Joint wideband source localization and acquisition based on a grid-shift approach

Tzagkarakis, Christos; Kleijn, W. Bastiaan; Skoglund, Jan

DOI
10.1109/WASPAA.2017.8169999

Publication date
2017

Document Version
Final published version

Published in
2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2017

JOINT WIDEBAND SOURCE LOCALIZATION AND ACQUISITION BASED ON A GRID-SHIFT APPROACH

Christos Tzagkarakis[1], *W. Bastiaan Kleijn*[1,2,3], *Jan Skoglund*[3]

[1]Circuits and Systems, Delft University of Technology, The Netherlands
[2]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
[3]Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043

ABSTRACT

This paper addresses the problem of joint wideband localization and acquisition of acoustic sources. The source locations as well as acquisition of the original source signals are obtained in a joint fashion by solving a sparse recovery problem. Spatial sparsity is enforced by discretizing the acoustic scene into a grid of predefined dimensions. In practice, energy leakage from the source location to the neighboring grid points is expected to produce spurious location estimates, since the source location will not coincide with one of the grid points. To alleviate this problem we introduce the concept of grid-shift. A particular source is then near a point on the grid in at least one of a set of shifted grids. For the selected grid, other sources will generally not be on a grid point, but their energy is distributed over many points. A large number of experiments on real speech signals show the localization and acquisition effectiveness of the proposed approach under clean, noisy and reverberant conditions.

Index Terms— wideband acoustic sources, localization, acquisition, off-grid sparse recovery

1. INTRODUCTION

In the current work, the focus is on joint acoustic source localization and acquisition based on a set of microphones randomly placed in the acoustic scene. Specifically, we cast the localization problem as a sparse recovery problem [1] by discretizing the acoustic scene using a grid of a certain size. We assume that the center of each grid cell, termed a *grid point*, corresponds to a possible source position leading to the sparsification of the problem since only a few grid points will be non-zero due to the presence of sources. Here, we assume that the acoustic sources correspond to speech signals, and thus a time-frequency separability property can be adopted, i.e., the sources do not overlap in each time-frequency region [2].

Sparsity-based localization emerged as an alternative approach in solving the source localization problem compared to the traditional methods of beamforming [3], [4]. The motivation behind sparsity enforcing techniques [1], [5], [6], [7], [8] was to alleviate the problem of poor localization performance due to a limited number of data snapshots, low signal-to-noise ratio (SNR) levels and correlation between the emitted sources. Compressed sensing (CS) [9], [10] was adopted in order to guarantee that under certain conditions the sparsity-based localization problem can be solved efficiently.

The main hypothesis used in the sparsity-based localization framework states that there is an exact match between the real physical model and the assumed one, reflected in the grid structure. This implies that the acoustic sources lie exactly on the grid points. Based on this assumption, wideband acoustic source localization is examined in [11] under no reverberation effects, while in [12] the authors consider that the acoustic transfer functions are known. In [13], [14] wideband acoustic source localization is extended to source separation for speech recognition.

In contrast to the aforementioned assumptions we expect that the grid points will not coincide with the actual source locations. This causes energy leakage from the *off-grid* source position to the neighboring grid points, leading to errors in the sparse solutions. Many algorithms have been proposed to solve the off-grid problem within the CS theory. In [15] a semidefinite program is used to solve the line spectrum sparse recovery problem via atomic norm minimization. An atomic norm-regularized least-squares problem is adopted in [16], while an off-grid narrowband direction-of-arrival estimation problem is studied in [17].

A second class of methods employs a Bayesian framework, where the off-grid narrowband source localization task is examined in [18]. Another category of off-grid sparse optimization problems falls into the so-called perturbed matrix theory, which is based on allowing some perturbation of the matrix corresponding to the grid structure. Specifically, a perturbed version of adaptive matching pursuit is examined in [19] for narrowband source localization, while a perturbation-based orthogonal matching pursuit is proposed in [20] and [21] for an imaging and radar application, respectively. An additional subcategory of perturbed-based techniques using structured total least squares was considered in [22], [23] for dealing with narrowband source localization and cognitive radio sensing applications.

Contributions. We introduce a wideband joint acoustic source localization and acquisition approach using a sparse optimization framework based on a grid-shifting procedure. In particular, we are interested in studying the effect of *estimating the acoustic source locations in a joint fashion with source acquisition in the presence of misalignment between the grid points and the actual source locations. An orthogonal matching pursuit-based grid-shift scheme is proposed to solve the problem*. Since we cannot arbitrarily increase the grid size in order to achieve a better coverage of the acoustic scene (this will cause the violation of the restricted isometry property as described in the next section), the core idea is to consider a specific grid structure which is "shifted" across the acoustic scene. It is expected that each source will be located close to a grid point in at least one of the set of shifted grids. We then combine (based on K-means clustering) the sparse solutions corresponding to the (shifted) grids to obtain the source location estimates. The estimated source positions are used as side information to obtain the original source signals.

This work was supported by Google Inc.

978-1-5386-1631-4/17/$31.00 ©2017 IEEE

The novelty of the current paper is twofold. Firstly, a more realistic solution is employed by assuming that the sources are not located near the grid points as compared to the majority of the off-grid techniques, where typically it is assumed that the grid points are close enough to the true source locations. Secondly, a greedy sparse recovery algorithm such as orthogonal matching pursuit [24] is adopted to obtain a method with low computational complexity.

2. BACKGROUND INFORMATION

Before proceeding to the description of the proposed approach we provide a short overview of the sparse recovery framework. Let us assume that a signal $\mathbf{x} \in \mathbb{C}^N$ can be represented as $\mathbf{x} = \mathbf{\Psi s}$, where $\mathbf{\Psi} \in \mathbb{C}^{N \times N}$ is a *transform basis* and $\mathbf{s} \in \mathbb{C}^N$ denotes the *transform coefficients vector*. If $\mathbf{s}$ has only $K \ll N$ non-zero components, then $\mathbf{x}$ is called K-sparse.

Let us also consider an $M \times N$ matrix $\mathbf{\Phi}$ corresponding to the measurement process of signal $\mathbf{x}$ with $M < N$, where the rows of $\mathbf{\Phi}$ are *incoherent* with the columns of $\mathbf{\Psi}$. It is possible to obtain directly a *compressed set of measurements* $\mathbf{y}$ if the signal $\mathbf{x}$ is sparse in $\mathbf{\Psi}$, as follows:

$$\mathbf{y} = \mathbf{\Phi x} \overset{\mathbf{x}=\mathbf{\Psi s}}{=} \mathbf{\Phi \Psi s} \overset{\mathbf{A}:=\mathbf{\Phi \Psi}}{=} \mathbf{As}, \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{M \times N}$ corresponds to the *sensing matrix*. In real world applications the compressed measurements can be corrupted by noise $\mathbf{n} \in \mathbb{C}^M$, leading to noisy measurements of the form $\mathbf{y} = \mathbf{As} + \mathbf{n}$. Given the compressed measurements $\mathbf{y}$ and the sensing matrix $\mathbf{A}$ an ℓ_1-minimization problem can be solved to recover the sparse vector $\mathbf{s}$ as follows:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \|\mathbf{s}\|_1 \,, \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{As}\|_2 \leq \varepsilon \,, \tag{2}$$

which provides a recovery consistent with the observed measurements with an approximation error $\mathbf{y} - \mathbf{As}$ up to the noise level ε. The problem (2) is guaranteed to provide an accurate sparse solution with high probability if the *restricted isometry property (RIP)* of order K

$$(1 - \delta_K)\|\mathbf{s}\|_2^2 \leq \|\mathbf{As}\|_2^2 \leq (1 + \delta_K)\|\mathbf{s}\|_2^2 \tag{3}$$

is satisfied for sufficiently small values $\delta_K > 0$ [25]. However, it is computational intractable to explicitly verify the RIP as expressed in (3) especially for large size matrices [26], and thus it is preferable to use properties of the sensing matrix per se that are easily computable towards providing recovery guarantees. The *coherence* [27] of a matrix is one such property defined as

$$\mu = \max_{1 \leq i,j \leq N} \frac{|\boldsymbol{a}_i^H \boldsymbol{a}_j|}{||\boldsymbol{a}_i||_2 \, ||\boldsymbol{a}_j||_2} \,, \tag{4}$$

where $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ denote the i-th and j-th column, respectively, of the sensing matrix $\mathbf{A}$. It is important to notice that the lower the coherence the higher the probability to obtain a correct estimation of the sparse vector $\mathbf{s}$, which translates into linear independence among the columns of $\mathbf{A}$. However, it is obvious that a trade-off exists between the sparse recovery accuracy and the coherence violation. In other words, if the sparsity level is increased by increasing the dimension of the vector $\mathbf{s}$ (leading to a higher number of columns in $\mathbf{A}$), then the sparse recovery estimation is likely to provide decreased performance due to larger inter-column linear dependence. In addition, as stated in the introduction, inadequate discretization of the acoustic scene can lead to spectral leakage phenomena as a result of the so-called basis mismatch between the continuous physical model and the discretized assumed model reflected in $\mathbf{A}$. Next, we aim to efficiently tackle the aforementioned issues by introducing the concept of grid-shift.

3. PROPOSED METHOD

As a first step towards joint wideband acoustic source localization and acquisition we focus on computing the source positions. The location estimation problem is formulated in terms of a sparse recovery problem as described in the previous section.

3.1. Source localization based on spatial sparse recovery

Let us assume M microphones and K acoustic sources, with $K < M$, placed at random within an acoustic scene that corresponds to a box-shaped room of arbitrary dimensions. Let us also denote the position of the m-th microphone and the i-th source as $\mathbf{q}_m \in \mathbb{R}^3$ and $\mathbf{p}_i \in \mathbb{R}^3$, respectively. The physical model of the signal propagation between source i and microphone m accounting for the reflections onto the walls can be expressed via the Green's acoustic transfer function [28], [29], where the acoustic transfer function for each source-microphone pair is estimated based on the Image-Source model [28]. The time-frequency representation of the received signal at the m-th microphone can be written as follows:

$$y_m(t, \omega_l) = \sum_{i=1}^{K} A_{m,i}(\omega_l) x_i(t, \omega_l) + n_m(t, \omega_l) \tag{5}$$

for all $m = 1 \ldots, M$, where $t = 1, \ldots, T$ and $l = 1, \ldots, F$ is the time frame and angular frequency index, respectively. The total number of analysis frames and the number of frequency bins is denoted as T and F, respectively. The acoustic transfer function between microphone m and source i for a specific angular frequency $\omega_l = 2\pi f_l$, with f_l denoting the frequency in Hz, is $A_{m,i}(\omega_l)$. The i-th source and noise at microphone m are given by $x_i(t, \omega_l)$ and $n_m(t, \omega_l)$, respectively.

Let us assume that the area of interest is discretized into $N \gg K$ grid points, where each acoustic source could be possibly located at one out of N grid points. Under the previous assumption, an activation vector $\mathbf{s}(t, \omega_l) \in \mathbb{C}^N$ can be introduced as

$$s_n(t, \omega_l) = \begin{cases} x_i(t, \omega_l), & \text{if source } i \text{ is active at grid point } n \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

indicating if an acoustic source is active at a specific grid point or not. It is obvious that $\mathbf{s}(t, \omega_l)$ is a K-sparse vector. Using matrix form notation, and combining (5)-(6), the observation vector can be written as

$$\mathbf{y}(t, \omega_l) = \mathbf{A}(\omega_l)\mathbf{s}(t, \omega_l) + \mathbf{n}(t, \omega_l), \tag{7}$$

where $\mathbf{y}(t, \omega_l) \in \mathbb{C}^M$ denotes the complex-valued data vector from the observations at the M microphones, $\mathbf{s}(t, \omega_l) \in \mathbb{C}^N$ is the unknown vector of the complex source amplitudes at all N grid points of the grid of interest and $\mathbf{n}(t, \omega_l) \in \mathbb{C}^M$ is the additive noise error term. The sensing matrix

$$\mathbf{A}(\omega_l) = [\mathbf{a}(\mathbf{p}_1)|\mathbf{a}(\mathbf{p}_2)|\ldots|\mathbf{a}(\mathbf{p}_N)] \in \mathbb{C}^{M \times N} \tag{8}$$

acts as a mapping between $\mathbf{s}(t, \omega_l)$ and $\mathbf{y}(t, \omega_l)$ whose columns are the propagation vectors at all N grid points.

According to the description above, the acoustic source localization problem can be translated into the recovery of the sparse activation vector $\mathbf{s}(t, \omega_l)$ given the observation vector $\mathbf{y}(t, \omega_l)$ and the sensing matrix $\mathbf{A}(\omega_l)$. In principle, only a few sources generate the acoustic field, and thus we can assume that $K < M \ll N$ which means that problem (7) is underdetermined. Applying the sparse recovery framework as described in the previous section (as formulated in (2)), an estimate of the activation vector $\mathbf{s}(t, \omega_l)$ can be obtained by solving the following sparse optimization problem

$$\hat{\mathbf{s}}(t, \omega_l) = \arg\min_{\mathbf{s}} \|\mathbf{s}\|_1 \,, \text{ s.t. } \|\mathbf{y}(t, \omega_l) - \mathbf{A}(\omega_l)\mathbf{s}\|_2 \leq \varepsilon, \tag{9}$$

where ε is the noise threshold. Then, each source location can be easily inferred by the index of each non-zero element in $\hat{\mathbf{s}}(t,\omega_l)$. We adopt the orthogonal matching pursuit (OMP) [24] algorithm to solve (9).

As stated in the introduction, we are interested in wideband acoustic source localization as part of a generic joint acoustic source localization and acquisition system. This led us to incorporate and utilize the sparse recovery model under a real-time framework. As a result, the RIP should not be violated through a large number of grid points N, while at the same time the positions of the actual acoustic sources are assumed to rarely be in close proximity to the grid points especially in real-life scenarios (i.e., energy leakage will be observed from the actual source locations to the neighboring grid points).

Algorithm 1: Acoustic source localization using grid-shift

Input: $\mathbf{y}(t,\omega_l), t=1,\dots,T, l=1\dots,F$
grids $\boldsymbol{G}_r, r=1,\dots,R$, where R is the total number of grid-shifts
tolerance ε, number of sources K
Output: estimated source locations

1 **for** $r=1$ **to** R **do** // loop over grid shifts
2 **for** $t=1$ **to** T **do** // loop over time frames
3 **for** $l=1,\dots,F$ **do** // loop over frequency bins
4 build matrix $\mathbf{A}_r(\omega_l)$ using the current grid $\boldsymbol{G}_r$ according to (10)
5 $\hat{\mathbf{s}}(t,\omega_l) = \arg\min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{, s.t. } \|\mathbf{y}(t,\omega_l) - \mathbf{A}_r(\omega_l)\mathbf{s}\|_2 \leq \varepsilon$
6 $\mathbf{L}_r(t,\omega_l,:) \leftarrow \hat{\mathbf{s}}(t,\omega_l)^T$
7 **end**
8 **end**
9 $\mathbf{L}_r \in \mathbb{C}^{T\times F\times N}$
10 compute the magnitude of all $T\times F$ sparse solutions: $|\mathbf{L}_r|$
11 compute the mean $\boldsymbol{\mu}_r \in \mathbb{R}^{N\times 1}$ corresponding to all the $T\times F$ sparse solutions $|\mathbf{L}_r|$
12 find the indices $\{\hat{i}_1^r,\dots,\hat{i}_K^r\}$ of the top-K (maximum) values of the mean $\boldsymbol{\mu}_r$
13 find the current source position estimates $\hat{\mathbf{Q}}_r = [\boldsymbol{G}_r(:,\hat{i}_1^r),\dots,\boldsymbol{G}_r(:,\hat{i}_K^r)] \in \mathbb{R}^{3\times K}$
14 **end**
15 stack all the estimated source positions into one matrix $\hat{\mathbf{Q}} = [\hat{\mathbf{Q}}_1^T \dots \hat{\mathbf{Q}}_R^T]^T \in \mathbb{R}^{3\times RK}$
16 estimate source locations $\hat{\mathbf{P}} \in \mathbb{R}^{3\times K}$ by K-means clustering

To address the fore-mentioned problems, we introduce the concept of *grid-shift*. In particular, let us assume that a sensing matrix

$$\mathbf{A}_r(\omega_l) = [\mathbf{a}(\mathbf{p}_1^r)|\mathbf{a}(\mathbf{p}_2^r)|\dots|\mathbf{a}(\mathbf{p}_N^r)] \in \mathbb{C}^{M\times N} \tag{10}$$

corresponds to the r-th grid

$$\boldsymbol{G}_r = \left[\mathbf{p}_1^r|\mathbf{p}_2^r|\dots|\mathbf{p}_N^r\right] \in \mathbb{R}^{3\times N}, \tag{11}$$

where $\mathbf{p}_n^r \in \mathbb{R}^3$ (with $n=1,\dots,N$) are the coordinates of the grid points corresponding to the r-th grid. The grid size N is defined a-priori and a shifting procedure is followed to "scan" the entire box-shaped room. The process is iterative, and during each iteration (9) is solved for each grid $\boldsymbol{G}_r$ and for the time-frequency bin (t,ω_l), since we are dealing with wideband signals.

Algorithm 1 summarizes the wideband acoustic source localization procedure solving the ℓ_1-norm optimization problem as defined in (9) using the concept of grid-shift to compensate for the off-grid source locations. In line 6, $\mathbf{L}_r(t,\omega_l,:)$ denotes an N-dimensional vector, for fixed t and ω_l, in three-dimensional matrix $\mathbf{L}_r$, while in line 13 the notation $\boldsymbol{G}_r(:,\hat{i}_k^r)$ corresponds to the $\hat{i}_k^r$-th column of the matrix $\boldsymbol{G}_r$. To lower the source localization complexity in our practical implementation, a peak picking algorithm is applied to estimate the maximum spectral components of each time frame during the grid-shift process. Finally, the source location estimates $\hat{\mathbf{P}}$, provided by the K-means clustering, are used as input parameters during the source acquisition method.

3.2. Source acquisition based on estimated sparse solutions

After the estimation of the source locations we proceed with the acquisition of the original sources. Towards source acquisition the estimated source locations $\hat{\mathbf{P}}$ are exploited as side information (under the sparsity model (6)) in combination with the time-frequency separability property of speech signals [2]. As a result, we can address the frequency bin assignment problem which occurs when for example two different sources might be swapped at different frequency bins leading to ambiguity during the separation process (the interested reader is referred to [30]).

Specifically, to obtain the original source signals we need to solve a sparse optimization problem of the form (9) for all the time-frequency bins focusing on the estimated source positions $\hat{\mathbf{P}}$ which were provided by the K-means clustering process in Algorithm 1. During the first step, a grid of the same size N is built located in the center of the room. The grid points that are closest to the estimated source locations are replaced by the corresponding estimated locations' three-dimensional coordinates.

Algorithm 2: Acoustic source acquisition

Input: estimated source locations $\hat{\mathbf{P}}$
grid $\boldsymbol{G}$ of size N centered in the middle of the acoustic scene
tolerance ε, number of sources K
Output: separated sources (in the time domain)

1 find the indices $g_1,\dots,g_K$ of the grid points in $\boldsymbol{G}$ which are closest to the estimated source locations $\hat{\mathbf{P}}$
2 obtain a new grid $\tilde{\boldsymbol{G}}$ (from grid $\boldsymbol{G}$) by replacing the K grid points found in the previous step with the source location estimates $\hat{\mathbf{P}}$
3 **for** $t=1$ **to** T **do** // loop over all time frames
4 **for** $l=1,\dots,F$ **do** // loop over all frequency bins
5 build matrix $\tilde{\mathbf{A}}(\omega_l)$ using the grid $\tilde{\boldsymbol{G}}$
6 $\hat{\mathbf{s}}(t,\omega_l) = \arg\min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{, s.t. } \left\|\mathbf{y}(t,\omega_l) - \tilde{\mathbf{A}}(\omega_l)\mathbf{s}\right\|_2 \leq \varepsilon$
7 $\mathbf{L}(t,\omega_l,:) \leftarrow \hat{\mathbf{s}}(t,\omega_l)^T$
8 **end**
9 $\mathbf{L} \in \mathbb{C}^{T\times F\times N}$
10 **end**
11 **for** $k=1,\dots,K$ **do** // loop over each source
12 compute the inverse short-time Fourier transform (STFT) of $\mathbf{L}(:,:,g_k)$ to obtain the k-th source time-domain signal
13 **end**

Then, the optimization problem (9) is solved to obtain the estimated short-time Fourier transform (STFT) representation of each source. The inverse STFT is then applied for each source to compute the corresponding time-domain signal. Algorithm 2 summarizes the source acquisition process. The notation $\mathbf{L}(:,:,g_k)$ denotes the g_k-th page of the three-dimensional matrix $\mathbf{L}$.

4. EXPERIMENTAL RESULTS

In this section, we examine the localization performance of the proposed method in combination with the source acquisition quality. We used the Image-Source method [29] to simulate a box-shaped room of dimensions $6\times 3.5\times 3$ meters and produce signals of omni-directional speech sources at a reverberation time of 259 msec. We considered 10 microphones spread uniformly at random using all the space of dimensions $1.2\times 0.7\times 0.6$ meters located in the center of the room. In each simulation, the acoustic sources were speech recordings of 3 seconds sampled at 8 kHz and had equal power. The speech recordings were randomly selected from the VOICES

corpus, which is available from OGIs CSLU [31], consisting of 12 speakers (7 male and 5 female). To simulate different SNR values we added white Gaussian noise at each microphone, uncorrelated with the noise at other microphones and the source signals. During the localization/acquisition process a grid of size $7 \times 5 \times 6$ was used. The step size between each pair of neighboring grid points equals $\Delta\mathbf{p} = [0.75, 0.58, 0.42]$ meters along x-axis, y-axis and z-axis, respectively. This specific step size satisfies the need of appropriate coverage of the room as well as achieving low computational complexity without coherence violation. Towards this threefold direction the total number of grid-shifts R was set to eight during the localization.

We considered three scenarios of two, three and four sources placed uniformly at random within the acoustic scene. For processing, we used frames of 640 samples with 50% overlap and an FFT size of 1024. During the localization procedure we applied a peak picking process to estimate the ten maximum spectral components of each time-frame. First, we aimed at showing the effectiveness of the proposed grid-shift (GS) approach in terms of localization accuracy compared to the perturbed OMP (POMP) [21] method under the off-grid CS assumption. POMP adapts the signal dictionary to the actual measurements by performing perturbations of the parameters governing the signal dictionary. Here, we extended the POMP to a wideband version for anechoic acoustic source localization but due to lack of space we omit the details. Under these assumptions, in this experiment, we allowed for a large deviation in the off-grid offset of each source, where each off-grid distance was drawn from a uniform distribution over the range $[0.05\Delta\mathbf{p}, 0.5\Delta\mathbf{p}]$. It should be noted that the off-grid source locations are considered with respect to the non-shifted grid placed with regard to the center of the room. Table 1 depicts the localization root mean square error (RMSE) for both the GS and the compared POMP approach in the case of an anechoic environment, where white Gaussian noise of 5, 10, 15, 20 and 25 dB SNR is added at the microphones. It is obvious from the results that the proposed GS approach performs better in most cases.

Table 1: RMSE localization errors (in meters) for the grid-shift (GS) and perturbed-OMP (POMP) in the case of an anechoic environment.

SNR (dB)	two sources		three sources		four sources	
	GS	POMP	GS	POMP	GS	POMP
5	**0.2727**	0.3864	**0.4804**	0.5519	**0.3526**	0.6528
10	**0.3526**	0.5626	0.6642	**0.6067**	**0.4023**	0.7309
15	**0.3017**	0.3908	**0.3742**	0.4963	**0.4440**	0.5609
20	0.3517	**0.2828**	**0.3675**	0.3775	**0.3797**	0.4385
25	0.3244	**0.2464**	**0.3414**	0.3744	**0.3560**	0.6497

In the second experiment, we are interested in examining the localization efficiency of the proposed GS method under reverberant conditions against an MVDR-based beamformer[1] described in [32]. The beamformer is based on the computation of local angular spectra applied to each microphone pair and the resulting contributions (of all microphone pairs) are then aggregated following a pooling process. It is important to notice that the emergence of virtual sources is expected under a reverberant scenario. As a result, the energy leakage from the off-grid source locations to the neighboring grid points in combination with the emergence of virtual sources should lead us to develop a more careful analysis of the off-grid sparse recovery problem when large off-the-grid distances appear, and thus we choose that the off-grid offset of each source will be drawn from a uniform distribution over the range $[0.05\Delta\mathbf{p}, 0.2\Delta\mathbf{p}]$.

[1] http://bass-db.gforge.inria.fr/bss_locate/#mbss_locate

It is also important to notice that GS can provide an estimate of the radius of each estimated source location, while the MVDR beamformer gives an estimate only of the azimuth and elevation level of each source. Table 2 depicts the localization efficiency of the proposed GS approach against the MVDR beamformer in terms of angle-of-arrival error achieving better results in all noisy and reverberant cases.

Table 2: RMSE angle-of-arrival errors (in degrees) for the grid-shift (GS) and MVDR in the case of a reverberant environment.

SNR (dB)	two sources		three sources		four sources	
	GS	MVDR	GS	MVDR	GS	MVDR
5	**3.5230**	4.8109	**4.3875**	6.5245	**5.1718**	6.1696
10	**3.7004**	4.7953	**3.8504**	5.1604	**5.7962**	6.1664
15	**3.8226**	5.7409	**4.2013**	5.8668	**4.9736**	7.4858
20	**3.7263**	5.0382	**5.0752**	6.5678	**4.7431**	7.1989
25	**3.2619**	4.8438	**4.4231**	6.5934	**4.4616**	7.4194

Source acquisition performance is evaluated using the Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR) of the separated signals with the BSSEval toolbox [33]. As it is mentioned in Section 3.2, all the frequency bins are used during the acquisition process. Here, we also assumed a reverberant framework, and thus the off-grid offset of each source was drawn from a uniform distribution over the range $[0.05\Delta\mathbf{p}, 0.2\Delta\mathbf{p}]$. Table 3 shows the separation quality results in the case of a noisy and anechoic environment achieving good results especially in terms of SIR. Figure 1 depicts the BSSEval metrics in the case of a noisy and reverberant environment which can be seen that the proposed approach is promising for source separation under the off-grid sparse recovery assumption even in adverse acoustic conditions.

Table 3: Separation performance: no noise, no reverberation.

metric (dB)	two sources	three sources	four sources
SDR	5.7017	2.9224	2.1949
SIR	18.6001	13.7273	11.0495
SAR	6.3163	4.3916	3.9460

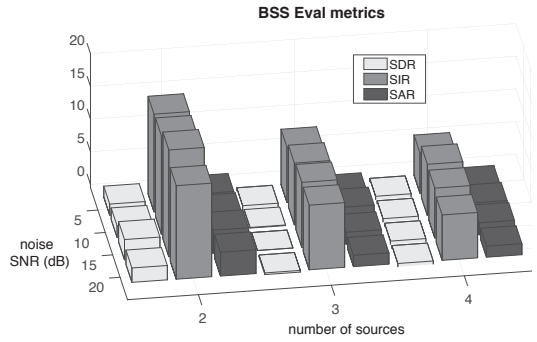

Figure 1: Separation performance: added white noise (SNR 5, 10, 15, 20 dB), reverberation time 259 msec.

5. CONCLUSIONS

In this work, we considered the joint problem of wideband acoustic source localization and acquisition in a microphone array of random arrangement under a sparse recovery framework. The concept of grid-shift was introduced to compensate for the displacement of the actual source positions with respect to the assumed grid point locations. An orthogonal matching pursuit-based method was adopted to speed up the location estimation process. Then, each acoustic source was acquired based on the microphone data and the estimated sparse location vectors. It was shown, through an experimental evaluation on real speech data, that the proposed method achieves an effective joint source localization and acquisition.

6. REFERENCES

[1] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. on Signal Processing*, vol. 53(8), pp. 3010–3022, August 2005.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52(7), pp. 1830–1847, July 2004.

[3] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. PRT Prentice Hall, Englewood Cliffs, NJ, 1993.

[4] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Proc. Magazine*, vol. 13(4), pp. 67–94, July 1996.

[5] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Trans. Signal Processing*, vol. 57(6), pp. 2275–2284, June 2009.

[6] G. F. Edelmann and C. F. Gaumond, "Beamforming using compressive sensing," *Journal of Acoust. Soc. of America (JASA)*, vol. 130, pp. 232–237, 2011.

[7] A. C. Gurbuz, J. H. McClellan, and V. Cevher, "A compressive beamforming method," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, March 2008, pp. 2617–2620.

[8] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *Journal of Acoust. Soc. of America (JASA)*, vol. 136, no. 1, pp. 260–271, 2014.

[9] E. Candés, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Information Theory*, vol. 52(2), pp. 489–509, February 2006.

[10] D. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52(4), pp. 1289–1306, April 2006.

[11] P. T. Boufounos, P. Smaragdis, and B. Raj, "Joint sparsity models for wideband array processing," in *Proc. SPIE*, August 2011.

[12] J. L. Roux, P. T. Boufounos, K. Kang, and J. R. Hershey, "Source localization in reverberant environments using sparse optimization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 4310–4314.

[13] A. Asaei, M. J. Taghizadeh, H. Bourlard, and V. Cevher, "Multi-party speech recovery exploiting structured sparsity models," in *Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH)*, Florence, Italy, August 2011.

[14] A. Asaei, M. E. Davies, H. Bourlard, and V. Cevher, "Computational methods for structured sparse component analysis of convolutive speech mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 2425–2428.

[15] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. on Information Theory*, vol. 59, no. 11, pp. 7465–7490, November 2013.

[16] X. Shen, J. Romberg, and Y. Gu, "Robust off-grid recovery from compressed measurements," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 3355–3359.

[17] A. Xenaki and P. Gerstoft, "Grid-free compressive beamforming," *Journal of Acoust. Soc. of America (JASA)*, vol. 137, no. 4, pp. 1923–1935, April 2015.

[18] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse Bayesian inference," *IEEE Trans. on Signal Processing*, vol. 61, no. 1, pp. 38–43, January 2013.

[19] T. Huang, Y. Liu, H. Meng, and X. Wang, "Adaptive matching pursuit with constrained total least squares," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 76, 2012.

[20] A. Fannjiang and H. C. Tseng, "Compressive radar with off-grid targets: A perturbation approach," *Inverse Problem*, vol. 29, no. 5, pp. 1–23, May 2013.

[21] O. Teke, A. C. Gurbuz, and O. Arikan, "Sparse delay-Doppler image reconstruction under off-grid problem," in *Proc. Sensor Array and Multichannel Signal Proc. Workshop*, A Coruña, Spain, June 2014, pp. 409–412.

[22] H. Zhu, G. Leus, and G. B. Giannakis, "Sparsity-cognizant total least-squares for perturbed compressive sampling," *IEEE Trans. on Signal Processing*, vol. 59, no. 5, pp. 2002–2016, May 2011.

[23] R. Jagannath and K. V. S. Hari, "Block sparse estimator for grid matching in single snapshot DoA estimation," *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 1038–1041, November 2013.

[24] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on Information Theory*, vol. 53(12), pp. 4655–4666, December 2007.

[25] E. Candés and M. B. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Magazine*, vol. 25(2), pp. 21–30, March 2008.

[26] W. U. Bajwa, R. Calderbank, and S. Jafarpour, "Why Gabor frames? Two fundamental measures of coherence and their role in model selection," *Journal of Communications and Networks*, vol. 12(4), pp. 289–307, August 2010.

[27] Y. C. Eldar and G. Kutyniok, Eds., *Compressed sensing: theory and applications*. Cambridge, New York: Cambridge University Press, 2012.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoust. Soc. of America (JASA)*, vol. 65, 1979.

[29] E. A. P. Habets, "Room impulse response generator," Ver. 2.0.20100920. [Available Online], 2010.

[30] G. R. Naik and W. Wang, *Blind source separation: advances in theory, algorithms and applications*. Springer, 2014.

[31] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.

[32] C. Blandin, A. Ozerov, and E. Vincent, "Multi source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, pp. 1950–1960, 2012.

[33] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.