Elliptically Contoured Alpha-Stable Representation for MUSIC-Based Sound Source Localization

Mathieu Fontaine, Diego Di Carlo, Kouhei Sekiguchi, Aditya Arie Nugraha, Yoshiaki Bando, Kazuyoshi Yoshii

Elliptically Contoured Alpha-Stable Representation for MUSIC-Based Sound Source Localization

Mathieu Fontaine* , Diego Di Carlo†, Kouhei Sekiguchi‡, Aditya Arie Nugraha‡,
Yoshiaki Bando§‡, and Kazuyoshi Yoshii¶‡

*LTCI, Télécom Paris, Institut Polytechnique de Paris, Paris, France, mathieu.fontaine@telecom-paris.fr.
†Inria, Université de Rennes 2, Rennes, France, diego.dicarlo89@gmail.com.
‡Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan, {kouhei.sekiguchi,adityaarie.nugraha}@riken.jp.
§National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan, y.bando@aist.go.jp.
¶Graduate School of Informatics, Kyoto University, Kyoto, Japan, yoshii@i.kyoto-u.ac.jp.

***Abstract*—This paper introduces a theoretically-rigorous sound source localization (SSL) method based on a robust extension of the classical multiple signal classification (MUSIC) algorithm. The original SSL method estimates the noise eigenvectors and the MUSIC spectrum by computing the spatial covariance matrix of the observed multichannel signal and then detects the peaks from the spectrum. In this work, the covariance matrix is replaced with the positive definite shape matrix originating from the elliptically contoured α-stable model, which is more suitable under real noisy high-reverberant conditions. Evaluation on synthetic data shows that the proposed method outperforms baseline methods under such adverse conditions, while it is comparable on real data recorded in a mild acoustic condition.**

***Index Terms*— sound source localization, MUSIC, α-stable theory, covariation**

I. Introduction

Sound source localization (SSL) aims at determining the source position in the space. It is essential for various machine listening applications such as sound event localization and detection [1], sound source separation [2] and speech enhancement [3]. Although it is a well-studied problem that benefits from decades of literature, the task of SSL still challenges today's technologies under adverse conditions due to noise [4], long reverberation and acoustic reflections [5], [6]. Most SSL state-of-the-art algorithms exploit the correlation between the observed data and build a function with respect to candidate source locations. For instance, the popular steered response power phase transform (SRP-PHAT) approach [7] computes correlation coefficients considering only the phase information of the signal. Then, the SSL is achieved by finding maxima through a grid search of potential locations.

Alternatively, the multiple signal classification (MUSIC) method aims to separate the noise and source subspaces via eigenvalue decomposition of the observed covariance matrix. Then, the so-called *pseudo spectrum* function exhibits peaks corresponding to the source positions. Therefore, most studies on SSL have investigated extensions of such methods to face real-world challenging scenarios. Under a low signal-to-noise ratio (SNR) condition, for instance, the work of [4] proposes a normalization along the frequency axis of the MUSIC pseudo-spectrum that increases the SSL accuracy. The work of [8] proposed an extension of the MUSIC method in the presence of multiple highly-correlated sources. Moreover, the generalized eigenvalue decomposition (GEVD-MUSIC) [9] introduced a notable improvement to deal with noise stronger than the source signal. This approach was later extended for real-time applications in [10], [11]. Finally, the work of [12] outlines recent development in deep learning-based approaches that provide robust data-driven extensions of such well-known baseline methods.

A basic assumption of the MUSIC approach is that the source and noise subspaces are orthogonal to each other. In practice, this concerns with the estimation of the underlying true mixture covariance matrix which may be strongly biased when dealing with non-stationary noise and reverberation. Therefore, another research direction comprises models to surrogate such matrix or its components. In [13], the authors propose a generalized covariance (GC) framework that summarizes a plethora robust estimators, including the covariation method [14] and new non-linear function-based GC (*e.g.*, hyperbolic tangent). Besides, a natural way to get a positive definite shape matrix was designed in [15] for various complex elliptically contoured distributions. Recent researches also focused on using complex multivariate elliptically contoured α-stable (α-EC) distributions for SSL where parameters are estimated through the empirical characteristic function (ECF) [16]. However, ECF-based algorithm can be tricky to estimate [17]. The α-stable distributions gather all random vectors that satisfies the reproductive property [18], while α-EC is a subclass of multivariate distribution parameterized by a positive definite shape matrix and a tail-index $\alpha \in (0, 2]$ measuring the heaviness of the distribution. A description of other α-based SSL algorithms can be found in [19], [20].

In this paper, we propose to exploit the α-EC through a covariation-based parameter estimation on the observation shape matrix that does not require to compute the ECF. Moreover, we include the proposed estimation in [21] of the tail-index α to capture the dynamic range of the signal and show the performance and time consumption in several SSL scenarios on both synthetic and real data.

All the code used to produce the results of this paper is available at `https://github.com/matfontaine/alphaMUSIC`.

II. MUSIC: The Basic Approach to SSL

This section outlines SSL methods based on the MUSIC algorithm. We first recall the pioneering one [22] in Section II-A and then overview the variants in Section II-B. Let us assume that the observed signal is captured by a microphone array composed of M sensors and represented in the short-time Fourier transform (STFT) domain as $\mathbf{x}_{ft} \triangleq [x_{1ft}, \ldots, x_{Mft}]^\top \in \mathbb{C}^M$, where $f \in \{1, \ldots, F\}$ and $t \in \{1, \ldots, T\}$ are the frequency and frame indexes, respectively. Here $\triangleq$ denotes the equality by definition and $^\top$ the transposition.

A. MUSIC Framework

Assuming the STFT window length is larger than the reverberation time, for all f, t the STFT of the recorded observed signals read [23]

$$\mathbf{x}_{ft} = \mathbf{A}_f \mathbf{s}_{ft} + \mathbf{n}_{ft}, \tag{1}$$

where $\mathbf{s}_{ft} \in \mathbb{C}^N$ are the $N < M$ punctual sources, $\mathbf{A}_f \in \mathbb{C}^{M \times N}$ is the mixing matrix, and $\mathbf{n}_{ft}$ is an additive white noise component with variance σ^2 uncorrelated to the signal $\mathbf{s}_{ft}$. We assume that the sources are located sufficiently far from the array to hold the plane wave assumption. In this context, the SSL task reduces to estimate the angle of arrival (AOA) of target sources. From Eq. (1), the covariance matrix $\mathbf{R}_{\mathbf{x},ft}$ of $\mathbf{x}_{ft}$ reads

$$\mathbf{R}_{\mathbf{x},ft} \triangleq \mathbb{E}\left[\mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}}\right] = \mathbf{A}_f \mathbf{R}_{\mathbf{s},ft} \mathbf{A}_f^{\mathsf{H}} + \sigma^2 \mathbf{I}_M, \tag{2}$$

where $^{\mathsf{H}}$ is the Hermitian transposition, $\mathbf{R}_{\mathbf{s},ft}$ is the covariance matrix of $\mathbf{s}_{ft}$ and $\mathbf{I}_M$ is the identity matrix of size $M \times M$. Assuming stationary sources, Eq. (2) is then time-averaged

$$\widehat{\mathbf{R}}_{\mathbf{x},f} \triangleq \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}}. \tag{3}$$

The full column rankness of $\mathbf{A}_f$, the positive definiteness of $\mathbf{R}_{\mathbf{s},ft}$ and $\widehat{\mathbf{R}}_{\mathbf{x},f}$ implies that the N largest eigenvalues of $\widehat{\mathbf{R}}_{\mathbf{x},f}$ are associated to the signal space and the $N - M$ other ones to the noise space. Let $\{\mathbf{v}_{p,f}\}_{p=1}^{N-M}$ be the noise eigenvectors spanning the noise space $\mathbf{Q}_{\mathbf{n},f} \triangleq [\mathbf{v}_{1,f}, \ldots, \mathbf{v}_{N-M,f}]$. Then

$$\mathbf{A}_f^{\mathsf{H}} \mathbf{v}_{l,f} = 0 \tag{4}$$

can be proved and means that the steering vectors associated to their AOAs are orthogonal to the noise eigenvectors. We obtain the so-called pseudo-spectrum function for an arbitrary angle of arrival θ and its steering vector $\mathbf{a}_f(\theta)$:

$$P_f^{\mathrm{MUSIC}}(\theta) = \frac{1}{\mathbf{a}_f^{\mathsf{H}}(\theta) \mathbf{Q}_{\mathbf{n},ft} \mathbf{Q}_{\mathbf{n},ft}^{\mathsf{H}} \mathbf{a}_f(\theta)}. \tag{5}$$

From Eq. (4), the denominator in Eq. (5) is supposed to be closed to zero when the angular direction θ is one of the sources resulting in a peaked function. Therefore, assuming that the number of sources N is known, the SSL task is performed by selecting the N highest peaks of $\sum_f P_f^{\mathrm{MUSIC}}(\theta)$ evaluated on a set of candidate sources' AOA $\{\theta_l\}_{l=1}^{L}$.

B. MUSIC Variants

The performance of MUSIC is known to decrease in case of coherent sources occurring in complex audio scenarios [8], and several works have tried to improve its robustness to different environments. NormMUSIC [4] reduces the incorrect response power estimation due to the SNR variations at different frequencies by arithmetic mean normalization of Eq. (5), *i.e.*,

$$P^{\mathrm{NormMUSIC}}(\theta) = \sum_{f=1}^{F} \frac{P_f^{\mathrm{MUSIC}}(\theta)}{\max\limits_{l=1,\ldots,L} P_f^{\mathrm{MUSIC}}(\theta_l)}. \tag{6}$$

GC-MUSIC [13] deals with impulsive noises that do not fit the Gaussian assumption in Eq. (1), and thus degrades the covariance matrix estimation. It replaces $\mathbf{R}_{\mathbf{x},ft}$ in Eq. (2) by a generalized covariance (GC) matrix $\mathbf{R}_{\mathbf{x},ft}^{(\mathrm{GC})}$ whose entries are defined as

$$\left[\mathbf{R}_{\mathbf{x},ft}^{(\mathrm{GC})}\right]_{mm'} = \mathbb{E}\left[\frac{g_1(x_{ftm}) g_2(x_{ftm'})}{h_1(x_{ftm}, x_{ftm'}) h_2(x_{ftm'}, x_{ftm'})}\right] \tag{7}$$

with g_1, g_2 are two single-variable functions while h_1, h_2 are two dual-variable functions. The covariance occurs when $g_1(u) = u, g_2(u) = u^\star, h_1 = h_2 = 1$, where * is the conjugate operator. GC includes covariation-based matrix that we discuss further in the next section.

III. α-MUSIC: The Proposed Approach to SSL

The α-stable random vectors are the one that preserve the law under a finite linear combination [18]. The heaviness of the distribution is controlled by a tail-index $\alpha \in (0, 2]$ for whose $\alpha = 2, \alpha = 1$ and $\alpha = 0.5$ represent the Gaussian, Cauchy and Levy case ranging from lightest to heaviest respectively. Most of those distributions however are not parameterized with a GC matrix but rather via a spectral measure [18]. For $1 < \alpha \leq 2$, the so-called *covariation* [18] provides a correlation information controlled by the spectral measure of two α-stable variables.

A complex multivariate elliptically contoured α-stable distribution (α-EC) [24], [25] vector $\mathbf{u}$ of size M is an α-stable vector which naturally designs a so-called positive definite shape matrix $\mathbf{R}^{(\alpha)} \in \mathbb{C}^{M \times M}$ and will be denoted $\mathbf{u} \sim \mathcal{S}_{\mathbb{C}}^{\alpha}\left(\mathbf{R}^{(\alpha)}\right)$. Contour plot of α-EC are represented in Fig. 1. A link between the shape matrix coefficients $\left[\mathbf{R}^{(\alpha)}\right]_{mm'}$ and the covariation coefficients denoted $r_{mm'}^{(\alpha)}$ was established [18]:

$$\left[\mathbf{R}^{(\alpha)}\right]_{mm'} = \begin{cases} 2\left(r_{mm}^{(\alpha)}\right)^{2/\alpha} & \text{if } m = m' \\ 2^{\alpha/2} r_{mm'}^{(\alpha)} \left[\mathbf{R}^{(\alpha)}\right]_{m'm'}^{-\frac{\alpha-2}{2}} & \text{otherwise} \end{cases} \tag{8}$$

Our main purpose is to build a natural positive definite matrix estimator essential for MUSIC to work in practice. As recently shown in [16], a MUSIC-based α-EC model as follows:

$$\mathbf{x}_{ft} \sim \mathcal{S}_{\mathbb{C}}^{\alpha}\left(\mathbf{R}_{\mathbf{x},ft}^{(\alpha)}\right), \tag{9}$$

$$\mathbf{s}_{ft} \sim \mathcal{S}_{\mathbb{C}}^{\alpha}\left(\mathrm{Diag}\left[\sigma_{1ft}^{\alpha}, \ldots, \sigma_{Nft}^{\alpha}\right] \triangleq \mathbf{R}_{\mathbf{s},ft}^{(\alpha)}\right), \tag{10}$$

$$\mathbf{n}_{ft} \sim \mathcal{S}_{\mathbb{C}}^{\alpha}\left(\sigma_{ft}^{\alpha} \mathbf{I}_M\right) \tag{11}$$

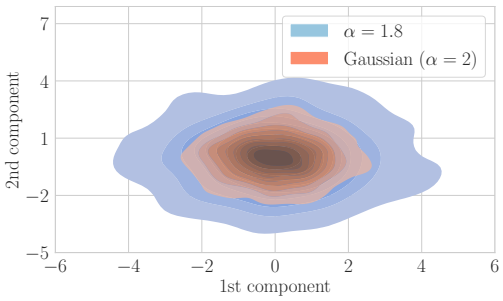

Fig. 1. Contour plot from Gaussian (in red) and α-EC (in blue) samplings.

combined with Eq. (1) and Eq. (9 - 11) leads to

$$\mathbf{R}^{(\alpha)}_{\mathbf{x},ft} = \mu_{ft}\left(\mathbf{A}_f \mathbf{R}^{(\alpha)}_{\mathbf{s},ft}\mathbf{A}_f^{\mathsf{H}} + \sigma^2\mathbf{I}_M\right) \tag{12}$$

with $\mu_{ft} \triangleq 2^{\frac{2-\alpha}{2}}\left[\sigma^{\alpha}_{ft} + \sum_{n=1}^{N}\sigma^{\alpha}_{nft}\right]^{\frac{2-\alpha}{\alpha}}$. In [16], Eq. (12) is estimated using the empirical characteristic function (ECF). We rather propose a covariation-based estimation in order to avoid drawing representative sampling issue in ECF computation [17].

From Eq. (8), the shape matrix $\mathbf{R}^{(\alpha)}_{\mathbf{x},ft}$ estimation is equivalent to compute covariation coefficients between all $\mathbf{x}_{ft}$ entries. The sources immobile assumption motivates to use a time-independent covariation estimator $\widehat{r}^{(\alpha)}_{f,mm'}$ between all x_{ftm} and $x_{ftm'}$. The estimation is processed through the following fractional lower order moment technique [17]:

$$\widehat{r}^{(\alpha)}_{f,mm'} = \begin{cases} \exp\left[\alpha^{-1}(T^{-1}\sum_{t=1}^{T}\ln|x_{ftm}| - \eta)\right] & \text{if } m = m' \\ \frac{\sum_{t=1}^{T} x_{ftm}|x_{ftm'}|^{p-2}x^{\star}_{ftm'}}{\sum_{t=1}^{T}|x_{ftm'}|^{p}}\widehat{r}_{f,m'm'} & \text{otherwise} \end{cases} \tag{13}$$

where $1 \leq p < \alpha$ and $\eta = \gamma(\alpha^{-1} - 1) - \ln 2$ with $\gamma \approx 0.577$ the Euler constant. An empirical estimator $\widehat{\mathbf{R}}^{(\alpha)}_{\mathbf{x},f} \triangleq \left[\widehat{r}^{(\alpha)}_{f,mm'}\right]_{m,m'}$ is obtained by combining Eq. (13) and Eq. (8) along the frequency axis. We force $\widehat{\mathbf{R}}^{(\alpha)}_{\mathbf{x},f}$ to be Hermitian by considering $\widehat{\mathbf{R}}^{(\alpha)}_{\mathbf{x},f} \leftarrow \frac{1}{2}\left(\widehat{\mathbf{R}}^{(\alpha)}_{\mathbf{x},f} + \left[\widehat{\mathbf{R}}^{(\alpha)}_{\mathbf{x},f}\right]^{\mathsf{H}}\right)$. Due to an α-EC observed model, the tail-index α can be estimated as in [21]. The positive definiteness of $\widehat{\mathbf{R}}^{(\alpha)}_{\mathbf{x},f}$ and $\mathbf{R}_{\mathbf{s},ft}$ leads to a fractional noise space $\mathbf{Q}^{(\alpha)}_{n,f}$ and the following pseudo-spectrum function

$$P^{\alpha-\text{MUSIC}}(\theta) = \frac{1}{F}\sum_{f=1}^{F}\frac{1}{\mathbf{a}_f^{\mathsf{H}}(\theta)\mathbf{Q}^{(\alpha)}_{n,f}\left(\mathbf{Q}^{(\alpha)}_{n,f}\right)^{\mathsf{H}}\mathbf{a}_f(\theta)}. \tag{14}$$

The proposed α-MUSIC can be easily extend to α-NormMUSIC by combining Eq. (14) and Eq. (6). All MUSIC algorithms variants are outlined in Algorithm 1.

IV. Experiments

We investigate the SSL performances of both proposed α-MUSIC and α-NormMUSIC with MUSIC [22], NormMUSIC [4] and SRP-PHAT [26] as baselines.[1] Two types of dataset are used: a *synthetic dataset* made with LibriSpeech utterances [28] to generate microphones recordings through `pyroomacoustics` simulator [27] corrupted by various real noise coming from DEMAND dataset; and a *real dataset* coming from LOCATA Challenge. The tail-index α in proposed SSL algorithms is estimated with the same setting as in [21] and $1 \leq p < 2$ in Eq. (13) is set to $\frac{\alpha+1}{2}$. The SSL performances are evaluated in terms of angular error in degree. As a proof of concept, only the azimuth estimation for single and multiple sources are considered. In case of $N > 1$, the mean error is computed first over the N sources and then averaged over all the observations.

[1] as available in `pyroomacoustics` library [27].

Algorithm 1 MUSIC algorithms and variants

1) **Input**
   a) Number of sources N and observed signal $\mathbf{x}_{ft}$;
   b) Candidate AOAs $\{\theta_l\}_{l=1}^{L}$ and related steering vectors $\mathbf{a}_f(\theta_l)$.
2) **(fractional) pseudo-spectrum function estimation**
   a) *Optional*: estimation of α using $\mathbf{x}_{ft}$ as in [21];
   b) Compute $\widehat{\mathbf{R}}_f$ according to Eq. (2) or Eq. (13) for MUSIC or α-MUSIC respectively;
   c) Estimate the noise eigenspace $\mathbf{Q}_{\mathbf{n},f}$ (or $\mathbf{Q}^{(\alpha)}_{\mathbf{n},f}$);
   d) Compute $\forall l, P(\theta_l)$ according to Eq. (5) or Eq. (14) for MUSIC or α-MUSIC respectively;
   e) *Optional*: Apply Eq. (6) to get "Norm" extension.
3) **Apply pick detection on** $P(\theta_l)$.

A. Experiments on the Synthetic Data

The *synthetic dataset* exploits 4 settings summarized in Table I for whose one of the following key scenario conditions is varying and the other ones are fixed: the number of microphones M, reverberation time (RT60), signal-to-noise ratio (SNR) and energy of acoustic reflections.[2]

Otherwise specified, we considered a shoebox room of size $6 \times 5 \times 3$ m in mild acoustic conditions (SNR = 10 dB, RT60 = 0.5 s). Here $N \in \{1, 2, 3\}$ sound sources are deployed at a fixed distance of 1 m and elevation of 0° with respect to a linear uniform array with an inter-microphone spacing of 8 cm. For a fixed value N, 180 observations are generated by randomly drawing sources' azimuth and the free scenario parameter. The sampling rate of microphone signals is 16 kHz. The STFT of the data is computed with a window length of 32 ms, overlap of 50% and 513 real frequencies. Only the frequency range 500 - 4000 Hz was considered for a more robust estimation. Finally, the resolution of candidate azimuths grid is 1°.

At first, we study the influence of the tail-index α for more insight on this parameter. The results in Fig. 3 show that α increases when the RT_{60} increases in $\mathcal{D}^{\text{RT}_{60}}$ and the SNR decreases in $\mathcal{D}^{\text{SNR}}$, respectively. Since no significant trend is ex-

[2] If RT60, room size, and array position are fixed, then increasing the reflections energy corresponds to "sending" the sources closer to the room reflectors. The source-to-array distance is then used as a proxy for studying the robustness to acoustic reflections.

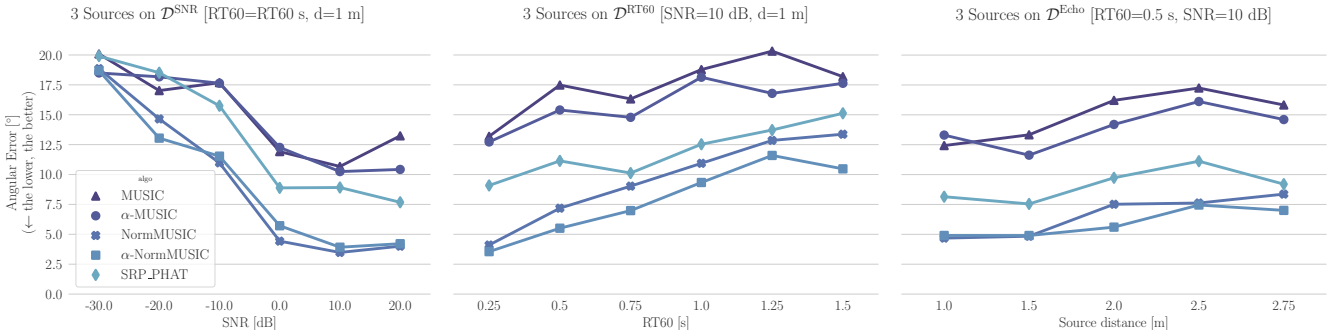

Fig. 2. Angular Error (in degree) as function of different acoustic parameters (SNR, RT60 and sources-to-microphone distance) involving 3 sources and a linear microphone array of 4 sensors. Details of the dataset used to study each parameter are reported in Tab. I.

TABLE I
DATASET CONDITION FOR SYNTHETIC EVALUATION.

	M mics	RT60 [s]	SNR [dB]	Sources' dist. [m]
$\mathcal{D}^{\text{mics}}$	[4, 6]	0.5	10	1
$\mathcal{D}^{\text{RT60}}$	4	$\mathcal{U}$[0.25:0.25:2]	10	1
$\mathcal{D}^{\text{SNR}}$	4	0.5	$\mathcal{U}$[-30:5:20]	1
$\mathcal{D}^{\text{Echo}}$	4	0.5	10	$\mathcal{U}$[1:0.5:3]

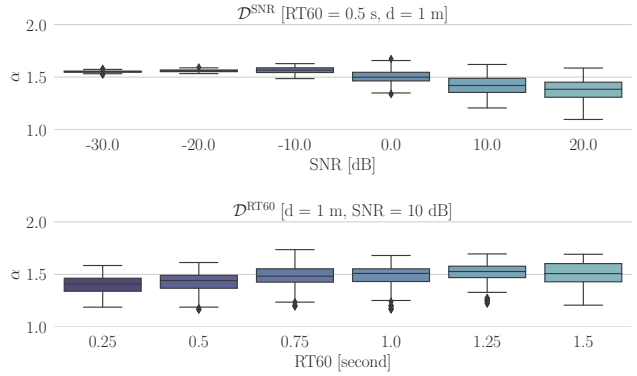

Fig. 3. Value of α as function of the SNR (top) and RT60 (botton).

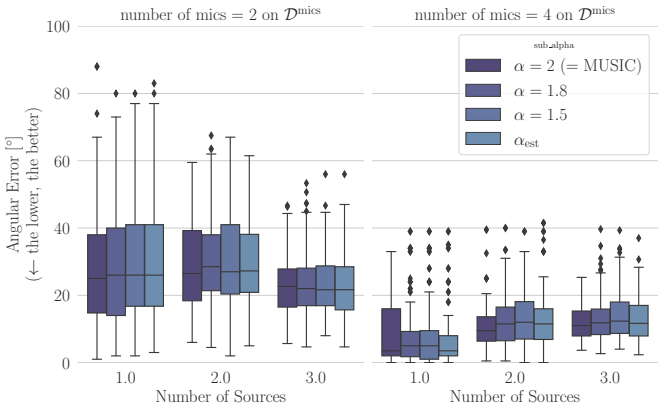

Fig. 4. Angular Error (in degree) for $N \in \{1, 2, 3\}$ sources and $M \in \{2, 4\}$ microphones in the uniform linear array.

TABLE II
TIME ELAPSED IN SECONDS. FASTEST IS BOLDED.

	MUSIC	NormMUSIC	SRP-PHAT	α-MUSIC	α-NormMUSIC
$M = 2$	**0.008**	**0.008**	0.018	0.016	0.016
$M = 4$	**0.015**	**0.015**	0.028	0.030	0.030
$M = 6$	**0.024**	**0.023**	0.041	0.049	0.049

hibited when varying the echo energy in the $\mathcal{D}^{\text{Echo}}$, the figure is not reported. Previous studies in [29], [30] pointed out that the clean speech is better modelled with $\alpha = 1.2$, following our results. A small variability of the α values is noted regarding extreme SNR scenarios and could be explained by a low dynamic range of the observed signal.

Fig. 4 show the results in terms of MAE for the proposed α-MUSIC with either a tail-index estimation or a fixed $\alpha \in \{1.5, 1.8, 2\}$ for $N \in \{1, 2, 3\}$ sources in the $\mathcal{D}^{\text{mics}}$ setting. Note that when $\alpha = 2$, it corresponds to the vanilla MUSIC case. The estimated α-MUSIC version is slightly better than other variants in terms of 95% confidence interval and median value in multiple speaker scenarios. The gap between MUSIC and α-MUSIC is growing as the number of microphones increases. As the SSL performances increase with the number of microphones, no significant difference between the different methods is, hence, not reported here.

Results in term of MAE of retrieved azimuths for the different acoustic parameters of $\mathcal{D}^{\text{SNR}}$, RT_{60} and $\mathcal{D}^{\text{Echo}}$ for $N = 3$ sources and $M = 4$ microphones are shown in Fig. 2. In general, it can be noticed that the proposed approach slightly outperforms the standard MUSIC approaches in case of strong adverse conditions (SNR $<$ - 10 dB, RT60 $>$ 0.25, distance $>$ 1.5 m). While SRP-PHAT outperforms both MUSIC and α-MUSIC in terms of MAE of the retrieved angles, it fails at recovering more than 2 sources in more than 57% of the observations. Conversely, all MUSIC-based algorithms always estimated N different sources. Nevertheless, a noticeable difference is reported in the RT_{60} scenario, confirming that α-NormMUSIC is better suited in case of strong reverberation.

Finally,the empirical time elapsed in seconds is reported in Tab. II. The proposed extension is two times slower than the baselines and scales linearly with the number of sensors, mainly due to the α estimation step. Nevertheless, as the overall latency is in the order of milliseconds, the proposed algorithm is suitable for real-time applications.

TABLE III

MEAN ANGULAR ERROR ± STANDARD DEVIATION FOR TASK #1 (SINGLE STATIC SOURCE) AND TASK #2 (MULTIPLE STATIC SOURCES) ON THE LOCATA DATASET

Task	Algorithm	DICIT array	Robot Head	Hearing aids
1	MUSIC	3.09 ± 7.91	2.14 ± 1.66	7.77 ± 16.62
	NormMUSIC	$\mathbf{2.74 \pm 6.85}$	2.00 ± 6.64	$\mathbf{1.25 \pm 3.46}$
	SRP_PHAT	4.42 ± 10.92	1.84 ± 2.13	1.77 ± 5.06
	α-MUSIC	4.07 ± 9.57	2.91 ± 3.43	6.77 ± 15.80
	α-NormMUSIC	3.34 ± 9.40	$\mathbf{1.83 \pm 6.25}$	1.27 ± 3.71
2	MUSIC	10.86 ± 13.73	15.61 ± 15.93	24.56 ± 15.26
	NormMUSIC	$\mathbf{10.39 \pm 14.45}$	16.35 ± 14.71	22.30 ± 15.68
	SRP_PHAT	16.19 ± 14.18	20.29 ± 16.24	$\mathbf{17.35 \pm 16.09}$
	α-MUSIC	12.48 ± 13.87	$\mathbf{15.39 \pm 15.15}$	24.31 ± 15.23
	α-NormMUSIC	10.81 ± 14.12	17.43 ± 15.45	23.04 ± 15.38

B. Experiments on the Real Data

The LOCATA dataset comprises recordings from a room of size $7.1 \times 9.8 \times 3$ m with RT60$= 0.55$ s affected by some low noise. We use the development tasks # 1 and #2 featuring single and multiple static speakers, respectively, for the DICIT array, Robot head and Hearing aids. As in LOCATA challenge, the scores are computed on speech-only frames, using the voice activity detection annotations provided within the ground-truth data. The data are processed with the same parameters as above.

Results are reported in Table III. We observed that the proposed methods do not consistently outperform the baselines. Interestingly, this trend is common to the normalized-version of MUSIC. Therefore, these results confirm that our methods should be tested on a real dataset featuring a stronger level of noise and reverberation, such as the one envisioned in [31].

V. CONCLUSION

This paper proposes α-MUSIC, a theoretically justified adaptive sound source localization (SSL) method based on a variant of the classical multiple signal classification (MUSIC) method with the complex multivariate elliptically contoured α-stable model. We show that in case of multiple static sources and high reverberation, high distance or low SNR α-MUSIC is more robust. Future work includes α-MUSIC with time-varying tail-indexes and a real-time implementation.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Nos. 19H04137, 20K19833, and 20K21813.

REFERENCES

[1] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proc. DCASE Workshop*, 2021, pp. 125–129.

[2] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *Int. J. Neural Syst.*, vol. 25, no. 1, pp. 1 440 003:1–1 440 003:21, 2015.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[4] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 581–585, 2014.

[5] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. 26503, pp. 1–19, 2006.

[6] D. Di Carlo, A. Deleforge, and N. Bertin, "Mirage: 2D source localization using microphone pair augmentation with echoes," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 775–779.

[7] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[8] Y. Gao, W. Chang, Z. Pei, and Z. Wu, "An improved music algorithm for doa estimation of coherent signals," *Sensors and Transducers*, vol. 175, no. 7, pp. 75–82, 2014.

[9] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. Int. Conf. Intell. Rob. Syst.*, 2009, pp. 664–669.

[10] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *Proc. Int. Conf. Intell. Rob. Syst.*, 2012, pp. 694–699.

[11] F. Grondin and J. Glass, "Fast and robust 3-D sound source localization with DSVD-PHAT," in *Proc. Int. Conf. Intell. Rob. Syst.*, 2019, pp. 5352–5357.

[12] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *arXiv preprint arXiv:2109.03465*, 2021.

[13] S. Luan, M. Zhao, Y. Gao, Z. Zhang, and T. Qiu, "Generalized covariance for non-Gaussian signal processing and GC-MUSIC under Alpha-stable distributed noise," *Digital Signal Processing*, vol. 110, p. 102923, 2021.

[14] P. Tsakalides and C. L. Nikias, "The robust covariation-based music (roc-music) algorithm for bearing estimation in impulsive noise environments," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1623–1633, 1996.

[15] S. Fortunati, A. Renaux, and F. Pascal, "Robust semiparametric doa estimation in non-gaussian environment," in *Rad. Conf.*, 2020, pp. 1–6.

[16] M. Asghari, M. Zareinejad, S. M. Rezaei, and H. Amindavar, "ECF-MUSIC: An empirical characteristic function based direction of arrival (DOA) estimation in the presence of impulsive noise," *Digital Signal Processing*, vol. 123, p. 103440, 2022.

[17] C. L. Nikias and M. Shao, *Signal processing with alpha-stable distributions and applications*. Wiley-Interscience, 1995.

[18] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance: Stochastic Modeling*. Routledge, 2017.

[19] M. Fontaine, C. Vanwynsberghe, A. Liutkus, and R. Badeau, "Sketching for nearfield acoustic imaging of heavy-tailed sources," in *Int. Conf. on Latent Var. Ana. and Sig. Sep.*, 2017, pp. 80–88.

[20] P. G. Georgiou, P. Tsakalides, and C. Kyriakakis, "Alpha-stable robust modeling of background noise for enhanced sound source localization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 6, 1999, pp. 3085–3088.

[21] M. Fontaine, K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Alpha-stable autoregressive fast multichannel nonnegative matrix factorization for joint speech enhancement and dereverberation," *Proc. Interspeech*, pp. 661–665, 2021.

[22] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[23] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation priors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 12, pp. 2453–2465, 2016.

[24] J. P. Nolan, "Multivariate elliptically contoured stable distributions: theory and estimation," *Comput. Stat.*, vol. 28, no. 5, pp. 2067–2089, 2013.

[25] S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable multichannel audio source separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 576–580.

[26] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[27] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 351–355.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[29] M. Fontaine, A. Liutkus, L. Girin, and R. Badeau, "Explaining the parameterized wiener filter with alpha-stable processes," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 51–55.

[30] U. Şimşekli, A. Liutkus, and A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2289–2293, 2015.

[31] D. D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dEchorate: A calibrated room impulse response dataset for echo-aware signal processing," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, pp. 1–15, 2021.