Introduction to microphone array processing and beamforming algorithms

Péter Rucz, Bence Csóka, Péter Fiala, Gergely Firtha

Spring 2025

1 Introduction

Array processing is a general term for techniques that rely on the processing of signals attained by the time-synchronous recording of multiple sensors of the same kind. In case of acoustical beamforming by microphone arrays, the sensors are microphones, sampling the sound field in well-defined locations.

From a farther perspective, acoustical beamforming is a remote sensing method, which allows for the detection and separation of sound sources located far away from our sensor. Such techniques are beneficial in many applications, where the sources of sound cannot be measured in their near vicinity. Microphone array processing allows for implementing different engineering tasks, such as 1) source localization (estimation of the direction or position of sound sources) and tracking of moving sources, 2) source separation (i.e., decomposition of a complex, spatially extended source into components, such as differentiating between aerodynamic and structure-borne noise of the pass-by noise of a high-speed train), 3) acoustical focusing and noise suppression (enhancing useful signals and suppressing interference by creating a highly directive sensor).

Looking at a somewhat wider aspect, we can realize that there are many related sensing methods, which also involve array processing. Without aiming at completeness, these related techniques include MIMO radars, sonar systems, ultrasonic imaging methods, antenna array processing, among many others. While the targeted frequencies and ranges may be significantly different from that of acoustic microphone arrays, all these methods share the same theoretical background [1].

This document is devoted to summarizing the basic concepts and theory of microphone array processing. Starting with the simplest arrangement of uniform line arrays (ULA), the detection principles and the most important properties of microphone array sensors are introduced. The material is structured as follows. Section 2 discusses delay & sum (conventional) beamforming and introduces the concepts of array measurements in the time and frequency domains. One of the main properties, i.e., the directivity of the array is introduced in Section 3 along with strategies that allow for designing the directivity of a microphone array. The task of source localization is addressed in Section 4, where point spread functions are also defined. While sections 2 to 4 deal with one-dimensional line arrays, the concept is extended to two-dimensional (planar) arrays Section 5. Finally, some beamforming algorithms often applied in practice are discussed in Section 6.

2 Delay & sum beamforming

Let's consider a plane wave pressure impulse, which arrives to the microphone array located along the x axis at an angle of incidence of θ_0. This arrangement is depicted in Figure 1. If the sound pressure at the origin is $p_0(t)$, then, at an arbitrary location x, the sound pressure is found as the time-shifted impulse

$$p(x,t) = p_0\left(t + \frac{x}{c}\sin\theta_0\right), \tag{1}$$

with c denoting the speed of sound. Compared to the $x = 0$ position, in case of positive θ_0 angles and positive x, this means a positive time advancement (negative delay), while for negative x, a positive

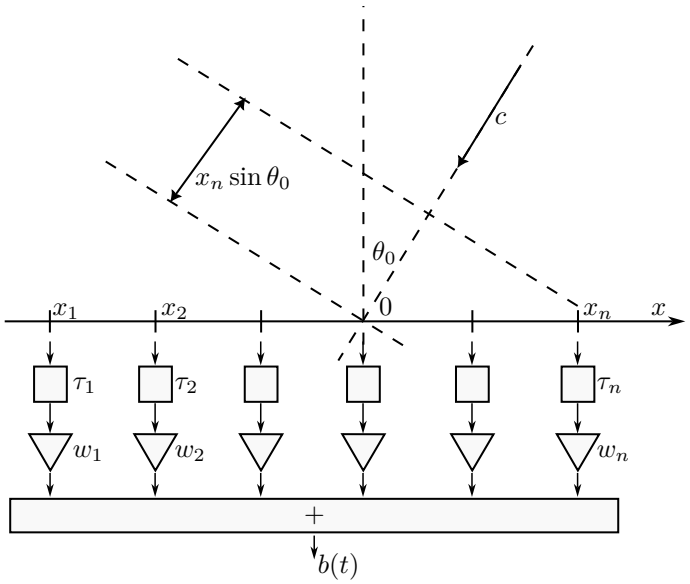

Figure 1: Delay and sum beamforming by a line array

delay is found. Naturally, for an incidence angle of $\theta_0 = 0$, the delay is zero for all x, while in case of $\theta_0 \pm \pi/2$, the delay is maximal.

The *core idea* of the processing *is to compensate the delays* resulting from sound propagation. The sound field is sampled by the microphones at discrete locations x_n $(n = 1, 2, \ldots, N)$. Hereafter, N denotes the number of elements of the microphone array. By the superimposition of the signals after the respective time delays τ_n, the summed signal $b(t)$ reads

$$b(t) = \frac{1}{N} \sum_{n=1}^{N} w_n p_n(t - \tau_n). \tag{2}$$

The factor w_n is the amplification for the nth microphone. The result of the sum $b(t)$ is the so-called *focused*, or *beamformed* signal. The above idea is called *delay & sum* beamforming.

Naturally, if $\tau_n = x_n \sin(\theta_0)/c$ and $w_n = 1$ are taken, propagation delays are perfectly compensated, and $b(t) = p_0(t)$ results; or, in other words, the array is focused to the θ_0 angle of incidence. Plausibly, focusing to an arbitrary θ direction is obtained by the choice

$$\tau_n = \frac{x_n}{c} \sin\theta \qquad \text{and} \qquad w_n = 1. \tag{3}$$

What happens if the direction of focus differs from the actual angle of incidence, i.e., $\theta \neq \theta_0$? In the ideal case, we expect perfect transmission for $\theta = \theta_0$, and aim for maximal suppression for $\theta \neq \theta_0$. The *fundamental task of beamforming* is achieving this behavior by the proper choice of the sampling locations x_n, amplifications w_n, and time delays τ_n.

For the sake of easier analysis, it is useful to transform the above results into the frequency domain. Here, we adopt the $\mathrm{e}^{\mathrm{j}\omega t}$ time dependence convention, i.e., we suppose that all signals are time-harmonic with the angular frequency ω, and we derive the time signals $f(t)$ from their frequency domain representation $\hat{f}$ as

$$f(t) = \Re\left\{\hat{f}\mathrm{e}^{\mathrm{j}\omega t}\right\} = F\cos(\omega t + \phi). \tag{4}$$

The complex peak value $\hat{f} = F\mathrm{e}^{\mathrm{j}\phi}$ contains both the magnitude F and phase information ϕ. Hereafter, the hat notation is used for denoting quantities in the frequency domain.

Then, the vector containing the complex pressure magnitudes of all microphones is written as

$$\hat{\mathbf{p}}(\omega) = \begin{pmatrix} \hat{p}_1(\omega) \\ \vdots \\ \hat{p}_n(\omega) \\ \vdots \\ \hat{p}_N(\omega) \end{pmatrix} = \hat{p}_0(\omega) \cdot \hat{\mathbf{a}}(\omega, \theta_0), \tag{5}$$

where the *propagation vector* $\hat{\mathbf{a}}(\omega, \theta_0)$ contains the phase shifts corresponding to the plane wave incoming from the θ_0 direction:

$$\hat{a}_n(\omega, \theta_0) = \mathrm{e}^{\mathrm{j}\omega x_n \sin(\theta_0)/c} = \mathrm{e}^{\mathrm{j}k x_n \sin\theta_0}. \tag{6}$$

Recall that the wave number k is found as $k = \omega/c = 2\pi/\lambda$, where c is the speed of sound and λ is the wavelength. The complex peak of the beamformed signal $b(t)$ is found as

$$\hat{b}(\omega) = \frac{1}{N}\hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{p}}. \tag{7}$$

The complex weight vector $\hat{w}_n = w_n \mathrm{e}^{\mathrm{j}\omega\tau_n}$ contains both amplifications and delays. Importantly, the superscript H denotes the transposed conjugate. Notice that due to the latter transposition, the vector operation of (7) is a scalar product that includes the summation, while the conjugation negates the phase. If we need to delay the signal of a microphone, then the phase of the corresponding $\hat{w}_n$ complex value must be positive.

2.1 Directivity

For a given $\hat{\mathbf{w}}$ weighting vector we can define the *directivity* (directional sensitivity) of the microphone array by means of (7) and (5) as:

$$\hat{D}(\omega, \theta_0) = \frac{\hat{b}(\omega)}{\hat{p}_0(\omega)} = \frac{1}{N}\hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{a}}(\omega, \theta_0). \tag{8}$$

The directivity $\hat{D}$ tells the amplification of the plane wave with θ_0 angle of incidence. Of course, the directivity depends on the weight vector $\hat{\mathbf{w}}$.

2.2 Conventional beamforming

Many algorithms exist for constructing the weight vector $\hat{\mathbf{w}}$. The simplest one is referred to as *conventional beamforming* (CBF), and is based on the idea:

$$\hat{\mathbf{w}}(\omega, \theta) = \hat{\mathbf{a}}(\omega, \theta), \tag{9}$$

that is, the weighting compensates the delays of a plane wave with θ angle of incidence (due to the complex conjugate appearing in the scalar product); or, in other words, the array is focused to the θ direction. Thus, the directivity of CBF is found as

$$\hat{D}_{\mathrm{CBF}}(\omega, \theta_0; \theta) = \frac{1}{N}\hat{\mathbf{a}}^{\mathrm{H}}(\omega, \theta)\hat{\mathbf{a}}(\omega, \theta_0). \tag{10}$$

Notice that conventional beamforming is frequency domain equivalent of the delay and sum approach introduced above.

2.3 Equidistant line arrays

Equidistant line arrays (also referred to as uniform line arrays or ULAs) constitute the simplest array arrangements with respect to analysis. Microphone locations are given by $x_n = n \cdot d$, where d is the distance between two neighboring microphones. For the sake of simplicity, we consider arrays with $N + 1$ elements, having a total diameter of Nd, and the index n runs from $-N/2$ to $N/2$. Using the above relations, the directivity of a ULA focusing to the θ direction reads

$$\hat{D}_{\mathrm{CBF}}(\omega, \theta; \theta_0) = \frac{1}{N+1}\sum_{n=-N/2}^{N/2} \mathrm{e}^{\mathrm{j}knd(\sin\theta_0 - \sin\theta)}. \tag{11}$$

The directivity is illustrated in Figure 2, with $N = 10$ and $d = 6\,\mathrm{cm}$ at different frequencies, with the speed of sound taken as $c = 340\,\mathrm{m/s}$. The microphone array is focused to the $\theta = 0$ direction in all

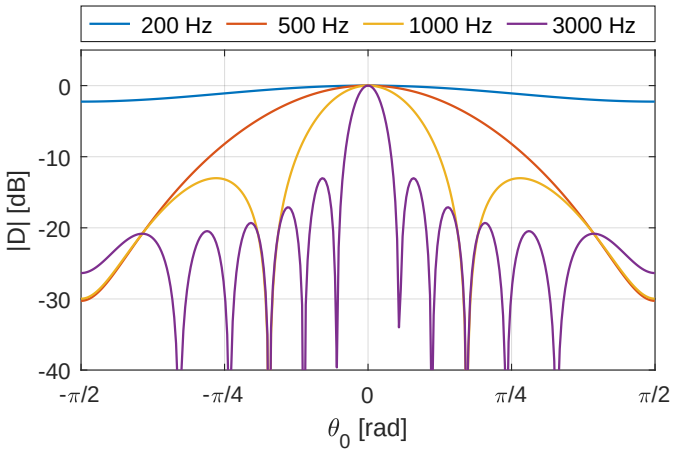

Figure 2: Directivity patterns of a ULA ($N = 10$, $d = 6\,\text{cm}$) at different frequencies, focused to the $\theta = 0$ direction.

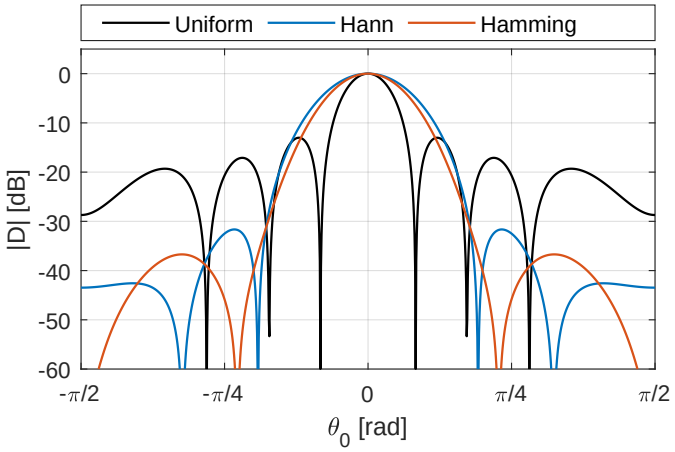

Figure 3: Directivity patterns of a ULA ($N = 10$, $d = 6\,\text{cm}$) at $f = 2\,\text{kHz}$ frequency, focused to the $\theta = 0$ direction, using different weighting functions.

cases (corresponding to the choice $\hat{w}_n \equiv 1$) and we see the amplification of plane waves arriving from different angles of incidence θ_0. At low frequencies, where the acoustic wavelength is much greater than the diameter of the array, the directivity is essentially uniform: the array has unit amplification in the main direction, but waves incoming from the sides are not suppressed. By increasing the frequency, the main lobe is compressed, its width decreases, and more and more sidelobes appear in the directivity pattern. The side lobe suppression (i.e., the difference of the magnitudes of the first side lobe and the main lobe) of the ULA with uniform weights is $-13\,\text{dB}$ independent of the frequency.

The suppression of the sidelobes can be increased by using nonuniform weighting. Weighting functions are also referred to as window functions. If nonuniform weighting is applied, the normalization of the directivity in (8) changes correspondingly as

$$\hat{D}(\omega, \theta_0) = \frac{\hat{b}(\omega)}{\hat{p}_0(\omega)} = \frac{1}{\sum w_n} \hat{\mathbf{w}}^{\mathrm{H}} \hat{\mathbf{a}}(\omega, \theta_0). \tag{12}$$

Two common weighting functions, i.e., Hann and Hamming are compared to uniform weighting in Figure 3. While the nonuniform weighting efficiently reduces the magnitude of the side lobes, the main lobe becomes wider.

Focusing into different θ directions, a.k.a. *electronic steering* is achieved by the proper choice of the $\hat{\mathbf{w}}$ weights that include the appropriate phase shifts. Figure 4 depicts the effect of steering the array. While for small $|\theta|$ angles the steered directivities resemble that at $\theta = 0$, at higher steering angles, the main lobe gets wider and becomes distorted.

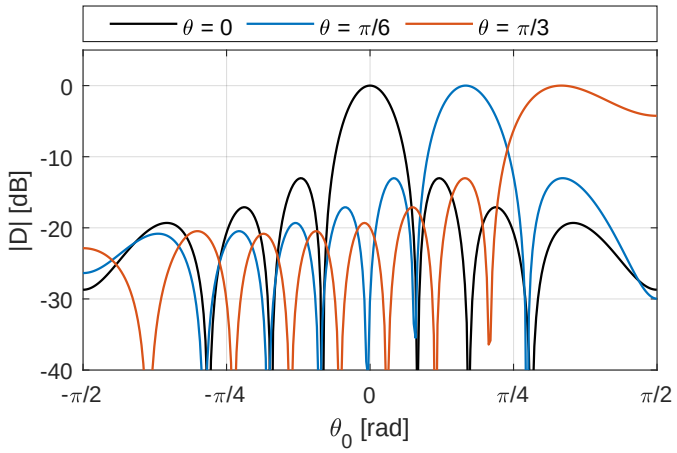

Figure 4: Directivity patterns of a ULA ($N = 10$, $d = 6\,\mathrm{cm}$) at $f = 2\,\mathrm{kHz}$ frequency, focused to the different directions.

3 Designing the directivity

3.1 Methods based on the Fourier transform

Let's write the definition of the directivity (8) with expanding the propagation vector $\hat{a}_n$:

$$\hat{D}(\omega, \theta_0) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \hat{w}_n^* \mathrm{e}^{\mathrm{j}x_n k \sin\theta_0}. \tag{13}$$

Then, taking a limit with a large number of microphones $N+1$ and a small distance d between them, such that $N \cdot d = L$ is the width of the array, we find:

$$\hat{D}(\omega, \theta_0) = \frac{1}{L} \sum_{n=-N/2}^{N/2} \hat{w}_n^* \mathrm{e}^{\mathrm{j}x_n k \sin\theta_0} d \quad \rightarrow \quad \frac{1}{L} \int_{-L/2}^{L/2} \hat{w}^*(x) \mathrm{e}^{\mathrm{j}xk\sin\theta_0} \mathrm{d}x. \tag{14}$$

We can observe that the directivity $\hat{D}$ is the Fourier transform of the weight function $\hat{w}(x)$. This Fourier transform does not convert from the time domain into the frequency domain, but from the x space domain into the $k_x = k \sin\theta_0$ wave number domain (i.e., spatial frequency domain).

In the special case of the $\hat{w}(x) = 1$ rectangular window function, the directivity results in:

$$\hat{D}(\omega, \theta_0) = \frac{1}{L} \int_{-L/2}^{L/2} \mathrm{e}^{\mathrm{j}xk\sin\theta_0} \mathrm{d}x = \mathrm{sinc}\left(k \sin\theta_0 L/2\right), \tag{15}$$

where $\mathrm{sinc}(x) = \sin(x)/x$. As expected, and in accordance with the diagrams above, uniform weighting results in a sinc directivity. The second local maximum of the sinc function is found at the argument value of $3\pi/2$, and hence the side lobe suppression is $1/(3\pi/2) \approx -13\,\mathrm{dB}$. Zeros are found at arguments of $n\pi$ ($n = \pm 1, \pm 2, \dots$), therefore the zeros of the directivity are found at

$$\sin\theta_0 = n\frac{2\pi}{kL} = n\frac{\lambda}{L}, \tag{16}$$

with $\lambda = c/f = 2\pi/k$ denoting the wavelength.

The width of the main lobe is defined by the angles corresponding to $-6\,\mathrm{dB}$ from the main lobe peak. As $\mathrm{sinc}(0.6\pi) \approx 1/2 \approx -6\,\mathrm{dB}$, the main lobe width Θ reads

$$\Theta = 2 \arcsin\left(0.6\frac{\lambda}{L}\right). \tag{17}$$

This result shows that the wider the microphone array, and the higher the frequency, the narrower the main lobe, and hence, the line array is more focused. However, a side lobe suppression better than $13\,\mathrm{dB}$ cannot be achieved by uniform weighting. As discussed above, nonuniform weighting must be used at a cost of slightly increasing the main lobe width.

3.2 The effect of discretization

When discretizing the $\hat{w}(x)$ continuous weight function by a finite number of microphones, the sampling theorem must be respected: the inverse of the sampling distance d must be greater than twice the maximal spatial frequency, i.e.:

$$\frac{2\pi}{d} \geq 2\left(k \sin\theta_0\right)_{\max}. \tag{18}$$

Taking into account that $-1 \leq \sin\theta_0 \leq 1$, the upper frequency limit (a.k.a. cut-off frequency) of the microphone array is

$$k \leq \frac{\pi}{d}, \quad \omega \leq \frac{\pi c}{d}, \quad d \leq \frac{\lambda}{2}. \tag{19}$$

Thus, using a microphone array with $d = 6\,\mathrm{cm}$, the upper frequency limit is $\approx 2860\,\mathrm{Hz}$. Above this frequency limit, a periodically repeated "copy" of the main lobe of the directivity can be overlapped into the $[-\pi/2, \pi/2]$ interval. The end result depends on both the frequency and the direction of focus.

3.3 Adaptive microphone array

So far, we considered methods by which the directivity (the main lobe direction and width, side lobe suppression, zero locations) are determined independent of the incoming sound pressure signals. In contrast, adaptive methods shape the weight vector and hence the directivity based on the received sound signals. A typical measurement task is to have unit directivity at a known angle θ_0 (assuming that the useful signal is arriving from the θ_0 direction), and suppressing signals (interference or noise) coming from other directions. If the unwanted signals arrive from characteristic directions, a proper choice is setting the zeros of the directivity to these angles, i.e., angles from where most of the power of disturbing sources comes from.

We can express this idea formally as well. Let the energy of the beamformed signal $\hat{b}$ minimal, while the directivity in the θ_0 angle is unit:

$$\mathrm{E}\left(|\hat{b}|^2\right) \to \min, \qquad \hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{a}}(\theta_0) = 1. \tag{20}$$

In the condition (20), E denotes the expectation (expected value) with respect to time, i.e., $\mathrm{E}\left(|\hat{b}|^2\right)$ is the mean of the square of the magnitude of the complex peaks taken in subsequent time slices, which is proportional to the power of the beamformed signal. The minimization can be expressed using (7) if the incident sound pressure and the weight vector are known as:

$$\mathrm{E}\left(|\hat{b}|^2\right) = \mathrm{E}\left(\hat{b}\hat{b}^*\right) = \mathrm{E}\left(\hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{p}}\left(\hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{p}}\right)^*\right) = \mathrm{E}\left(\hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{H}}\hat{\mathbf{w}}\right) = \hat{\mathbf{w}}^{\mathrm{H}}\mathrm{E}\left(\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{H}}\right)\hat{\mathbf{w}} = \hat{\mathbf{w}}^{\mathrm{H}}\mathbf{C}\hat{\mathbf{w}} \to \min, \tag{21}$$

where we exploited the identity $(\hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{p}})^* = \hat{\mathbf{p}}^{\mathrm{H}}\hat{\mathbf{w}}$. The newly introduced matrix $\mathbf{C}$ is the *covariance matrix* of the sound pressure signals received by the microphone array.

In the following, the covariance matrix $\mathbf{C}$ will play a very important role. Element (i, j) of matrix $\mathbf{C}$ is defined as $\hat{C}_{ij} = \mathrm{E}\left(\hat{p}_i\hat{p}_j^*\right)$. By this definition, we see that the covariance matrix is a self-adjoint matrix: $\mathbf{C} = \mathbf{C}^{\mathrm{H}}$. The elements in the main diagonal are real valued, and are proportional to the power received by each microphone: $\hat{C}_{ii} = \mathrm{E}\left(|p_i|^2\right)$. Outside the main diagonal are the cross-powers of the microphone pairs. These are complex values and as a result of the complex conjugation they also contain the phase delay between microphone i and j.

The optimization task can be formulated as

$$\hat{\mathbf{w}}^{\mathrm{H}}\mathbf{C}\hat{\mathbf{w}} \to \min, \qquad \hat{\mathbf{w}}^{\mathrm{H}}\hat{\mathbf{a}}(\theta_0) = 1. \tag{22}$$

The optimum solution $\hat{w}_{\mathrm{opt}}$ is found by the Lagrange multiplier technique which gives the result

$$\hat{\mathbf{w}}_{\mathrm{opt}} = \frac{\mathbf{C}^{-1}\hat{\mathbf{a}}_0}{\hat{\mathbf{a}}_0^{\mathrm{H}}\mathbf{C}^{-1}\hat{\mathbf{a}}_0}, \tag{23}$$

where $\hat{\mathbf{a}}_0 = \hat{\mathbf{a}}(\theta_0)$ is the propagation vector corresponding to the predefined θ_0 angle. The adaptive microphone array must continuously measure (estimate) the covariance matrix $\mathbf{C}$ of the incident

pressure field, and in order to find the optimal weights, it must apply formula (23). This means that a system of linear equations, with the number of unknowns being equal to the number of microphones, must be solved continuously (practically, time window by time window).

3.4 Estimation of the covariance matrix

We can estimate the covariance matrix by simple time averaging

$$\mathbf{C} = \mathrm{E}\left(\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{H}}\right) \approx \frac{1}{K}\sum_{k=0}^{K-1}\hat{\mathbf{p}}_k\hat{\mathbf{p}}_k^{\mathrm{H}}, \tag{24}$$

where the index k refers to the kth time window before the current one. In the typical implementation, the time-varying pressure signals are transformed into the frequency domain, with taking time windows of them and applying the fast Fourier transform (FFT). Then, a single frequency bin at the detection frequency ω is taken, and by a dyadic product, the spectral covariance matrix of one time window is found. Notice that as this matrix contains a single dyad, its rank is only 1. Finally, the dyads are averaged over K time windows to get an estimate of the covariance matrix. For $K \geq N$ a full-rank covariance matrix $\mathbf{C}$ can be attained. The time windows can also overlap in this approach.

Alternatively, instead of the moving average represented by (24), exponential averaging with a time constant of KT_{win} can also be applied, with T_{win} denoting the length of one time window. Exponential averaging is easy to implement by a simple recursive formula. It is not necessary to apply FFT for evaluating the complex peak at the given frequency ω. By means of a modulation by a signal of the form $\mathrm{e}^{-\mathrm{j}\omega t}$ and then applying a narrow low-pass filter gives the time-varying complex peaks, on which time-averaging can be applied.

4 Source localization

Source localization is an important task regarding the application of microphone arrays. We exploit the electronic steering of the main lobe of the array: the domain of possible angles of incidence θ is "scanned", and the energy of the beamformed signal b is evaluated. Notice that "scanning" is not implemented step-by-step, yet, all directions are evaluated in parallel. To the directions where significant energy is found, sources are assigned.

Assume that M sources are located in the space, in different directions θ_i $(i = 1, 2, \ldots, M)$. All sources emit sound in the narrow band centered to ω, and these narrow band signals are represented by the slowly varying complex peaks $\hat{q}_i(t)$. Thus, information is contained in the slow variation of the magnitude and phase of $\hat{q}_i$, while the frequency ω can be regarded as the "carrier" of the signals. We suppose that the sources are uncorrelated, that is $\mathrm{E}\left\{\hat{q}_i(t)\hat{q}_j^*(t)\right\} = 0$, if $i \neq j$. The microphone array receives the superimposed signals, and thus the signal vector becomes

$$\hat{\mathbf{p}}(t) = \sum_{i=0}^{M}\hat{q}_i(t)\hat{\mathbf{a}}(\theta_i). \tag{25}$$

We apply the conventional beamforming, focusing into the θ direction, and evaluate the beamformed signal:

$$\hat{b}(t|\theta) = \mathbf{w}^{\mathrm{H}}\hat{\mathbf{p}}(t) = \hat{\mathbf{a}}^{\mathrm{H}}(\theta)\hat{\mathbf{p}}(t) = \sum_{i=1}^{M}\hat{q}_i(t)\hat{\mathbf{a}}^{\mathrm{H}}(\theta)\hat{\mathbf{a}}(\theta_i). \tag{26}$$

Then, taking the power of the signal:

$$\mathrm{E}\left(|\hat{b}(t|\theta)|^2\right) = \sum_{i=1}^{M}\sum_{j=1}^{M}\mathrm{E}\left(\hat{q}_i\hat{q}_j^*\right)\hat{\mathbf{a}}^{\mathrm{H}}(\theta)\hat{\mathbf{a}}(\theta_i)\hat{\mathbf{a}}^{\mathrm{H}}(\theta_j)\hat{\mathbf{a}}(\theta). \tag{27}$$

By exploiting that the sources are uncorrelated, the double sum boils down to a simple sum:

$$\mathrm{E}\left(|\hat{b}(t|\theta)|^2\right) = \sum_{i=1}^{M}\sigma_i^2|\hat{\mathbf{a}}^{\mathrm{H}}(\theta)\hat{\mathbf{a}}(\theta_i)|^2 = \sum_{i=1}^{M}\sigma_i^2|D_{\mathrm{CBF}}(\theta,\theta_i)|^2. \tag{28}$$

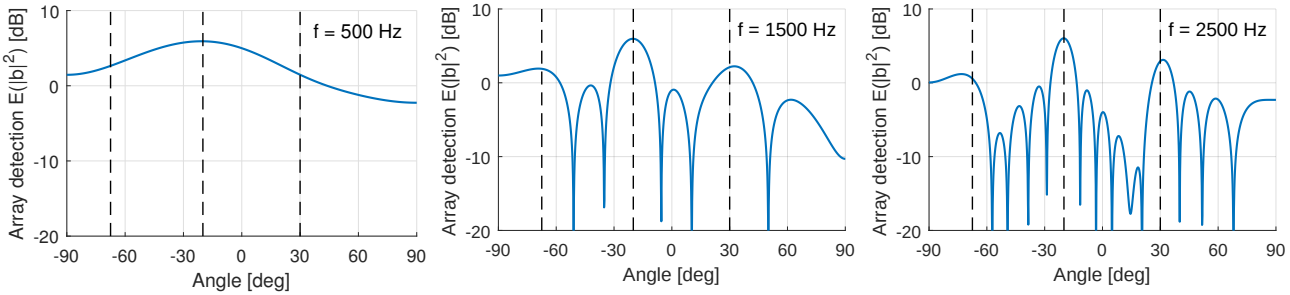

Figure 5: Detection of three uncorrelated sources ($-67.5°$, $-20°$, and $+30°$) by a ULA ($N = 16$, $d = 6\,\mathrm{cm}$) at different frequencies. The dashed lines represent the true directions of the sources.

The resulting formula expresses that by focusing into the θ direction, the power of the beamformed signal is attained as a superposition of the powers σ_i^2 of the uncorrelated sound sources, where the weights in the linear combination are the squared directivities. The latter quantity is often referred to as the *point spread function* (PSF) of the microphone array:

$$\mathrm{PSF}(\theta, \theta_0) = |D(\theta, \theta_0)|^2. \tag{29}$$

The PSF is interpreted as the spatial impulse response of the microphone array. By scanning the whole θ domain, the power of the beamformed signal attained by conventional beamforming results as the convolution of the spatially distributed source power and the point spread function. This naturally means, that at low frequencies, where the point spread function is almost a constant, the source powers are spread out by the convolution and localization becomes impossible. However, at high frequencies, where the main lobe of the PSF is narrow, the local peaks of the beamformed power correspond to the real source locations. When evaluating the source power, we still have to take into account that the power in the beamformed signal is a superposition that contains power from all sources. Evaluating the exact source powers is a deconvolution task.

The phenomenon is illustrated in the diagrams of Figure 5, where the detection of three uncorrelated sources (powers: 1, 4, and 2 units, respectively) is exemplified at different frequencies. At low frequencies the main lobe of the directivity is too wide which prevents the separation of the sources. With increasing the frequency, the sources can be distinguished, but the side lobes can mask sources of smaller powers. The best results are found near but below the cut-off frequency of the array, where we are close to the frequency limit imposed by the sampling theorem (19). Above this limit, the overlapping images of the main lobe will impede the localization.

5 Microphone arrays

For the sake of simplicity, we considered microphone line arrays, restricted to one dimension (microphones aligned on the x axis and the angle represented by θ). Both focusing and source localization are easily extended to planar microphone arrays for three-dimensional problems, represented by an azimuth and elevation angle. Conventional beamforming, attaining the focused signal, the algorithm of adaptive beamforming introduced above, point spread functions, and the theory of source localization depend solely on the propagation vector $\hat{\mathbf{a}}$ which can be generalized in a straightforward way.

If the position of the microphone is in the $z = 0$ plane, given as (x_n, y_n), and the incident plane wave is represented by the usual spherical angles θ and ϕ, the time delay compared to the reference point in the origin reads

$$\Delta t = \frac{1}{c}\left(x_n \cos\phi \sin\theta + y_n \sin\phi \sin\theta\right), \tag{30}$$

from which the elements of the propagation vector are found as

$$\hat{a}_n = \mathrm{e}^{\mathrm{j}k(x_n \cos\phi \sin\theta + y_n \sin\phi \sin\theta)} = \mathrm{e}^{\mathrm{j}k_x x_n + \mathrm{j}k_y y_n}, \tag{31}$$

where the wave numbers $k_x = k \sin\theta \cos\phi$ and $k_y = k \sin\theta \sin\phi$ were introduced.

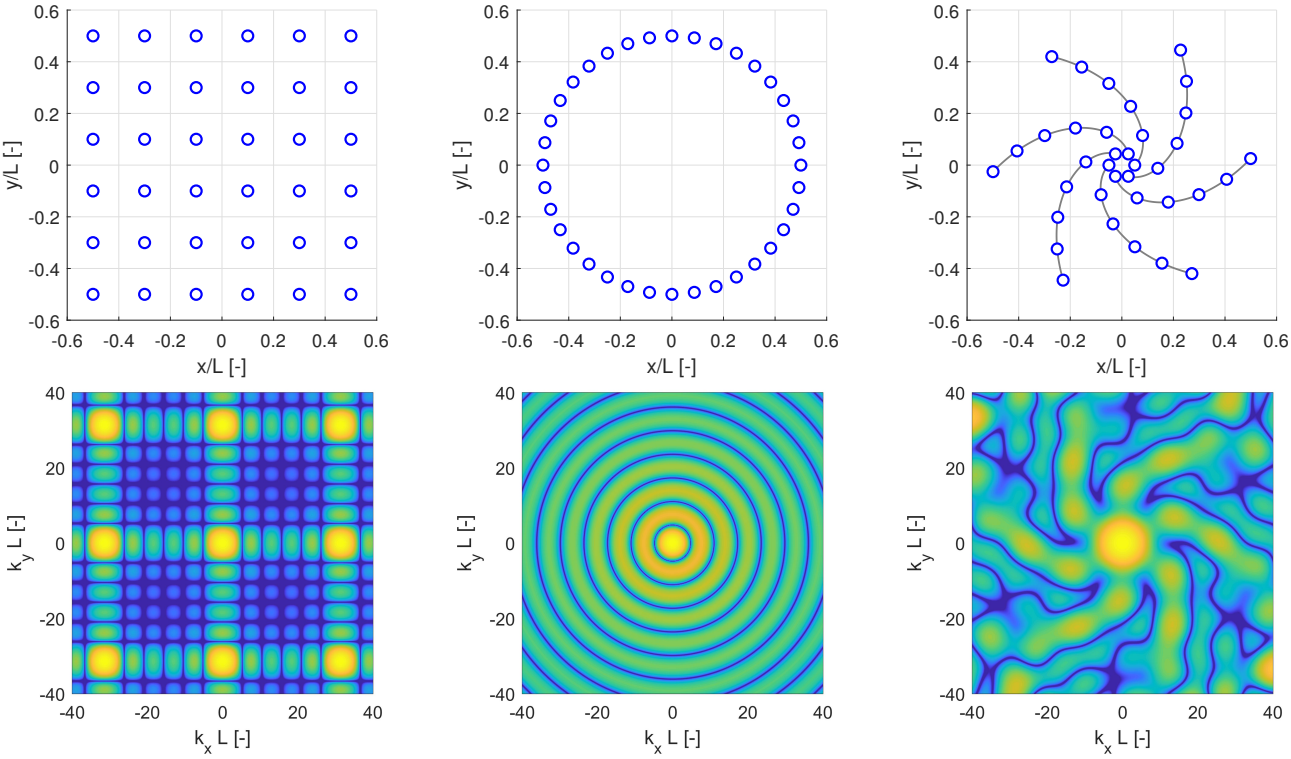

Figure 6: Microphone array arrangements and the corresponding PSFs.

Figure 6 depicts a few typical microphone arrangements (rectangular grid, circle, and multi-arm spiral) and the corresponding PSFs, for a source located at zero azimuth and elevation angle. Notice that all arrays in the figure are built up from 36 microphones. As it can be observed, the resulting PSF reflects the shape of the array. The PSFs are displayed in the (k_x, k_y) wave number domain, and the wave numbers are normalized by the diameters of the arrays. Therefore, these PSFs are independent of the frequency. If the diameter of the array is known, the actual frequency and direction dependent PSFs are found from the diagrams by using the definition (31). Thus, at low frequencies we find the point spread functions by taking disks of radius $k = \omega/c$ of the wavenumber spectra. By increasing the detection frequency, the radius of the disk increases with more and more side lobes appearing. Finally, mirror images of the main lobe also appear above the cut-off frequency of the array.

6 Beamforming and deconvolution algorithms

In the previous section we saw that the point spread function of the microphone array appears in the detection: even if a point source is considered, the estimated power resulting from conventional beamforming is blurred by the PSF. Advanced beamforming and deconvolution algorithms aim at estimating the source power distribution with removing (or minimizing) the effect of the convolution by the PSF.

For introducing the algorithms discussed in this section using a unified nomenclature and notations, let's consider the arrangement depicted in Figure 7. As in the previous section, the microphone array, sometimes referred to es *acoustic antenna* is assumed to be located in the x–y plane. We don't estimate the source sound power distribution as a continuous function of the azimuth and elevation angles $\boldsymbol{\theta}$, yet we *assign source powers to a discrete set of points* in the space. Most often, we create a grid (either planar or spherical), and assign an estimated source power to each position (node) of the grid. As the assigned power distribution is usually visualized by colors over the grid (*acoustical image*), the grid is often referred to as the *acoustical canvas*, while the grid points are also known as *virtual source points*. The number of grid points in the canvas is denoted by N_c hereafter. For the localization and identification of noise sources, it is often useful to project the acoustical image onto an optical image of the same space, allowing for a visual localization of the sources as well. That's why microphone

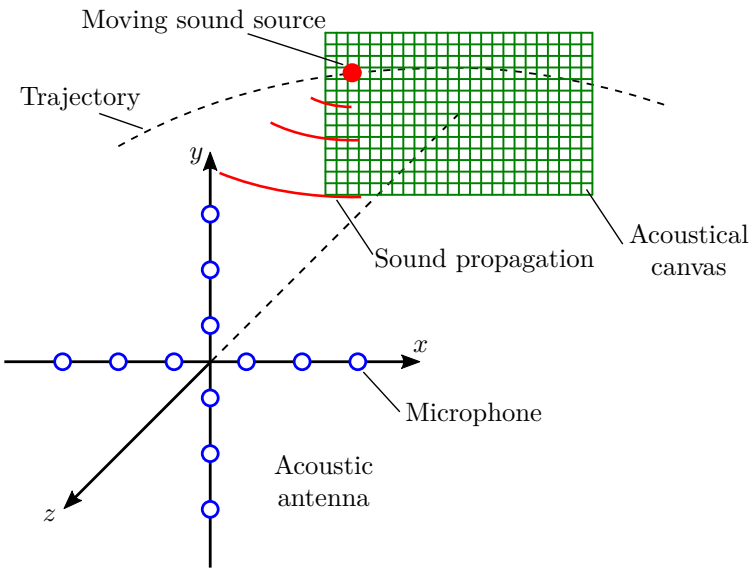

Figure 7: General arrangement for the detection of a moving sound source by a microphone array.

arrays are also called acoustic cameras.

The task is to assign a source power $P(\boldsymbol{\theta})$ to each grid point, based on the signals received by the microphone array. For convenience, the source power over the canvas will be denoted by the vector $\mathbf{P}$, where $P_j = P(\boldsymbol{\theta}_j)$, with $\boldsymbol{\theta}_j$ $(j = 1, 2, \ldots, N_\mathrm{c})$ representing the direction of the i-th point of the grid. Algorithms discussed in the following paragraphs make use of the propagation vectors associated with each point of the canvas, denoted by $\hat{\mathbf{a}}_j = \hat{\mathbf{a}}(\boldsymbol{\theta}_j)$ as above. The acoustical image attained by the conventional beamforming will be represented by the vector $\mathbf{S}$, with

$$S_j = E\left\{\left|\hat{b}(t|\boldsymbol{\theta}_j)\right|^2\right\} = \sum_{i=1}^{M} \sigma_i^2 \mathrm{PSF}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i) \qquad j = 1, 2, \ldots, N_\mathrm{c} \tag{32}$$

resulting from M uncorrelated sources with powers σ_i^2 $(i = 1, 2, \ldots, M)$ in the directions $\boldsymbol{\theta}_i$.

6.1 MUSIC algorithm

The MUSIC (Multiple Signal Classification) algorithm [2] uses linear algebraic tools for the estimation of the sound source distribution. The key idea of the algorithm is the eigenvalue-eigenvector analysis of the covariance matrix $\mathbf{C}$. Assume that the microphone array receives the sound signal from M uncorrelated sources, located at the directions $\boldsymbol{\theta}_i = (\theta_i, \phi_i)$, respectively. As above, the resulting signal of the array $\hat{\mathbf{p}}$ is written in the frequency domain as

$$\hat{\mathbf{p}}(t) = \sum_{i=1}^{M} \hat{q}_i(t)\hat{\mathbf{a}}(\boldsymbol{\theta}_i). \tag{33}$$

This means that the signal of the microphone array with N microphones spans an M-dimensional subspace of the total N dimensions, with the base vectors determined by the propagation vectors $\hat{\mathbf{a}}(\boldsymbol{\theta}_i)$. The covariance matrix, assuming that there is no additional noise, reads

$$\mathbf{C} = \mathrm{E}\left\{\hat{\mathbf{p}}\hat{\mathbf{p}}^\mathrm{H}\right\} = \sum_{i=1}^{M} \sigma_i^2 \left|\hat{\mathbf{a}}(\boldsymbol{\theta}_i)\hat{\mathbf{a}}^\mathrm{H}(\boldsymbol{\theta}_i)\right|^2, \tag{34}$$

which is the sum of M dyads. This also means that the maximum rank of the covariance matrix $\mathbf{C}$ is M in the noiseless case.

As the covariance matrix is self-adjoint, its eigenvalues are real and the corresponding eigenvectors are orthogonal. Since the rank of the matrix is maximum M, $N - M$ from the total of N eigenvalues must be zero. Denote the eigenvectors corresponding to the smallest $N - M$ eigenvalues by $\boldsymbol{\varphi}_n$,

$n = 1, 2, \ldots, N - M$. These eigenvectors span the $N - M$-dimensional subspace, i.e., a subspace orthogonal to all the propagation vectors $\hat{\mathbf{a}}(\boldsymbol{\theta}_i)$, which is called the noise subspace of the covariance matrix. These eigenvectors are arranged into an $N \times (N - M)$ type matrix as:

$$\boldsymbol{\Phi} = \left[\boldsymbol{\varphi}_1 | \boldsymbol{\varphi}_2 | \ldots | \boldsymbol{\varphi}_{N-M}\right]. \tag{35}$$

Then, source localization is performed by assigning a source strength P_j to all virtual source locations on the canvas $\boldsymbol{\theta}_j$ by evaluating the projection of the propagation vector $\hat{\mathbf{a}}_j = \hat{\mathbf{a}}(\boldsymbol{\theta}_j)$ onto the noise space. The source strength is found as the inverse of the squared norm of the projected vector:

$$P_j = \frac{\|\hat{\mathbf{a}}_j\|^2}{\|\hat{\mathbf{a}}_j^{\mathrm{T}} \boldsymbol{\Phi}\|^2} \qquad j = 1, 2, \ldots, N_{\mathrm{c}}. \tag{36}$$

Notice that this assigned source power is *not* the real power of the source at the direction $\boldsymbol{\theta}_j$. As the propagation vector is projected onto the noise space (in the noiseless case, the null-space) of the covariance matrix, if $\boldsymbol{\theta}_j$ matches one of the true source directions, the assigned power P_j can theoretically be infinite.

It has to be noted that applying the MUSIC algorithm necessitates an *a priori* knowledge of the number of sources to be detected. However, the number of sources can also be estimated based on the analysis of the eigenvalues of the covariance matrix. If the signal to noise ratio of the measurement is high, a significant difference of the eigenvalues corresponding to signal and noise results. Then, the size of the signal and noise subspaces can be chosen based on the number of significantly large eigenvalues.

6.2 CLEAN algorithms

CLEAN algorithms are iterative approaches that aim at attaining an "clean image" $\mathbf{P}$ from an original "dirty image" (e.g. $\mathbf{S}$ resulting from CBF). Here we discuss the point spread function-based CLEAN-PSF variant [3]. The procedure starts with an initially empty clean image and repeats the following steps in each iteration

1. We look for the virtual source index $j_{\max}$, where the magnitude of the dirty image is maximal.
2. We compute the corresponding point spread function $\mathrm{PSF}_{\max} = \mathrm{PSF}(\boldsymbol{\theta}, \boldsymbol{\theta}_{j_{\max}})$.
3. We subtract this PSF from the dirty image with the appropriate magnitude, and onto the clean image we insert a single lobe (without side lobes) into the position $j_{\max}$. In an extreme case, the inserted lobe can be a single point.
4. We repeat the above steps until a convergence criterion is reached. Such criterion can be reaching the maximal allowed number of iterations, the decrease of the total power of the dirty image below a threshold value, the power contained in the PSF to be subtracted being smaller than a reference value, or a certain combination of these.

The algorithm can be formalized as

1. Let $\mathbf{D}$ be the measured covariance matrix $\mathbf{C}$ of the microphone array. The estimated distribution of source powers is denoted by $\mathbf{P}$, and i stands for the iteration index.

$$\begin{aligned} i &\leftarrow 0 \\ \mathbf{D}^{(i)} &\leftarrow \mathbf{C} \\ \mathbf{P} &\leftarrow \mathbf{0} \end{aligned}$$

2. Optionally, we can drop the diagonal components of the covariance matrix. By this operation, we disregard the autocorrelation of the microphones which does not hold information about the relations of the received signals. This reduced matrix is denoted by $\bar{\mathbf{D}}$

$$\bar{\mathbf{D}}^{(i)} \leftarrow \mathbf{D}^{(i)} - \mathrm{diag}\mathbf{D}^{(i)} \tag{37}$$

3. Using conventional beamforming, estimate the source power distribution as

$$S_j^{(i)} = \frac{\hat{\mathbf{a}}_j^{\mathrm{T}} \bar{\mathbf{D}}^{(i)} \mathbf{a}_j}{N^2} \qquad \text{for} \quad j = 1, 2 \ldots N_p \tag{38}$$

4. Find the maximal magnitude $S_{\max}$ and the corresponding index $j_{\max}$.

$$S_{\max} = \max_j S_j^{(i)}$$
$$j_{\max} = \arg\max_j S_j^{(i)}$$

5. Update matrix $\mathbf{D}$ making use of $S_{\max}$, $j_{\max}$ and the propagation vector $\hat{\mathbf{a}}_{\max}$ corresponding to the direction of $j_{\max}$:

$$\mathbf{D}^{(i+1)} \leftarrow \mathbf{D}^{(i)} - \alpha S_{\max} \hat{\mathbf{a}}_{\max} \hat{\mathbf{a}}_{\max}^{\mathrm{H}} \tag{39}$$

   α is a relaxation coefficient that controls the amount of power removed from the dirty image in a single operation. By the choice $\alpha = 1$, all the power found at $j_{\max}$ is subtracted at once. However, this may not be the best choice, as it leads to a false result if the local maximum resulted from the superposition of a main lobe and one or more side lobes. Therefore, in practice we choose $\alpha \approx 0.6 \ldots 0.8$. For smaller α values, too many iterations are needed for cleaning the dirty image. Similar to above, setting the diagonal entries of matrix $\mathbf{D}$ to zero is optional.

6. Let's examine the convergence criteria. A possible choice is to examine the elements of matrix $\mathbf{D}$. If the sum of the magnitudes of the entries of $\mathbf{D}$ is greater than that in the previous iteration, i.e.,

$$\sum_{m,n} |D_{mn}^{(i+1)}| > \sum_{m,n} |D_{mn}^{(i)}| \tag{40}$$

   then, we did not remove power from the image in Step 5. In this case, we do not continue the iteration.

7. Otherwise, we place a lobe of strength $\alpha S_{\max}$ onto the output image $\mathbf{P}$, and we continue with the next, $i+1$-th iteration from Step 3. This way, we can compress the power of a beam into a single point of the acoustical canvas, but other strategies (e.g. putting a narrower beam or a single main lobe onto the canvas) are also possible. To achieve the correct total power on the output image $\mathbf{P}$, the total power contained in the new beam must equal $\alpha S_{\max}$.

The CLEAN algorithm has further variants, such as the CLEAN-SC that is capable of the identification of coherent aeroacoustic sources [4], or the time domain CLEAN-T version, which allows for the detection of moving sound sources [5].

6.3 DAMAS

The DAMAS (*Deconvolution approach for the mapping of acoustic sources*) algorithm solves a deconvolution problem. According to formula (28) the acoustical image attained by conventional beamforming is the result of a convolution of the source powers and the corresponding PSFs of the microphone array. The deconvolution task is finding the real source power distribution based on the acoustical image an the analytically evaluated PSFs. Similar to the CLEAN algorithm, DAMAS solves the deconvolution problem by an iterative approach [6]. The deconvolution cannot be solved directly, as the system is largely underdetermined and badly conditioned. Hence, DAMAS performs the deconvolution by introducing additional constraints. The new constraint is that the source power over the acoustical canvas $\mathbf{P}$, must be non-negative in all points as the sources cannot radiate negative sound power. Thus, the mathematical constraint is easy to interpret physically.

The system of equations

$$\mathbf{S} = \mathbf{A}\mathbf{P} \tag{41}$$

needs to be solved, where, for the sake of simplicity, the matrix $\mathbf{A}$ contains the PSF corresponding to source location j in its jth column. Vector $\mathbf{S}$ is known, as this is the result of applying conventional

beamforming. The vector $\mathbf{P}$ must be determined; however, the inverse $\mathbf{P} = \mathbf{A}^{-1}\mathbf{S}$ cannot be evaluated as $\operatorname{rank}\mathbf{A} \ll N_c$. Beside (41) we also suppose that

$$P_i \geq 0, \quad i = 1, 2, \ldots, N_c. \tag{42}$$

The n-th equation of (41) can be expanded as

$$A_{n1}P_1 + A_{n2}P_2 + \cdots + A_{nn}P_n + \cdots + A_{nN}P_N = S_n. \tag{43}$$

By exploiting that $\mathbf{A}$ contains the point spread functions, we can rightfully assume that $A_{nn} = 1$, and hence the equation can be rearranged as:

$$P_n = S_n - \sum_{n'=1}^{n-1} A_{nn'}P_{n'} - \sum_{n'=n+1}^{N} A_{nn'}P_{n'}. \tag{44}$$

We use (44) for the deconvolution, introducing the solution in the i-th iteration, denoted by $\mathbf{P}^{(i)}$. (Similar to the case of the CLEAN-PSF algorithm, we start from $\mathbf{P}^{(0)} = \mathbf{0}$.)

$$P_n^{(i)} = S_n - \sum_{n'=1}^{n-1} A_{nn'}P_{n'}^{(i)} - \sum_{n'=n+1}^{M} A_{nn'}P_{n'}^{(i-1)}, \qquad n = 1, 2, \ldots N_c \tag{45}$$

We also apply the constraint on non-negative source powers in each step, that is, if (45) results in a negative power for $P_n^{(i)}$, we set $P_n^{(i)} = 0$ instead. Notice that when using (45) in case of $n = 2, 3, \ldots, N_c$, the previously found elements of the vector $\mathbf{P}^{(i)}$ are already made use of during the evaluation. The convergence of the iteration can be improved by reversing the order of the evaluation in formula (45) when stepping to the next iteration. That is, in the first iteration ($i = 1$) we follow the order $n = 1, 2, \ldots, N_c$, while in the second iteration ($i = 2$) we chose the order as $N_c, N_c - 1, \ldots, 1$, and so on.

The great advantage of the DAMAS algorithm is its robustness: after performing enough iterations, the resulting acoustical image becomes clean. The disadvantage is that it is difficult to predict how many iterations will be needed. Furthermore, the iterative algorithm is computationally expensive, which limits its applicability in real-time evaluation. It's worth mentioning that the size of the matrix $\mathbf{A} \in \mathbb{R}^{N_c \times N_c}$ (and the corresponding storage capacity) increases with the number of virtual source positions N_c squared, which also increases the memory consumption of the algorithm. Alternatively, as we only need to use a single row of matrix $\mathbf{A}$ at once, we can compute only one or a few of its rows at a time. This can be done as the evaluation of the PSFs is not very demanding computationally.

6.4 Further algorithms

There are many more algorithms for beamforming, with some of them related to specific applications, such as the ROSI beamformer [7] developed for the identification of noise sources in rotating systems, such as fan blades. Above, we aimed at introducing a few of the more general approaches, which are not specific to any application. We focused on methods that work in the frequency domain, which are useful for capturing narrowband signals, often buried in wideband noise. All the above algorithms used an acoustical canvas (grid), on which the source power is calculated. While this property is useful for visualization, it is not necessary for the estimation of the source direction (DoA – Direction of Arrival). As the development of algorithms for array processing is an active field of research in signal processing, new algorithms appear continuously in the literature.

Recently, array processing techniques relying on the compressive sensing principle became more popular [8]. The compressive sensing principle, similar to the idea of DAMAS, solves the deconvolution problem by imposing further constraints. However, instead of the non-negative source power constraint, compressive sensing minimizes the number of grid points where non-zero source power is assigned, i.e., it keeps the L_0-norm of the vector $\mathbf{P}$ minimal. In other words, the method maximizes the sparsity of the resulting vector $\mathbf{P}$, relying on the assumption that the real number of sources is much smaller than the number of grid points on the canvas, i.e., $M \ll N_c$.

References

[1] H. L. van Trees. *Optimum array processing*. New York: Wiley & Sons, 2002.

[2] R. Schmidt. "Multiple emitter location and signal parameter estimation." In: *IEEE Transactions on Antennas and Propagation* 34.3 (1986), pp. 276–280. DOI: 10.1109/TAP.1986.1143830.

[3] J. A. Högbom. "Aperture synthesis with a non-regular distribution of interferometer baselines." In: 15.41 (1974), pp. 417–426.

[4] P. Sijtsma. *CLEAN based on spatial source coherence*. Tech. rep. NLR-TP-2007-345. National Aerospace Laboratory NLR, 2007. URL: https://reports.nlr.nl/xmlui/bitstream/handle/10921/408/TP-2007-345.pdf.

[5] R. Cousson, Q. Leclère, M.-A. Pallas, and M. Bérengier. "A time domain CLEAN approach for the identification of acoustic moving sources." In: *Journal of Sound and Vibration* 443 (2019), pp. 47–62. DOI: 10.1016/j.jsv.2018.11.026.

[6] T. F. Brooks and W. M. Humphreys. "A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays." In: *Journal of Sound and Vibration* 294 (2006), pp. 856–879. DOI: j.jsv.2005.12.046.

[7] P. Sijtsma, S. Oerlemans, and H. Holthusen. "Location of rotating sources by phased array measurements." In: *7th AIAA/CEAS Aeroacoustics Conference and Exhibit*. DOI: 10.2514/6.2001-2167.

[8] A. Xenaki, P. Gerstoft, and K. Mosegaard. "Compressive beamforming." In: *Journal of the Acoustical Society of America* 136.1 (2014), pp. 260–271. DOI: 10.1121/1.4883360.