Localization of Sound Sources in a Room with One Microphone

Helena Peić Tukuljac, Hervé Lissek and Pierre Vandergheynst

Signal Processing Laboratory LTS2
École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

ABSTRACT

Estimation of the location of sound sources is usually done using microphone arrays. Such settings provide an environment where we know the difference between the received signals among different microphones in the terms of phase or attenuation, which enables localization of the sound sources. In our solution we exploit the properties of the room transfer function in order to localize a sound source inside a room with only one microphone. The shape of the room and the position of the microphone are assumed to be known. The design guidelines and limitations of the sensing matrix are given. Implementation is based on the sparsity in the terms of voxels in a room that are occupied by a source. What is especially interesting about our solution is that we provide localization of the sound sources not only in the horizontal plane, but in the terms of the 3D coordinates inside the room.

Keywords: resonant frequency, room mode, room transfer function, sparsity, sound source localization

1. INTRODUCTION

In the last decade the theory of compressed sensing[1,2] has arised in the domain of acoustic signal processing. There was always a need for finding a structure in the high dimensional acoustical data that was cumbersome to handle. In 2015 Boche et al.[3] provided a detailed state of the art for the application of compressed sensing in the domains of image and acoustic signal processing.

The origins of sparsity in acoustical data include, but are not limited to: voxels (directions of arrival) occupied by sound sources which is usually exploited for the localization of the sound sources in free field[4] by designing a Fourier domain dictionary, or in rooms for the estimation of the sound pressure distribution.[5] The sparsity in the image-source model has been mainly used for the estimation of the room shape[6] and the estimation of the direction of arrivals of early echoes,[7] the sparsity of plane wave representation of the sound pressure is used for the characterization of the sound fields inside the room.[8] The sparsity of the room modes may also be exploited in the low-frequency range of the room transfer functions (RTF).[9] By RTF we denote the ration between the received and emitted signal in Fourier domain.

Instead of estimating the position of the sound sources from time difference of arrival between different microphones in an array,[10,11] we aim to rely only on one microphone and combine the sparsity that exists in the term of the voxels of a room occupied by the sound sources and the low-frequency room modes in the RTF toward successfull localization. To this end, we will analyze the transfer functions below the so called Schröeder frequency, which is defined as: $f_s = 2000\sqrt{\frac{RT_{60}}{V}}$, where V is the volume of the room and RT_{60} is the reverberation time.[12] This combination should result in a fast localization of sound sources by only one microphone as will be further explained.

The remainder of the paper is organized as follows: In Section 2 we discuss the sparsity that exists in the low frequency domain of room transfer functions. Section 3 gives a general introduction to compressed sensing and its application to the localization of sound sources. The design and the limitations of the sensing matrix for our case is given in Section 4 and final remarks and conclusions are given in Section 6.

Further author information:
Helena Peić Tukuljac: helena.peictukuljac@epfl.ch
Hervé Lissek: herve.lissek@epfl.ch
Pierre Vandergheynst: pierre.vandergheynst@epfl.ch

2. MODAL REPRESENTATION OF THE SOUND PRESSURE AND ITS LOW-FREQUENCY PROPERTIES

In the further development of our approach, we are going to rely on two facts: the room shape is known and the microphone position is known. These assumptions imply that we know the resonant frequencies of the room and the room modes related to the microphone's positions.

The solution of the wave equation for the sound pressure at a given receiver position $\mathbf{r_{mic}}$ and for a given source position $\mathbf{r_{ss}}$ in a room in the Fourier domain (with the angular frequency ω) is given by:[12]

$$H_\omega(\mathbf{r}_{\text{mic}}, \mathbf{r}_{\text{ss}}) = \rho_0 c^2 \omega Q \sum_n \frac{\Xi_n(\mathbf{r}_{\text{mic}})\Xi_n(\mathbf{r}_{\text{ss}})}{K_n[2\delta_n\omega_n + i(\omega^2 - \omega_n^2)]} \tag{1}$$

where ρ_0 is the density of the propagating medium (air), c is the sound celerity, Q is the volume flow velocity of the sound source, $\Xi_n(\cdot)$ are the eigenfunctions, K_n is the gain, δ_n is the damping coefficient and ω_n are the resonant frequencies. We can notice an interesting underlying symmetry that exists in this equation: position of the microphone $\mathbf{r}_{\text{mic}}$ and position of the sound source $\mathbf{r}_{\text{ss}}$ are interchangeable, meaning that if we exchange these positions, the expression will remain the same.

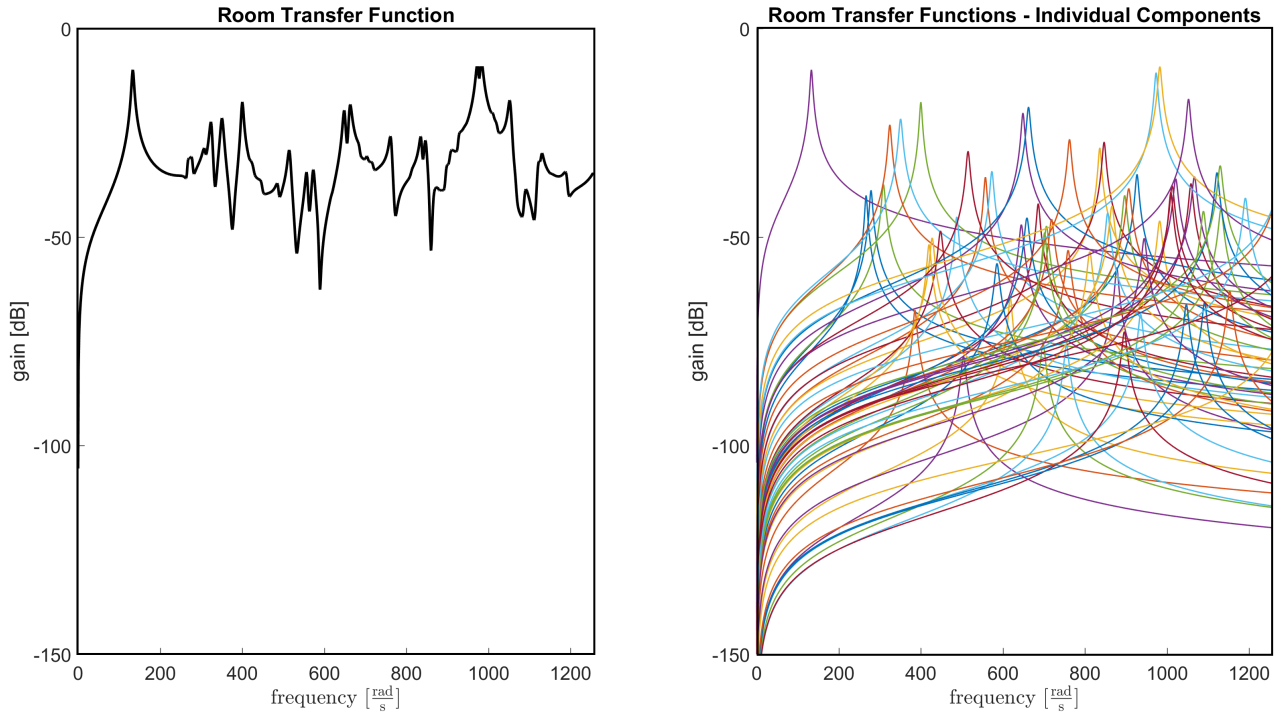

Figure 1. Individual components of the RTF are called room modes. As illustrated, room modes can be simply modeled as second order bandpass filters.

In Figure 1 we can see a segment of RTF for an arbitrary set of positions $\mathbf{r}_{mic}$ and $\mathbf{r}_{ss}$ up to $200Hz$ ($1200\frac{rad}{s}$) and its decomposition into the room modes. The sharpness of the peaks of room modes is dependent on the damping properties of walls of the room. Peaks of the room modes are aligned with the resonant frequencies of the room.

The angular eigenfrequencies for a rectangular room of size $L_x \times L_y \times L_z$ are given by the expression: $\omega_r = \pi c \sqrt{\left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_y}{L_y}\right)^2 + \left(\frac{n_z}{L_z}\right)^2}$ where $(n_x, n_y, n_z) \in \mathbb{N}_0^3 \setminus (0,0,0)$.

2.1 Room Transfer Function at Different Positions Across the Room

In 1985, Richardson et al.[13] have proposed a curve fitting algorithm allowing the reconstruction of the RTF curve from discrete measurements using room mode shaped functions as basic fitting elements. Each RTF is characterized by a set of parameters: resonant frequencies (eigenfrequencies) which are aligned with the position of the peaks of room modes, and with damping, attenuation and phase of these room modes.

For different positions of the microphones/sound sources across the room, some parameters stay the same - *common parameters*: eigenfrequencies which depend on the room shape, and the room mode damping which depends on the damping of the wall. The attenuation and the phase of the room modes are position dependent parameters - *specific parameters*.

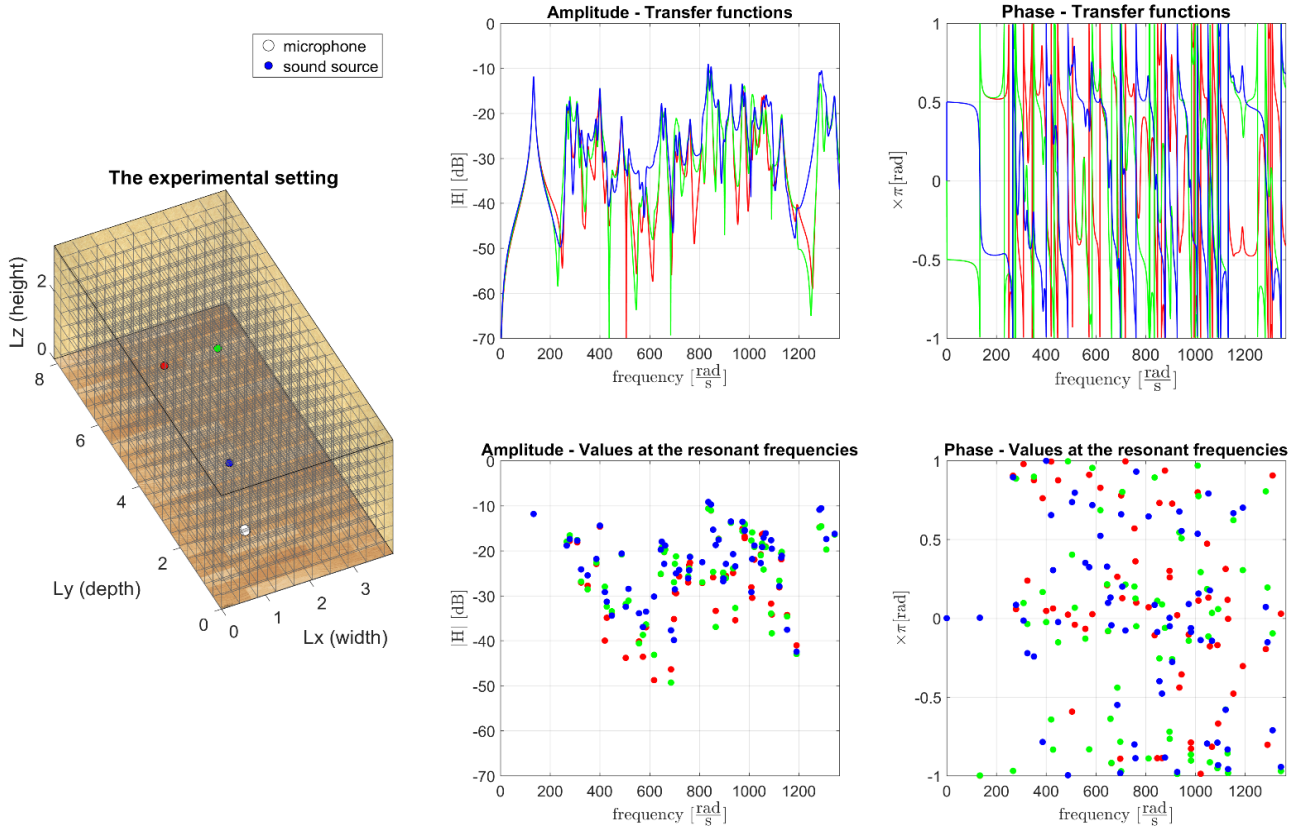

Figure 2. Values of the RTF across the room vary in the terms of attenuation and phase value at the resonant frequencies. We exploit only the difference in the attenuation because in our target experimental setting there exists only one microphone and the sources will emit white noise.

Figure 2 illustrates the difference between the attenuation and the phase of the RTFs across the room at the resonant frequencies. White point shows the fixed and known position of the microphone and colorful points are the positions of sound sources that should be estimated. As can be observed, although that all the positions of the sound sources result in the peaks at the same set of frequencies (the resonant frequencies of the room), the set of the heights of these peaks seems unique (this will be further observed in the next section). This means that each pair of the positions of a sound source and a microphone could potentially result in a unique set of attenuation factors at the resonant frequencies.

Although that there exists uniqueness of phase for each room mode, since we plan to use only one microphone and white noise sources, this is irrelevant for our case but has a potential for some other type of room characterization. We have decided to investigate the potential of unique representation of position of the sound source

within the room with the set of attenuations of RTF at resonant frequencies. Therefore we have established a valuable reasoning for the design of our sensing matrix.

2.2 Relation Between Room Modes and Plane Waves

In a rectangular room, each eigenfunction (eigenmode of the Laplacian operator) represents a sum of 8 plane waves that share a wave number:

$$\Xi(\mathbf{k}_n, \mathbf{r}_\mathrm{m}) = \sum_{i=1}^{8} a_i e^{j(\mathbf{S}(:,j)\odot\mathbf{k}_n)\cdot\mathbf{r}_m} \tag{2}$$

where $\odot$ is a Hadamard product, $\mathbf{S}_{3\times 8}$ is a sign matrix whose columns alternate from $[1,1,1]^T$ to $[-1,-1,-1]^T$, $\mathbf{k}_n = (\frac{n_x\pi}{L_x}, \frac{n_y\pi}{L_y}, \frac{n_z\pi}{L_z})$, $(n_x, n_y, n_z) \in \mathbb{N}_0^3 \setminus (0,0,0)$, is the eigenvalue of the wave equation for the n^{th} room mode (wave vector), and $\mathbf{r}_m$ is a position inside the room.

As can be seen in Figure 3, these wave vectors are just corners of a parallelepiped ($\mathbf{k} = [\pm k_x, \pm k_y, \pm k_z]^T$). We can also notice the periodicity of the wave vector grid: $\frac{\pi}{L_x}$, $\frac{\pi}{L_y}$, $\frac{\pi}{L_z}$, along each of the axes.

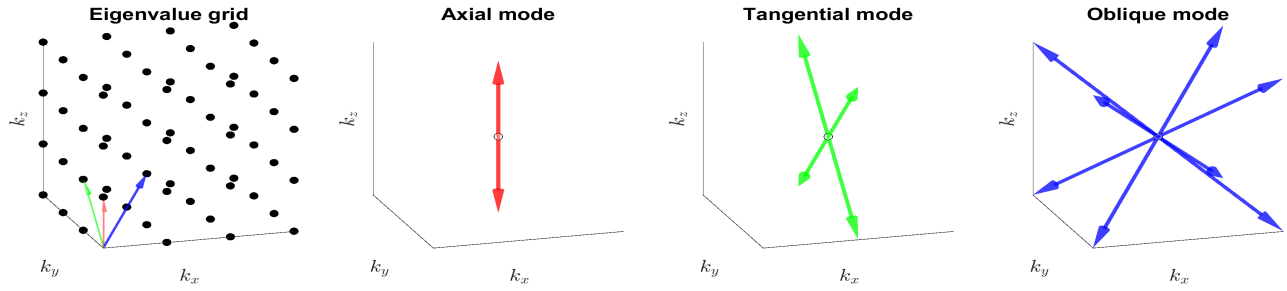

Figure 3. Eigenvalue space of a rectangular room with rigid walls. The left-hand side shows just one octant because of the symmetry that exists (there are 8 plane waves for each wave number). The length of the wave vector is proportional to the eigenvalue of the Laplacian.

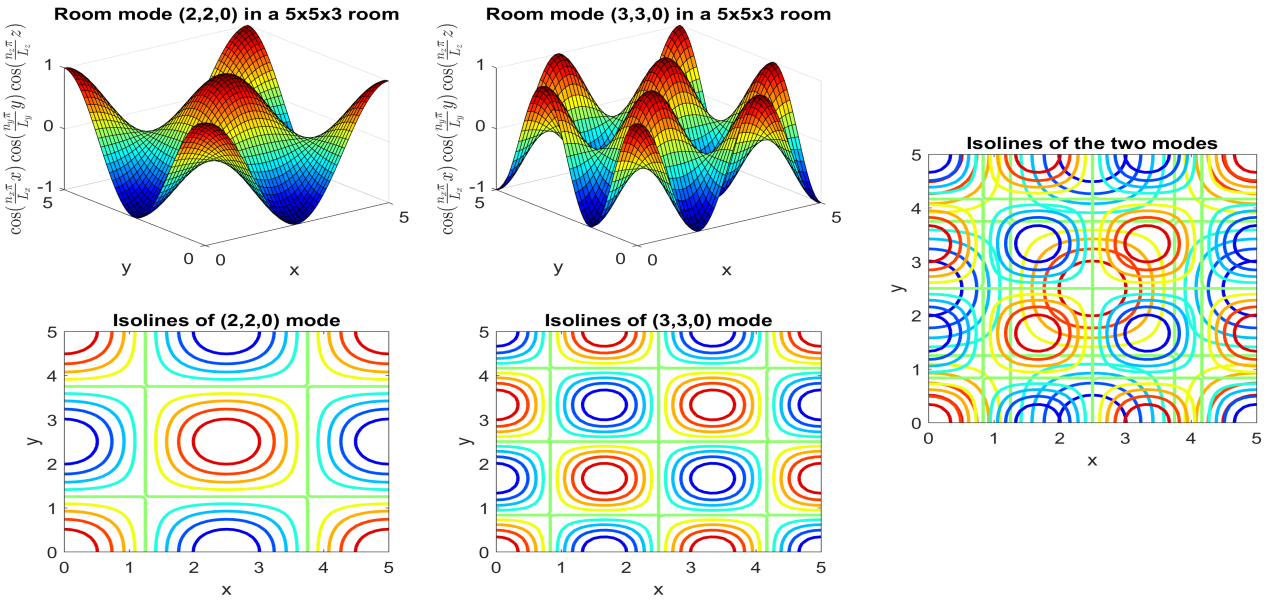

Figure 4. An example of $(n_x, n_y, n_z) \in \{(2,2,0), (3,3,0)\}$ room modes in a 5m × 5m × 3m room with rigid walls. We can notice that the isolines of different modes intersect in just a few locations, which supports our assumption of different height of sets of peaks in the RTF.

In a theoretical case for which all walls of a room are perfectly rigid, all plane waves have the same expansion coefficient ($\forall i,\ a_i = a$), so our sum of the 8 plane waves can be represented as a product of cosine functions:

$$\Xi(\mathbf{k}_n, \mathbf{r}_m) \sim a \cos\left(\frac{n_x \pi}{L_x}\mathbf{r}_m(x)\right) \cos\left(\frac{n_y \pi}{L_y}\mathbf{r}_m(y)\right) \cos\left(\frac{n_z \pi}{L_z}\mathbf{r}_m(z)\right). \tag{3}$$

where $\mathbf{r}_m(x),\ \mathbf{r}_m(y),\ \mathbf{r}_m(z)$ are the Cartesian coordinates of position $\mathbf{r}_m$ and a is a constant. An example of room mode for a room with rigid walls is given in Figure 4.

2.3 Ambiguities that Exist in the Terms of Uniqueness of the Attenuation Across the Room

We will observe the basic axial modes in Figure 5 in order to illustrate that relying only on them would not be sufficient to have a unique position representation. Although that the sound pressure value function is in 5D, we can successfully visualize only values in 3D. First row shows the x- and y-axial modes (everything that will be said applies analogously to z-axial modes as well). We can see that these two modes form pairs of points that result in a unique location identifier. But, since we have decided to explore the special case with only one microphone, we need to neglect the phase of the RTF, therefore we can just observe the absolute value of the RTF. As seen in the second row of the same figure, this introduces ambiguity - there exists a unique representation, but only in $\frac{1}{8}$ of the room.

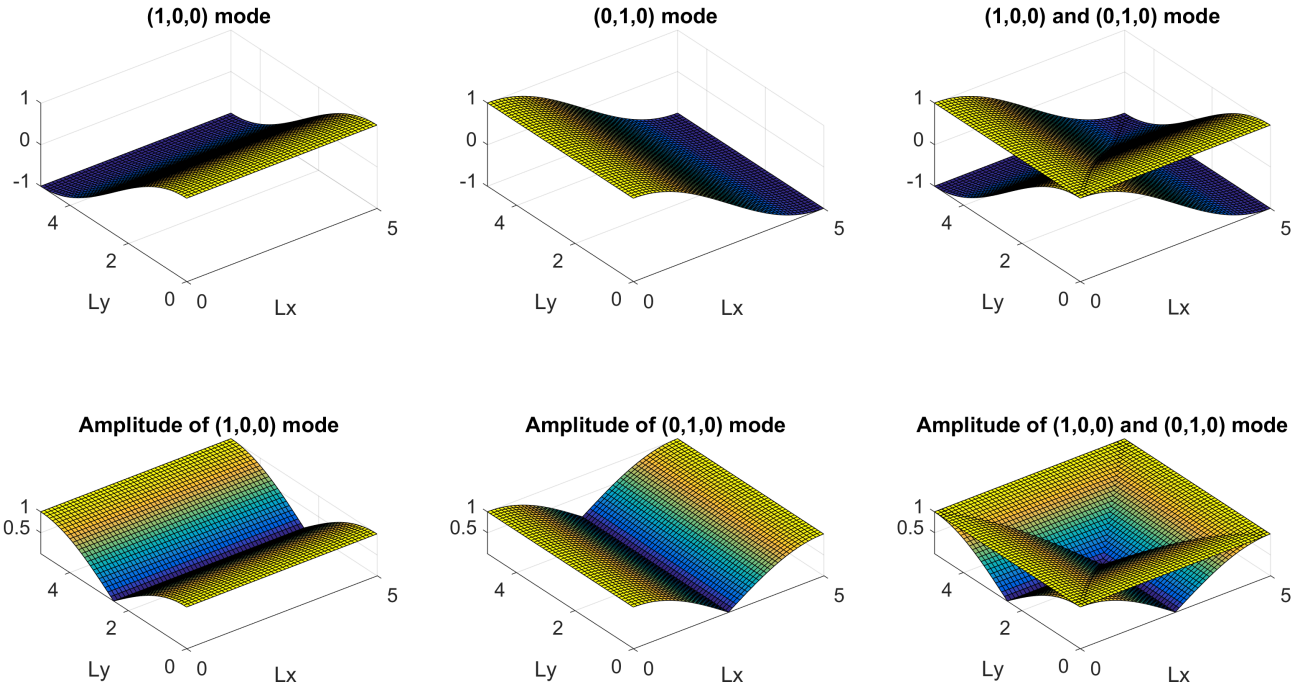

Figure 5. Basic modes and their attenuation values.

This ambiguity is illustrated in Figure 6. Axis represent the dependency between frequency and the real and the imaginary part of the RTF. Here we see 3 modes of two different positions in a room. The small mode in the middle has the same amplitude and phase and the other two modes have an opposite phase, which can not be seen when we project it to neglect the phase. This means that we can not rely only on the basic axial modes for the sound source localization.

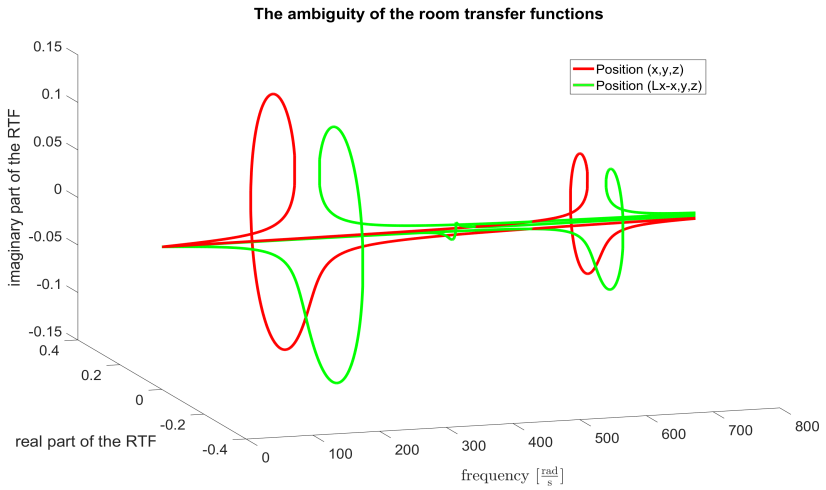

Figure 6. Ambiguities that exist in the term of the uniqueness of the RTF.

3. COMPRESSED SENSING AND SOUND SOURCE LOCALIZATION

3.1 Sparse Representation of the Position of Sources

In sound source localization problems the domain of interest is usually divided into an angular grid such that the sources occupy just a few of these angles. Since our sources are positioned inside a room, we will divide the room into voxels and assume that the number of voxels occupied by a source is small. We recognize that this is a problem with underlying sparsity. These problems are usually solved by using the theory of compressed sensing.

3.2 Compressed Sensing

Our signal of interest $\mathbf{y}$ is the measurement of sound pressure at a known location inside a known room:

$$\mathbf{y} = \mathbf{\Psi}\mathbf{\Phi}\mathbf{x} \tag{4}$$

where $\mathbf{y} \in \mathbb{R}^N$ are the sound pressure measurements, $\mathbf{\Psi}_{N\times N}$ is the inverse Fourier Transform (represents the change of domain), $\mathbf{\Phi}_{N\times M}$ is a representational dictionary (columns of this matrix are called atoms) with the RTFs as columns and $\mathbf{x} \in \mathbb{R}^M$ are the sparse expansion coefficients. The product $\mathbf{A} = \mathbf{\Psi}\mathbf{\Phi}$ is usually referred to as the sensing matrix. $\mathbf{x}$ is K-sparse, which means that it contains at most K non-zero elements and $K \ll N$. Since $M > N$ we are facing an underdetermined system of equations. The problem in this form is non-convex. Introducing the assumption on the sparsity of x makes the problem well-posed.

These types of problems are usually solved using one of the 5 groups of approaches listed in,[14] where the most common ones are the convex relaxation and the greedy pursuit. The convex relaxation,[15,16] which is also known as the Basis Pursuit, relies on the relaxation of the minimization of the ℓ_0-norm to ℓ_1-norm which favors the sparse solutions, although at a cost of requiring higher number of measurements.[17] There also exists a relaxation to ℓ_2-norm, but this norm favors the minimization of the energy of the signal rather than finding a sparse solution. In practice convex relaxation approach is usually used for smaller and medium size problems, because large scale data causes computational issues.

In our solution we will rely on the greedy approaches such as Orthogonal Matching Pursuit (OMP)[18] and Compressive Sampling Matching Pursuit (CoSaMP).[19] These methods select up to K atoms of a dictionary that give the least approximation error. CoSaMP is a faster contemporary method which works by selecting multiple atoms at every iteration. The main drawback of these methods is that the sparsity of the signal has to be known upfront.

Regardless of the approach, one of the main advantages of compressed sensing technique is the robustness to noise since we project our signal to the vectors that span the signal space, and therefore we neglect the residual related to the existing noise as long as the noise is not highly correlated with the signal.

3.3 Conditions for Dictionary Design

3.3.1 Spark and coherence of the dictionary

Spark of a matrix $\mathbf{\Phi}$ is the smallest number of linearly dependent columns of matrix $\mathbf{\Phi}$. The requirement for the sensing matrix $\mathbf{\Phi}$ in compressed sensing is that the following holds:

$$\text{spark}(\mathbf{\Phi}) > 2K \tag{5}$$

where K is the level of sparsity. In other words: To achieve an injective mapping we need to assure that there are no two K-sparse vectors that map to the same measurements. This implies that the rank of our sensing matrix has to be at least $2K$ which is tightly related to the restriction on the coherence of the dictionary.

According to the theory of compressed sensing, we have to ensure the appropriate coherence parameter of a dictionary: $\mu = \max_{1 \leq i < j \leq n} \frac{|\langle \varphi_i, \varphi_j \rangle|}{\|\varphi_i\|_2 \|\varphi_j\|_2} = \max_{1 \leq i < j \leq n} \cos \sphericalangle(\varphi_i, \varphi_j)$. Coherence is the cosine of the acute angle between the closest pair of atoms in a given dictionary. We want our dictionary to be incoherent, so μ should be the smallest possible. The best case is the case where we have orthogonal atoms with the coherence parameter equal to zero between different atoms of the dictionary.

3.3.2 Restricted Isometry Property

The restricted isometry property guarantees that the distances (lengths) are preserved when moving from one space to another. Let Φ be an $M \times N$ matrix and let $1 \leq K \leq N$ be an integer. Suppose that there exists a constant $\delta_K \in (0, 1)$ such that, for every $M \times K$ submatrix $\mathbf{\Phi}_K$ of $\mathbf{\Phi}$ and for every K-sparse vector $\mathbf{y}$,

$$(1 - \delta_K)\|\mathbf{y}\|_2^2 \leq \|\mathbf{\Phi}_K \mathbf{y}\|_2^2 \leq (1 + \delta_K)\|\mathbf{y}\|_2^2. \tag{6}$$

Then, the matrix $\mathbf{\Phi}$ is said to satisfy the K-restricted isometry property with restricted isometry constant δ_K. In most cases it is hard to check whether this property holds or not.

3.4 Sound Source Localization in a Room

The following question rises: How to tailor a simple incoherent dictionary for fast localization of sources inside the room? In order to have a well-posed problem we introduce the following assumptions:

1. the shape of the room and the reverberation time are known,
2. the position of the microphone is known, and
3. all the sound sources have a flat spectrum in the observed frequency range.

In most cases the Restricted Isometry Property is hard to check. We know that random matrices, which were used as the dictionaries in the early stages of compressed sensing, satisfy this property. Therefore we will choose the potential position of the sources on uniformly at random on the regular grid.

For each of the potential positions of sound sources and a fixed position of the microphone we have one atom in the dictionary which consists out of the height of the peaks in the RTFs at the resonant frequencies. The height of the dictionary is proportional to the number of the resonant frequencies in the observed frequency range. The number of resonant frequencies below a given frequency f_s[12] can be computed by: $N(f_s) = \frac{4}{3}\pi V \left(\frac{f_s}{c}\right)^3 + \frac{1}{4}\pi S \left(\frac{f_s}{c}\right)^2 + \frac{1}{2} L \frac{f_s}{c}$, where $V = L_x L_y L_z$, $S = 2(L_x L_y + L_y L_z + L_z L_x)$ and $L = L_x + L_y + L_z$. The width of the dictionary is proportional to the number of observation points on the predefined grid.

In order to localize the sources, we search for a subset of atoms that give the best fitting for the signal recorded by the microphone. Once we discover which atoms of our sensing matrix have the highest expansion coefficients in the sparse representation, we can easily recover the position of the sound sources in the room, because we know which atom corresponds to which position since we have tailored the dictionary ourselves.

4. DESIGNING AN EFFICIENT SENSING MATRIX

4.1 Coherence

Coherence of a dictionary can be seen from the maximum off-diagonal element of the coherence Gram matrix $\mu = \max_{i \neq j} \mathbf{G}_{ij}$. In our case where $\mathbf{\Psi}_{N \times N}$ is the inverse Fourier Transform and $\mathbf{\Phi}_{N \times M}$ is the matrix with the RTF coefficients, the Gram matrix has the following form:

$$\mathbf{G} = |\mathbf{A}^H \mathbf{A}| = |(\mathbf{\Psi}\mathbf{\Phi})^H \mathbf{\Psi}\mathbf{\Phi}| = |\mathbf{\Phi}^H \mathbf{\Psi}^H \mathbf{\Psi}\mathbf{\Phi}|. \tag{7}$$

Since the Fourier matrix has orthonormal atoms up to a scaling constant $\mathbf{\Psi}^H\mathbf{\Psi} = \frac{1}{N}\mathbf{I}$, we have:

$$\mathbf{G} = \frac{1}{N}\mathbf{\Phi}^H\mathbf{\Phi}. \tag{8}$$

Therefore we observe the coherence of the sensing matrix by focusing on the discretization of the room transfer function. Since our exponentials in the plane wave representation are not equidistant, we can not apply the Dirichlet kernel sum to our case to simplify the expression (an approach common for many solutions[4,20,21]).

For a uniform case, the off-diagonal elements of our Gram matrix are proportional to:

$$\mathbf{G}_{ij} \sim cos(k_x r_x)cos(k_x(r_x \pm m\Delta x)) + cos(k_y r_y)cos(k_x(r_y \pm n\Delta y)) + cos(k_z r_z)cos(k_x(r_z \pm o\Delta z)). \tag{9}$$

It results in a complex form of the elements of Gram matrix. Some observations have shown that we are dealing with highly correlated atoms. Therefore we need to find a workaround in order to have a successful source localization. Due to the smoothness of cosine function, the points on the potential sound source position grid that lay close, result in similar heights of the peaks in RIR.

4.2 Battle of the Grids

Our problem has two degrees of freedom and both of them represent a selection process of the nodes on a uniform grid. We have a grid of wave vectors - *features* and a grid of potential positions of sound sources - *samples*. In Figure 7 the grid on the left-hand side repeats in all 6 directions and the one on the right-hand side repeats in 3 directions.

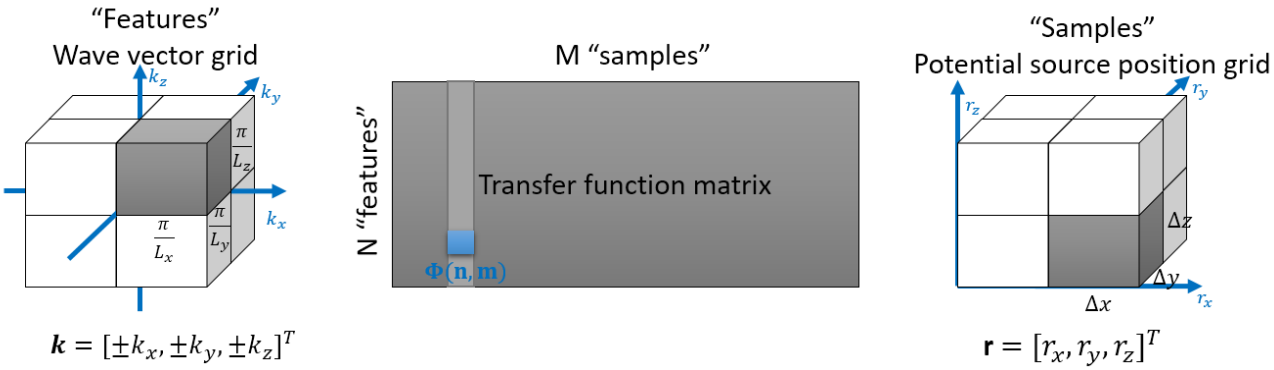

Figure 7. Two grids that represent two degrees of freedom that we have for designing the sensing matrix.

We will observe the room transfer function in a matrix form at the resonant frequencies. If we go back to equation (1) and introduce $\omega = \omega_n$, we get that each of the elements of our sensing matrix $\mathbf{\Phi}_{N \times M}$ is of the following form:

$$\mathbf{\Phi}(n,m) = \frac{\rho_0 c^2 Q_m}{2K_n \delta_n} \Xi(\mathbf{k}_n, \mathbf{r}_{\text{mic}}) \Xi(\mathbf{k}_n, \mathbf{r}_m) \tag{10}$$

which corresponds to n^{th} wave vector and m^{th} potential sound source position. The only coefficients that differ among the atoms of the dictionary are represented in blue. The difference due to the volume velocity of the

sound source Q, will not affect our approach since we assume that we are observing our sound sources in a linear regime. This parameter has an effect only on the expansion coefficients of the sparse representation. Therefore we focus on the sound sources' position that produces different attenuation of room modes.

So the RTF matrix has the following decomposition:

$$\mathbf{\Phi} = \frac{\rho_0 c^2}{2} \begin{bmatrix} \frac{\Xi(\mathbf{k}_1, \mathbf{r}_{\text{mic}})}{K_1 \delta_1} & \dots & \frac{\Xi(\mathbf{k}_1, \mathbf{r}_{\text{mic}})}{K_1 \delta_1} \\ \vdots & \ddots & \vdots \\ \frac{\Xi(\mathbf{k}_N, \mathbf{r}_{\text{mic}})}{K_N \delta_N} & \dots & \frac{\Xi(\mathbf{k}_N, \mathbf{r}_{\text{mic}})}{K_N \delta_N} \end{bmatrix} \odot \begin{bmatrix} Q_1 \Xi(\mathbf{k}_1, \mathbf{r}_1) & \dots & Q_M \Xi(\mathbf{k}_1, \mathbf{r}_{\text{M}}) \\ \vdots & \ddots & \vdots \\ Q_1 \Xi(\mathbf{k}_N, \mathbf{r}_1) & \dots & Q_M \Xi(\mathbf{k}_N, \mathbf{r}_{\text{M}}) \end{bmatrix}. \tag{11}$$

Just to recall, our rigid wall room modes are of the form: $\Xi(\mathbf{k}_n, \mathbf{r}_{\text{m}}) = \sum_{i=1}^{8} e^{j(\mathbf{S}(:,j) \odot \mathbf{k}_n) \cdot \mathbf{r}_m}$, where $\mathbf{k}_n$ belongs to the positive octant of the left-hand side grid.

5. RESULTS

5.1 The Recovery of Signal's Support in a Highly Coherent Dictionary

Candès et al.[22] discuss the potential of recovery of data that has a sparse representation in a coherent dictionary. Coherent dictionaries can give guarantees only on the recovery of the sparse signal, but not on the recovery of the set of indices of atoms in sparse representation. That is because if we have pairs of atoms that are extremely coherent (almost collinear), e.i. we are far away from satisfying $\mu \leq \frac{1}{3(s-1)}$, we can not tell which one of them will be used for our sparse representation when projecting to a lower-dimension space. Schnass et al. have approached this problem by introducing a complementary dictionary of the same size, but with low coherence, which maintains the sparse support of the measurements.[23] Our approach will be in the spirit of random subdictionary selection.[24] There have been some approaches with subsampling of dictionaries over rows and columns in order to increase the speed of the convergence of greedy methods,[25,26] but using such subsampling methods for coherent dictionaries is still unexplored. Authors of these papers named one of these methods as StoCoSaMP (Stochastic CoSaMP).

We restate our problem in the following manner: Recover sparse signal $\mathbf{x}$ from the following:

$$\mathbf{S}_{rf} \mathbf{\Psi}^* \mathbf{y} = \mathbf{S}_{rf} \mathbf{\Phi} \mathbf{S}_{sp} \mathbf{x} \tag{12}$$

where $\mathbf{y}$ is the measured signal, $\mathbf{S}_{rf}$ is a resonant frequency selector that defines which points on the wave vector grid we observe, $\mathbf{S}_{sp}$ is a sound source position selector that defines which points on the potential source position grid we observe and $\mathbf{\Psi}^*$ is the Fourier transform. Both matrices, $\mathbf{S}_{rf}$ and $\mathbf{S}_{sp}$, are just submatrices of an identity matrix - the first one is constructed from selected rows and the second one is constructed from selected columns. We could characterize our case as a highly sparse case, since the number of sources to be localized is going to be small (only one or a few).

Support of $\mathbf{x}$ shows which of the positions on the grid are the most probable positions of the sources. Without subsampling of the coherent dictionary, this support is usually wrongly estimated due to the ill-conditioness of the problem coming form the high coherence of the dictionary.

Here is the description of the algorithm ($\mathbf{I}$ is the identity matrix):

Algorithm 1 Localization of sound sources in a room with one microphone

input : Highly coherent room mode dictionary $\mathbf{\Phi}_{N\times M}$, uniform grid of potential points of the sound sources and ground truth positions of the sound sources (including the measured signal in Fourier domain $\mathbf{y^F} = \mathbf{\Psi}^*\mathbf{y}$).

output : Reconstructed positions of the sound sources.

do

Generate random subsampling matrices $\mathbf{S}_{rf} \underset{row}{\subset} \mathbf{I}_{N\times N}$ and $\mathbf{S}_{sp} \underset{column}{\subset} \mathbf{I}_{M\times M}$.

Subsample the dictionary: $\Phi_{ss} = \mathbf{S}_{rf}\mathbf{\Phi}\mathbf{S}_{sp}$ and the measured signal $\mathbf{y^F_{ss}} = \mathbf{S}_{rf}\mathbf{y^F}$.

Try to estimate the positions of the sound sources by estimating the support of $\mathbf{x}$ on $\mathbf{\Phi}_{ss}$ using CoSaMP for the given measured signal $\mathbf{y^F_{ss}}$ knowing the level of sparsity.

while CoSaMP[19] sparse representation does not converge (has norm of the residual a lot greater than zero)

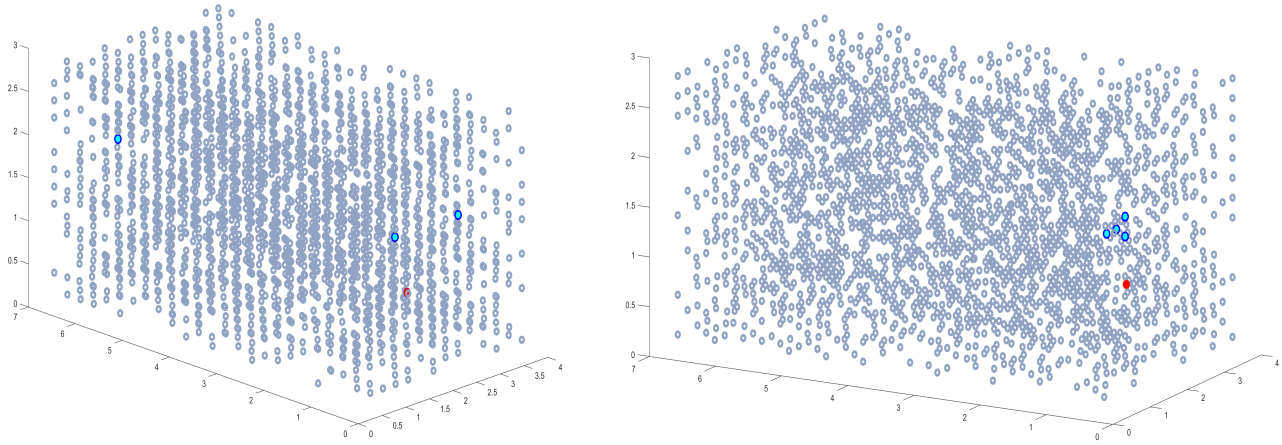

Figure 8. These are the results for localization of 3 sound sources inside a $4m\times7m\times3m$ room for a uniformly undersampled $10 \times 15 \times 10$ grid.

Figure 8. shows a reconstruction example for a case with 3 sound sources. Grey circles are the potential positions taken into account in the current iteration, blue circles are the true positions and light blue points are the reconstructed positions. The red point represents the known position of the sound source. This algorithm has no problems with identifying position of sources that are close, as can be seen from the right hand side of the figure.

We will observe how different subsampling schemes effect the success and speed of our sparse support estimation.

We have performed 100 Monte Carlo simulations for each set of parameters and for the estimation of the position of two sound sources. Experiments were performed on a single core of Intel Xeon processor at 2.8GHz of a computer with 16GB of RAM. If the algorithm did not converge within 300 iterations, we would consider that to be a failure. If we do not bound the number of iterations, the algorithm always converges but sometimes it needs a few thousands of iterations. Reconstruction time does not include the time needed for constructing the dictionary.

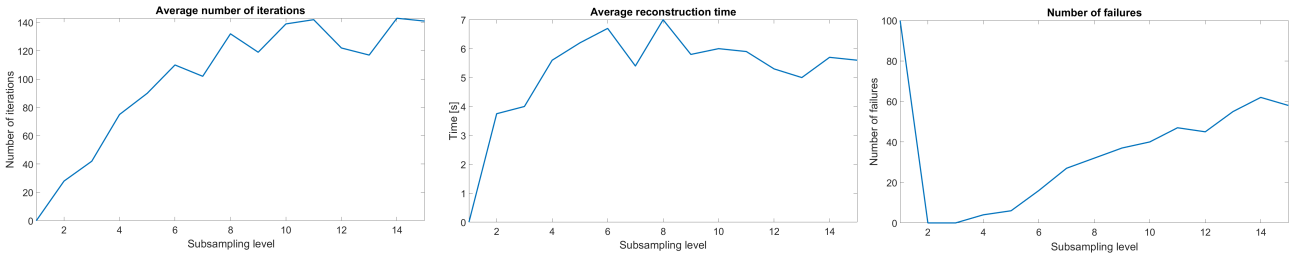

Figure 9. Here we see results for different potential sound source position grid subsampling (from no subsampling up to subsampling 15 times).

In Figure 9 we can see results for no subsampling over resonant frequencies (first 63 resonant frequencies were taken into account - room modes between $(1,0,0)$ and $(3,3,3)$) and different subsamplings over the potential sound source positions. There were no successful reconstruction attempts when the whole grid was taken into account. Subsampling 2 or 3 times showed the best performance with the convergence within the predefined 300 iterations. Average number of iterations and average reconstruction time were computed only for the successful quick reconstructions.

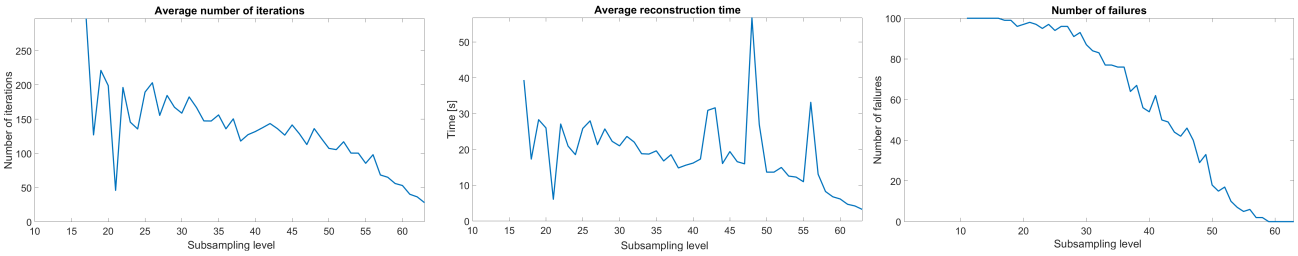

Figure 10. Here we see results for different resonant frequency grid subsampling (from selecting 11 up to selecting all 63 resonant frequencies).

In Figure 10 we can see results for subsampling level of 2 over the potential positions of sound sources and different subsets of resonant frequencies have been taken into account (from 11 up to 63 out of 63). If we choose a subset of below 17 resonant frequencies, the algorithm never converges. If we had average results over more than 100 simulations, the curves in the results would have been smoother. We see that we can not subsample a lot such a small set of resonant frequencies.

Therefore, we have to subsample the sound source position grid since we are dealing with a highly coherent dictionary. By increasing the level of subsampling over columns of the dictionary, we decrease the probability that the atoms that we are searching for are present in the subset. On the other hand, the resonant frequency grid should not be too oversampled in order to achieve a quick convergence (below predefined 300 iterations or similar).

5.2 Precision and Basis Mismatch

Due to the smoothness of the room mode functions, there is a small variation in the value between the close points. This supports the idea of similarity of the atoms of the dictionary of the spatially close positions.

Compressed sensing usually assumes the existence of a grid with finite density and our signals of interests can fail to coincide with the nodes of the predefined grid, especially in the case of moving sources. As shown in[21] this can cause that sparse signals appear as incompressible. The work we have observed before[4] has an extension to a continuous case[27] by applying the semi-definite programming.[28] In our observations we have assumed that our grid of the potential positions of the sources is dense enough to avoid the spectral leakage and continuous approaches will be left for future work.

5.3 Requirements and Limitations

In a setting where we have multiple sound sources and a microphone, the sound received is equal to the linear combination of the convolution of sounds emitted by the sound sources and the transfer functions that correspond to their positions. Therefore we need the following assumption: we can efficiently localize sources which have a flat spectrum in the observed frequency band, since they result in a nearly constant Fourier coefficients of emitted signals spectrum. Otherwise we have to know upfront the signals that will be emitted by sources.

In order to avoid ill-conditioness the microphone should lie off the planes of symmetry.

6. CONCLUSION

By observing the sound source localization problem through the theory of compressed sensing, we have enabled localization of multiple sound sources in a room using only one microphone. Unlike most of the localization algorithms, this approach guaranties the localization in 3D, without neglecting the elevation angle, which is rarely estimated. The simplicity of our solution lays in the low required prior knowledge about the room - only the height of the peaks in the RTF at the resonant frequencies should be know.

Our solution has the potential of being applied to the optimization of the quality of the hearing aids - once the location of source is estimated we can introduce weighting on the reception side, as well as in robotics for monoaural localization. The emerging field of virtual reality would be just another domain of potential application.

Future work will include estimation *off the grid* in order to avoid the basis mismatch and the challenging computational costs. Removal of the assumption on the level of sparsity should also be investigated further.

7. SUPPLEMENTARY MATERIALS

matlab code used for generating each of the figures in this paper as well as the acoustical room mode framework is available for download on the following link: https://github.com/epfl-lts2/room_transfer_function_toolkit. *python* version of the toolkit is also available.

ACKNOWLEDGMENTS

The work of H. Peić Tukuljac was supported by the Swiss National Science Foundation under Grant No. 200021_169360 for the project *"Compressive Sensing applied to the Characterization and the Control of Room Acoustics"*. We would like to thank Adrien Besson for fruitful discussions and valuable suggestions during the preparation of this manuscript.

REFERENCES

[1] Donoho, D. L., "Compressed sensing," *IEEE Trans. Information Theory* **52**(4), 1289–1306 (2006).

[2] Candes, E. J., Romberg, J., and Tao, T., "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theor.* **52**, 489–509 (Feb. 2006).

[3] Boche, H., Calderbank, R., Kutyniok, G., and Vybral, J., [*Compressed Sensing and Its Applications: MATHEON Workshop 2013*], Birkhäuser Basel, 1st ed. (2015).

[4] Xenaki, A., Gerstoft, P., and Mosegaard, K., "Compressive beamforming," *The Journal of the Acoustical Society of America* **136**(1), 260–271 (2014).

[5] Kitić, S., Bertin, N., and Gribonval, R., "Hearing behind walls: Localizing sources in the room next door with cosparsity," in [*2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 3087–3091 (May 2014).

[6] Dokmanić, I., Parhizkar, R., Walther, A., Lu, Y. M., and Vetterli, M., "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences* **110**(30), 12186–12191 (2013).

[7] X. Falourd, L. Rohr, M. R. and Lissek, H., "Spatial echogram analysis of a small auditorium with observations on the dispersion of early reflections," *Inter-Noise 2010 - noise and sustainability* (June 2010).

[8] Koyano, Y., Yatabe, K., Ikeda, Y., and Oikawa, Y., "Physical-model based efficient data representation for many-channel microphone array," in [*2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 370–374 (March 2016).

[9] Mignot, R., Chardon, G., and Daudet, L., "Low frequency interpolation of room impulse responses using compressed sensing," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* **22**, 205–216 (Jan. 2014).

[10] Marmaroli, P., Carmona, M., Odobez, J. M., Falourd, X., and Lissek, H., "Observation of vehicle axles through pass-by noise: A strategy of microphone array design," *IEEE Transactions on Intelligent Transportation Systems* **14**, 1654–1664 (Dec 2013).

[11] Marmaroli, P., Odobez, J. M., Falourd, X., and Lissek, H., "A bimodal sound source model for vehicle tracking in traffic monitoring," in [*2011 19th European Signal Processing Conference*], 1327–1331 (Aug 2011).

[12] Kuttruff, H. and Mommertz, E., [*Room Acoustics*], 239–267, Springer Berlin Heidelberg, Berlin, Heidelberg (2013).

[13] Richardson, M. H. and Formenti, D. L., "Global curve fitting of frequency response measurements using the rational fraction polynomial method," (1985).

[14] Tropp, J. A. and Wright, S. J., "Computational methods for sparse solution of linear inverse problems," *Proceedings of the IEEE* **98**, 948–958 (June 2010).

[15] Tropp, J. A., "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory* **52**, 1030–1051 (March 2006).

[16] Boyd, S. and Vandenberghe, L., [*Convex Optimization*], Cambridge University Press, New York, NY, USA (2004).

[17] Candès, E. J., Wakin, M. B., and Boyd, S. P., "Enhancing sparsity by reweighted l 1 minimization," *Journal of Fourier Analysis and Applications* **14**(5), 877–905 (2008).

[18] Tropp, J. A., Gilbert, A. C., and Strauss, M. J., "Algorithms for simultaneous sparse approximation: Part i: Greedy pursuit," *Signal Process.* **86**, 572–588 (Mar. 2006).

[19] Needell, D. and Tropp, J., "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis* **26**(3), 301 – 321 (2009).

[20] Blu, T., Dragotti, P. L., Vetterli, M., Marziliano, P., and Coulot, L., "Sparse sampling of signal innovations," *IEEE Signal Processing Magazine* **25**, 31–40 (March 2008).

[21] Chi, Y., Pezeshki, A., Scharf, L., and Calderbank, R., "Sensitivity to basis mismatch in compressed sensing," in [*2010 IEEE International Conference on Acoustics, Speech and Signal Processing*], 3930–3933 (March 2010).

[22] Candès, E. J., Eldar, Y. C., and Needell, D., "Compressed sensing with coherent and redundant dictionaries," *CoRR* **abs/1005.2613** (2010).

[23] Schnass, K. and Vandergheynst, P., "Dictionary preconditioning for greedy algorithms," *IEEE Transactions on Signal Processing* **56**, 1994–2002 (May 2008).

[24] Tropp, J. A., "On the conditioning of random subdictionaries," *Applied and Computational Harmonic Analysis* **25**(1), 1 – 24 (2008).

[25] Peel, T., Emiya, V., Ralaivola, L., and Anthoine, S., "Matching pursuit with stochastic selection," in [*2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*], 879–883 (Aug 2012).

[26] Pal, D. K. and Mengshoel, O. J., "Stochastic cosamp: Randomizing greedy pursuit for sparse signal recovery," in [*Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*], 761–776 (2016).

[27] Xenaki, A. and Gerstoft, P., "Grid-free compressive beamforming," *CoRR* **abs/1504.01662** (2015).

[28] Vandenberghe, L. and Boyd, S., "Semidefinite programming," *SIAM Rev.* **38**, 49–95 (Mar. 1996).