Proceedings of the 2007 IEEE/RSJ International
Conference on Intelligent Robots and Systems
San Diego, CA, USA, Oct 29 - Nov 2, 2007

Broadband Variations of the MUSIC High-Resolution Method for Sound Source Localization in Robotics

Sylvain Argentieri and Patrick Danès

Abstract— **The MUSIC algorithm (MUltiple SIgnal Classification) is a well-known high-resolution method to sound source localization. However, as it is essentially narrowband, several extensions can be envisaged when dealing with broadband sources like human voice. This paper presents such extensions and proposes a comparative study w.r.t. specific robotics constraints. An online beamspace MUSIC method, together with a recently developed beamforming scheme, are shown to constitute a mathematically sound and potentially efficient solution.**

I. INTRODUCTION

"Robot Audition" is a key paradigm to natural human-robot communication. Enabling a human to interact through voice requires to endow the robot with capabilities of speaker localization, voice extraction and speech recognition. Numerous techniques coming from the Array Signal Processing community can be assessed to precisely localize sound sources. Generally, they rely on the farfield assumption, i.e. sound sources are assumed far enough so that planar wavefronts can be considered. However, in real operations with human speakers, the voice signal, traditionally restricted to the bandwidth $[300\text{Hz}; 3\text{kHz}]$, together with the use of a small-size microphone array, imply spherical wavefronts. Beamforming can be straightly adapted to this nearfield case by expressing the array pattern as an explicit function of the distance r to the source. But classical delay-and-sum beamforming is shown to have a bad resolution in the bearing and range estimation [1]. Consequently, we have proposed in [2] an improved nearfield beamforming method. As it presupposes the knowledge of r, other methods must be envisaged for simultaneous azimuth and range estimation. For instance, recent robotics papers combine classical delay-and-sum beamforming with other cues or algorithms to estimate the sound source positions. A successful range and bearing estimation based on cross-correlation techniques mixed with a particle-based tracking algorithm is proposed in [3]. A 2D sound source mapping system which benefits from the movement of the robot is presented in [4]. A real-time tracking of multiple sound sources by integration of multiple microphone arrays is performed in [5]. Ref. [1] shows that the well known subspace method MUSIC (for MUltiple SIgnal Classification) can be successfully applied offline to get a good range and bearing estimation. To our knowledge, this has been the only robotics implementation based on any high-resolution algorithm. In fact, MUSIC cannot be used in realtime because of its very heavy computational cost. Nevertheless, very recent broadband extensions can drastically reduce the load. Some of them are hereafter presented and compared with respect to robotics constraints.

Sylvain Argentieri and Patrick Danès are with Université de Toulouse, France: LAAS-CNRS, 7 avenue du Colonel Roche, 31077 Toulouse, and Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse. sargenti@laas.fr, danes@laas.fr

The paper is organized as follows. The notations are first defined in Section II. A recall on the classical narrowband MUSIC algorithm follows. Next, a first immediate extension of the previous algorithm to the broadband case, similar to [1], is proposed in Section III. Its computational cost makes it unsuited to robotics. So, Section IV describes a recent broadband beamspace extension which, combined with our beamforming synthesis method proposed in [2], forms an efficient scheme. A conclusion ends the paper.

II. THE NARROWBAND MUSIC ALGORITHM

In the whole paper, $(.)^T$ and $(.)^H$ respectively term the transpose and Hermitian transpose operators. Normal lowercase letters and capitals depict signals in the temporal and frequency domains, respectively. Bold letters term vectors made of such signals while underlined bold letters relate to matrices. The $N \times N$ identity and $N \times M$ zero matrices are denoted $\underline{\mathbb{I}}_N$ and $\underline{\mathbb{O}}_{N,M}$. $\mathbb{E}[.]$ is the expectation operator.

A. Hypotheses and Problem statement

It is assumed that the 3D space is homogeneous and isotropic when no acoustic perturbation is present, and that the linear acoustics hypotheses hold. Let D pointwise independent zero-mean stationary sources be positioned into this environment. The wavefield they create is spatially sampled by an array of $N > D$ omnidirectional sensors. Each d^{th} source position, $d = 1, \ldots, D$, is depicted by its spherical coordinates vector $\mathbf{r}_d = (r_d, \theta_d, \phi_d)^T$ in a frame $\mathscr{F} = (O, \vec{x}, \vec{y}, \vec{z})$. The wave propagation velocity is supposed constant and equal to $c = 340\text{m.s}^{-1}$. Though the localization methods hereafter presented do not need this property, the antenna is linear, with evenly spaced elements distributed along the $\vec{z}$-axis at coordinates z_n, $n = 1, \ldots, N$, so that θ_d and ϕ_d term the d^{th} source azimuth and elevation angles.

The received complex signal at $\mathscr{F}$'s origin is thus the sum of the contributions $s_d(t)$ relative to each d^{th} source, $d = 1, \ldots, D$. In the so-called narrowband sources case, each $s_d(t)$ reads as $s_d(t) = S_d(k) e^{2j\pi ft} = S_d(k) e^{jkct}$, with f and $k = \frac{2\pi f}{c}$ the sources common temporal and spatial frequencies. Define the vector $\mathbf{S}(k) = (S_1(k), \ldots, S_D(k))^T$. The complex envelopes $X_n(k)$ of the signals $x_n(t)$ perceived at the array, $n = 1, \ldots, N$, gathered into the vector $\mathbf{X}(k) = (X_1(k), \ldots, X_N(k))^T$, then satisfy [6]

$$\mathbf{X}(k) = \underline{\mathbf{V}}(\mathbf{r}_1, \ldots, \mathbf{r}_D, k)\mathbf{S}(k) + \mathbf{B}(k), \tag{1}$$

with $\underline{\mathbf{V}}(\mathbf{r}_1, \ldots, \mathbf{r}_D, k) = \left(\mathbf{V}(\mathbf{r}_1, k) \mid \ldots \mid \mathbf{V}(\mathbf{r}_D, k) \right)$ the $N \times D$ matrix built up with the steering vectors $\mathbf{V}(\mathbf{r}_d, k)$ relative to each d^{th} source and $\mathbf{B}(k) = (B_1(k), \ldots, B_N(k))^T$ an additive noise

on the sensors. This noise is assumed zero-mean, stationary, temporally and spatially white, of known equal power on each microphone, and independent of the sources, so that

$$\mathbb{E}[\mathbf{B}\mathbf{B}^H] = \sigma^2_{\mathscr{N}}\underline{\mathbb{I}}_N \text{ and } \mathbb{E}[\mathbf{B}(\underline{\mathbf{V}}\mathbf{S})^H] = 0. \quad (2)$$

This independence assumption, though a priori criticizable for robotics, will be proved nonrestrictive thanks to a new problem statement in Section IV. Due to the symmetry of the problem around the $\vec{z}$-axis, each dependency on $\mathbf{r}_d$ can be reduced to a dependency on (r_d, θ_d). Recalling that z_n terms the $\vec{z}$-coordinate in $\mathscr{F}$ of the n^{th} sensor, the n^{th} entry of any steering vector $\mathbf{V}(\mathbf{r},k)$, with $\mathbf{r} = (r,\theta,\phi)^T$ a dummy vector of spherical coordinates, takes the form $V_n(r,\theta,k)$ where

$$V_n(r,\theta,k) = re^{jkr}\frac{e^{-jk\sqrt{r^2+z_n^2-2rz_n\cos\theta}}}{\sqrt{r^2+z_n^2-2rz_n\cos\theta}},\ n=1,\ldots,N. \quad (3)$$

If $r = ||\mathbf{r}||$ tends to $+\infty$, then $\mathbf{V}(r,\theta,k)$ reads as the farfield steering vector $\mathbf{V}^\infty(\theta,k)$ of coordinates

$$V_n^\infty(\theta,k) = \lim_{r\to+\infty} V_n(r,\theta,k) = e^{jkrz_n\cos\theta},\ n=1,\ldots,N. \quad (4)$$

In fact, $\mathbf{V}(r,\theta,k)$ can be approximated by $\mathbf{V}^\infty(\theta,k)$ as soon as r exceeds the Rayleigh distance $\mathscr{R} = 2L_0^2/\lambda$, with L_0 the array length and $\lambda = c/f = 2\pi/k$ the wavelength.

A meaningful problem is as follows: *the perceived signals vector* $\mathbf{x}(t) = (x_1(t),\ldots,x_N(t))^T$ *being given, how can the number D of sources and their ranges and azimuths* (r_d,θ_d), $d = 1,\ldots,D$, *be determined?*

B. Mathematical foundations of the MUSIC method

The so-called "high-resolution" parametric spectral analysis method MUSIC constitutes an efficient solution to the above problem. In order to alleviate the notation, the dependencies of variables upon the single involved wavenumber k will be temporarily omitted.

MUSIC is based on the eigendecomposition of the covariance—or interspectral—matrix $\underline{\mathbf{C}}_X = \mathbb{E}[\mathbf{X}\mathbf{X}^H]$ of $\mathbf{X}$, which describes the second-order statistics of the signals perceived at the array. From (1) and (2), $\underline{\mathbf{C}}_X$ also satisfies

$$\underline{\mathbf{C}}_X = \underline{\mathbf{V}}(\mathbf{r}_1,\ldots,\mathbf{r}_D)\underline{\mathbf{C}}_S\underline{\mathbf{V}}^H(\mathbf{r}_1,\ldots,\mathbf{r}_D) + \sigma^2_{\mathscr{N}}\underline{\mathbb{I}}_N, \quad (5)$$

with $\underline{\mathbf{C}}_S = \mathbb{E}[\mathbf{S}\mathbf{S}^H]$ the $D\times D$ covariance matrix of the sources and $\underline{\mathbf{C}}_B = \mathbb{E}[\mathbf{B}\mathbf{B}^H] = \sigma^2_{\mathscr{N}}\underline{\mathbb{I}}_N$ the noise covariance matrix. The $N\times N$ matrix $\underline{\mathbf{C}}_Y = \underline{\mathbf{V}}(\mathbf{r}_1,\ldots,\mathbf{r}_D)\underline{\mathbf{C}}_S\underline{\mathbf{V}}^H(\mathbf{r}_1,\ldots,\mathbf{r}_D)$ is Hermitian symmetric, positive semidefinite, and thus admits N real nonnegative eigenvalues λ_n which can be associated to orthogonal right eigenvectors $\mathbf{U}_n$, $n = 1,\ldots,N$. As the sources are assumed mutually independent and as $\underline{\mathbf{V}}(\mathbf{r}_1,\ldots,\mathbf{r}_D)$ is assumed full rank whatever $\mathbf{r}_1,\ldots,\mathbf{r}_D$, $\underline{\mathbf{C}}_Y$ has rank D so that its eigenvalues can be ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D > \lambda_{D+1} = \ldots = \lambda_N = 0$. Note that the vectors $\mathbf{U}_1,\ldots,\mathbf{U}_D$ span the range of $\underline{\mathbf{V}}(\mathbf{r}_1,\ldots,\mathbf{r}_D)$, i.e. the D-dimensional subspace $\mathscr{S}$ of $\mathbb{C}^N$ generated by the steering vector evaluated at the sources locations, and henceforth termed "signal space". From $\underline{\mathbf{C}}_X = \underline{\mathbf{C}}_Y + \sigma^2_{\mathscr{N}}\underline{\mathbb{I}}_N$, it follows:

$$\underline{\mathbf{C}}_X = (\underline{\mathbf{U}}_{\mathscr{S}} \mid \underline{\mathbf{U}}_{\mathscr{N}}) \begin{pmatrix} \lambda_1+\sigma^2_{\mathscr{N}} & & \underline{\mathbb{O}} & | & \\ & \ddots & & | & \underline{\mathbb{O}} \\ & & \lambda_D+\sigma^2_{\mathscr{N}} & | & \\ \text{--} & \text{--} & \text{--} & | & \text{--} \\ & \underline{\mathbb{O}} & & | & \sigma^2_{\mathscr{N}}\underline{\mathbb{I}}_{N-D} \end{pmatrix} (\underline{\mathbf{U}}_{\mathscr{S}} \mid \underline{\mathbf{U}}_{\mathscr{N}})^H; \quad (6)$$

- $\underline{\mathbf{U}}_{\mathscr{S}} = (\mathbf{U}_1 \mid \ldots \mid \mathbf{U}_D) \in \mathbb{R}^{N\times D}$ is the matrix of the D aforementioned eigenvectors, now associated to the eigenvalues $\lambda_n + \sigma^2_{\mathscr{N}}$, $n = 1,\ldots,D$, and still generating the signal space;
- $\underline{\mathbf{U}}_{\mathscr{N}} = (\mathbf{U}_{D+1} \mid \ldots \mid \mathbf{U}_N) \in \mathbb{R}^{N\times(N-D)}$ is the matrix of the $(N-D)$ remaining eigenvectors, associated to the eigenvalues equal to $\sigma^2_{\mathscr{N}}$, and whose range is henceforth termed the "noise space" $\mathscr{N}$. Remind that $(\underline{\mathbf{U}}_{\mathscr{S}} \mid \underline{\mathbf{U}}_{\mathscr{N}})^H(\underline{\mathbf{U}}_{\mathscr{S}} \mid \underline{\mathbf{U}}_{\mathscr{N}}) = \underline{\mathbb{I}}_N$.

Consequently, under the aforementioned statistical hypotheses, as soon as the covariance matrix $\underline{\mathbf{C}}_X$ is exactly computed, (6) can enable the recovery of the number of sources—which is N minus the number of repetitions of $\sigma^2_{\mathscr{N}}$— and of their locations—for their associated steering vectors are orthogonal to $\underline{\mathbf{U}}_{\mathscr{N}}$.

C. Estimation of the covariance matrix

In practice, $\underline{\mathbf{C}}_X$ is not known, as only one time record of $\mathbf{x}(t) = (x_1(t),\ldots,x_N(t))^T$ is available. Moreover, the complex envelopes vector $\mathbf{X}(k)$ cannot be exactly determined. So these quantities need to be approximated. One common strategy consists in computing such approximations on time snapshots. If the localization has to be computed at time indexes $t = T_0, 2T_0\ldots$, then one can proceed in two steps. On the one hand, $\mathbf{x}(t)$ is sampled—provided the Shannon theorem holds—at a rate equal to $\frac{L}{T_0}$, $L\in\mathbb{N}$, so that at each time t, $\mathbf{X}(k)$ is approximated by $\hat{\mathbf{X}}_t(k)$ from a L-point Discrete Fourier Transform (DFT) onto the snapshot $\{t(1-\frac{L-1}{L}),\ldots,t(1-\frac{1}{L}),t\}$. On the other hand, $\underline{\mathbf{C}}_X$ is estimated at time t by replacing in its definition the expectation by a weighted sum over a sliding window of W samples with interspace of T_0 , e.g. by defining[1]

$$\hat{\underline{\mathbf{C}}}_X = \frac{1}{W}\sum_{l=\frac{t}{T_0}-(W-1)}^{\frac{t}{T_0}} \hat{\mathbf{X}}_{lT_0}(k)\hat{\mathbf{X}}^H_{lT_0}(k). \quad (7)$$

Though this estimation process looks trivial at first glance, it can condition the performance of the whole method. For instance, the expectation in $\underline{\mathbf{C}}_X$ is all the better mimicked as the estimated $\hat{\mathbf{X}}_l(k)$ over the W-length window are independent, which precludes the use of overlapping snapshots [6].

D. The MUSIC algorithm

As aforementioned, the steering vector $\mathbf{V}(\mathbf{r}) = \mathbf{V}(r,\theta,k)$ gets orthogonal to the noise space $\mathscr{N}$ if and only if it is evaluated at the location of one source. Then, it follows that[1] $\forall(r,\theta)\in\{(r_1,\theta_1),\ldots,(r_D,\theta_D)\}$,

$$\sum_{i=D+1}^{N} |\mathbf{V}^H(r,\theta)U_i|^2 = \mathbf{V}^H(r,\theta)\underline{\Pi}_N\mathbf{V}(r,\theta) = 0, \quad (8)$$

where $\underline{\Pi}_N = \underline{\mathbf{U}}_{\mathscr{N}}\underline{\mathbf{U}}^H_{\mathscr{N}}$ is called the "projector in the noise space". As the genuine covariance matrix $\underline{\mathbf{C}}_X$ and its estimate $\hat{\underline{\mathbf{C}}}_X$ at time t differ, the genuine and approximated projectors $\underline{\Pi}_N$ and $\hat{\underline{\Pi}}_N$ do not perfectly match. So, the locations are established by isolating the maximum values of the pseudo-spectrum

$$h(r,\theta) = \frac{1}{\mathbf{V}^H(r,\theta)\hat{\underline{\Pi}}_N\mathbf{V}(r,\theta)}. \quad (9)$$

The whole MUSIC algorithm is depicted in Algorithm 1.

[1]The reference to t is not made explicit in order to alleviate the notation.

for *each localization time* t **do**
- turn the values of $\{\mathbf{x}(t(1-\frac{L-1}{L})),\ldots,\mathbf{x}(t(1-\frac{1}{L})),\mathbf{x}(t)\}$ into the approximation $\hat{\mathbf{X}}_t(k)$ of $\mathbf{X}(k)$ by a L-point DFT;
- estimate $\underline{\mathbf{C}}_X$ by $\underline{\hat{\mathbf{C}}}_X$ from the knowledge of $\hat{\mathbf{X}}_{t-(W-1)T_0}(k),\hat{\mathbf{X}}_{t-(W-2)T_0}(k),\ldots,\hat{\mathbf{X}}_t(k)$ through (7);
- detect the number of sources as N minus the number of eigenvalues of $\underline{\hat{\mathbf{C}}}_X$ approximately equal to $\sigma^2_{\mathscr{N}}$;
- from the spectral decomposition of $\underline{\hat{\mathbf{C}}}_X$, compute the projector $\underline{\hat{\Pi}}_N$ on the noise space;
- isolate the sources locations as the maxima of the pseudo-spectrum $h(r,\theta)$ defined in (9).

end

Algorithm 1: MUSIC narrowband algorithm

III. Towards a broadband extension of MUSIC

A. Generalities

In many practical cases, the sources cannot be considered as narrowband. This is so in the context of Human-Robot Interaction, where any filtering of human voice should not reject components from the frequency bandwidth $[300\text{Hz};3\text{kHz}]$ in order to ensure the intelligibility of the message. Extending the MUSIC algorithm to broadband generally follows a "frequency-based" approach, i.e. a dedicated processing is first applied to narrow frequency intervals—or "bins"— coming from a partition of the whole frequency range, prior to turning the obtained informations into a "composite" pseudo-spectrum. B such frequency bins are henceforth considered, each being centered on k_b, $b=1,\ldots,B$.

B. A naive extension

The straightest strategy follows as far as possible the lines of the narrowband algorithm. In other words, the Fourier transform $\mathbf{X}(k)$ is still approximated from the DFT of the perceived signals vector $\mathbf{x}(t)$ on a L-point snapshot. Likewise, the covariance matrices $\underline{\mathbf{C}}_X(k_1),\ldots,\underline{\mathbf{C}}_X(k_B)$ of $\mathbf{X}(k_1),\ldots,\mathbf{X}(k_B)$ are estimated by applying to each bin a scheme similar to (7). From the subsequent spectral decomposition of each $\underline{\hat{\mathbf{C}}}_X(k_b)$, separate pseudo-spectra $h_b(r,\theta)$ are determined, $b=1,\ldots,B$. The localization is then performed by looking for the maxima of the average pseudo-spectrum

$$h_{naive}(r,\theta)=\frac{1}{B}\sum_{b=1}^{B}h_b(r,\theta). \tag{10}$$

This naive extension of MUSIC to broadband is summarized in Algorithm 2.

for *each localization time* t **do**
- **for** *each frequency bin* k_b **do**
  - collect the approximate Fourier transforms $\hat{\mathbf{X}}_t(k_b)$ then compute the estimates $\underline{\hat{\mathbf{C}}}_X(k_b)$, $b=1,\ldots,B$, as was done in Algorithm 1;
  - from the spectral decomposition of $\underline{\hat{\mathbf{C}}}_X(k_b)$, compute the projector $\underline{\hat{\Pi}}_N(k_b)$ and the pseudo-spectrum $h_b(r,\theta)$ related to the b^{th} bin;
- **end**
- isolate the sources locations as the maxima of the pseudo-spectrum $h_{naive}(r,\theta)$ defined in (10).

end

Algorithm 2: MUSIC naive extension to broadband

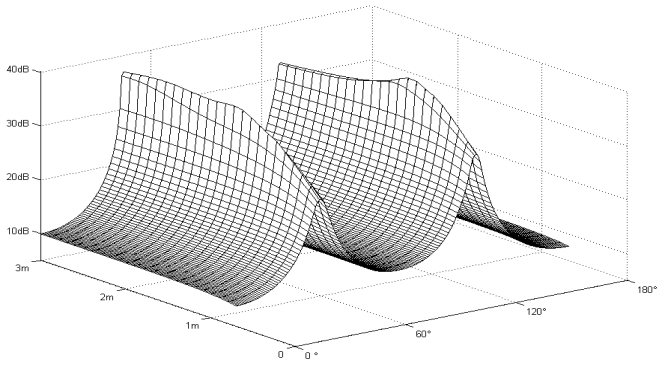

Fig. 1. Localization result obtained with the naive broadband extension

C. Localization results

In all the following, a $N=8$-microphone array with even interspace $d=\lambda_{3kHz}/2=5.66$cm is considered. $D=2$ pointwise sound sources emit independent voice signals from the positions $(2\text{m},45^\circ)$ and $(1.5\text{m},130^\circ)$. Spatially and temporally white noise is then added to each microphone perception, with a Signal to Noise Ratio (SNR) set to 10dB. As indicated in the previous subsection, a spatial covariance matrix $\underline{\mathbf{C}}_X(k_b)$ is defined for the microphone signals on each b^{th} frequency bin and for each localization time t, and is estimated through a time-average similar to (7). More precisely, the localization period being set to $T_0=1024/f_e$, with $f_e=15$kHz, the approximation $\hat{\mathbf{X}}_t(k_b)$ defined in subsection II-C is obtained from a 1024-point Fast Fourier Transform (FFT) of samples acquired at f_e, leading to 513 evenly spaced bins ranging brom 0 Hz to $f_e/2$. A number $W=16$ of FFT results, computed over successive disjoint 1024-point rectangular temporal windows, is then collected to estimate the covariance matrix $\underline{\mathbf{C}}_X$. In all the following, only the frequencies within $[300\text{Hz};3\text{kHz}]$ are used for localization, so that about $B=180$ frequency bins are considered in order to form the MUSIC pseudo-spectrum (10).

The naive broadband extension depicted in Algorithm 2 has been implemented under MATLAB comprehensible non-optimized code for performances comparison. Simulations produce 58 MUSIC pseudo-spectra $h_{naive}(r,\theta)$ all along the time horizon. One of these is reported on Figure 1:

- Two sharp peaks, with angular positions at the sources azimuths θ_1 and θ_2 whatever the scanned range r, can be exhibited. This nice property shows that the estimation of the sources bearings can be performed without any knowledge of their distances, even when the farfield hypothesis does not hold—typically for distances closer than 1.2m for the lowest frequency 300Hz. This is a significant improvement compared to beamforming techniques, which must be adapted to the source distance for correct azimuth estimation in the nearfield [2].
- Source ranges can be estimated by detecting the pseudo-spectrum maxima. Figure 1 shows in fact two wide lobes whose maxima are located close to the source simulated distances r_1 and r_2. More precisely on the 58 simulated pseudo-spectra, the estimation error on the two sources ranges does not exceed 30 cm.

The above interesting properties are similar to these shown in [1], and could lead to the conclusion that this wideband extension of the MUSIC algorithm is well suited to robotics applications. However, one has to keep in mind that Figure 1 has been obtained by averaging about 180 independently computed pseudo-spectra, each one corresponding to a bin in the bandwidth [300Hz; 3kHz]. Such independent decompositions require very heavy computational resources—say 16 1024-point FFTs, 180 eigendecompositions of 8×8 matrices being the most critical and complex part of the implementation [7], and 180×10 matrix products corresponding to each hypothesized range and bearing values—so that the real time implementation of Algorithm 2 on an embedded plateform would be prohibitive. Consequently, other wideband extensions of the MUSIC algorithm must be envisaged for real time practical implementation.

IV. Broadband MUSIC in beamspace

In light of the above section, a less costly strategy must be envisaged to extend MUSIC to the localization of wideband sources. This implies a reduction in the number of eigendecompositions to be performed at each localization time. One basic possibility might consist in averaging over the B bins the covariance matrices $\underline{\mathbf{C}}_X(k_b)$, $b = 1, \ldots, B$, prior to computing the eigendecomposition of the single resulting matrix. Unfortunately, such an approach is not sound, for the subspaces of $\mathbb{C}^N$ spanned by the noise eigenvectors matrices $\underline{\mathbf{U}}_{\mathscr{N}}(k_b)$ relative to each b^{th} bin differ, a property called "misalignment". The difficulty is then how to coherently combine the signal (resp. noise) spaces at each bin into a single signal (resp. noise) space endowed with algebraic properties which depend on the sources number and locations.

A. Fundamentals of alignment

In [8], Wang and Kaveh develop a mathematically principled solution to this problem in the planar waves case. Interestingly, under a mild constraint, it allows the sources to be correlated and to emit with a low SNR in an arbitrary noise field. To make the approach cope with nearfield sources, the starting point remains (1), the matrix $\underline{\mathbf{V}}(\mathbf{r}_1, \ldots, \mathbf{r}_D, k)$ being again defined from (3) and column full-rank. However, $\underline{\mathbf{C}}_S(k) = \mathbb{E}[\mathbf{S}(k)\mathbf{S}^H(k)]$ may be singular due to sources correlation, and the noise wavefield —though independent of the sources— is allowed to have the form $\mathbb{E}[\mathbf{B}(k)\mathbf{B}^H(k)] = \sigma^2_{\mathscr{N}}\underline{\mathbf{C}}(k)$, with $\underline{\mathbf{C}}(k)$ known. Note that such a problem statement can cope with real reverberant environments encountered in robotics applications, in that it suffices to consider the multipath propagation of a source as the propagation of several correlated—mirrored—sources.

First, a reference frequency k_0 is selected. Then, "focalization matrices" $\underline{\mathbf{T}}(r, k_b)$, $b = 1, \ldots, B$, are defined so that

$$\forall (r,\theta),\ \underline{\mathbf{T}}(r,k_b)\mathbf{V}(r,\theta,k_b) = \mathbf{V}(r,\theta,k_0), \tag{11}$$

i.e. so as to transform the array vector at any frequency bin k_b into its value at k_0. Computing the second order statistics of $\underline{\mathbf{T}}(r,k_b)\mathbf{X}(k_b)$ for each b^{th} bin, then summing the obtained matrices, one gets $\underline{\Gamma}_X(r) = \sum_{b=1}^{B} \underline{\mathbf{T}}(r,k_b)\underline{\mathbf{C}}_X(k_b)\underline{\mathbf{T}}^H(r,k_b)$, which also satisfies

$$\underline{\Gamma}_X(r) = \mathbf{V}(r,\theta,k_0)\underline{\Gamma}_S\mathbf{V}^H(r,\theta,k_0) + \sigma^2_{\mathscr{N}}\underline{\Gamma}_{\mathscr{N}}(r), \tag{12}$$

with $\underline{\Gamma}_S = \sum_{b=1}^{B} \underline{\mathbf{C}}_S(k_b)$ and $\underline{\Gamma}_{\mathscr{N}}(r) = \sum_{b=1}^{B} \underline{\mathbf{T}}(r,k_b)\underline{\mathbf{C}}(k_b)\underline{\mathbf{T}}^H(r,k_b)$. $\underline{\Gamma}_X(r)$ and $\underline{\Gamma}_{\mathscr{N}}(r)$ are respectively called the "focalized covariance matrices" of the array and noise signals.

The generalized eigenvalues μ_n and eigenvectors $\mathbf{U}_n$, $n = 1, \ldots, N$, of the matrix pencil $(\underline{\Gamma}_X(r), \underline{\Gamma}_{\mathscr{N}}(r))$, reordered so that $\mu_n \geq \mu_{n+1}$, then satisfy, with $\underline{\mathbf{U}}_{\mathscr{S}}(r) = (\mathbf{u}_1 \mid \ldots \mid \mathbf{u}_D)$ and $\underline{\mathbf{U}}_{\mathscr{N}}(r) = (\mathbf{u}_{D+1} \mid \ldots \mid \mathbf{u}_N)$,

$$\mu_{D+1} = \cdots = \mu_N = \sigma^2_{\mathscr{N}} \quad \text{and} \quad \mathbf{V}^H(r,\theta,k_0)\underline{\mathbf{U}}_{\mathscr{N}}(r) = 0. \tag{13}$$

As before, a pseudo-spectrum in the array signals space

$$h_{\text{array}}(r,\theta) = \frac{1}{\mathbf{V}^H(r,\theta,k_0)\hat{\underline{\mathbf{U}}}_{\mathscr{N}}(r)\hat{\underline{\mathbf{U}}}^H_{\mathscr{N}}(r)\mathbf{V}(r,\theta,k_0)} \tag{14}$$

can be defined from the estimates $\hat{\underline{\Gamma}}_X(r)$ and $\hat{\underline{\Gamma}}_{\mathscr{N}}(r)$ of the focalized covariance matrices, and is maximum at the sources locations. Importantly, these focalized covariance matrices estimates not only require the computation of $\hat{\underline{\mathbf{C}}}_X(k_b)$, $b = 1, \ldots, B$, over time snaphots, but also the approximation of the genuine focalization matrices $\underline{\mathbf{T}}(r,k_b)$. Ref. [8] suggests to use some prior knowledge about the locations to be estimated. Ref. [9] proposes an interesting alternative based on modal analysis, which does not necessitate any prior information. Rather than going into their details, another less costly approach developed in [10] is presented hereafter, relying on modal analysis and beamforming.

B. A broadband beamspace algorithm

1) Modal representation of beampattern: The response —or beampattern— $D_q(r,\theta,k_b)$ of any beamformer indexed by q to a point source at polar coordinates (r,θ) and frequency k_b is given by [6]

$$D_q(r,\theta,k_b) = \mathbf{W}_q^H(k_b)\mathbf{V}(r,\theta,k_b), \tag{15}$$

where $\mathbf{W}_q(k_b) = (w_{q,1}(k_b), \ldots, w_{q,N}(k_b))^T$ is the array weights vector at frequency k_b. One can show that (15) can be turned into

$$D_q(r,\theta,k_b) = \sum_{m=0}^{\infty} \alpha_{q,m}(k_b)R_m(r,k_b)Y_m(\theta), \text{ where} \tag{16}$$

$$R_m(r,k_b) \triangleq re^{jkr}h_m^{(2)}(k_b r),\ Y_m(\theta) \triangleq \sqrt{\frac{2m+1}{4\pi}}P_m(\cos\theta),$$

$P_m(.)$ and $h_m^{(2)}(.)$ terming the Legendre and the spherical Hankel functions, respectively. The set $\{\alpha_{q,m}(k_b)\}$ is composed of the complex "modal coefficients". In fact, (16) defines an orthogonal transform pair analogous to the familiar Fourier series, so that $\{\alpha_{q,m}(k_b)\}$ fully characterize $D_q(r,\theta,k_b)$. More details can be found in [2].

2) A beamspace processor: In beamspace processing, Q beamformers —with $D \leq Q < N$— exploit the N microphones outputs to form the Q-dimensional vector

$$\mathbf{Z}(k_b) = \big(Z_0(k_b), \ldots, Z_{Q-1}(k_b)\big)^T = \underline{\mathbf{W}}^H(k_b)\mathbf{X}(k_b), \tag{17}$$

with $\underline{\mathbf{W}}(k_b) = (\mathbf{w}_{0(k_b)} \mid \ldots \mid \mathbf{w}_{Q-1(k_b)})$. Thus, $\underline{\mathbf{W}}(k_b)$ can be seen as an operator from a N-dimensional microphones elementspace to the Q-dimensional output beamspace. The key idea of the beamspace broadband MUSIC method [10]

is to perform such a transform by selecting beamformers which lead to a focusing property similar to (11). In fact, [10] champions the use of the set of beamformers

$$D_q(r,\theta,k_b) = Y_q(\theta), \tag{18}$$

which are therefore invariant w.r.t. frequency and range. This last property can be easily obtained by considering (16). Let $\underline{\mathbf{W}}^{\infty}(k_b)$ be the set of coefficients producing the farfield response $D_q(\infty,\theta,k_b) = Y_q(\theta)$. The set of coefficients $\underline{\mathbf{W}}(r,k_b)$ leading to the same responses at distance r, $D_q(r,\theta,k_b) = Y_q(\theta)$, are [10]

$$\underline{\mathbf{W}}(r,k_b) = \underline{\mathbf{W}}^{\infty}(k_b)\underline{\mathbf{R}}_b^H(r), \tag{19}$$

with $\underline{\mathbf{R}}_b^H(r) = \mathrm{diag}\big[j/(k_b R_0(r,k_b)),\ldots,j^Q/(k_b R_{Q-1}(r,k_b))\big]$. In other words, (19) simply indicates how the coefficients $\underline{\mathbf{W}}^{\infty}(k_b)$ achieving the farfield beampattern $Y_q(\theta)$ must be modified so as to steer the beamspace processor with the same response at distance r.

3) Broadband beamspace MUSIC algorithm: Suppose that the array weight matrix $\underline{\mathbf{W}}^{\infty}(k_b)$ related to Q farfield beamformers defined in (18) has been synthesized for each b^{th} frequency bin, $b = 1,\ldots,B$. Setting $\mathbf{Z}(k_b) = \underline{\mathbf{W}}^H(r,k_b)\mathbf{X}(k_b)$, the $Q\times Q$ "beamspace covariance matrix" $\underline{\mathbf{C}}_Z(r,k_b) = \mathbb{E}[\mathbf{Z}(k_b)\mathbf{Z}^H(k_b)]$ comes as

$$\underline{\mathbf{C}}_Z(r,k_b) = \underline{\mathbf{D}}\underline{\mathbf{C}}_S(k_b)\underline{\mathbf{D}}^H + \sigma^2_{\mathscr{N}}\underline{\mathbf{C}}_W(r,k_b), \tag{20}$$

with $\underline{\mathbf{D}} = \underline{\mathbf{D}}(\mathbf{r}_1,\ldots,\mathbf{r}_D,k_b) = \underline{\mathbf{W}}^H(r,k_b)\underline{\mathbf{V}}(\mathbf{r}_1,\ldots,\mathbf{r}_D,k_b)$, and $\underline{\mathbf{C}}_W(r,k_b) = \underline{\mathbf{W}}^H(r,k_b)\underline{\mathbf{C}}(k_b)\underline{\mathbf{W}}(r,k_b)$ the "beamspace noise covariance matrix". Through a procedure similar to section IV-A, and by defining the two matrices $\underline{\Gamma}_Z(r)$ and $\underline{\Gamma}_W(r)$ as

$$\underline{\Gamma}_Z(r) = \sum_{b=1}^{B}\underline{\mathbf{C}}_Z(r,k_b) \quad\text{and}\quad \underline{\Gamma}_W(r) = \sum_{b=1}^{B}\underline{\mathbf{C}}_W(r,k_b), \tag{21}$$

the generalized eigenvalue decomposition of the matrix pencil $(\underline{\Gamma}_Z(r),\underline{\Gamma}_W(r))$ leads to a pseudo-spectrum $h_{array}(r,\theta)$ in the vein of (14), from which the localization can be performed. This beamspace broadband extension of MUSIC is summarized in Algorithm 3; the key point is to notice that a total of B eigendecompositions of $N\times N$ complex matrices is traded for a single generalized eigendecomposition of a —lower dimension— $Q\times Q$ complex matrix.

for *each time localization time t* **do**
  for *each distance r* **do**
    for *each frequency bin k_b* **do**
      - collect the approximate Fourier transforms $\hat{\mathbf{X}}(k_b)$ then compute the estimates $\hat{\underline{\mathbf{C}}}_X(k_b)$, $b = 1\ldots,B$.;
      - compute the matrix $\underline{\mathbf{W}}(r,k_b)$ through (19) by using offline optimization results;
      - compute the noise covariance matrix $\underline{\mathbf{C}}_W(r,k_b) = \underline{\mathbf{W}}^H(r,k_b)\underline{\mathbf{C}}(k_b)\underline{\mathbf{W}}(r,k_b)$ and the estimate $\hat{\underline{\mathbf{C}}}_Z(r,k_b)$ of the beamformer's output covariance matrix $\underline{\mathbf{C}}_Z(r,k_b) = \underline{\mathbf{W}}^H(r,k_b)\underline{\mathbf{C}}_X(k_b)\underline{\mathbf{W}}(r,k_b)$;
    end
    - compute the true/approximated focalized noise and beamformer's output covariance matrices $\underline{\Gamma}_W(r)$ and $\hat{\underline{\Gamma}}_Z(r)$ along (21);
    - compute the generalized eigenvalue decomposition of the matrix pencil $(\hat{\underline{\Gamma}}_Z(r),\underline{\Gamma}_W(r))$;
    - from the signal space $\mathscr{S}$ and the noise space $\mathscr{N}$, define the two matrix $\hat{\underline{\mathbf{U}}}_{\mathscr{S}}$ and $\hat{\underline{\mathbf{U}}}_{\mathscr{N}}$, and compute the MUSIC pseudo-spectrum $h_{array}(r,\theta)$;
  end
end

Algorithm 3: Wideband beamspace MUSIC

C. Localization results

As outlined in the previous subsection, the localization algorithm is based on two successive steps. The first step consists in synthezising Q farfield beampatterns $D_q(\theta,k) = Y_q(\theta)$ for all frequencies k in the frequency bandwidth of interest. Such beampatterns are then used in a second step to obtain the MUSIC pseudo-spectrum $h_{array}(r,\theta)$. We recently proposed in [2] a new beampattern synthesis method based on convex optimization, which benefits from the modal form (16) to reduce the computational cost. Such a method is particularly indicated to obtain the reference beampattern (18) needed for the localization algorithm. This subsection is organized as follows. The synthesis method presented in [2] is first recalled together with some synthesis outcomes. Next, localization results are proposed.

1) Synthesis method: The aim is to determine the vector $\mathbf{W}_q^{\infty}(k_b) = (w_{q,1}(k_b),\ldots w_{q,N}(k_b))^T$ which enables to approximate the reference farfield beampattern $\widetilde{D}_q^{\infty}(\theta,k_b) = Y_q(\theta)$ described by its complex modal coefficients $\widetilde{\alpha}_{q,0}(k)\ldots\widetilde{\alpha}_{q,M-1}(k)$. This is dealt with through the following convex optimization problem

$$\begin{aligned}&\text{minimize } \varepsilon \text{ subject to}\\ &\quad ||\ \alpha_{q,m}(k_b) - \widetilde{\alpha}_{q,m}(k_b)\ || \le \varepsilon,\ \forall m\in\{0..M-1\},\end{aligned} \tag{22}$$

which minimizes the "distance" between the first M modal coefficients of the reference and actual beampatterns. These of the actual beampattern, denoted $\alpha_{q,m}(k_b)$, are shown to be a function of the array sensor positions z_n, $n = 1,\ldots,N$ and of the unknown vector weights $\mathbf{W}_q^{\infty}(k_b)$; they verify

$$\alpha_{q,m}(k_b) = \gamma_m(k_b)\ \mathbf{W}_q^{\infty}(k_b)^H\mathbf{J}_m(k_b), \tag{23}$$

where $\mathbf{J}_m(k) = (j_m(kz_1),\ldots,j_m(kz_N))^T$, with $j_m(x) \triangleq \sqrt{\frac{\pi}{2x}}J_{m+\frac{1}{2}}(x)$ the spherical Bessel function, and $\gamma_m(k) = -2ik\sqrt{\pi(2m+1)}$. Notice that in the case considered here, the reference beampatterns $\widetilde{D}_q(\theta,k)$, $q = 0,\ldots Q-1$, are the spherical harmonics $Y_q(\theta)$. Consequently, according to (16), their modal coefficients $\widetilde{\alpha}_{q,m}(k_b)$ at frequency k_b can immediately be written as

$$\widetilde{\alpha}_{q,m}(k_b) = \frac{1}{R_m(r,k_b)}\delta_{q,m}, \tag{24}$$

where $\delta_{q,m}$ is the Kronecker delta.

2) Synthesis results: The optimization problem (22)-(23)-(24) is solved by means of the solver SDPT3 coupled with YALMIP[1] under MATLAB for each frequency k_b and each spherical harmonics $Y_q(.)$, $q = 0,\ldots,Q-1$. Considering the small size of the array (about 40cm long) and the reduced microphone number ($N = 8$), we assert that the $Q = 4$ first spherical harmonics can be synthesized without any visible error. In the case considered here, about 180 frequency bins in the bandwith [300Hz; 3kHz] are considered, so that $180\times Q = 720$ farfield beamformers must be determined offline. Such syntheses take only about 10 minutes thanks to the low number $M = 14$ of constraints involved in (22) for each frequency, and lead to the results shown in Figure 2. The $Q = 4$ resulting beampatterns are clearly frequency

[1] http://www.math.nus.edu.sg/~mattohkc/sdpt3.html
http://control.ee.ethz.ch/~joloef/yalmip.php

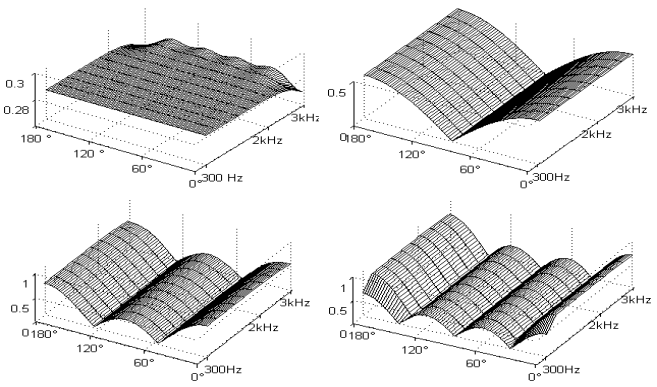

Fig. 2. Synthesis result

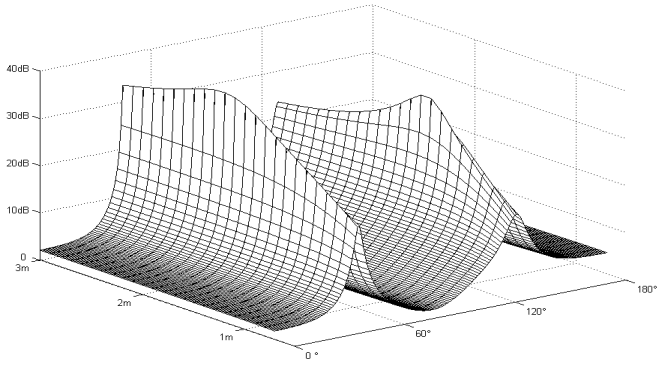

Fig. 3. Localization result

invariant on the bandwidth of interest. In fact, the synthesized spherical harmonic $Y_3(\theta)$ shows small gain variations for the lowest frequency, which have fortunately no effect on the forthcoming localization algorithm. In conclusion, the optimization process we have proposed in [2] clearly benefits from the modal representation of beampatterns and is consequently well suited to this broadband beamspace extension of the MUSIC method.

3) Localisation results: In all the following, all the parameters proposed in Section III-C remain unchanged, so that an estimation $\hat{\underline{\mathbf{C}}}_X(k_b)$ of the spatial covariance $\underline{\mathbf{C}}_X(k_b)$ is computed at each localization time t through (7). According to Algorithm 3, this estimation is then combined with the previous offline synthesis results to compute at each time t the two covariance matrices $\hat{\underline{\mathbf{C}}}_Z$ and $\underline{\mathbf{C}}_W$. Then, simulations produce 58 MUSIC pseudo-spectra $h_{array}(r,\theta)$ all along the time horizon. One of these, corresponding to the same time snapshot as in Figure 1, is reported on Figure 3. The two pseudo-spectra $h_{naive}(r,\theta)$ and $h_{array}(r,\theta)$ are very similar, so that the conclusions itemized in subsection III-C still hold. The search for the source locations can be performed by two one-dimensional searches: the first step consists in estimating the bearing in the farfield while the second step is achieved by scanning over r using the former bearing estimation. But though the two functions $h_{naive}(r,\theta)$ and $h_{array}(r,\theta)$ look identical, the ways they are obtained are fundamentally different. The first naive broadband extension is based on the meaningless average of multiple pseudo-spectra independently computed for each frequency k_b, while the broadband beamspace extension exploits the spectral alignment property to compute directly the final MUSIC spatial spectrum. Consequently, we assert that this broadband beamspace extension clearly outperforms the naive method in terms of computing cost and ease of implementation. For instance, the 58 pseudo-spectra corresponding to each localization time t are computed with Algorithm 2 in about 75 minutes under MATLAB, while Algorithm 3 produces quite the same results in only 5 minutes. Since simulations of the classical narrowband Algorithm 1 —already implemented in real time with DSP [7] or FPGA-based [11] system— take about 2 minutes, our first implementation tests let us think that Algorithm 3 will be used in real time.

V. Conclusion

A theoretical study and a performances comparison of two broadband methods has been presented for sound sources localization in robotics. The first one, called "naive extension", is based on multiple independent narrowband MUSIC computations and therefore requires important computational ressources. The second one benefits from frequency and range invariant beamformers to obtain a "focalization property" which enables the direct computation of a single pseudo-spectrum for all frequencies of interest. This beamspace extension appears to be well suited to robotics applications thanks to its lower computational requirements and because multipath propagation in reverberant environments can be handled. In view of the promising results obtained in simulation, we plan to work on the practical implementation and assessment of the method onto our experimental FPGA-based prototype.

References

[1] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot jijo-2," in *IEEE Int. Conference on Multisensor Fusion and Integration for Intelligent Systems*, 1999.

[2] S. Argentieri, P. Danès, and P. Souères, "Modal analysis based beamforming for nearfield or farfield speaker localization in robotics," in *IEEE Int. Conference on Intelligent Robots and Systems*, 2006.

[3] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *IEEE Int. Conf. on Acoustics, Speech, and Sign. Processing*, 2006.

[4] Y. Sazaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *IEEE International Workshop on Intelligent Robots and Systems*, 2006.

[5] K. Nakadai, H. Nakajima, M. Murase, H. G. Okuno, Y. Hasegawa, and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *IEEE Int. Conference on Intelligent Robots and Systems*, 2006.

[6] H. L. Van Trees, *Optimum Array Processing*, ser. Detection, Estimation, and Modulation Theory. John Wiley & Sons, Inc., 2002, vol. IV.

[7] M. Erić and B. Igrić, "Practical implementation and performance estimation of MUSIC method implemented on signal processor TMS 320c30," *Scientific Technical Review*, vol. 54, no. 1, 2004.

[8] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 823–831, 1985.

[9] T. Abhayapala and H. Bhatta, "Coherent broadband source localization by modal space processing," in *International Conference on Telecommunications*, vol. 2, March 2003, pp. 1617– 1623.

[10] D. B. Ward and T. D. Abhayapala, "Range and bearing estimation of wideband sources using an orthogonal beamspace processing structure," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, May 2004, pp. 109–112.

[11] M. Kim, K. Ichige, and H. Arai, "Implementation of FPGA based fast DOA estimator using unitary music algorithm," in *IEEE Vehicular Technology Conference*, vol. 1, 2003, pp. 213–217.