Direction and Distance Estimation of Sound Sources using Microphone Arrays

Bence Csóka, Péter Fiala, and Péter Rucz

***Abstract*—This paper is concerned with the estimation of the direction and distance of sound sources using the MUSIC (Multiple Signal Classification) beamforming algorithm, and their tracking with the help of Kalman filter. Direction-of-arrival (DOA) estimations are performed by combining acoustical focusing and beamforming. Distance estimation is usually not part of the process, but it is possible through an extension of the beamforming method. MUSIC is a relatively fast and simple method for locating sound sources. It is based on the separation of the cross-spectral matrix of the received signals into signal and noise subspaces. We use the Kalman filter and its extended non-linear version for tracking moving sound sources. We evaluate the performance of these methods by simulations in the MATLAB environment and measurements with unmanned aerial vehicles (UAV). DOA estimations and tracking are possible in both scenarios, but distance estimation is shown to be significantly more problematic in the latter case. We examine the possible causes of the observed errors and discuss possibilities for developing a more robust distance estimation method in the future.**

***Index Terms*—beamforming, Kalman filter, microphone arrays, MUSIC.**

I. Introduction

The position and trajectory of moving objects can be estimated using numerous remote sensing technologies, including optical and heat cameras or radar. If the object emits sound, it can also be localized by means of an acoustical sensor, i.e., an array of microphones. The latter approach also has its own advantages and drawbacks, and favorable environmental conditions for working reliably. Our objective is to estimate the positions of moving sound sources and track them accurately utilizing an acoustical camera. We implement an algorithm in the MATLAB environment for this task and test it by means of both simulations and measurements.

The MUSIC algorithm is discussed extensively in the literature. The paper by Xenaki et al. in 2014 [1] details MUSIC alongside conventional and more modern methods and compares them in two-dimensional simulations. Gupta and Kar created a version of the algorithm that is suitable for DOA estimation of coherent sources [2]. Yaning et al. improved the method by decreasing the computational complexity [3]. In general, beamforming algorithms are utilized in various fields, e.g. mining [4], detecting weak signals underwater [5], room acoustics and teleconference systems [6], navigation systems etc.

Estimating the position of sound sources usually only covers the approximation of the direction. For full 3D localization, the distance also needs to be estimated, which is a relatively novel concept in the field of acoustical beamforming. Cai et al. proposed a three-dimensional sound field reconstruction method which combines the use of beamforming and a binocular camera [7]. Valin et al. devised a 3D localization method for a video conference application which worked for up to 3 meters of distance [8]. In 2022, Merino-Martínez et al. presented a distance estimation that is based on asynchronous measurements with the same microphone array at multiple locations, which works for a quasi-stationary sound field [9]. Sarradj used a gridless version of orthogonal beamforming for 3D source-mapping to improve the resolution and reduce the computational cost [10]. Liaquat et al. developed a three-dimensional localization method for a low number of microphones [11]. In contrast, our goal is to devise a purely acoustical method that exploits beamforming with a grid for locating moving sources in a wider range of distances. We use the approach together with a 48-channel microphone array.

First, we discuss the basics of beamforming, its most important principles and concepts. After that, we move on to the MUSIC algorithm, briefly introducing the formulation used in our implementation, addressing its benefits and shortcomings compared to other beamforming methods. An overview of the extension of some beamforming concepts that make distance estimation possible is presented next, before moving on to the Kalman filter and its use in tracking moving sound sources. We conclude the article with presenting our findings from simulations and measurements and evaluating the performance of the proposed algorithm, while looking for potential future improvements.

Submission date: 2023.07.21. Acknowledgement: This work is supported by the Hungarian National Research, Development and Innovation Office under contract No. K–143436.

B. Csóka, P. Fiala and P. Rucz are with the Department of Networked Systems and Services, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics (BME), Budapest, Hungary (e-mail: csokab@hit.bme.hu, fiala@hit.bme.hu, rucz@hit.bme.hu)

DOI: 10.36244/ICJ.2024.2.6

II. The Basics of beamforming

The two main tasks that need to be solved using microphone arrays and beamforming algorithms are acoustical focusing and source localization. Acoustical focusing is based on the Delay-and-Sum method (Figure 1). The received signals of the microphones in the array (which are assumed to be omnidirectional) are amplified and delayed separately, resulting in the amplification of sound waves arriving from the focal direction. Due to the phase relations of the original signals, the sum of these "steered" signals has a higher amplitude in case of waves arriving from the focal direction [12]. Thus, even though the individual microphones have spherical characteristics, the array can have a highly selective directivity that can be designed to suit our needs. The directional characteristics can be further improved (e.g., by the suppression of sidelobes) by individually modifying the amplifications and the delays of the received signals. It is possible to focus on different directions virtually, by applying different delays on the signals, without physically rotating the array.

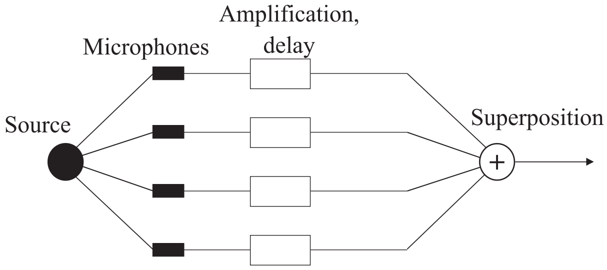

Fig. 1. The Delay-and-Sum method.

The other main task, source localization is solved by means of beamforming algorithms. A group of virtual source points is selected in three-dimensional space, making up the acoustical canvas (also called the scanning grid). By simulating sound propagation from all these virtual source points to the sensors of the array, the location where the similarity of the real (measured) and theoretical (simulated) sound fields is maximal, gives an estimated position. Focusing and source localization can be performed separately, but using them together is advantageous for the direction of arrival estimation of sound sources. By focusing on the virtual source positions one-by-one, an amplitude-like information representing the likelihood of the presence of a sound source at the given coordinate can be attained for each point of the grid using a beamforming algorithm. Thus, an amplitude map (or sound map) is attained, and we can estimate the direction of the source as the point on the canvas having the highest amplitude (likelihood). This way, source localization boils down to finding local maxima on sound maps (Figure 2).

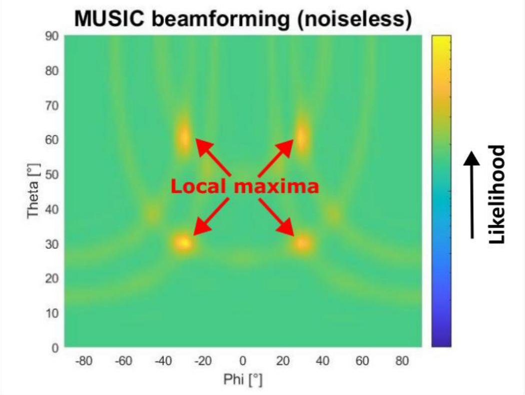

Fig. 2. Sound map created using a beamforming algorithm in a spherical coordinate system, with the azimuth denoted with Phi and the elevation denoted with Theta, having a phase domain of [-90°,90°] and [0°,90°], respectively. Source localization boils down to looking for local maxima on sound maps.

The points of the acoustical canvas are usually placed along an imaginary flat or spherical surface. The microphones of the array are most often in a cross, rectangular grid, circle, or spiral/multi-spiral formation. The distance between the array and the canvas is the focal distance, which can be taken as either finite or infinite (Figure 3). With an infinite focal distance, only the directions of the virtual source positions matter, and the wavefronts are assumed to be flat. The latter assumption holds only if the size of the array is negligible compared to the distance of the source, and hence the angles of incidence are roughly the same for each microphone. The received signals of the microphones also have the same amplitude and only differ in their phase. With a finite focal distance, the exact positions of the virtual sources must be defined, and spherical wavefronts are assumed. The received signals of the microphones differ in amplitude, phase, and angle of incidence.

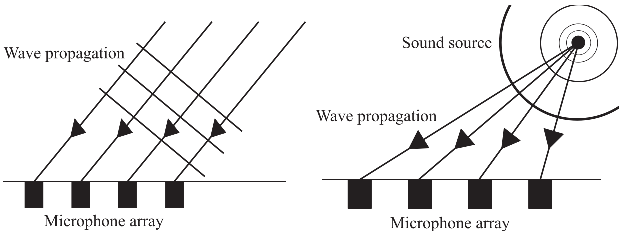

Fig. 3. Infinite and finite focal distance.

Beamforming algorithms in general can be used both in time domain and in frequency domain, and we chose the latter approach for our research. A narrow band is selected from the frequency spectra of short time windows of the received microphone signals, and the energy contained in this narrow band is the amplitude-like information calculated for each point of the acoustical canvas. Choosing the correct center frequency of the band is essential, because too low frequencies result in blurred amplitude maps, while at too high frequencies the principle of spatial sampling is violated, resulting in phantom sources at incorrect positions. This frequency f must satisfy

$$f < \frac{c}{2d} \qquad (1)$$

where c is the speed of the sound and d is the distance between adjacent microphones (which, in our case, are placed evenly).

III. MUSIC

Beamforming is used for estimating the source distribution vector ($\mathbf{q}$), which is done making use of the vector of the received signals ($\mathbf{p}$) and a sensing matrix ($\mathbf{S}$). The vectors $\mathbf{q}$ and $\mathbf{p}$ both contain information in the frequency domain, i.e., every element of the vector is the complex amplitude of a component of the emitted or received signal at a given frequency, during a short time window. The number of elements in $\mathbf{q}$ equals the number of virtual source positions on the acoustical canvas, while $\mathbf{p}$ has as many elements as the number of microphones in the array. The sensing matrix connects the two vectors as follows:

$$\mathbf{p} = \boldsymbol{Sq}. \tag{2}$$

The elements of $\mathbf{S}$ can be calculated in 3D space as

$$\mathbf{S}(i,j) = e^{-\mathrm{j}kd_{i,j}}\frac{1}{d_{i,j}}, \tag{3}$$

where j is the imaginary unit, k is the wavenumber (ratio of the angular frequency and the speed of sound), and $d_{i,j}$ is the distance between the i-th microphone and the j-th point of the canvas.

A very important concept for several beamforming algorithms is the cross-spectral matrix (CSM), which is the spectral cross-correlation between the received signals of the microphones. The CSM can be estimated with the received signal in the frequency domain:

$$\mathbf{G} = \frac{\mathbf{PP}^{\mathrm{H}}}{N}, \tag{4}$$

Here, the $\mathbf{P}$ matrix consists of the $\mathbf{p}$ vectors of the received signals from the last N number of time windows, that is, the estimated CSM is the average cross-correlation between the microphones during the most recent blocks.

MUSIC is a linear algebraic method that is based on the separation of the cross-spectral matrix of the received signals into signal and noise subspaces through eigenvalue decomposition [13]. The eigenvectors corresponding to the largest eigenvalues span the signal subspace, and the rest span the noise subspace. Then, the eigenvectors of the noise subspace $\mathbf{U}_{\mathrm{n}}$, and the sensing matrix $\mathbf{S}$ can be used for estimating the direction of the sound source. The computation is made by means of (4) [1],[2],[14],[15]:

$$\mathbf{P}_{\mathrm{MUSIC}}(j) = \frac{1}{\mathbf{s}(j)^{\mathrm{H}}\mathbf{U}_{\mathrm{n}}\mathbf{U}_{\mathrm{n}}^{\mathrm{H}}\mathbf{s}(j)} \tag{5}$$

where $\mathbf{s}(j)$ is the j-th column of the sensing matrix.

The advantages of the MUSIC algorithm are the higher resolution compared to conventional beamforming methods achieved without considerably larger computational cost and its relatively high noise tolerance. Its disadvantage is that the number of sources must be estimated beforehand. It is also important to mention that the algorithm only works if the sources are uncorrelated.

IV. DISTANCE ESTIMATION

It is important to choose the correct focal distance during acoustical focusing, because the clarity of the sound map greatly depends on it. The greater the difference between the distance of the source and the focal distance, the more blurred the image gets, therefore we can only estimate the position of the observed object with greater variance and uncertainty. This can be used to our advantage, because by extending the acoustical canvas into three dimensions and having virtual source positions at different distances, we are able to estimate the distance of the sound source on top of its direction.

Figure 4 shows the dependence of the amplitude map on the focal distance. In this simulation, the array consists of 48 microphones in a cross formation, the distance between the adjacent sensors is 6 cm, which means that with the speed of sound assumed to be around $340\ \mathrm{m/s}$ the upper frequency limit of spatial overlap is slightly over $2.8\ \mathrm{kHz}$ as per equation (1). The canvas consists of 91×181 points placed evenly on a quarter of an imaginary spherical surface, whose radius varies. Narrow band beamforming is performed with the center frequency being $2500\ \mathrm{Hz}$. There is one source $5\ \mathrm{m}$ from the centre of the microphone array that emits Gaussian white noise. We assume the ambient background noise to be negligible, so that the signal-to-noise ratio is high. As expected, the quality of the image becomes higher when the focal distance is closer to the real distance of the source. As the focal distance becomes less accurate, the peak on the image is more spread out, the sidelobes due to the cross arrangement of the microphones are more prominent, and the level of the background is higher relative to the peak.

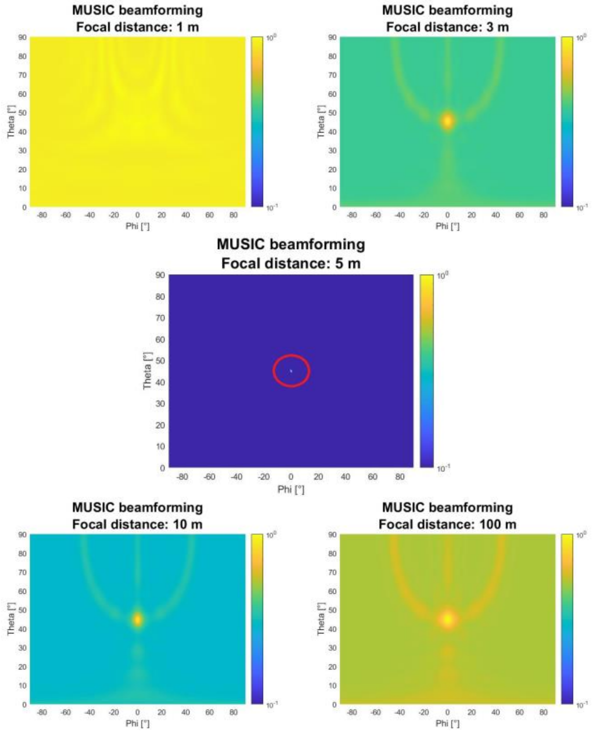

Fig. 4. Amplitude map of the same source distribution with different focal distances. The results of MUSIC beamforming are divided by their maxima in each case, and their logarithms are plotted as a two-dimensional function of the direction.

Extending the acoustical canvas into three dimensions can be done in different ways. One possible approach is to make an initial direction estimation on a primary canvas in the usual way discussed above. After the direction is determined, a secondary canvas is created, in which the virtual source positions are along a straight line in the estimated direction, but at different distances, thus creating a discretized line as the canvas (Figure 5). We apply beamforming on this secondary canvas, and the point corresponding to the local maximum of the secondary sound map gives the estimated distance of the source. The advantage of this method is that it allows for placing the points on the second canvas very densely without increasing the computational cost significantly.

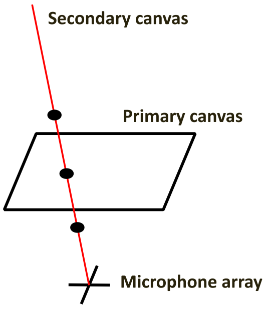

Fig. 5. Extending the initial canvas into three dimensions by multiplying the coordinates of the point that corresponds to the estimated direction.

Figure 6 shows a typical result of a distance estimation (after a successful direction estimation), where one source is located at 5 meters distance, the other is at 50 m. The secondary canvases are made up of 4500 points each, which are placed densely between 0.01 and 1000 m in a partially logarithmic manner (uniform between 0.01 m and 0.1 m, then uniform with a different resolution between 0.1 m and 1 m, and so on). The maxima are at 4.98 m and 49.4 m, respectively, which are the estimated distances. The peak corresponding to the farther source is less sharp than the one to the closer source, because the longer the focal distance, the more similar the situation becomes to infinite focal distance in the sense that the wavefronts arriving at the microphones are closer to planar. This means that the farther a source is from the array, the more difficult and inaccurate its distance estimation becomes.

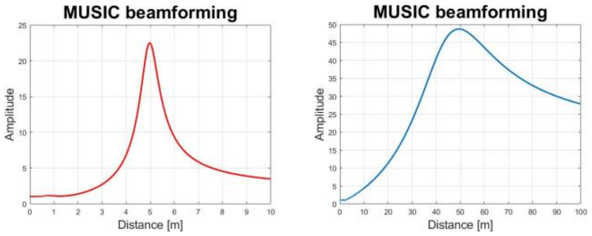

Fig. 6. Distance estimation with one source being at 5 meters (left) and the other being at 50 meters (right).

V. KALMAN FILTER

The Kalman filter estimates the state of temporally dynamic systems [16] (possible alternatives include Particle Filters, Dynamic Data Reconciliation and Double Exponential Smoothing predictors). In this case, the dynamic system is a moving sound source, and its state is its position and velocity, which can be predictively tracked. Not only the current measurements, but the earlier states of the system are also taken into consideration by the algorithm. This characteristic gives the filter higher accuracy compared to methods that only take the present state into account and opens the possibility of predicting the movement of the source. The traditional Kalman filter is an optimal estimator of linear systems, and it can be extended for nonlinear systems as well.

The Kalman filter starts from a state and an output equation, which are used in discrete time in our case:

$$\mathbf{x}(n+1) = F(x(n), u(n), n), \tag{6}$$

$$\mathbf{y}(n) = F(x(n), u(n), n). \tag{7}$$

Assuming that the system is linear and time-invariant, the state equation takes on the following form:

$$\mathbf{x}(n+1) = \mathbf{Ax}(n) + \mathbf{Bu}(n) + \mathbf{w}(n). \tag{8}$$

Here $\mathbf{x}(n)$ is the state vector in the n-th sampling moment, a vector which has six elements in three dimensions, those corresponding to the position (x, y, and z coordinates) and velocity (v_x, v_y, and v_z) of the sound source. The position can be measured by means of beamforming algorithms. $\mathbf{u}(n)$ is the input excitation vector, $\mathbf{A}$ and $\mathbf{B}$ are system matrices, and $\mathbf{w}(n)$ is the process noise on the input that represents the inaccuracies of the model.

The output vector $\mathbf{y}(n)$is defined by equation (9) in the linear and time-invariant case:

$$\mathbf{y}(n) = \mathbf{Cx}(n) + \mathbf{Du}(n) + \mathbf{v}(n), \tag{9}$$

where $\mathbf{C}$ and $\mathbf{D}$ are system matrices ($\mathbf{D}$ is negligible here because the input does not affect the output directly in our case), and $\mathbf{v}(n)$ is the noise vector on the measurements. The two noise vectors $\mathbf{w}(n)$ and $\mathbf{v}(n)$ are uncorrelated and normally distributed, with their expected value being zero and their covariance matrices being $\mathbf{Q}_n$ and $\mathbf{R}_n$, respectively.

The first step is an a-priori estimation of the state and output vectors for the *n*+*1*-th sampling moment (where the "-" upper index denotes that the estimation is a-priori):

$$\mathbf{x}^- = \mathbf{A}\tilde{\mathbf{x}}(n) + \mathbf{Bu}(n), \tag{10}$$

$$\tilde{\mathbf{y}}(n) = \mathbf{Cx}^-. \tag{11}$$

The difference between the measurement $\mathbf{y}(n)$ and the estimation $\tilde{\mathbf{y}}(n)$:

$$\mathbf{d}(n) = \mathbf{y}(n) - \tilde{\mathbf{y}}(n). \tag{12}$$

This difference can be used for an a-posteriori estimation (where the "+" upper index denotes that the estimation is a-posteriori):

$$\tilde{\mathbf{x}}(n+1) = \mathbf{x}^+ = \mathbf{x}^- + \mathbf{K_n d}(n), \tag{13}$$

where $\mathbf{K}_n$ is the Kalman gain matrix. The optimal gain $\mathbf{K}_n$ can be found as:

$$\mathbf{P}_n^- = \mathbf{AP}_{n-1}\mathbf{A}^{\mathrm{T}} + \mathbf{Q}_n, \tag{14}$$

$$\mathbf{P}_n^+ = (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_n^{-1}(\mathbf{I} - \mathbf{K}_n\mathbf{C})^{\mathrm{T}} + \mathbf{K}_n\mathbf{R}_n\mathbf{K}_n^{\mathrm{T}} =$$

$$= (\mathbf{P}_n^{-1} + \mathbf{C}^{\mathrm{T}}\mathbf{R}_n^{-1}\mathbf{C})^{-1} =$$

$$= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_n^-, \tag{15}$$

$$\mathbf{K}_n = \mathbf{P}_n^-\mathbf{C}^{\mathrm{T}}(\mathbf{CP}_n^-\mathbf{C}^{\mathrm{T}} + \mathbf{R}_n)^{-1} = \mathbf{P}_n^+\mathbf{C}^{\mathrm{T}}\mathbf{R}_n^{-1}. \tag{16}$$

$\mathbf{P}_n^-$ and $\mathbf{P}_n^+$ are the covariance matrices of the a-priori and a-posteriori state vectors, respectively.

While the above method is optimal for the estimation of the state of linear systems, observed systems are often nonlinear in real life. In our implementation, the model for the state vector is linear, but the output measurement $\mathbf{y}(n)$ is given in spherical coordinates, therefore, it is necessary to extend the Kalman filter algorithm for tackling such problems. One such extension is called Unscented Kalman Filter (UKF) [17].

With UKF, the first step is the calculation of $\mathbf{x}^-$ and $\mathbf{P}_n^-$ using equations (10) and (14), respectively. Next, we create $2N$ sigma points around $\mathbf{x}^-$, where N is the number of dimensions in the state space:

$$\mathbf{x}_i^{\sigma*}, \mathbf{x}_{N+i}^{\sigma*} = \mathbf{x}_\mathrm{n} \pm \boldsymbol{\sigma}_\mathrm{i},\ i = 1 \ldots N, \tag{17}$$

where $\boldsymbol{\sigma}_i$ is the i-th row of the matrix $\sqrt{N\mathbf{P}_n^-}$. This way, the statistical average and variance of the sigma points have the same values as the state vector and its covariance matrix (and this is why these sigma points are named after the Greek letter that denotes the deviation). The nonlinear output equation (7) is applied on these sigma points, and the obtained points ($\mathbf{y}_i^\sigma$) have an average of $\tilde{\mathbf{y}}$. The covariance and cross-correlation matrices are calculated using equations (18) and (19):

$$\mathbf{P}_\mathrm{yy} = \frac{1}{2N}\sum_{i=1}^{2N}(\mathbf{y}_i^\sigma - \tilde{\mathbf{y}})(\mathbf{y}_i^\sigma - \tilde{\mathbf{y}})^\mathrm{T}, \tag{18}$$

$$\mathbf{P}_\mathrm{xy} = \frac{1}{2N}\sum_{i=1}^{2N}(\mathbf{x}_i^{\sigma*} - \mathbf{x}^-)(\mathbf{y}_i^\sigma - \tilde{\mathbf{y}})^\mathrm{T}. \tag{19}$$

The Kalman gain matrix is found as

$$\mathbf{K}_\mathrm{n} = \mathbf{P}_\mathrm{xy}\mathbf{P}_\mathrm{yy}^{-1}. \tag{20}$$

This correction matrix can now be used for estimating the state vector and its covariance while taking current measurements and previous states into account:

$$\mathbf{x}_{n+1} = \tilde{\mathbf{x}}^+ = \mathbf{x}^- + \mathbf{K}_n(\mathbf{y}_n - \tilde{\mathbf{y}}), \tag{21}$$

$$\mathbf{P}_{n+1} = \mathbf{P}^+ = \mathbf{P}^- + \mathbf{K}_n\big(\mathbf{P}_\mathrm{yy} + \mathbf{R}\big)\mathbf{K}_n^\mathrm{T}. \tag{22}$$

VI. SIMULATION EXAMPLE

In this section we test the proposed methods of moving sound source localization with distance estimation by a simulation example.

During the simulations, the array consists of 48 microphones placed in a cross arrangement, and the distance between neighboring ones is 6 cm. This means that the upper frequency limit for the spatial overlap is the same as in section IV. The primary canvas consists of 20000 virtual source positions distributed evenly on a rectangular area, 15 m from the array (Figure 7). The secondary canvas lies along the initially estimated direction and consists of 4500 points distributed in a partially logarithmic manner between 0.01 and 1000 m. There is one simulated point source in the space that emits filtered white noise and moves along a straight line with constant velocity (1, 5, or 10 m/s), parallel to the plane of the sensor array, constantly at either 5, 25 or 50 m (where this distance is from the plane of the array). The SNR is 10 dB, and the time windows are 50 ms long (the SNR here means the ratio of the variances of the “useful” white noise emitted by the source, and the background white noise).

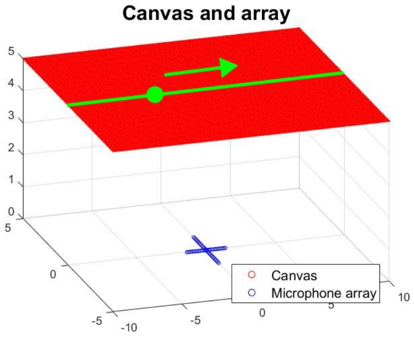

Fig. 7. The canvas (red), the microphone array (blue) and the sound source (green) in the simulation example. The source moves along a straight line with constant velocity.

The presented simulation covers both direction and distance estimation. Direction estimation is successful using the MUSIC algorithm, and the accuracy is further increased by the Kalman filter. Distance estimation is more difficult (because the main lobe of the beamforming is relatively wider along the secondary canvas, thus resulting in an estimation that varies more around the actual distance), but still successful in this simulation example (Figure 8). For farther sources, the method is less accurate, because when the size of the microphone array is much smaller than the distance of the source, the wavefronts are closer to planar and slight changes in distance result in only negligible changes in the angles of incidence. This can be evaluated by investigating the variance of the estimated distance, and this variance is proportionately greater for farther sources. In all three cases, the full 3D position estimation of MUSIC serves as the measurement data for the Kalman filter, and when its parameters are tuned properly, the estimation fluctuates considerably less around the correct distance (the MSE of beamforming is greater than the MSE of the Kalman filter, in most cases by roughly 120-150%, 50-100% and 20-60% for 5, 25 and 50 m, respectively). Thus, this example serves as a promising starting point when proving the adequacy of the distance estimation algorithm.

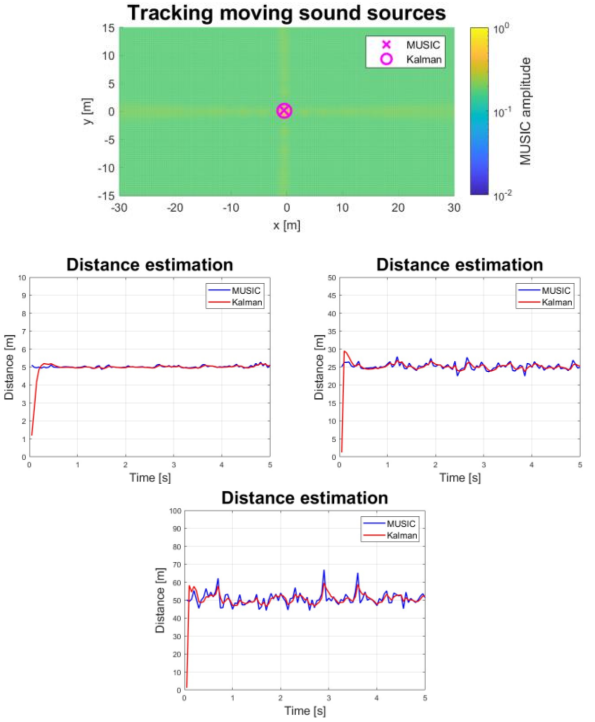

Fig. 8. Direction and distance estimation of a moving sound source using MUSIC and the Kalman filter; the distances are 5, 25 and 50 meters.

VII. OUTDOOR MEASUREMENTS

The validity of the algorithms was proven by simulations, and we demonstrate their results by processing measurements that are closer to real life situations in this section. We performed outdoor measurements as part of a larger project, where the position of unmanned aerial vehicles (rotary wing drones) was estimated by means of different principles. Because the drones were emitting sound, the acoustical method was viable. The two drones measured were of types Secop X8 and Tarot 680.

The microphone array used for the measurements consisted of 48 condenser microphones, placed in the same cross formation as in the previous simulations, stuck firmly in appropriately sized holes on a wooden plank. The array was connected to a PC through an analog/digital converter that sampled all channels simultaneously at rate of 48 kHz. A web camera was placed on the top of the wooden plank. Video capturing was made time-synchronized with the microphone array recordings. Hence, the video recordings of the movement of the drones could be fitted onto the sound maps, so that adequacy of DOA estimations could be confirmed visually.

Direction estimation using MUSIC and Kalman filter is successful in most of the time windows (one snapshot of each of the two drones is seen in Figure 9). The closer the drone is to the microphone array, and the louder its emitted sound is compared to the background noise, the more accurate the estimation becomes. The algorithm, however, does not give a correct estimation in every single time window, as there are a few moments when a strong background disturbance or ground reflection falsifies the result. Distance estimation, on the other hand, is unsuccessful. For Secopx8 we can give a rough estimation around 10 m (which is not nearly accurate enough), but for Tarot680, the estimated distance changes too erratically between time windows to accurately represent the actual distance of the drone.

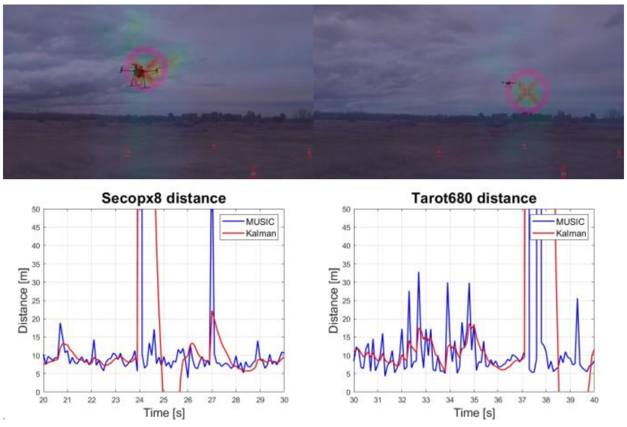

Fig. 9. Direction and distance estimation of Secopx8 (left) and Tarot680 (right).

VIII. Comparison and differences

The beamforming algorithm was capable of estimating direction both in simulations and measurements, but distance estimation was only successful in case of the simulation example. To ensure that the approach works for outdoor measurements, it is important to identify the critical differences, parameters, and conditions we neglected during the simulation that need to be accounted for in a more robust algorithm. Potential critical differences worth investigating include:

- The nature of the sound source. The simulated source is assumed to be a point source, it emits filtered white noise, with a spherical directivity. None of these three assumptions are true for an unmanned aerial vehicle.
- The nature of the background noise. In the simulation, we assumed additive white noise; however, in real measurements, the noise has a fluctuating amplitude and a time-variant spectrum.
- The presence of ground reflections was neglected in the simulation.
- The trajectory and velocity of the source. Drones did not move along a straight line with constant velocity.

So far, we have investigated the emitted sound and the effect of the center frequency (introduced in section II.), and ground reflections.

In the previous simulation, the sound source emitted filtered white noise. Because it is a wideband signal, many choices are possible under the upper frequency limit of spatial overlap for the narrow band detection. However, the noise emitted by a drone has strong tonal components, i.e., the energy in small time windows is concentrated around harmonically related spectral peaks. As the angular velocities of the rotors vary in time, the frequencies of the dominant spectral peaks also change. Therefore, in each time window, the analysis frequency should be fit onto the blade passing frequency or one of its overtones. Therefore, the performance of the distance estimation algorithm is expected to depend heavily on the frequency band.

In the next simulation, the sound source emits a signal that is a sum of sine waves, with a fundamental frequency of 300 Hz and overtones at its multiples up until 2100 Hz, with steadily decreasing amplitudes. The source moves along a straight line with constant velocity, 5 meters from the microphone array (that is placed in the same cross formation as before). The signal-to-noise ratio is set to 10 dB. Figure 10 shows the result of distance estimation depending on the analysis frequency (the analyzed narrow band is extracted with a 3rd order 20 Hz wide bandpass filter). As expected, when the analysis frequency is exactly the frequency of the overtone, the estimation is successful, and as we move farther, the variance of the result increases. At 1500 Hz, there are only minimal deviations from 5 m (at most 0.3 m), but at 1520 Hz, the algorithm is less accurate (the maximal deviation reaches 1 m). At 1540 Hz, 40 Hz from the overtone, the estimation is too unstable to be useful, even though direction estimation still works for most of the simulation. This means that distance estimation is more sensitive to the choice of the narrow frequency band analyzed in comparison to direction estimation.

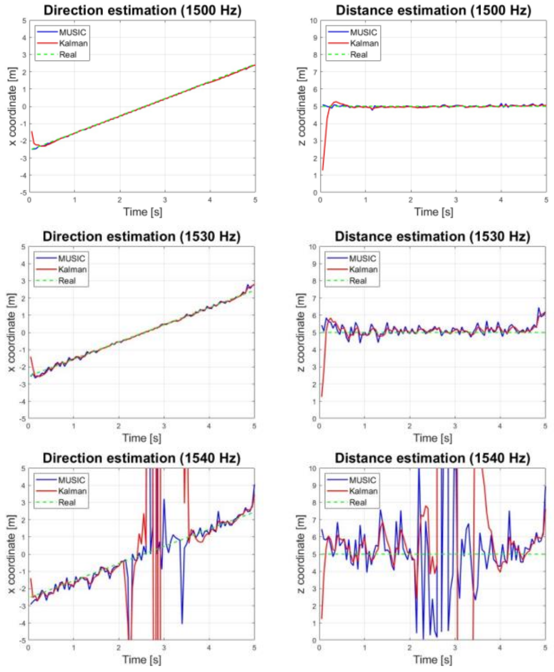

Fig. 10. Direction and distance estimation at different analysis frequencies of a simulated harmonic sound.

A more realistic sound source is incorporated into the simulation by using the sound extracted from a real recording of a moving Secop X8 drone, i.e., the sound signal received by one of the microphones in the array. In this particular recording, an overtone is present fluctuating around 640 − 650 Hz, so the analysis frequency is chosen as 640 Hz. In Figure 11 the comparison between the simulated drone sound and the real outdoor measurements can be seen. The simulated drone sound yields a worse result than the harmonic signal (probably due to the fluctuating frequency of the overtone), but the distance estimation is still much more stable than that in the outdoor measurement. It is worth mentioning that the background noise in the measurements becomes part of the emitted sound of the source in the simulation, which makes the estimation easier. From these results we can conclude that accounting for the waveform of the emitted sound by correctly choosing the analysis frequency does improve distance estimation, but in itself it is not sufficient for arriving at accurate distance estimation. Further critical differences need to be investigated.

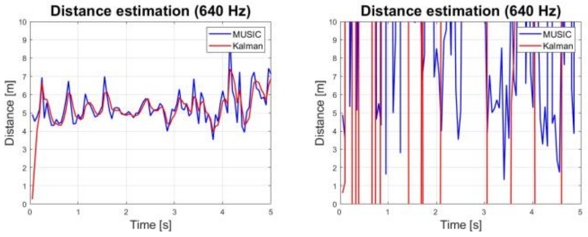

Fig. 11. Comparison of distance estimation during a simulation and a measurement on the same analysis frequency.

During outdoor measurements, the environmental conditions are not ideal, and this includes the sound of the drones being reflected from different surfaces, for example from the ground. If the reflected sound is strong enough, there is a false local maximum in its direction, which can impede correct localization. This issue was neglected in previous simulations, so it is worth evaluating how much negative effect ground reflections have on the performance of the distance estimation method.

The next measurement was carried out in a semi-anechoic chamber, with the same microphone array configuration as before. The sound source was the speaker of a stationary mobile the Secop X8 drone as in the previous simulation. The mobile phone was placed approximately 5 m from the microphone array, first on the reflective ground, to eliminate ground reflections, then on the top of a small table, around 46 cm from the ground, so that both direct and reflected sound waves can reach the microphones. In the former case, by placing the sound source as close as possible to the reflective floor, the difference between the distances of the actual and mirror sources to the same microphone is reduced close to zero, and thus the two signal paths have negligible phase differences.

Figure 12 shows the result of distance estimation, both without and with ground reflections. When reflections are not present, the estimation can be considered successful, even though it fluctuates a bit around the real distance. The measured distance is closer than 5 m, because the wooden plank holding the microphones has a small angle of inclination, and the source was somewhat closer to the plane of the microphone array than 5 m. When reflections are introduced, the peak of MUSIC beamforming on the secondary canvas becomes much flatter, and its magnitude also decreases, therefore the estimated distance has a larger variance. In this case, distance estimation produces a completely unusable result, similar to outdoor measurements. This means that ground reflections are indeed a critical condition that need to be accounted for in the algorithm.

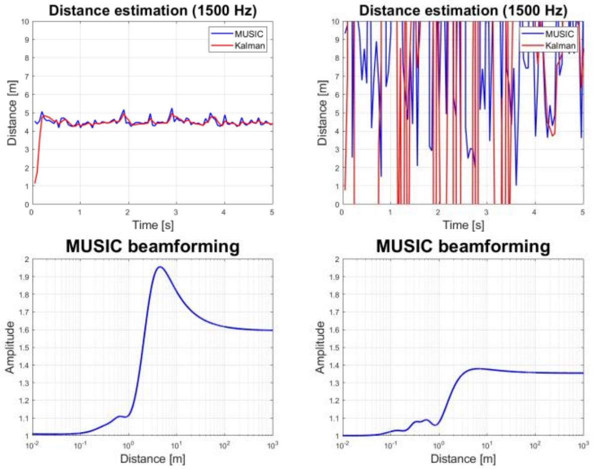

Fig. 12. Comparison of distance estimation during a measurement in a semi-anechoic chamber, without (left) and with (right) ground reflections. The lower figures show the direct, not normalized output of MUSIC on the secondary canvas, as a function of the distance.

To conclude, even though full three-dimensional position estimation is proven possible during measurements, the method still needs further refinement. Even in almost ideal conditions, in a semi-anechoic chamber, the estimation is slightly inaccurate, and the algorithm is not robust enough to handle unfavorable environmental conditions. So far, the only measurement where distance estimation was successful took place in a controlled environment with little to no disturbances.

IX. Conclusion and Future Plans

In this paper, we discussed the direction and distance estimation of sound sources using microphone arrays and beamforming algorithms. We chose the MUSIC algorithm for beamforming, which was extended by the Kalman filter method for tracking moving sound sources. During simulations and measurements, MUSIC was sufficient for direction estimation, and the Kalman filter improved the results further by smoothing out rapidly oscillating measurement data. However, distance estimation only worked initially during simulations.

Future goals are to investigate the critical differences between simulations and outdoor measurements, and to find the reason for the failure of the distance estimation algorithm. So far, we have investigated the waveform of the emitted sound and concluded that correctly choosing the analysis frequency improves the estimation. We also performed a measurement in a semi-anechoic chamber, where the absence of ground reflections made distance estimation possible, but their presence yielded similarly unstable estimations as those attained in case of outdoor measurements. These two conditions can be accounted for in the future by implementing an adaptive frequency tracking algorithm and using a method that is robust against ground reflections and correlated signals (for example the SAMV method [18]).

References

[1] A. Xenaki, P. Gerstoft, K. Mosegaard: "Compressive beamforming". *The Journal of the Acoustical Society of America*, Vol. 136 (1), 2014, pp. 260–271, DOI: 10.1121/1.4883360.

[2] P. Gupta, S. P. Kar: "MUSIC and improved MUSIC algorithm to estimate direction of arrival". *2015 International Conference on Communications and Signal Processing (ICCSP)*, Melmaruvathur, 2015, pp. 0757–0761, DOI: 10.1109/ICCSP.2015.7322593.

[3] L. Yaning, F. Juntao, R. Xinghao, M. Le: "An improved MUSIC algorithm for DOA estimation of non-coherent signals with planar array". *Journal of Physics: Conference Series*. 1060 012026, 2018, DOI: 10.1088/1742-6596/1060/1/012026.

[4] H. E. Camargo, R. A. Burdisso, P. A. Ravetta, A. K. Smith: "A comparison of beamforming processing techniques for low frequency noise source identification in mining equipment". *American Society of Mechanical Engineers*, 2009, pp. 1–7, https://www.cdc.gov/niosh/mining%5C/works/coversheet1134.html

[5] T. C. Yang: „Deconvolved Conventional Beamforming for a Horizontal Line Array". *IEEE Journal of Oceanic Engineering*, Vol. 43 (1), Jan. 2018, pp. 160–172, DOI: 10.1109/JOE.2017.2680818.

[6] M. Imran, A. Hussain, N. M. Qazi, M. Sadiq: „A methodology for sound source localization and tracking: Development of 3D microphone array for near-field and far-field applications". *2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Islamabad, 2016, pp. 586–591, DOI: 10.1109/IBCAST.2016.7429936.

[7] Y. Cai, X. Liu, Y. Xiong, X. Wu: "Three-Dimensional Sound Field Reconstruction and Sound Power Estimation by Stereo Vision and Beamforming Technology". *Applied Sciences*, 2021, 11(1), 92, DOI: 10.3390/app11010092.

[8] J.-M. Valin, F. Michaud, J. Rouat: "Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering". *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, France, 2006, pp. IV(841)–IV(844), DOI: 10.1109/ICASSP.2006.1661100.

[9] R. Merino-Martínez, B. von de Hoff, D. Morata, M. Snellen: "Three-dimensional acoustic imaging using asynchronous microphone-array measurements". *9th Berlin Beamforming Conference 2022*, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2022/papers/BeBeC-2022-S08.pdf.

[10] E. Sarradj: "Three-dimensional gridless source mapping using a signal subspace approach". *9th Berlin Beamforming Conference 2022*, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2022/papers/BeBeC-2022-S06.pdf.

[11] M. U. Liaquat, H. S. Munawar, A. Rahman, Z. Qadir, A. Z. Kouzani, M. A. P. Mahmud: "Sound localization for ad-hoc microphone arrays". *Energies 2021*, 14(12), 3446, DOI: 10.3390/en14123446.

[12] J. Novoa, R. Mahu, A. Díaz, J. Wuth, R. Stern, N. B. Yoma: "Weighted delay-and-sum beamforming guided by visual tracking for human-robot interaction". 2019, DOI: 10.48550/arXiv.1906.07298.

[13] R. Schmidt: "Multiple emitter location and signal parameter estimation". *IEEE Transactions on Antennas and Propagation* Vol. 34, 1986, pp. 276–280, DOI: 10.1109/TAP.1986.1143830.

[14] M. Mohanna, M. L. Rabeh, E. M. Zieur, S. Hekala: "Optimization of MUSIC algorithm for angle of arrival estimation in wireless communications". *NRIAG Journal of Astronomy and Geophysics*, Vol. 2 (1), June 2013, pp. 116–124, DOI: 10.1016/j.nrjag.2013.06.014.

[15] Q. Zhao, W. Liang: "A Modified MUSIC Algorithm Based on Eigen Space". In: Jin D., Lin S. (eds) Advances in *Computer Science, Intelligent System and Environment*. *Advances in Intelligent and Soft Computing*, Vol 104. Springer, Berlin, Heidelberg, 2011, DOI: 10.1007/978-3-642-23777-5_45.

[16] D. Simon: "Optimal State Estimation – Kalman, H∞, and Nonlinear Approaches". John Wiley & Sons, Inc., Hoboken, New Jersey (2006).

[17] Z. Belső, B. Gáti, I. Koller, P. Rucz, A. Turóczi: "Design of a nonlinear state estimator for navigation of autonomous aerial vehicles". *Repüléstudományi közlemények (Aviation scientific publications)* XXVII/3 pp. 255–276 (2015).

[18] H. Abeida, Q. Zhang, J. Li, N. Merabtine: "Iterative Sparse asymptotic minimum variance based approaches for array processing". *IEE Transactions on Signal Processing* 61(4), pp. 933–944, February 2013, DOI: 10.1109/TSP.2012.2231676.

Bence Csóka received his MSc degree in 2021 in Electrical Engineering from Budapest University of Technology and Economics. Currently he is a PhD student at the Laboratory of Acoustics and Studio Technologies at BME and his main areas of research are microphone arrays and beamforming algorithms.

Péter Fiala has received his MSc degree (2002) and PhD (2009) at the Laboratory of Acoustics, Dept of Networked Systems and Services, Budapest University of Technology and Economics. He is active in the fields of computational acoustics, acoustic signal processing and control of noise and vibration.

Péter Rucz received his PhD degree in electrical engineering from Budapest University of Technology (BME) in 2016. His professional interests are related to acoustics, signal processing, and numerical techniques. Currently, he is an associate professor at the Laboratory of Acoustics and Studio Technologies at BME.