Sound-source Localisation using a Microphone-array for NUbots Interim Report

Clayton Carlon, C3327986

April 30, 2023

Chapter 1

Introduction

Locating a source of sound is a long sought after ability in technology, especially in the study and development of robotics which seeks to emulate and compete with human senses. The main aspiration for this project is to develop an array of microphones along with an accompanying computation, preferably embedded, to locate a source of sound using signal-processing techniques. In the end, this system will be used on the football-playing robots in the NUbots team, the university's on-campus robotics team.

1.1 Background

NUbots is a multidisciplinary team under the University of Newcastle's robotics research group and has competed since 2002 in RoboCup, an international competition where humanoid robots play association-football*. The team is made up of both undergraduate and postgraduate students in computer science and engineering. At this time, the robots on the team are modified versions of the igus® Humanoid Open Platform; this modified platform is called the NUgus hardware platform and competes in the kid-size league.

The remit of the team is broad and not only competing in the RoboCup. Although it strives to win RoboCup, the team also does research more generally in areas related to robotics, such as computer-vision, robotic locomotion and control, etc. Therefore, the project does not necessarily have to be tied to the said competition. Furthermore, this project, as it evolves, does not necessarily have to be used in the application of NUbots nor robotics more generally although the original idea came from NUbots. Sound-source-localisation is a broad area of study, and the final prototype can just as easily be employed in other applications, such as a video-conference-system.

*otherwise known as football or soccer

1.2 Scope

Ideally, the project shall yield a working prototype fitted on one of the NUgus robots. At this time, since the project is more intended as a proof-of-concept, its goals are thus somewhat open-ended. However, such basic goals are that it should at the very least:

- estimate the three-dimensional direction, i.e. the azimuth and the elevation,
- locate a reasonably distinct sound in a moderate environment, e.g. a whistle, a lone voice, a loud thud,
- and handle the noise from the motors of the robot which will nearby.

Further goals are that it can:

- estimate the distance, essentially locate the source's full coordinates in three-dimensions,
- track the location over time using e.g. a Kalman filter,
- locate multiple simultaneous sources, as long as they have relatively distinct locations,
- spatially filter the localised sound,
- and work well enough if noisy and reverberant environments, e.g. in a hall full of people.

Although sound-source-localisation will be helpful in a scenario of football-playing robots, the final system does not have to directly help the robot's performance in a match. The ability for a robot to interact with humans is a broad and well sought after goal. Even a basic demonstration of the robot turning its head towards a human speaker can help the team's outreach and marketing in exhibition-shows, publicity-events, etc. Not only will this help sponsors and thus funding for the team, but even the basic proof of concept also contributes further to team's research in and remit of robotics

Chapter 2

Literature-Review

Since sound-source-localisation is a much widely researched problem, in e.g. robotics, drones, video-conferencing, the military, submarines, hearing aids, etc., the review of the literature can be quite broad. To narrow the search, and to also find examples that are most relevant to this project, especially for computation and dimensions, this review has mostly considered literature with a robotics context, rather than large-scale military outdoor application for example. Indeed, sound-source-localisation for drones shares a good deal with that for robotics, e.g. the influence of motors, but one important difference is the fact that applications for drones are mostly outdoors rather than indoors where the reverberation is an important factor.

To further help narrow the scope of the search, most methods considered are ones that use an array of at least two microphones and use some kind of signal-processing. Other families of methods exist such as binaural approaches with two microphones and head-models [3] and those that use machine-learning, such as convolutional neural networks [27]. It is not necessarily that these other approaches will be ignored outright but rather that the scope of the literature-review as well as the project as a whole will be kept relevant in order to ease the search.

2.1 Methods

2.1.1 Overview

There are already a few reviews of the literature, each going through a number of existing and studied methods for sound-source-localisation. A survey [3] attempted to give a state of the art of sound-localisation in robotics. It dealt with two main areas, namely binaural approaches and array-based approaches. Since an array of microphones will be used in this project, the latter area is most relevant. All the approaches of which that the survey explored were listed as such:

- MUSIC,
- TDOA, and
- beam-forming.

Another review [26] also classified methodologies as:

- one-dimensional single direction-of-arrival estimation,
- two-dimensional single direction-of-arrival estimation,
- multiple direction-of-arrival estimations, and
- distance-estimation.

Here, the review talks about correlation as a way to estimate a single direction of arrival and about beam-forming and MUSIC as a way to estimate multiple directions. It also discusses the potential use of correlation for multiple sources given that other sources are represented as secondary peaks.

To estimate the distance, the review proposes a number of ways, some of which are:

- the intersection of hyperbolic curves from multiple estimated TDOA,
- triangulation of multiple DOA at different positions of the robot using its mobility,
- triangulation of multiple DOA from different sub-arrays,

One important distinction about the geometry is the far field and the near field. Many methods approximate their algorithm and computation by assuming that the source is in the far field where the distance between the source and the array is comparably longer than the array's width. This is such that the sound-waves may be assumed to be planar at the array and such that the direct lines between the source and each microphone may be assumed to be parallel. Some methods, especially those of beam-forming, can estimate distance but only in the near field where the distance between the source and the array is comparable to the array's width. However, since the microphone-arrays in robotic applications are generally small, only the far field is an acceptable assumption. This may especially be the case for NUbots where candidate space for an array on the existing NUgus is sparse and a smaller array may have to be chosen.

2.1.2 Time-Difference of Arrival

A simple yet proven way to estimate the DOA of a sound-source using a microphone-array is to calculate the difference in time [3]. Commonly, the TDOA of two microphones is often found by the cross-correlation of two signals, the highest peak of which corresponds to the estimated TDOA.

$$R_{ij}(\tau) = \sum_{n=0}^{N-1} x_i[n]x_j[n-\tau] \tag{2.1}$$

where x_i and x_j are the two signals, from two microphones for example. Cross-correlation is useful in particular because it has an equivalent calculation in the frequency-domain:

$$R_{ij}(\tau) = \mathfrak{F}^{-1}\left(X_i[k]X_j[k]^*\right) \tag{2.2}$$

where $X_i[k]$ and $X_j[k]$ are the Fourier transforms of x_i and x_j respectively. Unlike the original calculation which relies on addition and has a complexity of $O(N^2)$, this frequency-domain equivalent relies on multiplication, can be made more efficient by FFT, and has a complexity of $O\left(N\ln(N)\right)$ [31].

An advanced form of this TDOA estimation is the generalised cross-correlation (GCC):

$$R_{ij}^{(w)}(\tau) = \mathfrak{F}^{-1}\left(\psi[k]X_i[k]X_j[k]^*\right) \tag{2.3}$$

where $\psi[k]$ is the frequency-weighting [3], [26]. A common weighting is the phase-transform (GCC-PHAT):

$$\psi[k] = \frac{1}{|X_i[k]||X_j[k]|} \tag{2.4}$$

The TDOA can therefore be estimated as the point in time where the GCC-PHAT is highest:

$$\Delta t_{ij} = \mathrm{argmax}_\tau\left(R_{ij}^{(w)}(\tau)\right) \tag{2.5}$$

Given the TDOA Δt_{ij}, and assuming that the source is far enough, the two-dimensional angle of arrival at a pair of microphones can be estimated from simple trigonometry of parallel lines:

$$\theta = \sin\left(\frac{c\Delta t_{ij}}{d_{ij}}\right) \tag{2.6}$$

where c is the speed of sound, and d_{ij} is the displacement between the microphones, and assuming that the source is in the far field.

The former expression is the most basic approach for a simple two-dimensional estimation of a single angle, often the azimuth; there are of course more sophisticated examples of finding the source three-dimensionally, even with distance. In fact, more generally, the given TDOA draws a hyperbola as the locus of possible positions on the two-dimensional plane (a hyperboloid in the three-dimensional case) where the two microphones are the foci of such a hyperbola. This hyperbola can be seen as a straight line in the far field.

Another important consideration is the resolution. Since the cross-correlation is a time-domain signal whose resolution is the sample-period of the two input signals, then the estimated TDOA is a multiple of the sample-period [3]. This thus yields a resolution on the estimated angle which is worst when the source is in line with the microphones. One way to improve the resolution is to interpolate the signal, i.e. growing the number of samples of the signal. Another way is to simply sample faster, but this may make the computation more burdensome.

Literature Examples

The usage of TDOA in sound-source-localisation in robotics is a widely researched. Despite the simplicity of estimating the azimuth, there are many more sophisticated methods of estimating the direction or even the position. For example, one method [31] solves a set of simultaneous equations from at least five microphones using the pseudo-inverse of a matrix; this yields a vector giving the three-dimensional direction. It also uses a different weighting than GCC-PHAT which gives more weight to areas of the spectrum where the SNR is higher. The experiment was performed with eight microphones in an open rectangular prism ($0.5 \times 0.4 \times 0.36$ m) on top of a mobile robot "in a room with a relatively high noise level mostly due to several fans in proximity" with "moderate" reverberation. The computation was done on a desktop PC and used about 15% of the CPU. Three sounds were tested, speech, snap of fingers, and tap of boot. The results showed an estimated direction that degraded as the source got nearer to the array; the mean-square-error was at best 0.6 and at worst 4.9 degrees with the distance being from 0.3 to 5 m. This degradation was most likely because of the far-field assumption. The system works "properly" between 3 and 5 m although the authors claimed this as a result of "the noise and reverberation conditions". They also said that broadband signals were detected better than narrowband ones like tones.

A similar approach [14] proposed a novel estimation of the TDOA, namely eigenstructure-based GCC (ES-GCC) which can handle an unknown number of multiple sources. This paper could compute as before the three-dimensional bearing but also the distance for the near-field case, i.e. when the source is near enough to the array. The K-means++ algorithm was used to cluster accumulated results. This approach was tested on an array of eight microphones forming a rhomboidal prism; the diagonal distances from the centre were 0.22 and 0.14 m. This was done in a real room for both a single source and multiple ones that were all 2.4 m away from the array. The worst SNR was 14.58 dB. The mean error for all experiments was less than 3 deg. The estimation of distance was not evaluated, only the azimuth and elevation. The computation was not discussed nor was reverberation.

However, the former methods only computes the bearing, at least for the far field which is more relevant for the case with a smaller array. Another example [11] estimates the full three-dimensional coordinates of the source by an iterative algorithm based on Newton's method to solve a set of spatial coordinate relations. This example also had a few improvements such as a partitioning process, a fast search-strategy of the GCC peak, a screening strategy, and a new weighting function in the GCC that dealt with reverberation. However, this proposed weighting function needed beforehand the parameters of the environment and the rough displacement of the source; the authors have admitted that this limits the universality. In simulations, the improved weighting function gave a better distinct peak in the GCC for a reverberation-time of 300 ms. This method was tried on an array of five microphones forming a rectangular pyramid 0.25 m wide and 0.125 m tall. The performance was tested at differ-

ent points around the array from 1 to 6 m and with a SNR of 45 dB. The error in distance increased as the source was further away; this was observed at best as 0.05 m and at worst as 0.25 m. The error in azimuth varied little in both distance and bearing and was observed as being within 1.5 deg. The performance was also tested at different levels of SNR, from 40 to 10 dB. In the worst case of 10 dB, the error in distance was observed at worst as 0.4 m. The paper neglected any evaluation of the estimated elevation, considering that the geometry of the array is not symmetrical. Furthermore, the computational needs were not discussed in depth.

A similar paper [7] solved a similar set of equations but using Lagrange multipliers instead. It also applied tracking filtering such as EKF. This approach was tested on an array of five microphones forming a double equilateral tetrahedron with sides 28 cm long. The SNR of the room was 15 db, and the room's reverberation-time at 60 dB was measured as 0.36 s. A number of sessions were tried where a speaking person moved along a different path. The mean squared error of the raw estimated position before filtering was observed to be at best $4.957 \times 10^{-3} m^2$ and at worst $27.21 \times 10^{-3} m^2$. However, all the sessions had the speaker within 1 m of the array; so, they may be thought of as in the near field. Also, the paper makes no evaluation of the computation.

Likewise, another paper [17] employed robust spatial filtering, namely a Kalman filter, but only to estimate the azimuth. It used GCC-PHAT to estimate the TDOA but also used a trained feed-forward network of a single hidden layer to quantify the reliability of the estimated TDOA.

Both two-dimensional and three-dimensional localisation was proposed in a paper [19] where a similar set of equations was solved, and the speed of sound could be estimated as a variable with at least six microphones. Where it was assumed to be constant, at least five microphones were needed instead. Only the two-dimensional case was evaluated in the preliminary testing. In which, the relative error in distance stayed below 1%, and the worst error was 7.6 mm at 1 m, but this experiment was only done in a workspace 1000×1000 mm and was done with ultrasonic pulses of 75 kHz. The authors did not reveal the geometry of the array and given how short the tested distances were, it may be assumed that the experiment was in the near field.

Potential unwanted noise from servo-motors, etc., may affect and disrupt how well the system localises wanted sources. A very recent study [20] set out an approach that estimated the azimuth and elevation and mitigated the noise from a drone's motors. Here, the authors computed the GCC-PHAT as an angular spectrum for many pairs of microphones and summed each pair's angular spectra; this method is similar to SRP-PHAT in that the azimuth and the elevation corresponding to the largest sum belong to the estimated direction. In order to mitigate the motors' noise, the authors subtracted the known angular spectrum of the drone's motors from the mixed angular spectrum before summing. This known angular spectrum was computed from noise-only recordings with specific parameters of motor's current and speed. The tested drone had fifteen pairs of microphones and was tested in a semi-anechoic chamber with a reverberation-time of 20 ms at 20 dB. The drone was held resting at a height

of 1 m above a circle of twelve sound-sources with a radius of 0.6 m. The results showed acceptable accuracy of the proposed method compared to that of GCC-PHAT and of MUSIC for a SNR of at least -30 dB. The experiments were performed for multiple scenarios such as for a single source and for multiple.

Distance

The estimation of distance is not very common in TDOA-based methods. Most examples only localise the direction, whether it is two-dimensional or three-dimensional. As said before, the TDOA yields a hyperbola or hyperboloid as the locus of the source. Therefore, a naive method would be to find the intersection given at least two microphones. However, error and noise of course make this hard as well as the computation in a robotics context.

Another way is to use two or more sub-arrays that yield single DOAs and are used to triangulate on a position [26]. However, such a method would need sub-arrays that are far apart enough which is hard for a small robot where space is a luxury.

Reverberation

Reverberation is a big influence on the estimation of TDOA, especially in a room, since it leads to delayed reflections of the signal which show up in the GCC-PHAT. A statistical analysis [13] showed that the PHAT is the best estimator for the TDOA. The numerical examples in the paper showed good results as long as the reverberation-time was more than 0.07 s, the displacement between the microphones is longer than 0.2 m, and the distance between the source and the microphones is longer than 1.5 to 2 m. The same examples also showed that the probability of outliers was "tolerable" for a SRR more than 0 dB.

Another analysis [9] simulated three methods of estimating the TDOA, namely GCC-PHAT, GCC-ML, and Biweight. For all three, both the percentage of anomalies and the RMS error worsened for a reverberation-time longer than 0.1 s.

However, although GCC-PHAT generally handles reverberation well, it does not handle noise well across the spectrum since the algorithm gives all frequencies equal weight [31]. It especially does not do well with narrowband signals such as tones or voice.

Multiple Sources

Although the localisation of multiple sources has been explored for other methods such as beam-forming, it has not been explored much in TDOA-based methods. This is since most examples in the literature have been built on top of the basic idea of a single peak in the GCC-PHAT. However, the relationship between secondary peaks in the cross-correlation and other sources has been discussed.

Only a few basic studies in the fundamental relationship were done. A convention paper [12] proposed a method in the context of audio-engineering

where multiple sources manifested themselves as distinct peaks in the GCC-PHAT. The results showed an accuracy of at least 85 % where the SNR is at least 46.5 dB. However, the paper only considered the estimation of the TDOA in the context of musical instruments, not localisation, and the experiment was performed with little reverberation in mind. Another analysis [18] derived the mathematical cross-correlation function based on GCC-PHAT for multiple sources but did not seem to give much experimental data.

Even so, no literature examples were found that exploited this relationship of secondary peaks for localisation in robotics [26]. However, some studies have lately developed algorithms albeit not in a robotics context. A paper [10] proposed a method where an acoustic map, such as GCF or SRP-PHAT, is used to find the most dominant source which is then de-emphasised by lessening the GCC-PHAT at the time-delay corresponding to that source; this is repeated for the next most dominant source until all other located. However, the paper only tested this method for the near field.

Another paper [8] proposed a method of a delay density map made up of cubic subvolumes each of which weighted by a likelihood that a source is in it. The tested system had a resolution of 0.2 m and an RMS error at worst of about 0.6 m at an SNR of 5 dB. The experiments were tested for both the near field and the far field as well as for different arrays. The reverberation-times at 60 dB in such experiments were tested from 0.11 to 0.55 s. It also compared its own method with that of GCF-D (GCF De-emphasised) where theirs performed better.

Even given a method that only locates a single source at a time, one other way to find multiple sources is to cluster multiple estimated DOAs over many time-windows. One paper [25] does such by tracking sources with a threshold on angle and smoothing such tracking with a Kalman filter. An adaptive variation of the K-mean++ algorithm has also been used to cluster multiple source in an aforesaid paper [14].

2.1.3 Beam-forming

Beam-forming is a common technique in signal-processing and in many applications for both sound and radio. It is a method of setting many sensors or transmitters in such an array that waves at particular angles and areas constructively interfere to form a beam focused either on a chosen spot or in a chosen direction. Traditionally, it has been used in telecommunications so that multiple radio antennae form a beam that selectively transmits only in one direction. Here, for sound-source-localisation, it is used to steer a beam or a focus in a chosen direction or on a chosen spot so that the energy only from that location is measured. If the beam is steered in enough directions, then an energy map as a function of direction, e.g. azimuth, is made.

The simplest kind of beamforming is delay-and-sum beamforming where the signal from each microphone is delayed such that the overall sum of all the delayed signals corresponds to a steered direction [26]. This sum is maximum when it is steered towards the source. The delay-and-sum beamformer's output

steered at the position $\vec{r}_0$ for M microphones is given as:

$$y_{\vec{r}_0}[t] = \sum_{m=1}^{M} x_m[t - \tau(vecr_0)] \tag{2.7}$$

where x_n is the signal from the m-th microphone, and $\tau(\vec{r}_0)$ is the TDOA corresponding at the position $\vec{r}_0$ [3]. In the far field, this output can be simplified to a direction θ_0 instead of a position:

$$y_{\theta_0}[t] = \sum_{m=1}^{M} x_m[t - \tau(\theta_0)] \tag{2.8}$$

This output can yield an energy-map as:

$$E_{\vec{r}_0} = \sum_{t=1}^{T} y_{\vec{r}_0}[t]^2 \tag{2.9}$$

or

$$E_{\theta_0} = \sum_{t=1}^{T} y_{\theta_0}[t]^2 \tag{2.10}$$

A more general kind of beamforming is filter-and-sum beamforming where the signal from each microphone is filtered by its own linear filter rather than delayed [3]. The beamformer's output is given as:

$$y_{\vec{r}_0}[t] = \sum_{m=1}^{M} w_m(\vec{r}_0)[t] x_m[t] \tag{2.11}$$

where $w_m(\vec{r}_0)[t]$ is the impulse-response of the m-th linear filter.

One advantage of beamforming over basic localisation through TDOA is that multiple sources can be easily located since the method spatially filters, i.e. it only receives signals from a particular direction or space [26]. However, if two sources are near to each other enough, then the resolution of the energy-map may not distinguish the two.

However, there are few considerations [3]:

- The more microphones there are, the fewer side lobes or beams there are which show up beside the main lobe or beam.
- The further apart the microphones are, the narrower the beam is; this is particularly a problem for robotics where the room on a robot for an array is small.
- The beam is wider at lower frequencies; this affects the resolution and precision for locating sources of lower frequencies.
- Copies of the main lobe appear for high frequencies; this is form of spatial aliasing. A Shannon spatial sampling theorem is given as $d < c/(2f_{\max})$ [3].

Literature Examples

A robotic implementation was done in 2004 by Valin et al where the beamformer's energy was calculated in the frequency-domain using the cross-correlation weighted similarly to the authors' work with estimating the TDOA. The authors claim that the spectral whitening before computing the beamformer's energy helps narrow the peaks. Here, the authors also proposed a spherical search-grid of 2562 points where the beamformer's energy of each is computed. This grid yields a resolution of about 2.5 deg. The direction, i.e. both the azimuth and the elevation, of the loudest source is found when the beamformer's energy is maximum. Thereafter, the cross-correlation is zeroed for that loudest source, and the search is repeated for the next loudest source. This is done for a predicted number of sources. If there are fewer sources than the set number, then a source is falsely detected. To handle this, the authors employed probabilistic post-processing to temporally smooth the estimations. Furthermore, in their experiments, they applied two estimators working together, namely a short-term estimator for two sources and a medium-term one for four. The experiments were performed with with eight microphones in an open rectangular prism ($0.5 \times 0.4 \times 0.36$ m) on top of a mobile robot in "a noisy environment with moderate reverberation". The computation was done on a desktop PC and used about 30% of the CPU. Firstly, the detection-rate was tested against distance for three kinds of sounds, namely hands clapping, speech, and a burst of white noise 250 ms long. The system was able to detect these sounds reliably up to 5 m, but drops around 7 m. The authors claim that narrowband signals such as tones and speech are detected worse, whilst those with a broader band, such as noise, can be detected much better at longer distances. Furthermore, the estimated azimuth was accurate for four moving speakers although struggled to detect seven. Similar accuracies for both the azimuth and the elevation were given when the robot was moving instead; the authors claim that this demonstrates the robustness against the noise of the motors. Lastly, the array was shown to still be able work when it was not completely open.

The same authors later used the same beamforming method but with a particle filter instead [32]. It also involved a refined search after detecting a source that also estimated distance, but this estimated range was found to be too unreliable. It did however improve the accuracy of the direction in the near field. This new approach was tested on two different arrays on a different mobile robot. The first array, C1, was an open cube of eight microphones 15.5 cm wide, whilst the second, C2, was a closed square of four about 40 cm wide on the robot's chest. The experiments were tested in two different environments; the first environment, E1, was a medium-sized room with a reverberation time of 350 ms at - 60 dB, whilst the second, E2, was a hall with a reverberation-time of 1.0 s. In the first environment E1, the open array C1 detected sources more reliably than the closed array C2 within seven metres; C2 struggled to detect hand-claps specifically. Again, in E1, the RMS error for both the azimuth and the elevation was at worst 1.10 deg for C1 and at worst 1.44 deg. As before, experiments where either multiple sources were moving or the robot was moving

were tested in both E1 and E2 as well as where the trajectories of two sources intersect. In such results, the system was deemed to track successfully. Unlike the one before, this paper also discusses the computation needed, namely for the cross-correlation. For 1024 samples at 48 kHz, eight microphones, and 2562 searched directions, the complexity is claimed to be only 48.4 Mflops after counting all time-frequency transformations.

A further study [5] compared different variations and strategies of this kind of beamforming as well as the classic estimation of TDOA through GCC-PHAT. Here, TDOA-estimation from the peak of the GCC-PHAT (PEAK) was compared against a classic beamformer, called the steered-response-power (SRP) by the authors. Two variations were also considered, namely spectral weighting (SW) against SNR as proposed in £ and direction-refinement (DR) as also proposed in £. The latter is where a local search with a finer resolution is done after the initial search. Furthermore, two search grids are compared, namely a spherical rectangular grid (R) tessellated at 3600 points and a triangular element grid (T) of 2562 points. A cubical array of eight microphones with dimensions of 32 by 32 by 36 cm was tested. The experiments were done in a room with a reverberation-time of 0.1 s. The source playing pre-recorded sequences of speech was set at five points of different angle and distance as well as at two heights, one level with the robot and the other at the height of a human. The SNR received was varied by lowering the volume of the source. Two distinct experiments were done; in the first, only the background noise affected the accuracy which was observed to be Gaussian; in the second, a source of classical music as noise was set at only one of the tested positions. In the experiments, the mean error was measured against the SNR, and anomalies where the error was more than 10 deg were counted as a percentage. The triangular grid was deemed better than the rectangular one given that the latter's resolution was not uniform and more concentrated at the poles of the sphere. Overall, the SRP with the SW performed worse than that without. The authors explained that this might have been the difficulty in estimating the noise spectrum. Furthermore, the SRP with the SW had more anomalies. The SRP was found to be better than the classic PEAK estimator.

A more recent paper [28] proposed diagonal unloading on a beamformer so that the system could work with high noise. This diagonal unloading was done by subtracting a diagonal matrix from the covariance matrix of the array's signal. In the robust design, the covariance matrix was estimated by the largest eigenvalue of the array's signal which was computed by the power method. This robust design was evaluated in simulation and compared against the suboptimal design using diagonal unloading, SRP-PHAT, and MUSIC. The array was a uniform circle of eight microphones with a radius of 20 cm, and the spatial resolution was 5 deg. When there was a single source, the RMS error in angle of the proposed design was at most about 2.5 deg for an SNR of at least -10 dB and grew for worse SNR. When there were two sources, the RMS error was at most about 4 deg for an SNR of at least -5 dB. In all cases, the proposed robust design worked better than SRP-PHAT and suboptimal design and did just as well as MUSIC. Furthermore, despite working just as well as MUSIC,

the authors claimed that their proposed design has much simpler computation of $O(M^2)$ instead of that of $O(M^3)$ where M is the number of microphones.

The size of the array affects the beamformer. Firstly, the smaller it is, the wider the beam is especially for low frequencies. Secondly, the bigger it is, the more spatial aliasing there is, i.e. side lobes. Another recent study [33] proposed a design to fix these two problems by employing an array of two concentric uniform circles of eight microphones for each. The first part of the design was frequency-classification-processing (FCP). Here, the signal's spectrum was split into bands, a lower and an upper one. A beamforming output was computed for each frequency bin. The smaller circle used the upper band of frequencies, and the bigger circle used the lower one. The two beamforming outputs were then summed together. The second part of the design is weighting each beamforming output; the best set of weights were found using particle-swarm-optimisation (PSO). The experimental array was made up a smaller circle of eight microphones with a radius of 10 cm and a bigger circle of eight with a radius of 20 cm. The experiment was done in an anechoic chamber and with two loudspeakers playing sounds of vehicle horns. Only the azimuth was tested. The proposed design worked better than the generic beamformer and had an error of 2 deg at worst.

Robotics is not the only application for localisation by beamforming. Autonomous drones are often imagined with the same capabilities of sound-source-localisation. One such study [6] proposed SRP-PHAT but with a modified PHAT weighting with a power factor to mitigate the effect of noise. In order to localise multiple neighbouring drones, it also proposed a system inspired by [10] of pruning the next dominant source in the computed cross-correlation. The paper evaluated both a passive method of localising other drones' engine-sounds and an active method of localising other drones' beacons. The passive method which is the most relevant was tested indoors with a resting drone localising a flying drone. Motion-tracking was used to measure the true positions. The authors disclaimed that the sound of cooling fans belonging to eight tracking cameras and two computers could be heard in the room. For a single flying drone, the angular RMS error was 1.39 deg. For multiple flying drones, the resting drone could localise at most three, and the precision worsen dramatically for the fourth. The array used for the passive method was a flat T-shape of four microphones. The authors did not say what the exact dimension were. This study is of particular interest to robotics given that the system seemed to have been implemented on an embedded system on a drone, but the authors did not discuss computation, etc.

One paper [29] tested beamforming on two geometries, namely three rings of eight microphones and a larger ring of eight. The authors also proposed a method of separating sound-sources in frequency-domain, named frequency band selection (FBS), where two different directions of the beamformer are compared. However, this method works as long as the two sources do not overlap too much in the frequency-domain. Since the beam is wider at lower frequencies, a bandpass filter was applied between 1 and 3 kHz. In the experiments, the array of three rings was found to be more accurate than that of a single ring. The

error in the azimuth of the first array was less than 3 deg with one sound-source and less than 5 deg with two, whilst the error in elevation was less than 6 deg with one source. Also, the array of three rings was also able to estimate distance but only within one metre; the error of which was mostly within 300 mm.

A very basic implementation for tracking only the azimuth was tested on a mobile robot [21]. Here, the robot was fitted with eight microphones around it and was able to track a sound-source reliably for frequencies tested from 200 Hz to 1.2 kHz.

Frequency and Width

As said before, the beam's shape is affected by the frequency of the received sound. Especially, the lower the frequency is the wider the beam is. This is a problem undergone in [29] where the band of observed frequencies has to be narrow, between 1 and 3 kHz. A frequency-invariant broadband beamformer using convex optimisation is proposed by a group for the far field [2], [1] and for the near field [4].

2.1.4 MUSIC

One branch of methods used to analyse and locate sound-sources separates the space of signals into subspaces. The most common kind of which is multiple-signal-classification (MUSIC). This method has a very high resolution but generally has a high computational burden, especially compared to that of TDOA or beamforming. Therefore, it is not explored as in depth as before, but the general performance across the literature is given.

The basic idea behind MUSIC is that the signals received by an array of microphones is split into subspaces, each representing either a signal or a noise. The model for the signal received is:

$$X = W_s S + V \tag{2.12}$$

where

- each row of $X \in \mathbb{C}^{M \times F}$ is the received signal of the m-th microphone x_m in the frequency-domain, i.e. X_m,
- each row of $S \in \mathbb{C}^{N \times F}$ is the n-th source's signal s_n in the frequency-domain, i.e. S_n,
- each row of $V \in \mathbb{C}^{M \times F}$ is the noise of the m-th microphones in the frequency-domain,
- $W_s \in \mathbb{C}^{M \times S}$ is a weighting that models the TDOAs of each source at each microphone for a given direction or position,

and where M is the number of microphones, F is the number of frequency-points, N is the number of sources, i.e. signals [26]. The weighting is written

as:

$$W_s[f] = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{-2\pi f \tau_{2,1}} & e^{-2\pi f \tau_{2,2}} & \cdots & e^{-2\pi f \tau_{2,N}} \\ e^{-2\pi f \tau_{3,1}} & e^{-2\pi f \tau_{3,2}} & \cdots & e^{-2\pi f \tau_{3,N}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-2\pi f \tau_{M,1}} & e^{-2\pi f \tau_{M,2}} & \cdots & e^{-2\pi f \tau_{M,N}} \end{bmatrix} \tag{2.13}$$

where $\tau_{m,n}$ is the TDOA of the n-th source at the m-th microphone.

Given this, the overall goal of MUSIC is to split the space of signals into two subspaces, namely one for signals and another for noise. This is done by eigen-decomposition of the sampled covariance matrix $R[f] \in \mathbb{C}^{N \times N}$ for a given frequency f.

The most basic form of MUSIC, known as standard eigen-value-decomposition (SEVD), finds the eigen-decomposition of the covariance matrix as the following:

$$R[f] = Q[f]\Lambda[f]Q^{-1}[f] \tag{2.14}$$

where $\Lambda[f]$ is a diagonal matrix of the M eigenvalues $\lambda_m[f]$, and each column of $Q[f]$ is corresponding eigenvector $q_m[f]$. This matrix of eigenvectors is often split into two subspaces, $Q[f] = [Q_s[f]|Q_n[f]]$, the former for signals, and the latter for noise [26]. Thereby, the spatial spectrum is found from the orthogonality between the steered direction and the eigenvectors for noise:

$$P(\theta_0, \phi_0)[f] = \frac{|A^*(\theta_0, \phi_0)A(\theta_0, \phi_0)|}{\sum_{m=\tilde{N}+1}^{M} |A^*(\theta_0, \phi_0)q_m[f]|} \tag{2.15}$$

where $\tilde{N}$ is the number of sources considered, and $A(\theta_0, \phi_0) \in \mathbb{C}^{M \times 1}$ is the steering vector of transfer-functions at each microphone for a given three-dimensional direction*, i.e. an azimuth θ_0 and an elevation ϕ_0 [24]. More simply, each row is often the lag $e^{-2\pi f \tau_m}$ where τ_m is the TDOA at the m-th microphone corresponding to the given direction [26].

This spatial spectrum however is narrowband for one point or bin of frequency. For a broadband response, the narrowband response is averaged over the given band of frequencies [16] [24].

An extension of SEVD is general eigen-value-decomposition (GEVD) where the noise is whitened before the decomposition [22] [23] [24]. Here, the eigen-decomposition is such:

$$K^{-1}[f]R[f] = Q[f]\Lambda[f]Q^{-1}[f] \tag{2.16}$$

where $K[f]$ is a freely chosen matrix but is often computed as $N[f]N^*[f]$ where $N[f]$ is the frequency-domain noise recorded when there are no signals. This is shown to be more robust than SEVD for a SNR less than 0 dB.

*Much like beamforming, a full three-dimensional position can more generally be considered rather than only a direction, but again like before, this only works in the near field. Since the far field is much more relevant to a small array on a robot, only direction has been considered in this example.

Literature Examples

One paper [15] proposed a form of broadband SEVD-MUSIC where the output was the average of all the narrowband responses over a frequency-range. This was done for a small array of fourteen microphones around a robot's neck. This proposed system estimated both the azimuth and the elevation on a discrete spherical grid with a resolution of about 5 deg. Since MUSIC needs a known number of sources, the authors proposed a fixed number of sources for the narrowband MUSIC response and a maximum number of sources from the broadband response. They also compare the magnitude against a threshold to find whether the peak was a source or not. Multiple sources were nevertheless found by sequentially subtracting a two-dimensional Gaussian centred where the next highest source was; this approach is similar to others [10], [6]. The paper evaluated this system for a range of parameters, namely number of FFT points, frequency-range, and value of the threshold. The system was tested in a variety of "noisy environments", namely an office where the main sources of noise were an air-conditioner and the robot's hardware and an outdoor shopping mall, but the authors did not tell udner what SNR and reverberation the system was tested. Nevertheless, the authors found that the system could run in real time given 64 FFT points for each frame which was about 4 ms long. Furthermore, the authors found that a frequency-range from 1 to 6 kHz, a threshold of 1.7, a fixed number of sources of 2, and a maximum number of sources of 5 were best. In most of the experiments, the accuracy was around 80 % successful detection-rate.

Another kind of MUSIC is called GEVD-MUSIC which is more robust to noise, especially that more powerful than the wanted signal itself. An early paper [22] that first proposed this kind of MUSIC in the context of robotics compared the accuracy of GEVD-MUSIC as a percentage to that of SEVD-MUSIC. The accuracy of the latter dropped below an SNR of around 5 dB whilst the accuracy of GEVD-MUSIC kept at 100 % at an SNR of around -7 dB. In a later paper [23], many of the same authors proposed the same method but with a audio-visual integration with a particle filter for tracking inactive sources and hierarchical Gaussian mixture-models. Again, the accuracy of SEVD-MUSIC dropped at an SNR of around -8 dB, whilst that of GEVD-MUSIC dropped at an SNR of around -14 dB.

To ease the computation, the same authors proposed a new kind of MUSIC called GSVD-MUSIC [24]. To further lessen computation, the authors proposed a hierarchical search from coarse to fine, and to improve the resolution, they linearly interpolated the transfer-functions in both the frequency- and time-domain. The accuracy of the proposed GSVD-MUSIC only began to drop at an SNR fo around -10 dB, about 5 dB less than that of GEVD-MUSIC. The average error in the azimuth was at best about 1 deg and at worst about 10 deg. The proposed method also worked well for a moving source. The array was a circle of eight microphones embedded in the robot's head. The authors claimed that their new method of GSVD-MUSIC lessened the computational cost by 40.6 % compared to GEVD-MUSIC and that their design of a hierarchical search

lessened it by a further 59.2 % for a single source. However, the computation was done on a laptop, and no comparison was made with other methods of sound-source-localisation such as SRP-PHAT.

2.1.5 Discussion of Literature Examples

Many examples and methods in the literature have been discussed. Here, a summary of the most relevant examples is given. The following attributes are compared:

- the kind of method, i.e. one of three categories discussed so far, namely TDOA, beamforming, and MUSIC,
- the specific method,
- the result or output of the algorithm, i.e. the type of localisation whether it is the direction only or the full three-dimensional coordinates (note that some methods could estimate distance but only in the near field; here, only the result of the far field is considered),
- the number of sources detectable, often either single or multiple,
- the accuracy or precision,
- the built-in resolution of the method,
- the number of microphones in the array,
- the array's dimensions (given a complex array, only the widest displacement is given),
- the distances tested,
- the noise under which the method was tested or for which it is rated, often given as the SNR,
- the reverberation under which the method was tested,
- the sampling rate,
- the length of the frame, window, or block of samples (may be in either number of samples or simply time),
- any comments on the computation, e.g. whether the method runs in real-time, what the method was computed on, etc.,
- and any comments about the design or method.

Nearly all methods only compute the three-dimensional angle. Only two attempt full three-dimensional position, i.e. direction along with distance. However, only Chen & Xu in 2019 [11] seemed to attempt this in the far field for distances more than 1 m; Bechler et al. in 2004 [7] only have tested within about 1 m which may considered to be the near field given the dimensions of their array.

A common theme throughout the literature-review is the inconsistent calculation of accuracy. Some papers used the RMS error (RMSE), the mean-squared error (MSE), or simply the average error itself. Some papers did not even give error but rather the percentage of successful localisation relative to the resolution. Furthermore, it is hard to summarise the accuracy of a method given that it varies depending on noise, distance, etc. Here, the accuracy at best and at worst is given as lower and upper bounds. Nonetheless, most methods tended to have an error of 2 deg.

Although some dimensions were not given at all, e.g. Basiri et al. in 2016 [6], the smallest array seemed to be that of Hu et al. in 2009 [14]. The biggest was that of Valin et al. in 2003 and 2004 (both papers use the same array but different methods) [31] [30].

As for noise, Manamperi et al. in 2022 [20] tested up to the worst SNR of -30 dB although this paper was specifically about mitigating noise in the context of drones. Whilst for reverberation, Bechler et al. in 2004 [7] had the longest reverberation-time of 0.36 ms at 60 dB. Again however, many papers neglect any information about noise or reverberation, some albeit with vague descriptions.

Most methods could run in real-time albeit on laptop and desktop computers. However, Basiri et al. in 2016 [6] seemed to be the only example implemented on an embedded system on the drone itself, namely an Atmel AVR32 microcontroller. This example is therefore of particular interest to this project given the want for an embedded implementation.

Chapter 3

Simulation

Bibliography

[1] S. Argentieri, P. Danes, and P. Soueres. Prototyping Filter-Sum Beamformers for Sound Source Localization in Mobile Robotics. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 3551–3556, April 2005. ISSN: 1050-4729.

[2] S. Argentieri, P. Danes, P. Soueres, and P. Lacroix. An experimental testbed for sound source localization with mobile robots using optimized wideband beamformers. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2536–2541, August 2005. ISSN: 2153-0866.

[3] S. Argentieri, P. Danès, and P. Souères. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112, 2015.

[4] Sylvain Argentieri, Patrick Danes, and Philippe Soueres. Modal Analysis Based Beamforming for Nearfield or Farfield Speaker Localization in Robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 866–871, October 2006. ISSN: 2153-0866.

[5] Anthony Badali, Jean-Marc Valin, François Michaud, and Parham Aarabi. Evaluating real-time audio localization algorithms for artificial audition in robotics. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2033–2038, October 2009. ISSN: 2153-0866.

[6] Meysam Basiri, Felix Schill, Pedro Lima, and Dario Floreano. On-Board Relative Bearing Estimation for Teams of Drones Using Sound. *IEEE Robotics and Automation Letters*, 1(2):820–827, July 2016.

[7] D. Bechler, M.S. Schlosser, and K. Kroschel. System for robust 3D speaker tracking using microphone array measurements. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 3, pages 2117–2122 vol.3, September 2004.

[8] Ritu Boora and Sanjeev Kumar Dhull. A TDOA-based multiple source localization using delay density maps. *Sādhanā*, 45(1):204, August 2020.

[9] M.S. Brandstein and H.F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 375–378 vol.1, April 1997. ISSN: 1520-6149.

[10] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer. Multiple source localization based on acoustic map de-emphasis. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:1–17, 2010. Publisher: Springer.

[11] Guoliang Chen and Yang Xu. A Sound Source Localization Device Based on Rectangular Pyramid Structure for Mobile Robot. *Journal of Sensors*, 2019:1–13, August 2019.

[12] alice clifford and joshua reiss. calculating time delays of multiple active sources in live sound. *journal of the audio engineering society*, November 2010.

[13] T. Gustafsson, B.D. Rao, and M. Trivedi. Source localization in reverberant environments: modeling and statistical analysis. *IEEE Transactions on Speech and Audio Processing*, 11(6):791–803, November 2003.

[14] Jwu-Sheng Hu, Chia-Hsing Yang, and Cheng-Kang Wang. Estimation of sound source number and directions under a multi-source environment. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 181–186, October 2009. ISSN: 2153-0866.

[15] Carlos T. Ishi, Olivier Chatot, Hiroshi Ishiguro, and Norihiro Hagita. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2027–2032, October 2009. ISSN: 2153-0866.

[16] Carlos T. Ishi, Dong Liang, Hiroshi Ishiguro, and Norihiro Hagita. The effects of microphone array processing on pitch extraction in real noisy environments. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 550–555, September 2011. ISSN: 2153-0866.

[17] Cheol-Taek Kim, Tae-Yong Choi, ByongSuk Choi, and Ju-Jang Lee. Robust estimation of sound direction for robot interface. In *2008 IEEE International Conference on Robotics and Automation*, pages 3475–3480, May 2008. ISSN: 1050-4729.

[18] Byoungho Kwon, Youngjin Park, and Youn-sik Park. Analysis of the GCC-PHAT technique for multiple sources. In *ICCAS 2010*, pages 2070–2073, 2010.

[19] A. Mahajan and M. Walworth. 3D position sensing using the differences in the time-of-flights from a wave source to various receivers. *IEEE Transactions on Robotics and Automation*, 17(1):91–94, February 2001.

[20] Wageesha Manamperi, Thushara D. Abhayapala, Jihui Zhang, and Prasanga N. Samarasinghe. Drone Audition: Sound Source Localization Using On-Board Microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:508–519, 2022.

[21] L. Mattos and E. Grant. Passive sonar applications: target tracking and navigation of an autonomous robot. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, volume 5, pages 4265–4270 Vol.5, April 2004. ISSN: 1050-4729.

[22] Keisuke Nakamura, Kazuhiro Nakadai, Futoshi Asano, Yuji Hasegawa, and Hiroshi Tsujino. Intelligent sound source localization for dynamic environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 664–669, October 2009. ISSN: 2153-0866.

[23] Keisuke Nakamura, Kazuhiro Nakadai, Futoshi Asano, and Gökhan Ince. Intelligent Sound Source Localization and its application to multimodal human tracking. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 143–148, September 2011. ISSN: 2153-0866.

[24] Keisuke Nakamura, Kazuhiro Nakadai, and Gökhan Ince. Real-time super-resolution Sound Source Localization for robots. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 694–699, October 2012. ISSN: 2153-0866.

[25] Caleb Rascon, Gibran Fuentes-Pineda, and Ivan Meza. Lightweight multi-DOA tracking of mobile speech sources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015, December 2015.

[26] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017.

[27] Saulius Sakavičius, Artūras Serackis, and Vytautas Abromavičius. Multiple Sound Source Localization in Three Dimensions Using Convolutional Neural Networks and Clustering Based Post-Processing. *IEEE Access*, 10:125707–125722, 2022.

[28] Daniele Salvati, Carlo Drioli, and Gian Luca Foresti. Power Method for Robust Diagonal Unloading Localization Beamforming. *IEEE Signal Processing Letters*, 26(5):725–729, May 2019.

[29] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi. Three ring microphone array for 3D sound localization and separation for mobile robot audition. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4172–4177, August 2005. ISSN: 2153-0866.

[30] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat. Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach. In *IEEE International Conference on*

Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, volume 1, pages 1033–1038 Vol.1, April 2004. ISSN: 1050-4729.

[31] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau. Robust sound source localization using a microphone array on a mobile robot. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, volume 2, pages 1228–1233 vol.2, October 2003.

[32] Jean-Marc Valin, François Michaud, and Jean Rouat. Robust Localization and Tracking of Simultaneous Moving Sound Sources Using Beamforming and Particle Filtering. *Robotics and Autonomous Systems*, 55:216–228, March 2007.

[33] Yuting Zhang, Hongwei Zhang, and Honghai Liu. An Improved Multiple Sound Source Localization Method Using a Uniform Concentric Circular Microphone Array. In *2021 11th International Conference on Intelligent Control and Information Processing (ICICIP)*, pages 392–397, December 2021.