Sketching for nearfield acoustic imaging of heavy-tailed sources

Mathieu Fontaine[1], Charles Vanwynsberghe[2], Antoine Liutkus[1], Roland Badeau[3]

[1]Inria, speech processing team, Nancy Grand-Est, France
[2] Institut Jean le Rond d'Alembert, Saint-Cyr l'École, France.
[3] LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, Paris, France.

Abstract. We propose a probabilistic model for acoustic source localization with known but arbitrary geometry of the microphone array. The approach has several features. First, it relies on a simple nearfield acoustic model for wave propagation. Second, it does not require the number of active sources. On the contrary, it produces a heat map representing the energy of a large set of candidate locations, thus imaging the acoustic field. Second, it relies on a heavy-tail α-stable probabilistic model, whose most important feature is to yield an estimation strategy where the multichannel signals need to be processed only once in a simple online procedure, called *sketching*. This sketching produces a fixed-sized representation of the data that is then analyzed for localization. The resulting algorithm has a small computational complexity and in this paper, we demonstrate that it compares favorably with state of the art for localization in realistic simulations of reverberant environments.

1 Introduction

Source localization has attracted a lot of research interest, notably in acoustics [5] and wireless communications [15]. It aims at identifying the position or *direction of arrival* (DoA) of *sources* that are captured by an array of sensors. It has many applications, notably for isolating the target signals. In this paper, we are focused on the acoustic application.

Popular approaches for localization largely exploit the geometry of the sensor array. When the positions of the sensors are known, we can indeed predict and exploit the time difference of arrival (TDOA) to all sensors. In a more realistic environment with echoes and reverberation, localization becomes a much more challenging inverse problem composed of two classical parts. First, the knowledge of the geometry of the sensor array along with physics provides us with a *direct model*. Then, localization tries to invert this direct path so as to estimate the most likely location of the sources based on the observations. As in any challenging inverse problem, the difficulties come from having less observations than unknowns, and/or from uncertainties in the direct model. Furthermore, localization should ideally work regardless of the particular source signals considered, which brings an additional difficulty.

Many methods for source localization have already been proposed in the past. Since we usually have a huge number of candidate locations for only a limited amount of sensors, they all attempt to reduce the number of parameters. One approach is to fix the number of sources to look for, yielding for instance high resolution methods [16] such as MUSIC [10], [14] that provide good performance when the microphone array is not too massive and obeys some geometry assumptions. Another approach for exploiting this relative *sparsity* of active sources' locations is to use greedy methods [17] that iteratively detect the most predominant source and then remove its influence from the observation using the direct model. Provided the amount of reverberation is not too large and the direct model is sufficiently good, these methods yield good performance. Another direction is grounded on a probabilistic setting [3], [11], [12] where a prior distribution such as a multivariate Gaussian is assigned to both the unknown source signals and the mixing model.

Apart from raw performance, one important issue of source localization methods is their computational complexity. For the purpose of imaging, the *steering response power* method (SRP) simply averages the power of beamformed outputs targeted at all candidate directions. Although very simple computationally, it yields a very poor contrast. See however [19] for an improvement involving hierarchical search. On the same topic of computational complexity, localization under the Gaussian model [12] involves a demanding Expectation-Maximization algorithm (EM) that requires going through the data many times and inverting many covariance matrices. To a lesser extent, the same goes for RELAX and CLEAN [17].

In this paper, we propose a new imaging technique, conceptually close to SRP because it only requires going through the recordings once. However, it is also grounded in a probabilistic setting but the source signals are no longer assumed Gaussian as in [12] but rather α-stable, which is a heavy-tailed distribution permitting to describe audio signals with very large dynamics using only a very small amount of parameters [7], [13]. Departing from the costly EM, estimation in this model is based on moment-fitting, appearing as one instance of the recently popularized sketching methodology [4]. We use a near-field acoustic model here and simulate challenging reverberant environments.

2 Mixture model and α-stable theory

2.1 Notation and convolutive model

Let $x \in \mathbb{C}^{F\times T\times K}$ be the Short-Term Fourier Transforms (STFT) of the observations, where F is the number of frequency bins, T the number of time frames and K the number of microphones. $\boldsymbol{x}\left(f,t\right) \in \mathbb{C}^K$ gathers its entries for Time-Frequency (TF) bin (f,t). Now, we assume this recording is the superposition of signals originating from L potential locations, corresponding to a grid in the 3D-space. Let $s \in \mathbb{C}^{F\times T\times L}$ denote the STFT of the L corresponding sources, with entries $s_l\left(f,t\right) \in \mathbb{C}$. Our objective becomes to estimate the power of the sources at all these L locations. Of course, we expect most of them to be inactive.

Then, the acoustic model defines the mixture as a superposition of filtered versions of the sources. In the frequency domain, this convolution may be approximated as a simple multiplication with steering vectors $\boldsymbol{A}_l(f) \in \mathbb{C}^K$:

$$\forall (f,t),\ \boldsymbol{x}(f,t) = \sum_{l=1}^{L} \boldsymbol{A}_l(f)\, s_l(f,t). \tag{1}$$

A particular direct model then consists in a specific choice for the steering vectors. In this study, we adopt the near field region assumption, thus taking the steering vectors $\boldsymbol{A}_l(f)$ as:

$$\forall l, f\ \boldsymbol{A}_l(f) = \left[\frac{1}{r_{1l}} \exp\left(-i\frac{\omega_f r_{1l}}{c_0}\right), \ldots, \frac{1}{r_{Kl}} \exp\left(-i\frac{\omega_f r_{Kl}}{c_0}\right)\right]^\top, \tag{2}$$

where $\cdot^\top$ stands for transposition, c_0 is the speed of sound in the air, r_{kl} the distance between the k^{th}microphone and the l^{th} source and ω_f is the angular frequency at frequency band f. Note that if applicable, actual measurements may be used instead of the model (2) to provide numerical values for $\boldsymbol{A}_l(f)$ at every candidate location l.

2.2 Independent isotropic α-stable model for the sources

We assume that all the L sources are independent α-stable harmonizable processes as defined in [7]. In practice, it means that all $s_l(f,t)$ are independent and distributed w.r.t. a complex symmetric α-stable ($S\alpha S_c$) distribution:

$$s_l(f,t) \sim S\alpha S_c(\Upsilon_l), \tag{3}$$

where $\alpha \in (0,2]$ is called the *characteristic exponent*, controlling the tail of the distribution: the closer it is to 0, the heavier the tails. The nonnegative scale parameters $\Upsilon_l \in \mathbb{R}_+$ are the central quantity of interest in our study. Gathering them together in the $L \times 1$ vector $\boldsymbol{\Upsilon} = [\Upsilon_1, \ldots, \Upsilon_L]^\top$, we call it the *discrete spatial measure*. Our objective is to estimate this measure, since it gives the scale of the signal present at each location.

A remarkable fact of the model (3) is that the entries of s_l are modeled as having the same distribution for all f and t. This is made possible thanks to the heavy-tailed nature of the $S\alpha S_c$ distribution. In contrast, the classical Gaussian model [8] requires variances to depend on (f,t) to fit well the data.

2.3 The Levy exponent and the spatial measure

Since the distributions (3) and the acoustic model (1) do not depend on time, neither does the distribution of $\boldsymbol{x}(f,t)$. For a given f, let φ_f be the characteristic function (chf.) of $\boldsymbol{x}(f,t)$ and let I_f be the *Levy exponent*, i.e. the logarithm of its opposite:

$$\forall \boldsymbol{\theta} \in \mathbb{C}^K,\ \varphi_f(\boldsymbol{\theta}) \triangleq \mathbb{E}\left[\exp\left(i\Re\left\langle \boldsymbol{\theta}, \boldsymbol{x}(f,t)\right\rangle\right)\right] \text{ and } I_f(\boldsymbol{\theta}) = -\log \varphi_f(\boldsymbol{\theta}), \quad (4)$$

with $\langle .,. \rangle$ the inner product on $\mathbb{C}^K$. In this study, the argument $\boldsymbol{\theta} \in \mathbb{C}^K$ of the chf. is called a *sketching frequency*. Combining the $S\alpha S_c$ model for the sources and the propagation model (1), it can be shown that we have:

$$\forall \boldsymbol{\theta}, \in \mathbb{C}^K,\ I_f(\boldsymbol{\theta}) = \sum_{l=1}^{L} \left|\left\langle \boldsymbol{\theta}, \boldsymbol{a}_l(f)\right\rangle\right|^{\alpha} \Upsilon_l, \quad (5)$$

where $\boldsymbol{a}_l(f) = \boldsymbol{A}_l(f) / \left\|\boldsymbol{A}_l(f)\right\|_2 \in \mathbb{C}^K$ are the normalized steering vectors.

Now, the approach undertaken in this study is to pick a set of L sketching frequencies and exploit relation (5). Even if we could pick any complex vector for $\boldsymbol{\theta}$, informal experiments shows that taking the normalized steering vectors $\boldsymbol{\theta} = \boldsymbol{a}_l(f)$ gives good performance. This yields L relations of the form (5), that can be expressed in compressed form as $\boldsymbol{I_f} = \boldsymbol{\Psi}_f \boldsymbol{\Upsilon}$, where:

$$\boldsymbol{I}_f \triangleq \left[I_f(\boldsymbol{a}_1(f)), \ldots, I_f(\boldsymbol{a}_L(f))\right]^{\top} \text{ and } \forall l, l'\ \left[\boldsymbol{\Psi}_f\right]_{ll'} = \left|\left\langle \boldsymbol{a}_l(f), \boldsymbol{a}_{l'}(f)\right\rangle\right|^{\alpha}. \quad (6)$$

Finally by gathering all $\boldsymbol{I}_f$ and $\boldsymbol{\Psi}_f$ into $\boldsymbol{I} \in \mathbb{R}^{FL}$ and $\boldsymbol{\Psi} \in \mathbb{R}_+^{FL \times L}$, respectively, we get:

$$\boldsymbol{I} = \boldsymbol{\Psi}\boldsymbol{\Upsilon}, \quad (7)$$

which is our main tool for estimating $\boldsymbol{\Upsilon}$. Indeed, $\boldsymbol{I}$ is estimated from the data and $\boldsymbol{\Psi}$ is given by combining our acoustic model for $\boldsymbol{a}_l(f)$ and (6).

3 Parameter estimation

3.1 Sketching for the Levy exponent

As noted above in (4), the Levy exponent is defined as the logarithm of the negative chf. A naive idea would be to simply replace $\varphi_f(\boldsymbol{\theta})$ in (4) by its empirical counterpart averaged over the different time frames. However, this may lead to numerical instability in case of negative empirical chf. To address this issue, a new unbiased estimator for the chf. specific to symmetric α-stable random vectors is proposed here:

$$\forall \boldsymbol{\theta} \in \mathbb{C}^K,\ \widehat{\varphi}_f(\boldsymbol{\theta}) = \left| \frac{1}{T} \sum_{t=1}^{T} \exp\left(i \frac{\Re\left\langle \boldsymbol{\theta}, \boldsymbol{x}(f,t)\right\rangle}{2^{1/\alpha}} \right) \right|^2. \quad (8)$$

As can be seen, this estimate is guaranteed to be nonnegative. Hence, no numerical instability is to be expected when considering the empirical Levy exponent $\widehat{\boldsymbol{I}}_f \in \mathbb{R}_+^L$, defined as:

$$\forall f,\, \widehat{\boldsymbol{I}}_f = \left[-\ln\left(\widehat{\varphi}_f\left(\boldsymbol{a}_1\left(f\right)\right)\right), \ldots, -\ln\left(\widehat{\varphi}_f\left(\boldsymbol{a}_L\left(f\right)\right)\right)\right]^\top. \tag{9}$$

Gathering them as $\widehat{\boldsymbol{I}} = \left[\widehat{\boldsymbol{I}}_1^\top, \ldots, \widehat{\boldsymbol{I}}_F^\top\right]^\top \in \mathbb{R}_+^{FL}$, we obtain a relation similar to (7):

$$\widehat{\boldsymbol{I}} \approx \boldsymbol{\Psi}\boldsymbol{\Upsilon}. \tag{10}$$

Interestingly enough, relation (10) provides us with a linear model where all factors except the desired spatial measure $\boldsymbol{\Upsilon}$ are either empirically estimated from the data ($\widehat{\boldsymbol{I}}$) or provided by the acoustic model ($\boldsymbol{\Psi}$). The fundamental fact here is that the observed data is only used once for estimating the Levy exponent in (8), through a very simple procedure, producing the $LF \times 1$ fixed-sized vector $\widehat{\boldsymbol{I}}$. This is reminiscent of the sketching strategy recently described, e.g. in [4].

3.2 A proposed NMF algorithm to determine $\boldsymbol{\Upsilon}$

The estimation method for $\boldsymbol{\Upsilon}$ is undertaken by a classical minimization of the divergence between the two terms of (10):

$$\widehat{\boldsymbol{\Upsilon}} \leftarrow \arg\min_{\boldsymbol{\Upsilon}\geq 0} d_\beta\left(\widehat{\boldsymbol{I}} | \boldsymbol{\Psi}\boldsymbol{\Upsilon}\right) + \lambda\|\boldsymbol{\Upsilon}\|_1, \tag{11}$$

where d_β depicts a data-fit cost function such as the β-divergence [1], and $\lambda\|\boldsymbol{\Upsilon}\|_1$ is an ℓ_1-regularization penalty term to enforce sparsity of $\boldsymbol{\Upsilon}$. Following classical multiplicative updates strategy, we can straightforwardly estimate $\boldsymbol{\Upsilon}$. The algorithm box below summarizes the whole process, which is of total complexity $\mathcal{O}\left(FTL^2\right)$.

Algorithm 1 Estimation of the spatial measure $\boldsymbol{\Upsilon}$

1. Input
   - Number L of possible locations, distances r_{lk} with the microphones.
   - Characteristic exponent α
   - β-divergence to use, number of iterations, regularization parameter λ.
2. Compute steering vectors $\boldsymbol{A}_l(f)$ as in (2).
3. Sketching: $\forall f,\, \widehat{\boldsymbol{I}}_f \leftarrow$(9)
   (the mixture $\boldsymbol{x}$ may only be streamed and not stored)
4. Analysis
   - Gather all $\boldsymbol{\Psi}_f$ and $\widehat{\boldsymbol{I}}_f$ to form $\boldsymbol{\Psi}$ and $\widehat{\boldsymbol{I}}$ (6)
   - Estimation of $\boldsymbol{\Upsilon}$: iterate $\widehat{\boldsymbol{\Upsilon}} \leftarrow \widehat{\boldsymbol{\Upsilon}} \cdot \frac{\boldsymbol{\Psi}^\top\left(\left(\boldsymbol{\Psi}\widehat{\boldsymbol{\Upsilon}}\right)^{\beta-2}\cdot\widehat{\boldsymbol{I}}\right)}{\boldsymbol{\Psi}^\top\left(\left(\boldsymbol{\Psi}\widehat{\boldsymbol{\Upsilon}}\right)^{\beta-1}\right)+\lambda}$.

4 Evaluation

We now compare the proposed approach with several baseline methods for wideband source localization. We consider $J = 5$ speech signals lasting 10s and taken for the CMU[1] dataset. They are sampled at 16 kHz and placed randomly in a simulated room of dimensions $5 \times 4 \times 3$ meters, featuring up to $K = 50$ omnidirectional microphones at random positions. The room impulse responses are obtained with the RIR[2] generator toolbox [2] by simulated a 0.4s reverberation time. Because of computational cost, the sources' positions are restrained to lie in a flat 2D surface that is 1.5m high. All source localization methods operate with a grid of 10cm step-size, located on the source plane, but which does not contain the exact sources' locations. This results in $L = 2091$ candidate locations. To optimize computational cost, the frequency range considered was reduced from 1 kHz to 3 kHz, since it proved sufficient for speech signals. The different techniques compared are the following ones:

DSM The Discrete Spatial Measure (proposed). We take $\alpha = 1$, corresponding to the Cauchy distribution [9], $\lambda = 1$ for sparsity regularization and we pick the Itakura-Saito divergence $\beta = 0$ as the NMF cost function.

SRP The Steering Response Power, also called delay-and-sum [18], is the most classical source localization approach. It is based on the near-field propagation model (2) and projects the STFT of observations on the steering vectors: $\forall l$, $SRP = \frac{1}{FT} \sum_{f,t} \frac{|\boldsymbol{A}_l^\star(f)\boldsymbol{x}(f,t)|}{\|\boldsymbol{A}_l(f)\|}$. We use the same frequency range for SRP as for the proposed method.

CLEAN is a greedy algorithm [17]: at one iteration, it successively identifies the strongest source in the grid with SRP, and removes its contribution. The algorithm is repeated until all sources are identified.

RELAX is an enhanced variation of CLEAN [6] presented in [17].

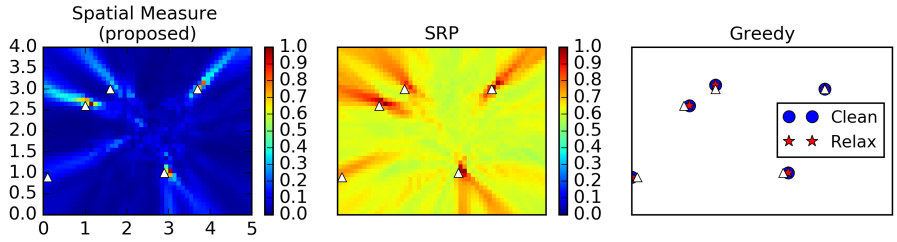

Fig. 1. Heat maps of spatial measure, SRP and both greedy algorithms.

[1] Carnegie Mellon University dataset : `http://www.festvox.org/cmu_faf/`

[2] `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`

A Monte Carlo simulation is carried out with arrays of $K = 5, 10, 20$ and 50 microphones. For each array configuration, we perform 50 trials with random positions of the sources on the 5×4m source plane. One trial for $K = 50$ is illustrated in Fig. 1, showing the estimated heat-maps. It first demonstrates that DSM is more accurate with better contrast than SRP, with only a slight increase in computational cost[3]. Indeed, the energy is focused on the ground truth positions and is close to 0 elsewhere, whereas the SRP map is noisier because of side lobes. Since CLEAN and RELAX exactly look for $J = 5$ sources, they result in the sparsest representations.

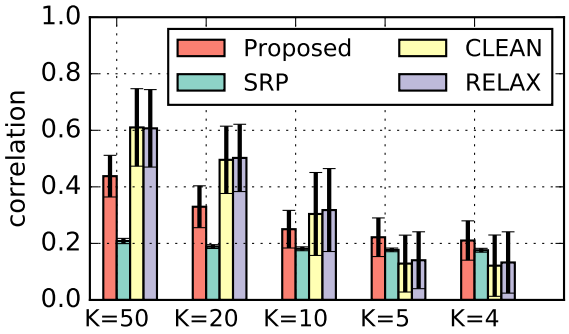

Fig. 2. Correlation with ground truth. Deviation is depicted with black whiskers.

The Monte Carlo experiment is evaluated by correlations between the estimated maps and the ground truth map. The latter is built by setting ones at ground truth positions, and zeros elsewhere, followed by a Gaussian smoothing with a 10cm length-scale. Correlation means and standard deviations along the 50 trials are depicted in Fig. 2. First, it shows that DSM outperforms SRP in all cases. For $K \geq 10$ microphones, CLEAN and RELAX have the highest correlation, notably thanks to the a priori on the source number J. However their performance decreases rapidly when K decreases. On the contrary, DSM performance appears more robust to a decrease of K. Lastly, the standard deviation of DSM is smaller than that of CLEAN/RELAX, showing that it also has a more stable behavior at different configurations.

[3] In the specific case where $J = 5$, the computation time of each method are 5.2 s for SRP, 54 s for DSM (comprising 24 s for computing $\boldsymbol{\Psi}$, which only needs to be done once). CLEAN and RELAX are implemented in GPU, and respectively need 0.45 s and 55 s. Note that the complexity of these two last methods depends on the a priori number of sources J and that our implementation for DSM did not exploit its highly parallelisable capabilities.

5 Conclusion

In this paper, we have introduced an acoustic imaging method for microphone arrays with known but arbitrary geometry. Interestingly, it requires going through the observed multichannel signals only once in order to compute a fixed amount of sufficient statistics called *sketch* from which the model parameters are estimated in a later analysis stage. This strategy has a linear complexity in terms of signal duration.

A fundamental feature of the probabilistic α-stable model we use is to describe the source emitting at each spatial location using a single scale parameter. This is possible because α-stable distributions correctly account for the marginal distribution of an acoustic signal in the Time-Frequency plane. Gathering all these location-specific scale parameters, we defined the Discrete Spatial Measure (DSM) and showed how it can be very easily estimated based on the sketch with a simple matrix factorization procedure.

In a very challenging simulation of heavily reverberant environments, the DSM method proved competitive with state-of-the-art methods, particularly when the number of microphones is comparable with the number of sources. Open directions include incorporating time-varying scale parameters and experimentally validating robustness to noise.

Acknowledgments. This work was partly supported by the research programme KAMoulox (ANR-15-CE38-0003-01) and EDiSon3D (ANR-13-CORD-0008-01) funded by ANR, the French State agency for research.

References

1. C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.
2. E.A.P. Habets. Room impulse response (RIR) generator, 2008.
3. Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani. Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 1153–1157. IEEE, 2016.
4. N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for large-scale learning of mixture models. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pages 6190–6194, May 2016.
5. T. Kundu. Acoustic source localization. *Ultrasonics*, 54(1):25–38, 2014.
6. J. Li and P. Stoica. Efficient mixed-spectrum estimation with applications to target feature extraction. *IEEE Trans. Signal Process.*, 44(2):281–295, 1996.
7. A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pages 266–270, April 2015.
8. A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Trans. Signal Process.*, 59(7):3155–3167, 2011.
9. A. Liutkus, D. Fitzgerald, and R. Badeau. Cauchy nonnegative matrix factorization. In *Proc. of IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics (WASPAA)*, pages 1–5, October 2015.

10. N. Ma and J. Goh. Ambiguity-function-based techniques to estimate DOA of broadband chirp signals. *IEEE Trans. Signal Process.*, 54(5):1826–1839, 2006.
11. Michael I Mandel, Ron J Weiss, and Daniel PW Ellis. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394, 2010.
12. J. Nikunen and T. Virtanen. Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 22(3):727–739, 2014.
13. G. Samoradnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.
14. R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.*, 34(3):276–280, 1986.
15. X. Sheng and Y. Hu. Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *IEEE Trans. Signal Process.*, 53(1):44–53, 2005.
16. P. Stoica and R. Moses. *Introduction to spectral analysis*, volume 1. Prentice hall Upper Saddle River, 1997.
17. Y. Wang, J. Li, P. Stoica, M. Sheplak, and T. Nishida. Wideband RELAX and wideband CLEAN for aeroacoustic imaging. *The Journal of the Acoustical Society of America*, 115(2):757–767, 2004.
18. E. Williams. *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.
19. D. N. Zotkin and R. Duraiswami. Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Trans. Speech Audio Process.*, 12(5):499–508, 2004.