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

Clayton Carlon, C3327986

April 10, 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.

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 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.

*otherwise known as football or soccer

Chapter 2

Literature-Review

Since sound-source-localisation is a much widely researched problem, in e.g. robotics, 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.

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 [21]. 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 [20] 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 can be 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 $O\left(N \ln(N)\right)$ [24].

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], [20]. 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:

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

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

For example, one method [24] 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 bearing. 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 13.38 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 [10] 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 different 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 [6] 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 from 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 [15] 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 [17] 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 $\times$ 1000 mm and was done with ultrasonic pulses of 75 kHz. The authors do not reveal the geometry of the array and given how short the tested distances are, it may be assumed that the experiment was in the near field.

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 [20]. 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 [8] 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 [24]. 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 [11] 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 [16] 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 [20]. However, some studies have lately developed algorithms albeit not in a robotics context. A paper [9] 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 [7] 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 [19] 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 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. Traditionally, it has been used so that two antennae form a beam that selectively transmits only in one direction. Here, for sound-source-localisation, it is used to steer a beam or focus in chosen direction or point so that the energy is measured. If the beam is steered in enough directions, then an energy map as a function of direction, e.g. azimuth, is made.

One advantage of beam-forming over basic localisation through TDOA is that multiple sources can be localised since the method spatially filters.

A robotic implementation was done in 2004 by Valin et al where the beam-former'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 beam-former's energy helps narrow the peaks. Here, the authors also proposed a spherical search-grid of 2562 points where the beam-former'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 beam-former'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 beam-forming method but with a particle filter instead [25]. 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] compares different variations and strategies of this kind of beam-forming as well as the classic estimation of TDOA through GCC-PHAT. Here, TDOA-estimation from the peak of the GCC-PHAT (PEAK) is compared against a classic beam-former, called the steered-response-power (SRP) by the authors. Two variations are 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. 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.

One paper [22] tested beam-forming 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 beam-former 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.

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 [22] where the band of observed frequencies has to be narrow, between 1 and 3 kHz. A frequency-invariant broadband beam-former using convex optimisation is proposed by a group for the far field [2], [1] and for the near field

[4].

Despite the beam's width, a basic implementation for tracking only the azimuth was tested on a mobile robot [18]. 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.

2.1.4 MUSIC

Some methods separate the space of signals into subspaces. The most common 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 cross-correlation or beam-forming. Therefore, it is not explored as in depth as before, but the general performance across the literature is given.

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] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] Spandan Dey, Srinivas Boppu, and M. Sabarimalai Manikandan. Design of a Real-Time Automatic Source Monitoring Framework Based on Sound Source Localization. In *2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC)*, pages 35–40, May 2019.

[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] 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.

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

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.