymbol{x}}_M^{\mathrm{T}}]^{\mathrm{T}}$ is a vector containing the compressed measurement of all M microphones, $\bar{\boldsymbol{x}}_m$ is the uncompressed signal data of m'th microphone and $\boldsymbol{y} \in \mathbb{R}^{NM}$ which can be considered as a complex vector with zero imaginal part, $\boldsymbol{\Phi}_m \in \mathbb{C}^{N\times F}$ is the sparse sensing matrix for the m'th microphone data to reduce the data volume, F is the frequency bins number and $N \ll F$ so the size of the microphone array data is reduced at a compression rate of $\frac{F-N}{F}$. $\boldsymbol{\Psi}$ denotes $F \times F$ dimensional inverse DFT matrix. Moreover, $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{X}}_1^{\mathrm{T}}, \cdots, \bar{\boldsymbol{X}}_M^{\mathrm{T}}]^{\mathrm{T}}$ is a vector containing the frequency data of the source signals and $\bar{\boldsymbol{X}}_m$ is the source signals of the m'th microphone. $\boldsymbol{V}$ is complex Gaussian noise and $\boldsymbol{V} \in \mathbb{C}^{NM}$.

To estimate the sound source location, $\bar{\boldsymbol{X}}$ must be reconstructed as the form of $\boldsymbol{X} = [\boldsymbol{X}_1^{\mathrm{T}}, \cdots, \boldsymbol{X}_f^{\mathrm{T}}, \cdots, \boldsymbol{X}_F^{\mathrm{T}}]^{\mathrm{T}}$ since $\boldsymbol{X}$ can be expressed using steering vector as (2), where $\boldsymbol{X}_f$ is fth frequency bin data of all microphones and $\boldsymbol{X}_f \in \mathbb{C}^M$. This $\boldsymbol{X}$ can be obtained by multiplying a permutation matrix $\bar{\mathbf{T}}$ with $\bar{\boldsymbol{X}}$, and $\bar{\boldsymbol{X}} = \mathbf{T}\boldsymbol{X} = \mathbf{T}\mathbf{A}\boldsymbol{S}$, where $\mathbf{T}$ is the inverse of $\bar{\mathbf{T}}$ and $\mathbf{A} = \mathrm{diag}[\mathbf{A}_1, \cdots, \mathbf{A}_F]$ is a block diagonal matrix, $\boldsymbol{S} = \left[\boldsymbol{S}_1^{\mathrm{T}}, \cdots, \boldsymbol{S}_f^{\mathrm{T}}\right]^{\mathrm{T}}$, $\boldsymbol{S} \in \mathbb{C}^{KF}$. Note that the only difference between $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$ is that $\boldsymbol{X}$ index by frequency blocks while $\bar{\boldsymbol{X}}$ index by channel blocks.

Assuming the same sparsity for each channel because signals received by all channels are generated by the same stationary sources, we have $\boldsymbol{\Phi}_m = \boldsymbol{\Phi},\ m = 1, \cdots, M$. Then, (14) can be rewritten as

$$\boldsymbol{y} = \tilde{\boldsymbol{\Psi}}\mathbf{T}\mathbf{A}\boldsymbol{S} + \tilde{\boldsymbol{\Psi}}\boldsymbol{V} = \mathbf{U}\boldsymbol{S} + \boldsymbol{W}, \tag{15}$$

where $\tilde{\boldsymbol{\Psi}} = \mathrm{diag}\left[\boldsymbol{\Phi}\boldsymbol{\Psi}, \cdots, \boldsymbol{\Phi}\boldsymbol{\Psi}\right]$ and $\mathbf{U} = \tilde{\boldsymbol{\Psi}}\mathbf{T}\mathbf{A}$ is a joint measurement matrix which utilizes the sparsity in both frequency and space, and $\boldsymbol{W} = \tilde{\boldsymbol{\Psi}}\boldsymbol{V}$ is the noise which can be considered as the circular, symmetric complex Gaussian noise.

B. Frequency estimation using Compressed Data

The linear model of (15) can be solved using the SBL method proposed in II-B. However, it is computationally expensive because $\boldsymbol{S}$ is a high-dimensional vector. To reduce the computational load, we can estimate the frequencies of the source first. Then estimate the source location using indexes of the highest power density frequencies.

Consider the compressed measurement of m'th channel,

$$\boldsymbol{\Phi}\bar{\boldsymbol{x}}_m = \boldsymbol{\Phi}\boldsymbol{\Psi}\bar{\boldsymbol{X}}_m = \boldsymbol{\Theta}(\bar{\boldsymbol{S}}_m + \tilde{\boldsymbol{S}}_m) = \boldsymbol{\Theta}\bar{\boldsymbol{S}}_m + \bar{\boldsymbol{N}}_m \tag{16}$$

where $\bar{\boldsymbol{X}}_m \in \mathbb{C}^F$ is the frequency date of the source signals received by the m'th microphone, and $\bar{\boldsymbol{S}}_m$ is sparse vector with sparsity $\bar{F}$, i.e., there are $\bar{F}$ elements in $\bar{\boldsymbol{S}}_m$ over exceeding a threshold. Furthermore, $\tilde{\boldsymbol{S}}_m$ contains the other smaller remaining elements and assume that $\tilde{\boldsymbol{S}}_m$ follows a circular, symmetric complex Gaussian distribution $\mathcal{CN}(0, \lambda_m\mathbf{I})$. As $\bar{\boldsymbol{N}}_m = \boldsymbol{\Theta}\tilde{\boldsymbol{S}}_m$ and $\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{H}} = a\mathbf{I}$, the noise $\bar{\boldsymbol{N}}_m$ follows a complex Gaussian distribution

$$p(\bar{\boldsymbol{N}}_m) = \mathcal{CN}(\bar{\boldsymbol{N}}_m|0, \lambda_m\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{H}}) = \mathcal{CN}(\bar{\boldsymbol{N}}_m|0, \sigma_m^2\mathbf{I}). \tag{17}$$

Then, considering all M channels,

$$\bar{\boldsymbol{y}} = \boldsymbol{\Phi}\boldsymbol{\Psi}\left(\tilde{\boldsymbol{X}} + \tilde{\boldsymbol{V}}\right) = \boldsymbol{\Theta}(\bar{\boldsymbol{S}} + \tilde{\boldsymbol{S}}) = \boldsymbol{\Theta}\bar{\boldsymbol{S}} + \bar{\boldsymbol{N}} \tag{18}$$

where, $\bar{\boldsymbol{y}} = [\boldsymbol{\Phi}\bar{\boldsymbol{x}}_1, \cdots, \boldsymbol{\Phi}\bar{\boldsymbol{x}}_M] \in \mathbb{R}^{N\times M}$ is the matrix of compressed measurements, $\tilde{\boldsymbol{X}} = [\bar{\boldsymbol{X}}_1, \cdots, \bar{\boldsymbol{X}}_M] \in \mathbb{C}^{F\times M}$ is a multidimensional vector in which each column is the frequency data vector of source signals,$\tilde{\boldsymbol{V}} = [\boldsymbol{V}_1, \cdots, \boldsymbol{V}_M]$ is a multidimensional vector in which each column is the noise of each microphone. $\bar{\boldsymbol{S}} = [\bar{\boldsymbol{S}}_1, \cdots, \bar{\boldsymbol{S}}_M] \in \mathbb{C}^{F\times M}$ is a row-sparse matrix, and $\boldsymbol{N} = [\bar{\boldsymbol{N}}_1, \cdots, \bar{\boldsymbol{N}}_M] \in \mathbb{C}^{N\times M}$ contains the noise. Then, similar to II-B, the frequency can be estimated using SBL. First, we define a hyper-parameter vector, $\boldsymbol{\zeta} = [\zeta_1, \cdots, \zeta_F]^T$, which represents the amplitude covariance of all frequency bins. The complex Gaussian prior is

$$p(\bar{\boldsymbol{S}}; \boldsymbol{\zeta}) = \prod_{m=1}^{M} \mathcal{CN}\left(\bar{\boldsymbol{S}}_m|0, \boldsymbol{\Gamma}\right) \tag{19}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}\left(\boldsymbol{\zeta}\right)$, and the likelihood function is

$$p(\bar{\boldsymbol{y}}|\bar{\boldsymbol{S}}; \sigma_m^2) = \prod_{m=1}^{M} \mathcal{CN}\left(\bar{\boldsymbol{x}}_m|\boldsymbol{\Theta}\bar{\boldsymbol{S}}_m, \sigma_m^2\mathbf{I}\right). \tag{20}$$

Similar to (9) and (10), the hyper-parameters $\boldsymbol{\zeta}$ and σ_m^2 can be updated using the following formula:

$$\hat{\zeta}_f^i = \hat{\zeta}_f^{i-1}\frac{\boldsymbol{\Theta}_f^{\mathrm{H}}\boldsymbol{\Xi}^{-1}\left(\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^{\mathrm{H}}\right)\boldsymbol{\Xi}\boldsymbol{\Theta}_f}{\boldsymbol{\Theta}_f^{\mathrm{H}}\boldsymbol{\Xi}^{-1}\boldsymbol{\Theta}_f} \tag{21}$$

$$\hat{\sigma}_m^2 = \frac{1}{M-\bar{F}}\mathrm{Tr}\left[\left(\mathbf{I}_{\mathrm{M}} - \boldsymbol{\Theta}_{\mathcal{M}}\boldsymbol{\Theta}_{\mathcal{M}}^{+}\right)\left(\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^{\mathrm{H}}\right)\right], \tag{22}$$

where

$$\boldsymbol{\Xi} = \mathrm{E}\left[\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^{\mathrm{H}}\right] = \boldsymbol{\Theta}\boldsymbol{\Gamma}\boldsymbol{\Theta} + \sigma_m^2\mathbf{I}. \tag{23}$$

Here, $\hat{\zeta}_f^i$ is the estimate of f'th element in $\boldsymbol{\zeta}$ at the i'th iteration, $\hat{\sigma}_m^2$ is the estimate of the noise variance, $\boldsymbol{\Theta}_f$ is the f'th column of $\boldsymbol{\Theta}$, $\bar{F}$ is the predefined number of frequency bins with non-zero coefficients, i.e., the sparsity of $\bar{\boldsymbol{S}}_m$, and $\mathcal{M}$ is the index set of the $\bar{F}$ largest parameters in $\boldsymbol{\zeta}^i$.

C. SBL-based Source Localization using Compressed Data

The model for compressed data was given in (15). Dividing the space into K uniformly distributed grid points as depicted in Figure 1 and assuming there is a source at each grid point, the complex amplitudes of all candidate sources $\boldsymbol{S} \in \mathbb{C}^{KF}$ is a high-dimensional vector, which can be processed as follows.

As all frequency bin blocks of $\boldsymbol{S} = \left[\boldsymbol{S}_1^{\mathrm{T}}, \cdots, \boldsymbol{S}_F^{\mathrm{T}}\right]^{\mathrm{T}}$ exhibits same sparsity, (15) can be rewritten as

$$\boldsymbol{y} = \sum_{f=1}^{F}\left\{\mathbf{U}_f\boldsymbol{S}_f + \boldsymbol{W}_f\right\}, \tag{24}$$

where $\mathbf{U}_f = \mathbf{A}_f \otimes \mathbf{\Phi}\boldsymbol{\psi}_f$ is a sub-matrix of $\mathbf{U}$, $\otimes$ denotes the Kronecker product, and $\boldsymbol{\psi}_f$ is the f'th column of $\mathbf{\Psi}$. Assuming that the sources are i.i.d complex Gaussian distributed at each frequency bin, we can process each frequency bin separately, while dealing with other frequency bins as noise. That is,

$$\boldsymbol{y} = \mathbf{U}_f \boldsymbol{S}_f + \bar{\boldsymbol{W}}_f, \ f = 1, 2, \cdots, F \tag{25}$$

where $\bar{\boldsymbol{W}}_f$ consists of all the noise at the f'th frequency bin and $p(\bar{\boldsymbol{W}}_f) = \mathcal{CN}(\bar{\boldsymbol{W}}_f|0, \sigma_f^2\mathbf{I})$, σ_f^2 is the noise variance at the f'th frequency bin.The equation (24) can be solved as follows:

Since most sound sources are sparse in frequency, we can estimate sound source location just using the non-zero frequency elements to reduce the the heavy computational load and the influence of the noise. Therefore, instead of using all frequency bins $\{1 : F\}$ as in (24), we use the index set $\mathcal{M}$ defined in (22). Since $\forall \boldsymbol{S}_f, \ f \in \mathcal{M}$ have the common sparsity according to our assumption, the hyper-parameters of each frequency bin $\hat{\boldsymbol{\xi}}_f$ can be combined as $\hat{\boldsymbol{\xi}}_{\mathcal{M}} = \frac{1}{F}\sum_{\mathcal{M}} \hat{\boldsymbol{\xi}}_f$, where $\hat{\boldsymbol{\xi}}_{\mathcal{M}}$ denotes the power of the sources at each grid point.

Similar to (12) and (13), the hyper-parameter $\boldsymbol{\xi}_{\mathcal{M}}$ and σ_f^2 are estimated using:

$$\hat{\xi}_{k,\mathcal{M}}^i = \hat{\xi}_{k,\mathcal{M}}^{i-1} \frac{\sum_{\mathcal{M}} \mathbf{U}_{kf}^{\mathrm{H}} \mathbf{\Sigma}_{yf}^{-1} \mathbf{\Upsilon}_y \mathbf{\Sigma}_{yf} \mathbf{U}_{kf}}{\sum_{\mathcal{M}} \mathbf{U}_{kf}^{\mathrm{H}} \mathbf{\Sigma}_{zf}^{-1} \mathbf{U}_{kf}} \tag{26}$$

$$\hat{\sigma}^2 = \frac{1}{M - \bar{K}} \mathrm{Tr}\left[\left(\mathbf{I}_{\mathrm{M}} - \mathbf{U}_{\mathcal{N}} \mathbf{U}_{\mathcal{N}}^{+}\right) \mathbf{\Upsilon}_y\right] \tag{27}$$

$$\mathbf{\Sigma}_{yf} = \mathrm{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{H}}\right] = \mathbf{U}_f \mathbf{\Sigma}_{Sf} {\mathbf{U}_f}^{\mathrm{H}} + \sigma_f^2 \mathbf{I}, \tag{28}$$

where $\mathbf{\Upsilon}_y = \boldsymbol{y}\boldsymbol{y}^{\mathrm{H}}$ is the cross-spectrum of the compressed array data, $\hat{\xi}_{k,\mathcal{M}}^i$ is the k'th element of $\hat{\boldsymbol{\xi}}_{\mathcal{M}}$ at the i'th iteration, and $\mathcal{N}$ is the set of the $\bar{K}$ largest elements in $\hat{\boldsymbol{\xi}}_{\mathcal{M}}^i$. The algorithm is summarized in Algorithm 1.

Algorithm 1 The proposed algorithm

Number of iterations for frequency estimation $i_f \leftarrow 0$;
Error of frequency estimate $e_f \leftarrow 1$;
Number of iterations for sound source localization $i_d \leftarrow 0$;
Error of sound source localization $e_d \leftarrow 1$;
while $i_f \leq i_{fmax}$ **and** $e_f \geq e_{fmin}$ **do**
  Update $i_f \leftarrow i_f + 1$;
  Compute $\mathbf{\Xi}$ using (23);
  Update $\hat{\boldsymbol{\zeta}}^i$ using (21);
  Find the frequency set $\mathcal{M}$;
  Update $\hat{\sigma}_m^2$ using (22);
  Update $e_f \leftarrow \frac{||\hat{\boldsymbol{\zeta}}^i - \hat{\boldsymbol{\zeta}}^{i-1}||_1}{||\hat{\boldsymbol{\zeta}}^i||_1}$;
end while
while $i_d \leq i_{dmax}$ **and** $e_d \geq e_{dmin}$ **do**
  Update $i_d \leftarrow i_d + 1$;
  Compute $\mathbf{\Sigma}_{yf}$ using (28);
  Update $\hat{\boldsymbol{\xi}}_{\mathcal{M}}^i$ using (26) and all frequencies in the set $\mathcal{M}$;
  Find the index set $\mathcal{N}$;
  Update $\hat{\sigma}^2$ using (27);
  Update $e_d \leftarrow \frac{||\hat{\boldsymbol{\xi}}_{\mathcal{M}}^i - \hat{\boldsymbol{\xi}}_{\mathcal{M}}^{i-1}||_1}{||\hat{\boldsymbol{\xi}}_{\mathcal{M}}^i||_1}$;
end while
return The hyper-parameter $\hat{\boldsymbol{\xi}}_{\mathcal{M}}$;
Amplitude of kth candidate source $\mathrm{P}_{\mathrm{SBL}}(k) = \hat{\xi}_{k,\mathcal{M}}$.

IV. Experimental Results

In this section, the performance of proposed algorithm in a simulated wireless acoustic sensor network in an acoustic environment resembling an indoor scenario is evaluated versus reverberation time, SNR and compression rate. The microphone array data is generated using the RIR Generator provided in [13], with a simulated room with dimensions $L \times W \times H = 7 \times 6 \times 6$ m. We consider the single source case and assume that the sound source and all microphones are located in the same horizontal plane ($H = 2$ m). We use an audio segment of violin provided in [14] as the sound source. In order to avoid silence segments at the beginning of the file, we choose the period from 10001'th point to 11024'th point. The onset of this period can be chosen using method proposed in [15]. The frequency of sampling is 48KHz, and the number of samples is 1024. The localization Root-Mean-Square-Error (RMSE) and accuracy are defined as $RMSE = \frac{1}{N_t}\sqrt{\sum_{i=1}^{N_t} \|\hat{\boldsymbol{P}}_i - \boldsymbol{P}_0\|}$ and $Accuracy = N_r/N_t$, respectively, where N_t is the number of Monte-Carlo experiments, $\hat{\boldsymbol{P}}_i$ is the estimation result at the i'th Monte-Carlo experiment, $\boldsymbol{P}_0$ is the true sound source location. A circular area (radius=0.2m) centered at the true source location is defined as the correct localization area, i.e., estimates located in this area are considered correct and otherwise wrong, N_r is the total number of correct estimates. Note that other state-of-art methods can not work under this scenario except SBL.

In the first experiment, the performance versus different reverberation times is tested. We use a wireless acoustic sensor network with randomly located microphones, as depicted in Figure 2 (a). The compression is 96 %, SNR=60 dB, the target space is uniformly divided as in Figure 1, and the distance between adjacent candidate sources is 0.2m. The results show that the proposed algorithm is robust to reverberation using the random distributed array as seen in Figure 2(b)-(d).

In the next experiment, the performance versus different SNRs is verified. The other parameters are set as in the first experiment except that the compressed rate is 80%. The results show that the algorithm is sensitive to the SNR. The localization accuracy of this algorithm is quite low under low SNR environment, see Figure 3(b), and high under high SNR environment, see Figure 3(c)-(d). The reason is that $\mathbf{\Phi}$ is a Gaussian matrix. Better performance under low SNR can be achieved if the compressed rate is lower.

In the final experiment, RMSE and accuracy versus reverberation time and compression rate are verified separately. The parameters are set the same as in the first and second experiments.The results show that the RMSE and the accuracy is robust to different reverberation times as it can be seen in Figure 4(a)-(b). Moreover, the proposed method yields low RMSE and high accuracy under high compression rates (96%) as it can be seen in Figure 4(c)-(d).

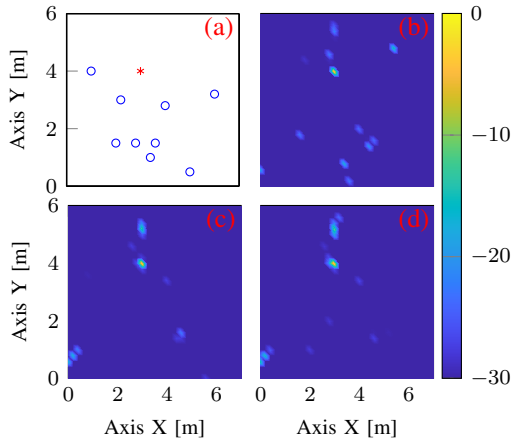

Fig. 2. (a) Locations of Microphone node and sound source . Localization result with (b) RT = 0.2 s, (c) RT = 0.4 s, and (d) RT=0.6 s. The red star in (a) denotes the sound source location.

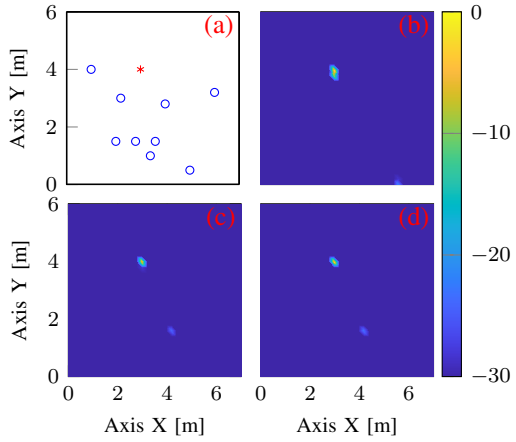

Fig. 3. (a) Microphone node and sound source locations. Localization result with (b) SNR=25dB, (c) SNR=30dB, and (d) SNR=35dB.

V. Conclusion

A sound source localization algorithm based on SBL and compressed data has been proposed in this paper. The method is intended for wireless acoustic sensor networks in indoor applications. The proposed method works by modeling the sound propagation from candidate locations to the microphones and exploits the sparsity of sound sources in both frequency and space. The proposed method reduces the amount of data that has to be exchanged via a measurement matrix and does not require any particular array geometry. The experimental results show that the proposed method performs well and is robust to reverberation and it is thus well-suited for indoor applications.

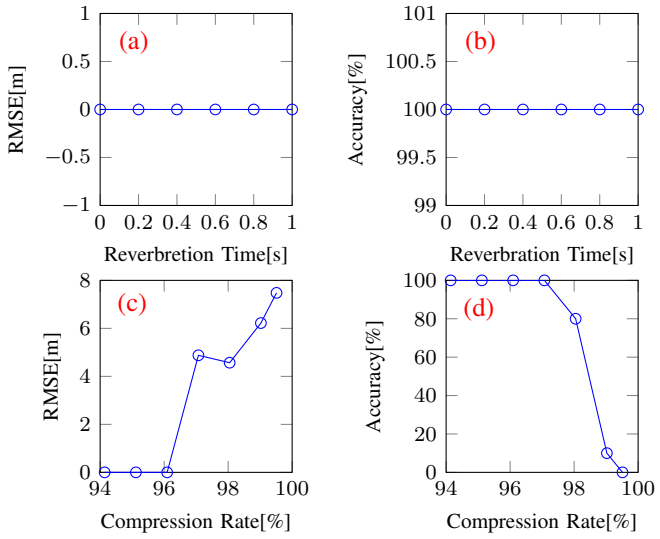

Fig. 4. (a) RMSE and (b) accuracy versus reverberation times, with a compression rate of 96%. (c) RMSE and (d) Accuracy versus compression rate for RT = 0.2 s and $N_t = 10$.

References

[1] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[2] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, Oct. 2017.

[3] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.

[4] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech & Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[5] P. Castellini and A. Sassaroli, "Acoustic source localization in a reverberant environment by average beamforming," *Mechanical Systems & Signal Processing*, vol. 24, no. 3, pp. 796–808, 2010.

[6] J.-Y. Lee, R. E. Hudson, and K. Yao, "Acoustic DOA estimation: An approximate maximum likelihood approach," *IEEE Systems Journal*, vol. 8, no. 1, pp. 131–141, 2014.

[7] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 620–633, 2014.

[8] K. L. Gembaa, S. Nannuru, P. Gerstoft, and W. S. Hodgkiss, "Multi-frequency sparse Bayesian learning for robust matched field processing," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3411–3420, 2017.

[9] S. Nannuru *et al.*, "Sparse Bayesian learning with uncertainty models and multiple dictionaries," in *The fifth IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, Nov. 2017, pp. 1190–1194.

[10] A. Xenakia, J. B. Boldt, and M. G. Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3912–3921, 2018.

[11] L. Shi, J. Jensen, J. Nielsen, and M. Christensen, "Multipitch estimation using block sparse bayesian learning and intra-block clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, United States, Apr 2018, pp. 666–670.

[12] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, USA: Wiley, 2018.

[13] E. Habets. (2015) Audiolabs-rir generator. [Online]. Available: http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

[14] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen. (2014) Smard. [Online]. Available: https://www.smard.es.aau.dk/

[15] C.-C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice," in *ISMIR*, 2008.