Acoustic DoA Estimation by One Unsophisticated Sensor

Dalia El Badawy[1], Ivan Dokmanić[2], and Martin Vetterli[1]

[1] École Polytechnique Fédérale de Lausanne
{dalia.elbadawy,martin.vetterli}@epfl.ch,
[2] University of Illinois at Urbana-Champaign
dokmanic@illinois.edu

Abstract. We show how introducing known scattering can be used in direction of arrival estimation by a single sensor. We first present an analysis of the geometry of the underlying measurement space and show how it enables localizing white sources. Then, we extend the solution to more challenging non-white sources like speech by including a source model and considering convex relaxations with group sparsity penalties. We conclude with numerical simulations using an unsophisticated sensing device to validate the theory.

Keywords: monaural localization, compressed sensing, direction of arrival, group sparsity, scattering, sound source localization

1 Introduction

Walking down a street, we (or a cat) are able to tell where a bird song is coming from. Perhaps it helps that we know birds live in trees, but it is the auditory scene analysis performed by the brain that enables us to almost instantaneously determine the direction of arrival (DoA), even for multiple sound sources [10]. In this paper, we study computational DoA estimation with a single sensor, a task usually referred to as monaural sound source localization. We begin by a brief review of the biological mechanisms from which we draw some inspiration.

First of all, we have two ears. Sound reaches each ear at a slightly different time and loudness providing us with so-called binaural cues. The shape of the outer ear as well as the shape of the head and torso additionally modify the sound as it reaches our ears, and thus provide us with *monaural* cues. These cues are encoded by the head-related transfer function (HRTF) [1]. Both types of cues are necessary for accurate localization. Indeed, obstructing one ear hurts the localization accuracy [7]. However, monaural localization is still possible, though it is known that monaurally deaf people usually require certain prior knowledge about the source to be localized [11]. We will see that (unless the sources are white), the same is true of algorithms we propose.

Consider a generalized *ear*, a sensor with the directional frequency response $\mathsf{a}(\omega;\theta)$ for sounds arriving from direction θ at a frequency ω. For J sources

emitting from directions $\Theta = \{\theta_1, \ldots, \theta_J\}$, what we measure at the single sensor is

$$\mathsf{y}(\omega) = \sum_{j=1}^{J} \mathsf{a}(\omega; \theta_j)\mathsf{s}_j(\omega) + \mathsf{n}(\omega), \tag{1}$$

where s_j are the source spectra and n is the measurement noise which will be ignored in the large part of the ensuing discussion.

DoA estimation is then an inverse problem concerned with mapping the measurement back to the directions Θ. The properties of the directional response a are key in determining whether it can be successful. For instance, if our sensor is omnidirectional, then $\mathsf{a}(\omega, \theta) = 1$ for all frequencies ω and directions θ, and no directional information is present. That is, the measurement remains the same even if the sources are rotated to different directions. Thus, we would prefer that a are diverse and act as distinguishable spectral signatures for their corresponding directions. Still, as can be seen from (1), the inverse problem is ill-posed since decomposing y back into a sum of products has infinitely many solutions. This ill-posedness can be resolved by a combination of scattering and proper source modeling.

Requiring that the responses a be diverse is similar to the HRTF case where for each ear, the frequency response differs with the angle of arrival. An especially interesting HRTF is that of a cat: it features prominent notches at frequencies that depend on the direction of arrival [10].[3] In fact, notches are one of two possibilities to get strong diversity, the other being resonances. They both enable localization of wideband sources, but while in enclosures such as rooms, they are easy to obtain and have been successfully used for localization [2], they otherwise require special design. For example, resonances were obtained in recent work [14] with a metamaterial-coated device which was then used to localize noise. Similarly, diversity of a was achieved in [9] using several microphone enclosures which were designed and tested for localizing a single sound source. In our prior work [3], we used a randomly shaped device to introduce random scattering and showed that noise can be localized without a source model. While all the latter work relies on the idea of a directional spectral signature, it was not made precise why or how such spectral signatures are good for DoA estimation. As we will show, whereas any *incoherence* of a is sufficient to localize noise sources, in order to compensate for the lack of diversity and to handle complex sound sources, an adequate source model is required, for example, a Hidden Markov Model [9] or a dictionary [14].

In this paper, we achieve the desired a by scattering by a very simple, haphazard structure. Unlike prior work, we show in Section 2 that the underlying principle that makes scattering useful requires neither a sophisticated sensing device nor a source model to localize noise. The geometry of the problem suggests a matched field processing approach [13] to DoA estimation, which has reasonable complexity for few sources. Then in Section 3, we turn to sparse reconstruction techniques with group sparsity penalties that can be optimized

[3] This is specific for localization in elevation.

efficiently. Beyond having controlled complexity for larger numbers of sources, this more sophisticated formulation also allows us to include more general source models like dictionaries. Finally, we present numerical results in Section 4.

2 Localization of Noise Sources

We assume having a set of D possible directions in the azimuth interval $[0, 2\pi)$ for which we know the sensing functions $\mathsf{a}_1, \mathsf{a}_2, \ldots, \mathsf{a}_D$. Further, we choose a set of F frequencies at which we examine the recorded signal; this can be done through a filterbank of F narrowband filters. We can then re-write (1) as

$$\mathsf{y} = \sum_{j \in \Theta} \mathsf{a}_j \odot \mathsf{s}_j + \mathsf{n}, \tag{2}$$

where $\mathsf{y} \in \mathbb{C}^F$, $\mathsf{a}_j \in \mathbb{C}^F$, $\mathsf{s}_j \in \mathbb{C}^F$, and $\odot$ denotes the Hadamard product. We think of y as corresponding to one audio frame.

2.1 Geometrical Structure

White In the presence of $J \geq 1$ independent white sources at locations Θ, the expected power of the frame y from (2) is

$$\mathbb{E}[|\mathsf{y}|^2] = \sum_{j \in \Theta} \sigma_j^2 |\mathsf{a}_j|^2, \tag{3}$$

where σ_j^2 is the power of the j^{th} source and we again set $\mathsf{n} = \mathbf{0}$. Thus, even if σ_j^2 are unknown, we see that the measured power spectrum is, in expectation, a positive linear combination of the power spectra of the sensing vectors, with coefficients being the source powers. Put differently, all power measurements arising from a certain configuration Θ lie in a cone characterized by the corresponding sensing vectors:

$$\mathbb{E}[|\mathsf{y}|^2] \in C_\Theta = \{\mathbf{w} \mid \mathbf{w} = \sum_{j \in \Theta} p_j |\mathsf{a}_j|^2,\ p_j \geq 0\}$$

as shown in Fig. 1. The entire space of measurements, for all possible configurations, is a union of those cones. It follows that if we can find the right cone, we will have identified the source locations. More precisely, the source localization task amounts to identifying which of the cones $\{C_\Theta \mid \Theta \text{ a set of } J \text{ directions}\}$ contains $\mathbb{E}[|\mathsf{y}|^2]$ or its empirical estimate. Without scattering, the measurement space is collapsed into a single cone.

Color Unlike white, colored sources will modulate the sensing functions and move them about in space as seen in (2):

$$\mathbb{E}[|\mathsf{y}|^2] = \sum_{j \in \Theta} \mathbb{E}[|\mathsf{s}_j|^2] \odot |\mathsf{a}_j|^2 = \sum_{j \in \Theta} \sigma_j^2 |\mathsf{b}_j|^2 \odot |\mathsf{a}_j|^2, \tag{4}$$

where $|\mathsf{b}_j|^2$ are the prototype power spectra for each source. Consequently, without knowing the modulation, we cannot identify the cones for localization. So if we know the sources' power spectral density or simply the time-varying power spectra, we are again in similar situation as for white sources except that the number of cones increases due to ambiguities in assigning $|\mathsf{b}_j|$ to directions.

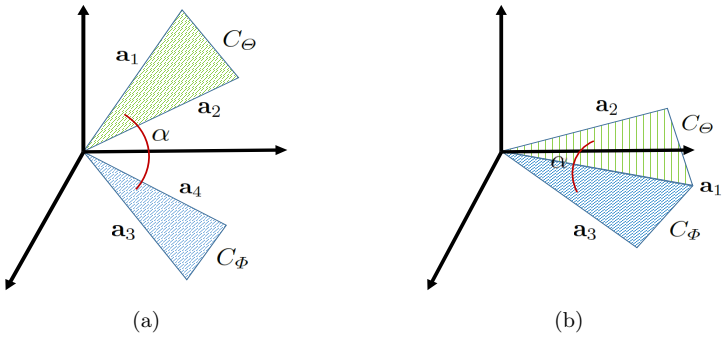


Fig. 1. Cones in the measurement space. (a) The angle between two cones formed by a pair of different sensing vectors. (b) Two distinct cones that share a sensing vector.

2.2 Structure Quality

Not all unions of cones are created equal. To ensure that we correctly identify the cone and hence solve the localization problem, we require adequate separation between the different cones. Thus, we examine the angles between every pair of cones (for a certain number of sources J) as illustrated in Fig. 1a. Consider two cones C_Θ and C_Φ for two sets of J directions Θ and Φ. The largest angle between them is

$$\breve{\alpha} = \max_{\substack{\mathbf{p}\in C_\Theta, \mathbf{q}\in C_\Phi \\ \|\mathbf{p}\|=\|\mathbf{q}\|=1}} \cos^{-1}\langle \mathbf{p}, \mathbf{q}\rangle. \tag{5}$$

For simplicity, instead of the inter-cone angle $\breve{\alpha}$, we will in the following look at the maximal angle between the smallest subspaces that contain the cones. For this to make sense, we need to assume that $J < F$ since otherwise cones will lie in the same subspace. We note that this relaxation will then give us sufficient conditions for localization.

Denote the orthonormal bases for the smallest subspaces containing C_Θ and C_Φ by $\mathbf{B}_\Theta$ and $\mathbf{B}_\Phi$, and define the largest angle as

$$\alpha = \cos^{-1}\sigma_{\min}(\mathbf{B}_\Theta^{\mathrm{T}}\mathbf{B}_\Phi), \tag{6}$$

where $\sigma_{\min}(.)$ denotes the smallest singular value. We do not consider the smaller angles because what matters is that the two cones are distinct i.e., the largest

angle is non-zero. If the smaller angles include zero, it means that the cones intersect; by definition the cones here indeed intersect at exactly the sensing vectors. For example as shown in Fig.1b, consider the following two cones $C_\Theta = \{\mathbf{w} = p\mathbf{a}_1 + q\mathbf{a}_2,\ p, q \geq 0\}$ and $C_\Phi = \{\mathbf{w} = p\mathbf{a}_1 + q\mathbf{a}_3,\ p, q \geq 0\}$, then $C_\Theta \cap C_\Phi = \{p\mathbf{a}_1,\ p \geq 0\}$.

The smaller the angle α, the more sensitive the sensing device is to noise. Hence, a good set of sensing functions are ones that result in large angles between every pair of cones. Thus, we are interested in the worst-case angle or alternatively the worst-case coherence between the cones which we define as

$$\mu_J = \max_{\Theta \neq \Phi} \sigma_{\min}(\mathbf{B}_\Theta^{\mathrm{T}} \mathbf{B}_\Phi), \tag{7}$$

where $\sigma_{\min}(.)$ denotes the smallest singular value. For the case of a single white source $J = 1$, (7) reduces to conventional coherence in the power domain

$$\mu_1 = \max_{i \neq j} \frac{\langle |\mathbf{a}_i|^2, |\mathbf{a}_j|^2 \rangle}{\||\mathbf{a}_i|^2\| \||\mathbf{a}_j|^2\|}. \tag{8}$$

The lower the coherence, the better. Nevertheless as we will see next, in the noiseless case, a sufficient condition for the accurate localization of any number of sources J is to simply have the corresponding coherence $\mu_J < 1$.

2.3 Conditions for Localization

We now turn our attention to the actual localization problem. Let $\mathbf{y} := |\mathrm{y}|^2$ denote the power spectrum of y. Based on the analysis in Section 2.1, we have $\mathbb{E}[\mathbf{y}] \in C_\Theta$ and accordingly $\mathbb{E}[\mathbf{y}] = \mathbf{B}_\Theta \mathbf{w}$. Then, we can write $\mathbf{y} = \mathbf{P}_\Theta \mathbf{y} + (\mathbf{I} - \mathbf{P}_\Theta)\mathbf{y}$ where $\mathbf{P}_\Theta$ denotes the projection onto range($\mathbf{B}_\Theta$). For a particular realization, the error vector $\mathbf{z}_\Theta = (\mathbf{I} - \mathbf{P}_\Theta)\mathbf{y}$ will be non-zero, but by the law of large numbers, its average over many frames will converge to zero.

Thus, a straightforward akin to matched field processing is to calculate the sample mean of N power frames and test it against every cone: perform an exhaustive search for the right match as determined by the minimum distance to the cone (more precisely, the corresponding subspace) of the empirical power mean. This procedure is summarized in Algorithm 1.

Algorithm 1 DoA estimation of J sources

Input: Number of sources J, bases $\mathbf{B}_\Theta\ \forall \Theta\ , |\Theta| = J$, N power frames $\mathbf{y}_n$ for $n = 1, \ldots, N$.

Output: Directions of arrival $\Theta^* = \{\theta_1^*, \ldots, \theta_J^*\}$.

Compute $\widetilde{\mathbf{y}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n$

$\Theta^* = \arg\min_\Theta \|(\mathbf{I} - \mathbf{B}_\Theta \mathbf{B}_\Theta^T)\widetilde{\mathbf{y}}\| = \arg\max_\Theta \|\mathbf{B}_\Theta^T \widetilde{\mathbf{y}}\|$

Algorithm 1 relies on the law of large numbers to justify using the empirical mean in lieu of the expectation, but the whole discussion has made no mention

of noise. The following proposition suggests that the localization will be correct even with measurement noise as long as a certain relationship holds between the signal-to-noise ratio (SNR) in the power domain and the worst-case coherence.

Proposition 1 (Correct localization). *Assuming J independent sources, let the source configuration be specified by the cone C_Θ and denote by $\widetilde{\mathbf{y}}$ the sample mean of N independent power frames. Consider further a zero-mean noise term n independent of the source signals, with power $\mathbf{n} := |\mathsf{n}|^2$. Then as long as the SNR in the power domain exceeds $\|\mathbb{E}[\sum_{j\in\Theta} \mathbf{a}_j \odot \mathbf{s}_j]\|/\|\mathbb{E}[\mathbf{n}]\| > \sqrt{2/(1-\mu_J)}$, localization by Algorithm 1 with input $\widetilde{\mathbf{y}}$ is correct with arbitrarily high probability for a sufficiently large N.*

Proof (sketch). Suppose first that we can measure the expected value of the power measurements $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{n}]$, where $\mathbf{x} := \sum_j \mathbf{a}_j \odot \mathbf{s}_j$. In Algorithm 1, we take subspace membership as a proxy to cone membership, meaning that we will have localized correctly as long as $\mathbb{E}[\mathbf{y}]$ is closer to the true subspace range($\mathbf{B}_\Theta$) than to any other range($\mathbf{B}_\Phi$); equivalently, we ask that $\langle \mathbb{E}[\mathbf{y}], \widehat{\mathbf{P}_\Theta \mathbb{E}[\mathbf{y}]} \rangle > \langle \mathbb{E}[\mathbf{y}], \widehat{\mathbf{P}_\Phi \mathbb{E}[\mathbf{y}]} \rangle$, where $\widehat{\mathbf{u}} := \frac{\mathbf{u}}{\|\mathbf{u}\|}$. This can be rewritten as (denoting $\overline{\mathbf{u}} := \mathbb{E}[\mathbf{u}]$ and setting $\widetilde{\mu} := \langle \widehat{\mathbf{P}_\Theta \overline{\mathbf{y}}}, \widehat{\mathbf{P}_\Phi \overline{\mathbf{y}}} \rangle$):

$$
\begin{aligned}
\langle \mathbf{P}_\Theta \overline{\mathbf{y}} + \overline{\mathbf{n}}, \widehat{\mathbf{P}_\Theta \overline{\mathbf{y}}} \rangle &> \langle\ \mathbf{P}_\Theta \overline{\mathbf{y}} + \overline{\mathbf{n}}, \widehat{\mathbf{P}_\Phi \overline{\mathbf{y}}} \rangle \\
&\Leftrightarrow \|\mathbf{P}_\Theta \overline{\mathbf{y}}\| + \langle \widehat{\mathbf{P}_\Theta \overline{\mathbf{y}}}, \overline{\mathbf{n}} \rangle > \|\mathbf{P}_\Theta \overline{\mathbf{y}}\| \widetilde{\mu} + \langle \mathbf{P}_\Phi \overline{\mathbf{y}}, \overline{\mathbf{n}} \rangle \\
&\Leftrightarrow \|\mathbf{P}_\Theta \overline{\mathbf{y}}\|(1-\widetilde{\mu}) > \langle \widehat{\mathbf{P}_\Phi \overline{\mathbf{y}}} - \widehat{\mathbf{P}_\Theta \overline{\mathbf{y}}}, \overline{\mathbf{n}} \rangle \\
&\Leftarrow \|\mathbf{P}_\Theta \overline{\mathbf{y}}\|(1-\widetilde{\mu}) \overset{(a)}{>} \|\widehat{\mathbf{P}_\Phi \overline{\mathbf{y}}} - \widehat{\mathbf{P}_\Theta \overline{\mathbf{y}}}\| \|\overline{\mathbf{n}}\| \\
&\Leftrightarrow \|\mathbf{P}_\Theta \overline{\mathbf{y}}\|(1-\widetilde{\mu}) \overset{(b)}{>} \sqrt{2}\sqrt{1-\widetilde{\mu}} \|\overline{\mathbf{n}}\| \Leftarrow \frac{\|\mathbf{P}_\Theta \overline{\mathbf{y}}\|}{\|\overline{\mathbf{n}}\|} \overset{(c)}{>} \sqrt{\frac{2}{1-\mu_J}}
\end{aligned}
$$

where in (a) we used the Cauchy-Schwarz inequality to upper bound the right-hand side, (b) we used the law of cosines and (c) we used the fact that the function $\frac{1}{\sqrt{1-t}}$ is increasing with $t \in [0,1)$ to replace $\widetilde{\mu}$ with the worst-case μ_J. Convergence in probability then follows from replacing expectations by empirical means and invoking the weak law of large numbers. □

Note that Proposition 1 does not quantify the number of frames N required to guarantee correct localization. However, the concentration of measure phenomenon for the Lipschitz $\|\cdot\|$ suggests that N is tightly controlled [6].

3 Algorithms

With the described matched field processing approach, it is not straightforward to use more complex source models such as overcomplete dictionaries. Moreover, for large D and J, computational complexity makes the search unfavorable. Therefore, in this section, we resort to convex relaxations for sparse recovery which can be optimized efficiently.

Let $\mathbf{a}_j \in \mathbb{R}_+^F$ denote the power spectrum of the j^{th} sensing function. Let $\mathbf{V} \in \mathbb{R}_+^{F\times K}$ be the source model such that $\mathbf{s}_j = \mathbf{V}\mathbf{x}_j$, e.g. a subspace basis or an overcomplete dictionary. Then we can write

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}, \tag{9}$$

where $\mathbf{A} \in \mathbb{R}_+^{F\times KD} = [\text{diag}(\mathbf{a}_1)\mathbf{V}, \ldots, \text{diag}(\mathbf{a}_D)\mathbf{V}]$, $\mathbf{x} \in \mathbb{R}_+^{KD}$ is a vector of concatenated source coefficients $\mathbf{x}_j \in \mathbb{R}_+^K$ and $\mathbf{z} \in \mathbb{R}_+^F$ is a term grouping all the cross-terms which arise when calculating the power of y (2).

Since the system of equations in (9) is underdetermined, we consider the solution of the following optimization problem

$$\min_{x\geq 0} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \Psi(\mathbf{x}), \tag{10}$$

where the first term is the data fidelity and Ψ is an appropriate regularization. The choice of Ψ is inspired by the underlying geometrical structure and is discussed in the following sections.

Once we solve for $\mathbf{x}$, localization amounts to finding the J direction indices corresponding to the $\mathbf{x}_j$ with the highest norms $\|\mathbf{x}_j\|_2$.

3.1 Subspace Model

The appropriate regularization for signals from a union of cones is to enforce group sparsity, i.e., only few $\mathbf{x}_j$ are non-zero. The ℓ_1/ℓ_2 penalty known to promote group sparsity [15] is defined as

$$\Psi(\mathbf{x}) = \lambda \sum_{j=1}^{D} \|\mathbf{x}_j\|_2, \tag{11}$$

where $\lambda > 0$ determines the weight of the penalty.

The source model for white sources is one-dimensional i.e., $\mathbf{V} = \mathbf{1}$ and thus Ψ reduces to the ℓ_1 penalty. We emphasize that in that case we do not need an explicit source model and only require knowledge of the sensing vectors where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_D]$.

3.2 Dictionary Model

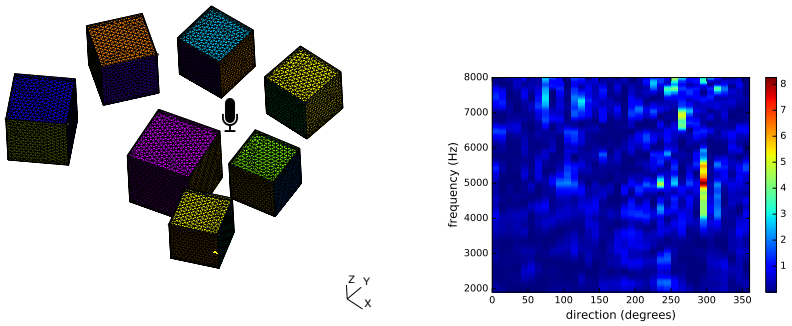
For colored sources, we consider using an overcomplete dictionary (i.e., $K > F$) to represent their time-varying power spectra. The dictionary is chosen such that every source admits a sparse representation and while we still have a union of cones structure, the elements in the union depend dynamically on the sources being localized and are not known a priori. Thus, to appropriately select the right subset, we add the ℓ_1 penalty

$$\Psi(\mathbf{x}) = \lambda \sum_{j=1}^{D} \|\mathbf{x}_j\|_2 + \gamma\|\mathbf{x}\|_1, \tag{12}$$

where $\lambda > 0$ and $\gamma > 0$ are the trade-off parameters determining the weights of their respective terms. This penalty (12) promotes sparsity across groups and within active groups. The corresponding objective is known as the sparse-group lasso [4] which we augment by the non-negativity constraint.

4 Numerical Results

In this section, we present numerical results for 2D DoA estimation in a 3D environment using a simulated 3D model of a randomly shaped sensing device. The sensing device consists of an omnidirectional sensor surrounded by 7 cubes of randomly chosen sizes (side lengths $\in [10, 14]$ cm) and orientations, spread over an area 60×60 cm^2 as shown in Fig.2a. The mesh was generated using Gmsh [5] and the directional frequency responses were calculated using the boundary element method package BEM++ [12]. Taking into consideration the sizes of the cubes, we use 193 frequencies between 2000 Hz and 8000 Hz which are most affected by the scattering. The power spectra of the sensing vectors for 36 directions equally spaced in the interval $[0°, 360°)$ are shown in Fig.2b.



(a) (b)

Fig. 2. The sensing device. (a) Illustration of the sensing device consisting of 7 cubes surrounding a microphone. (b) The corresponding transfer functions per direction.

To add modeling mistmatch in the simulations, the sources are randomly placed at a $\pm1°$ shift from the assumed model. We implemented consensus ADMM [8] to solve (10). Finally, we consider the localization successful when the estimate is the closest shift for all J sources.

4.1 White Sources

First we show that we can localize white sources without having an explicit dictionary or knowing the distribution parameters. The corresponding coherences

of our sensing device are $\mu_1 = 0.88$ and $\mu_2 = 0.93$ so Proposition 1 guarantees perfect localization of one and two sources. We simulate one and two white Gaussian (zero-mean unit-variance) and Bernoulli ($p_{\text{success}} = 0.5$) sources at all 36 directions. We solve the non-negative lasso (11) with $\lambda = 10$. The rate of successful localization averaged over 10 runs is shown in Table 1. We conjecture that any error is strictly due to the modeling mismatch.

4.2 Speech Sources

Next we show preliminary results for localization of one or two speakers with the help of a dictionary. Four speakers (two female, two male) were randomly chosen from the TIMIT speech corpus; the maximum amplitudes were normalized to 1 before computing the frequency representation. Every speaker emits 100 frames. As discussed in Section 2.1, colored sources require some prior knowledge. Thus, we assume knowing the power spectra for each speaker's frames where $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \mathbf{V}_4]$ and $\mathbf{V}_i \in \mathbb{R}_+^{F \times 100}$; this is similar to what was done in [14] and we leave for future work incorporating a more general learned speech dictionary. We solve the non-negative sparse group lasso (12) with $\lambda = 0.1$ and $\gamma = 0.1$. The average success rates are shown in Table 1. First, we note how it still possible to perfectly localize one speaker even at a very high coherence. In two-source cases, one source was almost always localized accurately (in 99 % of all cases). Second, the lower performance for localizing two sources compared to the white case is likely due to the higher coherence μ_2. In particular, the lower performance in localization of male speakers can probably be attributed to unfavorable interplay between the structure response and the source spectrum. It remains, however, to be completely explained.

Table 1. Success rates for DoA estimation of one or two sources

Type	Success rate
One Gaussian source	100%
One Bernoulli source	100%
Two Gaussian sources	86.7%
Two Bernoulli sources	86.7%
One female speaker	100%
One male speaker	100%
Two speakers (female)	75.9%
Two speakers (male)	41.7%
Two speakers (female & male)	41%

5 Conclusion

In conclusion, we demonstrated the potential of using a sensing device that introduces known scattering in the measurements for DoA estimation. In particular,

we showed that the scattering induces a union of cones structure which allows us to localize any number of white sources in the noiseless case granted the coherence is strictly less than 1. We then showed that with the proper modeling, in the form of an overcomplete dictionary and group sparsity penalties, we are able to localize more challenging sources like speech, all while using a single sensor with what may be considered a rather poor response, corrupted by scattering off of random clutter. Future work includes running a real-world experiment and using a general learned dictionary as well as extending the approach to handle reverberation.

References

1. Blauert, J.: Spatial hearing : the psychophysics of human sound localization. The MIT Press (1997)
2. Dokmanić, I.: Listening to Distances and Hearing Shapes: Inverse Problems in Room Acoustics and Beyond. Ph.D. thesis, École polytechnique fédérale de Lausanne (2015)
3. El Badawy, D.: Acoustic sensing using scattering microphones. Master's thesis, École polytechnique fédérale de Lausanne (July 2015)
4. Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. arXiv (2010)
5. Geuzaine, C., Remacle, J.F.: Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. International Journal for Numerical Methods in Engineering 79(11), 1309–1331 (2009)
6. Ledoux, M.: The concentration of measure phenomenon. Mathematical surveys and monographs, American Mathematical Society, Providence (R.I.) (2001)
7. Lessard, N., Pare, M., Lepore, F., Lassonde, M.: Early-blind human subjects localize sound sources better than sighted subjects. Nature 395(6699), 278–280 (Sept 1998)
8. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends in Optimization 1(3), 123–231 (2014)
9. Saxena, A., Ng, A.: Learning sound location from a single microphone. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 1737–1742 (2009)
10. Schnupp, J., Nelken, I., King, A.: Auditory neuroscience: making sense of sound. The MIT Press (2010)
11. Slattery, W.H.I., Middlebrooks, J.C.: Monaural sound localization: acute versus chronic unilateral impairment. Hearing Research 139(6), 38–46 (May 1994)
12. Śmigaj, W., Betcke, T., Arridge, S., Phillips, J.and Schweiger, M.: Solving boundary integral problems with BEM++. ACM Trans. Math. Softw. 41(2), 6:1–6:40 (2015)
13. Tolstoy, A.: Applications of matched-field processing to inverse problems in underwater acoustics. Inverse Problems 16(6), 1655 (2000)
14. Xie, Y., Tsai, T., Konneker, A., Popa, B., Brady, D.J., Cummer, S.A.: Single-sensor multispeaker listening with acoustic metamaterials. Proceedings of the National Academy of Sciences 112(34), 10595–10598 (August 2015)
15. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 68(1), 4967 (2006)