DOI: 10.61782/fa.2023.0310

TRACKING SOUND SOURCES WITH MICROPHONE ARRAYS AND BEAMFORMING ALGORITHMS

Bence Csóka[1*] **Péter Fiala**[1] **Péter Rucz**[1]

[1] Department of Networked Systems and Services, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary

ABSTRACT

Microphone arrays can be used for many purposes, this paper is concerned with the determination of the position and trajectory of moving sound sources. The estimation of the position is based on the Delay-and-Sum method, which can be used to focus on different points in space, and to create sound maps representing the source distribution. This method can be enhanced further with different beamforming algorithms, for example with Multiple Signal Classification (MUSIC), or Functional Beamforming (FB). These algorithms all have their strengths and weaknesses, and their main beam width, side lobe strength, computational cost etc. must be considered. Beamforming algorithms complemented with the Kalman filter algorithm are suitable for tracking non-stationary sources as well. Our goal is to evaluate these methods through simulations and measurements. The measurement data is acquired from Unmanned Aerial Vehicles (UAVs) acting as sound sources in an outdoor environment with far from ideal environmental conditions. Due to this, the results indicate the efficiency and limitations of the algorithms when used in real-life applications, so that improvements can be made to the methods for increased efficiency, accuracy, and robustness.

Keywords: *microphone arrays, beamforming, MUSIC, Kalman filter.*

1. INTRODUCTION

The estimation of the position of different objects can be done through several different methods, for example optical or heat cameras, radar technology etc. If the object observed is a sound source, then the localization can also be done using a microphone array and acoustical beamforming. The received signals of the microphones can be processed to visually represent the source distribution with sound maps, which can be used for position estimation. Usually only the direction is determined, but the distance can also be estimated if the method is extended into three dimensions.

First, we will discuss the basic principles of the Delay-and-Sum methods and two beamforming algorithms, which are MUSIC (Multiple Signal Classification) and FB (Functional Beamforming). MUSIC is a linear algebraic method that has been extensively discussed in the scientific literature with several extended versions. FB is a relatively new algorithm that was introduced in 2014. We then move on to distance estimation by extending beamforming into three dimensions. Next is the extension of the beamforming algorithms with Kalman filter and its use in the tracking of moving sound sources. Finally, we present the results of simulations in the MATLAB environment, and measurements of Unmanned Aerial Vehicles (UAVs), evaluate the performance of these algorithms, and propose potential future improvements.

2. METHODOLOGY

2.1 Delay-and-Sum method

The direction or position estimation of a sound source is done through two separate tasks: focusing and source localization. Focusing is the amplification of sound arriving from a specific direction and the suppression of other directions. The basic principle of focusing is the Delay-and-Sum method, which is the appropriate delay, amplification,

[*] ***Corresponding author**: csokab@hit.bme.hu*

and then superposition of the received signals, which depends on the differences of the arrival times for the different microphones (Figure 1). Focusing correctly on a sound source results in a superimposed signal with greater amplitude, while the signal from sources in other directions becomes attenuated [1]. This way, while the characteristic of just one microphone is uniform, the microphone array can have a characteristic designed at will. By modifying the delays appropriately, the microphone array can focus on an arbitrary direction/position.

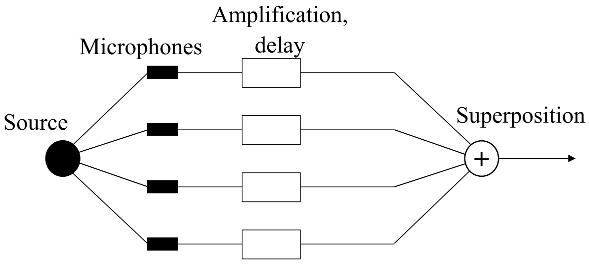

Figure 1. The Delay-and-Sum method.

Source localization is the actual estimation of the position of the sound source. During this process, we select a group of virtual source positions (usually along a flat or spherical surface), which together make up the so-called acoustical canvas or scanning grid. These points are all considered as the potential position of the source. Virtual sources are placed on each point one-by-one, and the position where the generated sound field resembles the real (measured) one the most, is where we assume the source to be located.
These two tasks can be solved together by performing the following steps:

1. We focus on every point of the scanning grid one-by one with the Delay-and-Sum method, during a short time-window.
2. For one point, we take a narrow band from the focused signal.
3. We calculate a value based on the energy of this band that represents the likelihood of a sound source being at that position.
4. We create a sound map assigning different colors to different likelihoods (expediently warmer colors to higher likelihoods).
5. Position estimation of sound sources can be then interpreted as looking for local maxima on the sound map (an example can be seen in Figure 2).

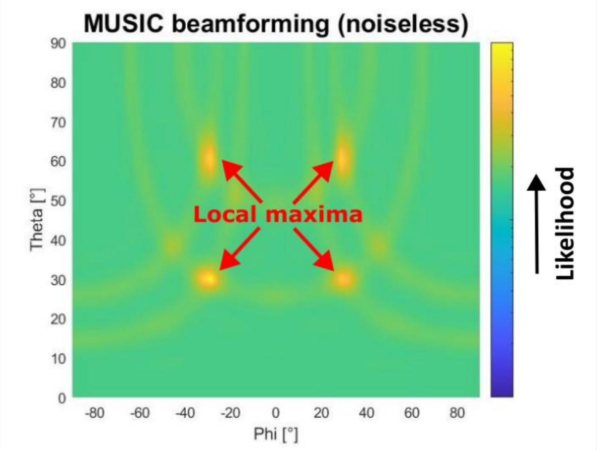

Figure 2. Sound map created by a beamforming algorithm. Position estimation can be interpreted as looking for local maxima on the map.

In the second and third steps, choosing the center frequency of the narrow band of the signal is crucial. Too low frequencies result in a blurred sound map, because with a larger wavelength, the phase differences between the delayed received signals will be negligible even farther from the correct direction. Too high frequencies, however, result in "phantom sources" in incorrect directions, because the principle of spatial sampling is violated. The upper frequency limit for the spatial overlap can be calculated using equation (1) (where "c" is the speed of sound, and "d" is the distance between neighboring microphones, assuming even distribution).

$$f = \frac{c}{2d} \tag{1}$$

2.2 MUSIC algorithm

The sound maps can be improved by using different beamforming algorithms for a more accurate and reliable estimation. These algorithms can decrease the main lobe width and suppress the sidelobes, at the cost of increased computational complexity. MUSIC is a simple beamforming algorithm that is based on eigenvalue-decomposition [2]. The cross-spectral matrix of the received signals is separated into signal and noise subspaces, where the signal subspace consists of the eigenvectors corresponding to the largest eigenvalues. The method then uses the eigenvectors in the noise subspace (denoted by $\boldsymbol{U_n}$) and the sensing matrix (denoted by $\boldsymbol{A}$) for the estimation (the sensing matrix containing the amplification and delay values between the scanning grid and the microphones), according to equation (2) [3]-[6]:

$$P_{MUSIC} = \frac{1}{A^H U_n U_n^H A} \tag{2}$$

2.3 Functional Beamforming

Functional Beamforming is another algorithm that is based on the eigenvalue-decomposition of the cross-spectral matrix of the received signals (denoted by **G**) [7]. This method is an improvement of conventional frequency domain beamforming by introducing the exponentiation of the CSM. This exponentiation is less involved computationally with eigenvalue-decomposition, because only the powers of the eigenvalues need to be calculated (equation (3)):

$$\boldsymbol{G} = \boldsymbol{U\Sigma U^H}, \boldsymbol{G^v} = \boldsymbol{U\Sigma^v U^H}. \tag{3}$$

It also suppresses the sidelobes and narrows down the main lobe, which is the main advantage of Functional Beamforming compared to conventional beamforming.

The energy values for each grid point are calculated using the CSM and the vector containing the amplifications and delays for that point (in short, steering vector, denoted by $\boldsymbol{a_j}$) [7]-[10]. In equation (4), there is also the *v* parameter which is the order of the functional beamforming map:

$$\boldsymbol{P_{FB}} = \left[\boldsymbol{a_j^H G^{\frac{1}{v}} a_j}\right]^v. \tag{4}$$

This is the most important parameter of FB, because it determines how much the algorithm can suppress the sidelobes and narrow the main lobe. If *v* equals 1, FB reduces to conventional beamforming, and the higher *v* is, the more these advantages can prevail. Its value is chosen typically between 20 and 300, but it cannot be increased at will, because inaccurate steering vectors, a coarse scanning grid or insufficient sensor calibration result in a suppressed main lobe at larger v values.

2.4 Distance estimation

The acoustical canvas / scanning grid usually constitutes a flat or spherical surface. This way, it is only suitable for direction estimation. However, distance estimation is necessary for full position estimation, and it can be achieved by extending the points of the grid into 3D.

One possible solution is to make an initial direction estimation on a primary canvas, and to create a secondary canvas consisting of points at different distances from the microphone array along the estimated direction. Basically, this means, that different focal distances are used in one specific direction (Figure 3). Beamforming is applied on this secondary canvas, and the distance of the sound source is estimated at the local maximum of the calculated energy.

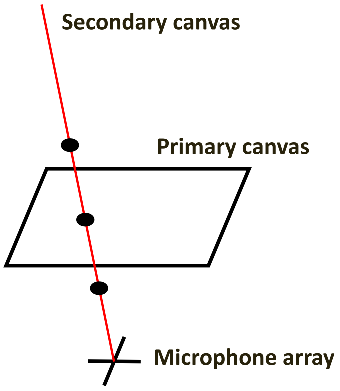

Figure 3. Distance estimation by extending the scanning grid into 3D.

2.5 Kalman filter

Beamforming algorithms are adequate for the position estimation of stationary sound sources. However, if we want to track and predict the movement of a non-stationary source, instead of simply taking snapshots of specific moments, we need to extend them.

One possible solution is the Kalman filter, which is an algorithm that can track the state of a system, where the state changes with time [11]. In this case, the system in question is a moving sound source, and its state is something that describes its movement, namely its position and velocity coordinates. The Kalman filter starts from the standard state equation (equations (5) and (6)), where the state vector (*x(n)*) consists of the three position and the three velocity coordinates (thus *x(n)* is a vector of 6 elements), and the excitation vector is denoted by *u(n)*. A, B, C and D are system matrices that determine how the next state vector and the output vector *y(n)* (in this case the measurement) depend on the current state vector and the excitation:

$$x(n+1) = Ax(n) + Bu(n) + w(n) \tag{5}$$

$$y(n) = Cx(n) + Du(n) + v(n) \tag{6}$$

The algorithm considers the measured position (which is the estimation of a beamforming algorithm), and the past states of the system, when making an a-priori estimation of the current state, and then makes a correction step based on the difference between the currently measured and estimated values; so, it uses more information than a beamforming algorithm. Another advantage of Kalman filter is that we can tune its parameters, for example the assumed process noise (*w(n)*) and measurement noise (*v(n)*), depending on how reliable the model and the measurement data are, so that the estimation of the filter follows the measurement data either quickly or slowly.

The traditional Kalman filter can only be used for linear systems, but it can be extended to handle nonlinear systems. One such extension is called Unscented Kalman Filter (UKF) [12]. UKF creates several points around the state vector (which are called sigma points), uses the nonlinear state and output equations on them, and the statistics (average and variance) of these transformed sigma points are used to update the state vector and the CSM.

3. RESULTS

3.1 Simulation example

The algorithms introduced in Section 2. were tested by means of simulations performed in the MATLAB environment. In the following example, the microphone array (blue on Figure 4) consists of 48 sensors placed in a cross formation, where the distance between adjacent microphones is 6 centimeters. This gives an upper frequency limit for the spatial overlap at around 2.8 kHz. The primary scanning grid (red on Figure 4) consists of 20000 points distributed evenly on a rectangular area. There is one point source located in the space, that emits filtered white noise, and it moves along a straight line with constant velocity, parallel with the plane of the microphone array and the acoustical canvas. The secondary canvas is always created after the initial direction estimation, and it consists of 4500 points, covering a distance range between 0.01 and 1000 meters. The distance of the primary canvas from the microphone array is 15 meters (the illustration is not proportional), the distance of the sound source is either 5, 25 or 50 meters. The signal-to-noise ratio is set to 10 dB, and the length of the time windows of CSM and sound map computation is 50 milliseconds.

Regardless of distance, both MUSIC and FB (ν = 20) are successful when it comes to direction estimation (Figure 5). They are also complemented by Kalman filter, which makes the noisy measurements slightly more accurate. The performance regarding the accuracy of the direction estimation of the two beamforming methods is similar in this example, but the sidelobes are significantly reduced during Functional Beamforming compared to MUSIC.

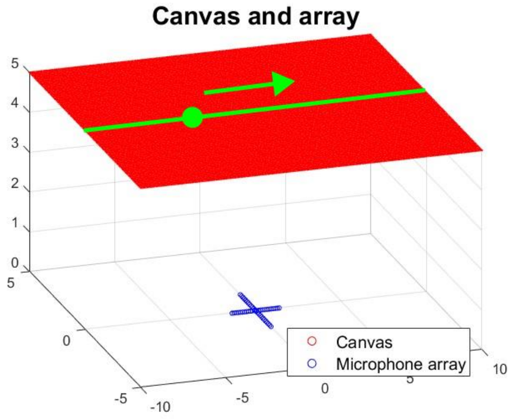

Figure 4. The canvas (red), the microphone array (blue) and the sound source (green) in the simulation example. The source moves along a straight line with constant velocity.

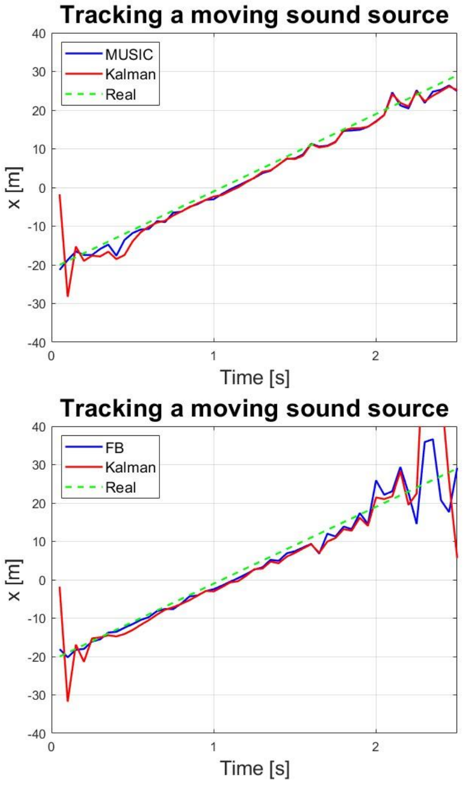

Figure 5. Localization and tracking of a moving sound source with different beamforming algorithms and Kalman filter.

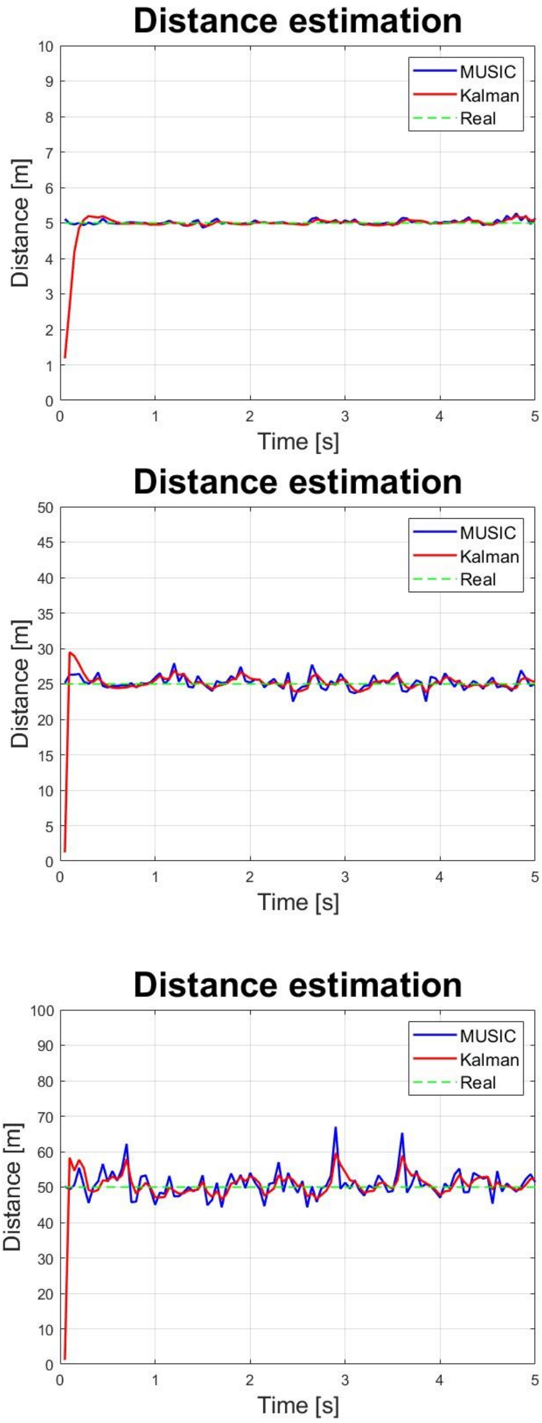

Figure 6. Distance estimation of a moving source with MUSIC and Kalman; the distances are 5, 25 and 50 meters.

Distance estimation proves to be more challenging especially for sources that are farther away from the microphone array (Figure 6). This is expected, because the farther the source is, the differences between the angles of incidence becomes smaller when slightly changing the distance. Again, neither algorithm proves to be better than the other in this regard (which is why only the MUSIC algorithm is presented on Figure 6). The estimation of the Kalman filter, when its parameters are tuned properly, is "smoother" than the beamforming algorithm, and it is closer to the actual distance of the source.

3.2 Measurements

During our work, we participated in outdoor measurements where unmanned aerial vehicles (or drones) served as sound sources. The measurements presented here are of two different vehicles: Secopx8 and Tarot680 (Figure 7) The microphone array formation was the same as the one used in the simulation. The acoustical canvas was also the same as in the simulation example: 20000 points distributed evenly on a rectangular area, 15 meters from the microphone array. The ¼ inch electret microphones were firmly fixed in holes drilled into a wooden board, such that their membranes were flush with the surface of the board. On the top of the board there was a web camera to capture a video recording of the drones' flight. This way, the recordings and the sound maps can be fitted onto each other to allow for a visual assessment of the tracking of the source by the beamforming algorithms. The sampling frequency of the microphones was 48 kHz, and the time window length for processing the received signals is 50 milliseconds.

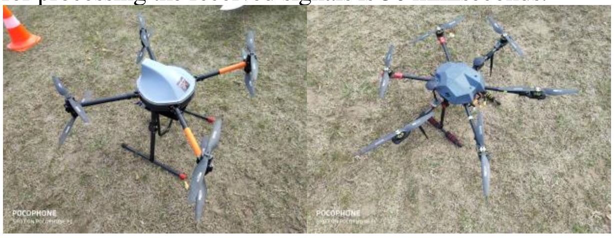

Figure 7. Photos of the Secopx8 (left) and Tarot680 (right) drones.

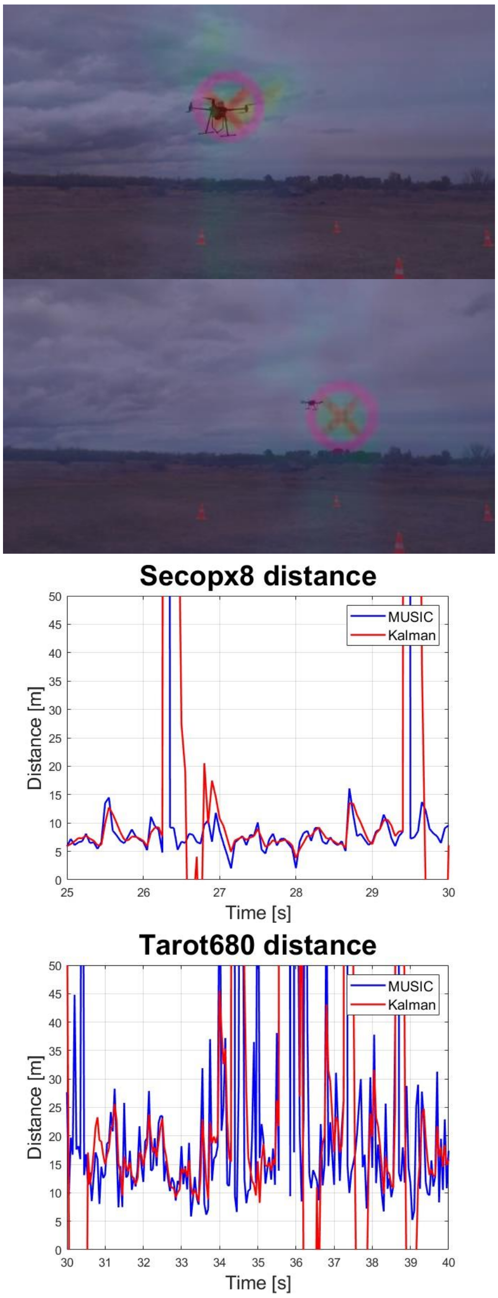

Figure 8. Direction and distance estimation of Secopx8 and Tarot680, with MUSIC (X) and Kalman filter (O).

The direction estimation with beamforming (this time with MUSIC) and Kalman filter is mostly successful (Figure 8). Because of the noisy background during the measurement, and potential ground reflections, there are some moments when the estimation is incorrect, but most of the time the algorithm can track the drone. As expected, the position estimation is more reliable when the vehicle is close to the microphone array and its sound is loud enough to stand out from the background. Distance estimation, however, is unsuccessful. In the case of secopx8, a distance between 5-10 meters can be assumed, but that is too inaccurate; for Tarot680, the distance estimation is far too unstable to be useful.

3.3 Comparison and differences

Direction estimation was successful both in simulations and measurements. Distance estimation, on the other hand, while showing promising results during simulations, is inadequate for real outdoor measurements as of now. The question naturally arises: what are the fundamental differences between simulations and measurements that cause the difference in performance? Simulations are simple and idealized compared to measurements, with many conditions and effects ignored.

Differences between simulations and measurements include:

- The waveform of the emitted sound. The simulated source emitted filtered white noise, while the sound of drones is more tonal; it also varies in a more unpredictable manner.
- Noise in the simulation was modelled as white noise, and its energy was defined in relation to the RMS of the useful signal. Disturbances during measurements are irregular.
- Ground reflections, which weren’t part of the simulation.
- The trajectory of the movement. In the simulation, the source moves along a straight line with constant velocity and distance from the array. The movement of a real UAV isn’t as simple.
- The finite extent of the sound source is neglected in the simulation example: it is simulated as a point source.

Unfiltered white noise is wideband, and the frequency for the Delay-and-Sum method (detailed in Section 2.1) can be chosen almost arbitrarily, because the signal has energy at any frequency. In the simulation, the emitted sound was white noise filtered by a bandpass filter, and the observed beamforming frequency conveniently fell into its band. This doesn’t automatically happen during measurements, where the useful sound doesn’t have much energy at every frequency. Therefore, it is useful to investigate the performance of distance estimation depending on the observed frequency.

In the next simulation, the sound source moves along a straight line 5 meters from the microphone array, with constant velocity. It emits a signal that is a sum of sine waves, with a base frequency of 300 Hz, harmonics up to 2100 Hz, and steadily decreasing amplitudes. This signal is somewhat closer to the emitted sound of a drone (though still idealized). Figure 9 shows the performance of distance estimation depending on the observed frequency, which is closest to the overtone at 1500 Hz. As expected, the closer the observed frequency to the overtone, the better the distance estimation. At 1500 Hz, it is nearly perfect, with only small deviations from 5 meters (in the case of MUSIC, no more than 0.3 meters). At 1520 Hz, these deviations are greater, at times rising to 1-2 meters, but the estimation is still useful most of the time. At 1540 Hz, even though direction estimation still works most of the time, distance estimation can no longer be considered successful, which is a similar situation to the measurements presented in this paper. It is worth noting, that at 1540 Hz MUSIC has worse performance than FB in the middle third of the simulation, around the time when the source is directly in front of the array (here even the direction estimation was wrong for a few time windows). FB produces greater deviations at 1500 and 1520 Hz towards the end of the simulation; neither for which the reason is known at the present.

To simulate a sound source one step closer to a real UAV, we can create one with the same trajectory as before, with the only difference that its emitted sound is extracted from real measurement data: here, the emitted sound is the received sound at one of the sensors when Secopx8 was being localized. Measurements show that Secopx8 has an overtone around 640-650 Hz. The frequency of this overtone fluctuates over time, but for now, the observed frequency will be constant (640 Hz). Figure 10 shows the comparison between simulations and measurements on different frequencies. Only the MUSIC algorithm is used here, as it performs better at low frequencies. Unfortunately, distance estimation still doesn't work for the measurement. It is better in the simulation, but it's still of worse quality than in the previous simulation. This could be due to the fluctuating frequency of the overtone.

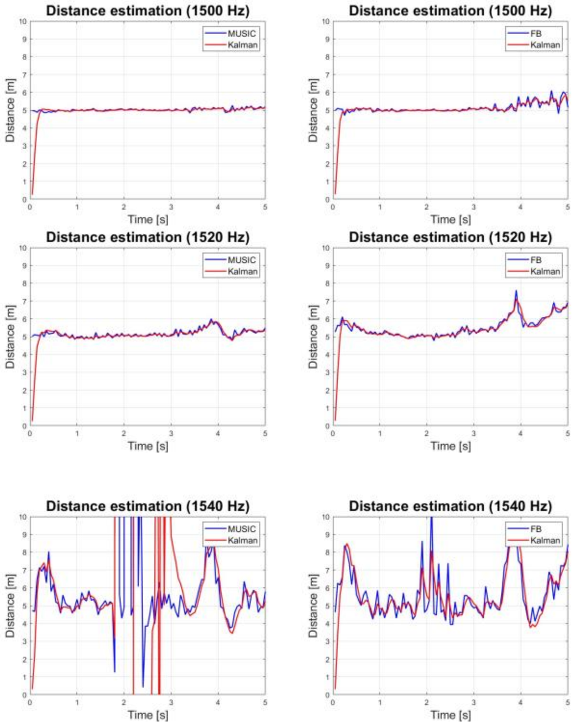

Figure 9. Distance estimation depending on the observed frequency, with simulated tonal sound.

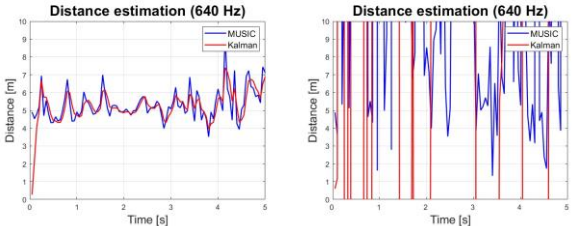

Figure 10. Comparison of distance estimation during simulations (left) and measurements (right).

From these results, we can conclude that while accounting for the waveform of the emitted sound by correctly choosing the observed frequency does improve the algorithm, and an adaptive frequency tracking method could improve it further; it most likely will not be enough to successfully estimate distance in a real measurement. For that, the other factors (ground reflections, irregular background noise, movement trajectory, the finite extent of the source) have to be investigated in the future.

4. CONCLUSION

In this paper, we discussed the determination of the position of sound sources using microphone arrays and beamforming algorithms. After introducing the theory of the used algorithms (namely, Multiple Signal Classification, Functional Beamforming, distance estimation and Kalman filter), we tested them by means of simulations and measurements. Direction estimation using both MUSIC and FB was successful, not just in simulations, but measurements as well, though in the latter case they weren't as reliable due to less favorable conditions. The Kalman filter algorithm improved the results further, by smoothing out the slightly inaccurate and thus rapidly oscillating measurement data from beamforming. Unfortunately, distance estimation only worked during simulations. Future goals include investigating the differences between the simulation example and real outdoor measurements, to the determine, and later eliminate the cause of inconsistent performance. The difference between simulated and measured waveforms can be countered in the future by implementing an adaptive frequency tracking method. Other differences, such as accounting for ground reflections, irregular background disturbances and movement trajectory, the finite extent of the sound source, are subject to further research.

5. ACKNOWLEDGEMENT

This work has been supported by the Hungarian National Research, Developement and Innovation Office under contract No. K–143436.

6. REFERENCES

[1] 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, https://doi.org/10.48550/arXiv.1906.07298.

[2] R. Schmidt: "Multiple emitter location and signal parameter estimation". IEEE Transactions on Antennas and Propagation Vol. 34, 1986, pp. 276–280.

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

[4] 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, https://doi.org/10.1016/j.nrjag.2013.06.014.

[5] 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, https://doi.org/10.1007/978-3-642-23777-5_45.

[6] 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, https://doi.org/10.1109/ICCSP.2015.7322593.

[7] R. P. Dougherty: "Functional Beamforming". 5th Berlin Beamforming Conference 2014, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2014/papers/BeBeC-2014-01.pdf.

[8] G. Battista, P. Chiariotti, P. Castellini: "Tuning of the Functional Beamforming Resolution for Wind Tunnel Measurements". 9th Berlin Beamforming Conference 2022, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2022/papers/BeBeC-2022-S05.pdf.

[9] R. P. Dougherty: "Enhancing Deconvolution with Functional Beamforming". 9th Berlin Beamforming Conference 2022, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2022/papers/BeBeC-2022-S01.pdf.

[10] R. P. Dougherty: "Robust Functional Beamforming". 9th Berlin Beamforming Conference 2022, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2022/papers/BeBeC-2022-S07.pdf.

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

[12] 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).