Integrating Smooth Motion Assumptions with RANSAC-based Sound Source Localization

Jens Gulin[‡⋆]
jens.gulin@sony.com

Kalle Åström[†]
karl.astrom@math.lth.se

Amir Aminifar[‡]
amir.aminifar@eit.lth.se

[†] Centre for Mathematical Sciences, Lund University, Lund, Sweden
[‡] Department of Electrical and Information Technology, Lund University, Lund, Sweden
[⋆] Sony Europe B.V., Lund, Sweden

***Abstract*—The momentary localization of a single sound source, in an environment with microphones distributed at known positions, can be done with multilateration, using time-difference-of-arrival (TDOA) estimates. However, TDOA estimates from cross-correlation are noisy in real environments, and a robust multilateration method must handle outliers. Assuming constraints on the smoothness of movement over time, the location estimate can be improved for both stationary and moving sources. In this work, the smooth motion assumption is explored in different stages of a RANSAC-based (Random Sample Consensus) implementation. The evaluation is done on real recordings from the public LuViRA dataset, giving the first 3D baseline result on the dataset. Each of the proposed steps is shown to reduce the localization error compared to the benchmark method.**

***Index Terms*—TDOA, GCC-PHAT, smooth motion, RANSAC, SSL**

I. Introduction

Sound source localization (SSL) from time-difference-of-arrival (TDOA) estimates is a well-known problem with many applications, see [1] for an overview. Here, the sound source is a singular, moving speaker playing music, recorded by several stationary, synchronized microphones with known locations. The setup can be seen as a distributed acoustic sensor network, such as a smart room, but we will not cover the related calibration and communication problems. The SSL task is thus to continuously locate the sound source without access to the original sound, over time forming a 3D trajectory.

A possible computational pipeline for accurate localization first obtains TDOA estimates using cross-correlation (GCC-PHAT) [2], then applies the RANSAC (Random Sample Consensus) [3] estimation of the sound source position in each single time frame. RANSAC can find good parameters to a known model even with outliers present. The method randomly samples the *selection set* repeatedly to find the estimate that fits the most samples, using the *voting set* when measuring consensus. Even for outlier-resistant multilateration [1], the presence of reverberation and noise can introduce ghost source estimates. Due to that, the measurements may not form a Gaussian distribution, and tracking and smoothing are additional mechanisms to stabilize the estimates.

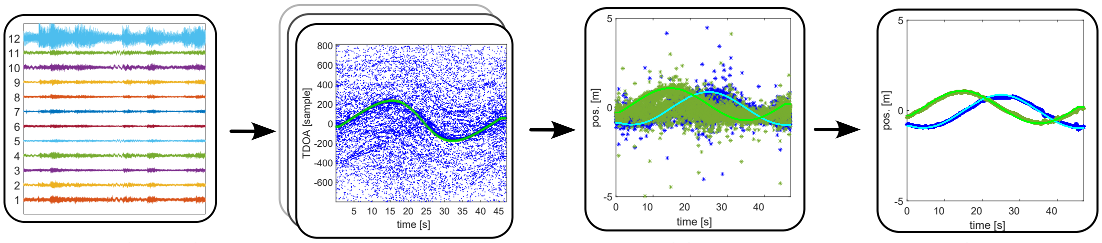
There are many related works focusing on other approaches to the problem, and we will only list a few. Sanitizing TDOA and grouping consistent measurements to aid multi-source estimations is introduced in [4] and later variants. When the microphones are in a relatively small and well-formed array, direction-of-arrival (DoA), is a simplification over full 3D localization. *Steered Response Power* approaches are robust [5], but the grid search makes them impractical for a 3D space. There are also a variety of machine learning methods, from TDOA segmentation [6] to full end-to-end solutions [7]. A comprehensive overview of the existing SSL techniques is provided in [8], [9].

A related task is object tracking, e.g. in 3D radar measurements. Multi-object tracking may be achievable through multi-hypothesis trackers [11]. The same idea can be useful to track and discard temporary ghost objects appearing from non-direct path TDOA in a single-source scenario. Tracking often depends on recognizable features, such as a visual appearance in a camera view, or a motion assumption. A Kalman filter is a way to track the likely position over time. Both the original and the extended Kalman filter assume a Gaussian model for the measurement error, as well as for the motion error [12]. Since this assumption is not always true, particle filters [13] and Gaussian mixture filters [14] can be useful to avoid that restriction. For sound source tracking, the context is again often reduced to DoA and microphone arrays, which is not directly applicable for our case.

Based on the available studies, we argue that the use of smoothness constraints in the context of 3D SSL is not well studied. Although RANSAC is frequently used to eliminate outliers within a time frame, it is rarely utilized to enforce temporal smoothness. We further consider *tracking* distinct from *smoothing*, as they can work together.

Our study aims to bridge this gap by leveraging the smoothness assumption on a realistic dataset, the challenging LuViRA dataset [15]. We examine how smooth motion constraints can improve SSL estimates and where they best integrate into the algorithm pipeline. To verify improvements, we extend the SFS2 framework [1], an open-source implementation of a multilateration algorithm already able to handle noise and reverberation fairly well. We use traditional (non-learning) methods because they are interpretable and do not rely on the availability of extensive training data.

ISBN: 978-9-46-459362-4



Input Synchronized microphone recordings

Extract TDOA estimates for each microphone pair

Use multilateration to calculate initial locations

Use initial estimates to calculate optimized locations

Fig. 1. The SFS2 pipeline illustrated for the RC2 trajectory in 2D. Used under CC-BY license from [10].

II. Background

A. Dataset

The LuViRA dataset [15] provides synchronized streams of video, radio and audio for a robot moving in an indoor area with no internal walls. Being an indoor use-case, noise and reverberation are apparent challenges. There are 11 stationary microphones that record the continuous sound from the moving sound source, mainly from a speaker playing back music, but also other noise in the room. The dataset provides extensive ground truth (GT) from a multi-camera motion tracking system, of which our experiments use the 3D positions. Because of how the microphones are spread out at different heights, the SSL is inherently a 3D localization problem. The speaker is mounted on top of the robot, approximately having a constant height, but we do not provide that information to the algorithm. Investigating the more general 3D problem is as well motivated by the slight height differences due to the sloping floor.

The robot movement trajectories are of different types and these experiments use the audio from the so-called ”random” trajectories. We will refer to the abbreviated names, e.g. RC1 and RC2 for the pre-programmed circular trajectories, RL1 and RU1 for the letter-shaped and RM1 to RM6 for the manually controlled, irregular trajectories. Note that RM1, RM2 and RM6 have people moving around in the measurement area, providing additional noise and line-of-sight interruptions. The RR1 and RR2 are excluded, since the audio played back on the speaker is not music, but chirps (as for the Grid trajectories). The RD2 is missing the processed ground truth and is discarded. The playback during RM6 faded after 48 seconds, so we use a shortened RM6x trajectory of that length. The single missing microphone recording of RM4 is replaced with silence to have a consistent setup.

B. Location estimation from TDOA estimates

The SFS2 framework[1] has a number of steps that can be combined to form an SSL pipeline. We will refer to these conceptual steps as A, B, C, etc. and describe them in more detail later. A pipeline is illustrated in Fig. 1: Input from 11 microphones (first panel) forms 55 unique pairs and TDOA is estimated using GCC-PHAT (second panel). The third panel (step A) shows the noisy results when robust multilateration combines the TDOA from all pairs for an initial location estimate for each time frame separately. The final panel shows the refined estimates (as x- and y-coordinates over time) after further localization improvement steps.

SFS2 has a background from [16], [17] and the *Structure from Sound* [18] concept. The Matlab implementation is a research platform for SSL and microphone self-calibration. The core structure of the pipeline is presented in [1], which focuses on developing a robust RANSAC scheme for initial estimates of 3D positions, through the selection of a set of inliers among the TDOA estimates. This is our step A, described in Section II-D below. In the reference, the outlier avoidance may have been limited to TDOA in the multilateration phase, but in our pipeline it is also a separate step B (Section III-B). The authors of [1] further consider improvements from smoothness constraints and suggest a nonlinear optimization (what we call C, see Section III-C).

Our implementation of the ABC pipeline was already used for a 2D benchmark on LuViRA data in [10]. However, the specific algorithm was not presented, nor was the context to motivate that smoothing, and we provide that contribution here. We will in addition extend the pipeline with a novel linear motion optimization (D) and experiment with TDOA-agnostic filtering (O). We study different combinations of steps and validate the effectiveness of such combinations on real data. The ABC method is taken as the baseline for this work.

C. Estimating TDOA

The pipeline processes synchronized audio streams from a fixed number of microphones (m). The audio is sectioned into n frames of 2048 samples, allowing overlap by starting a new frame every 960 samples. For each microphone pair GCC-PHAT provides a moving correlation measure, which estimates the audio time-difference of the pair. This lag is the TDOA if there is a single sound source and no reverberation. In practice, the maximum correlation will not always match the true TDOA. SFS2 intends to be resilient to such noise, by keeping several peaks from each pair as the possible TDOA for the time frame.

We will refer to the set of TDOA estimates for the sequence of frames as u. Each frame takes the four largest peaks as putative (noisy) measurements of the real TDOA, e.g. $\tau_{ij}(t_k, p) \in u$ being the pth peak ($p = 1, \ldots, 4$) of the kth time frame ($k = 1, \ldots, n$) for the microphone pair (i, j). Note

[1] https://github.com/kalleastrom/StructureFromSound2

that we expect only one estimate per frame to be the best one, and that it may still be slightly off from the true TDOA. For certain frames none of the estimates will be within an acceptable margin, and the remaining estimates are outliers.

D. Single-frame multilateration (A)

The audio processing and TDOA estimation can be considered an initial precondition. The next step is to combine the TDOA into the most likely 3D location for the window. This is known as multilateration, in SFS2 using the RANSAC paradigm where all microphone pairs are joined to vote for the best location. A vote is cast when a proposed position would match the TDOA of that pair within a threshold, using the known microphone positions and direct path geometry. The selection part could propose (random) positions directly, but in SFS2 the idea is to use a "minimal solver": From a random selection of three TDOA estimates, eight possible source positions can be quickly calculated [1]. This is again based on the direct path geometry and akin to solving a system of polynomial equations in a complex space. Three TDOA estimates cannot pinpoint a single solution and each RANSAC selection gives eight positions to vote for. Repeating the selection step several times, the goal is to find the true position, with the assumption that it is the most voted for location estimate. With random selection from a TDOA set with mostly outliers, many of the solutions are not valid. Mathematically, complex solutions could be discarded as invalid, but due to measurement noise and arithmetic imprecision, the current implementation keeps the real part as a safeguard and trusts voting to eliminate the invalid selections. If the number of votes for any position is below a set threshold, the estimate for that frame is considered invalid.

The output from A is (at most) a single 3D position per frame, a sequence we will refer to as s. The putative TDOA estimates u and the microphone positions are kept for use in later steps. Method A can be used to estimate the microphone positions, but we will read the stationary positions per trajectory from GT.

E. Local bundle optimization (L)

The estimate for the frame may be adjusted to better fit the TDOA inliers (those that voted) through least-squares minimization. This is a brief iterative local optimization (Equation (7) of [1]), which is not considered a separate step, but is included in other steps.

III. Smooth motion assumptions

Sound source motion is often smooth. Commonly used motion models are *stationary* and *constant speed*, with an added noise term that allows slow changes in e.g. position, speed or acceleration. Although a loss function for optimizing 3D positions can be derived from a motion and measurement model, solving the optimization problem is difficult, especially when finding the initial estimate.

Our way to make use of the smooth motion assumption is instead to assign the selection set and voting set from a wider window of time frames. A traditional zero-size window means that the minimal initialization is taken only from the current time frame, and only the current time frame gets to vote. In a low velocity setting, the true source location of adjacent time frames is likely very close to the true location for the current frame, thus the true TDOA peak of this frame should also be close to the TDOA of adjacent frames. Allowing a selection window of "size" 25 includes the current frame as well as 25 frames before and 25 frames after. With an expected inlier ratio of less than one in four, it does not increase the *chance* of a successful random selection, but the higher *amount* may still make it worthwhile as long as enough iterations are allowed. In addition, with the assumption that outliers are mostly random, increasing the voting window is directly influential to get a majority vote on a proper selection. This is in line with the promise of RANSAC. The same argument holds for higher velocity or low acceleration if the window size is small.

A. Widened multilateration (A')

Step A uses only the current time for both selection and voting. Would wider windows be worthwhile? The initial estimate is important to get right, but it is hard to apply any linear motion constraint on TDOA, since each pair sees a separate motion characteristic. It is likely more efficient to do so already in the TDOA detection step.

It is possible to increase the voting window, assuming a low velocity motion. It is still important to take many samples to get a fair coverage, and evaluating each hypothesis becomes slow with a large window. Preliminary analysis shows that the outlier avoidance in step B gives good results with the large voting window, even from a noisy initial estimate, so we continue to use the unmodified A.

B. Outlier avoidance (B)

We suggest step B as an outlier remedy, using the low velocity assumption in two ways. First, we increase the voting window, so that there is a stronger chance of votes in the vicinity of the true location. Second, the RANSAC selection is taken directly from the 3D estimates s and thus skips TDOA outliers and extraneous solutions from the minimal solver. With only one hypothesis per frame, even a large selection window has a low runtime penalty.

Two conditions control the results of this step, the availability of "good estimates" in the neighborhood and "proper voting" within the neighborhood. Good estimates means that at least some adjacent frame (including the current frame, of course) already has an estimate close to the GT of this frame. Proper voting requires that consensus approves the correct estimate, i.e. enough TDOA in the voting set matches GT and any other misguided voting cluster has fewer votes. There are no guarantees that B always has these beneficial conditions, but the conditions are likely, and empirically, using medium size windows shows promising results.

The current implementation simply copies the highest vote hypothesis to the current frame, which may cause clumping of estimates. It would be possible to instead infer a linear motion

TABLE I
RESULTS FOR THE MAIN METHODS ON EACH TRAJECTORY (ROWS). NOTE THAT THE PROPOSED METHOD (ABCD) CONSISTENTLY OUTPERFORMS THE OTHERS. THE RESULTS ARE BASED ON THE ABSOLUTE ERROR (3D DISTANCE IN CM) FROM GT TO THE ESTIMATED LOCATION.

Traj-	**GT**	**Method: A**			**Method: ABC**			**Method: ABCD**		
ectory	**# samples**	**Mean**	**SD**	**Median**	**Mean**	**SD**	**Median**	**Mean**	**SD**	**Median**
RC1	4691	107^{c}	85.6^{c}	92.6^{c}	21.4	**24.8**	11.0	**20.0**	25.2	**9.47**
RC2	4701	90.8^{b}	78.4^{b}	68.1^{b}	11.2	4.38	**10.7**	**11.0**	**2.80**	10.8
RD1	5002	133^{c}	101^{c}	123^{c}	30.1	39.8	10.2	**28.6**	**38.8**	**8.97**
RL1	2707	115^{c}	81.4^{c}	113^{c}	8.52	4.89	**6.90**	**8.45**	**3.57**	7.52
RM1	5532	100^{b}	84.3^{b}	79.9^{b}	34.9	45.0	12.8	**33.5**	**43.7**	**11.6**
RM2	5448	135^{c}	90.8^{c}	129^{c}	39.2	**41.9**	23.6	**36.1**	44.0	**18.2**
RM3	5265	73.3^{c}	70.3^{c}	49.9^{c}	10.2	8.24	7.64	**8.06**	**4.37**	**7.12**
RM4	4655	147^{c}	94.1^{c}	149^{c}	44.7	77.7	13.2	**40.4**^{a}	**69.6**^{a}	**11.8**^{a}
RM5	5563	149^{c}	**94.8**^{c}	154^{c}	57.3^{d}	114^{d}	14.2^{d}	**54.3**^{d}	111^{d}	**11.2**^{d}
RM6x	4799	119^{e}	98.6^{e}	96.6^{e}	25.4	17.9	18.1	**22.4**	**15.2**	**16.4**
RU1	5560	125^{c}	87.5^{c}	119^{c}	43.1	60.2	11.6	**40.2**	**57.5**	**10.0**
Total	53923	118^{c}	-	-	31.1^{b}	-	-	**28.9**^{b}	-	-

Invalid estimates excluded at the following rate: ^{a} 0.2%, ^{b} 1-2%, ^{c} 2-4%, ^{d} 6-7%, ^{e}10%.

TABLE II
RESULTS FOR THE ALTERNATIVE METHODS. NOTE THAT ALTHOUGH THE ABOCDO METHOD IS NOT CONSISTENTLY BETTER THAN ABCD, THE TOTAL SHOWS AN MAE IMPROVEMENT NEAR 15%. THE LOWEST VALUE FOR EACH METRIC IS MARKED IN BOLD, TAKING TABLE I INTO ACCOUNT.

Traj-	**Method: ABCDO**			**Method: ABOCDO**		
ectory	**Mean**	**SD**	**Median**	**Mean**	**SD**	**Median**
RC1	19.9	25.1	9.39	**10.3**	**5.03**	**9.34**
RC2	**10.9**	**2.76**	**10.6**	11.2	2.89	10.9
RD1	**28.5**	**38.8**	8.95	42.1	63.2	**8.94**
RL1	**8.36**	**3.51**	7.43	12.5	13.3	7.68
RM1	33.4	43.5	11.6	**19.9**	**22.7**	**11.1**
RM2	36.0	43.9	**18.1**	**24.4**	**16.3**	18.9
RM3	7.95	4.22	7.12	**7.88**	**3.76**	**7.03**
RM4	41.2	72.4	**11.5**	42.9	73.2	12.9
RM5	53.8^{d}	111^{d}	11.0^{d}	**28.8**	**47.0**	**10.9**
RM6x	**22.2**	**15.1**	16.1	22.6	24.7	**13.2**
RU1	**40.1**	57.9	**9.86**	41.4	61.3	10.8
Total	28.8^{b}	-	-	**24.6**	-	-

Invalid estimates excluded at the following rate: ^{b} 1-2%, ^{d} 6-7%.

assumption on the re-estimate, but the added value is not apparent for this dataset. Instead, the local bundle optimization (L) is given a chance to move estimates before the next step.

C. Smoothness optimization (C)

Step C does not use RANSAC. Using s for each frame, the inlier votes from u are calculated and the TDOA outliers discarded. The remaining TDOA and s are used in the smoothness optimization proposed in [1] (Equation (14)). It penalizes any rapid velocity changes over s and large measurement errors on the TDOA inliers.

D. Linear motion optimization (D)

In step D, we actually assume a linearization of the motion in a time window. The novel method makes the voting set optimize for a locally linear motion, iteratively starting from a zero-velocity assumption. The current time is naturally included in the voting set and the residual at the minima indicates the quality of the solution. With the current estimate as a starting point, a new estimate is found for the source location at the current time. Setting a larger (non-zero) selection window would allow taking the initial starting point from adjacent times instead, but preliminary results indicate that it is not needed. With a large voting set, a good linear fit should be relatively independent of a single point.

E. Standard filter method (O)

Considering the estimated locations (denoted s) as a sequence of points, they can be smoothed with standard filter methods without regard to the underlying TDOA measures. The approach has potential worst-case pitfalls, but it is worth evaluating as it should be fine for well-behaved data without too many outliers. We denote the filter O, to signal the difference from the A to D steps.

The outlier filter (O) is the Matlab function *filloutliers* with a linear replacement strategy and the *movmedian* outlier detector with a low *ThresholdFactor*. The effect is like a median filter, giving a new estimate in frame k as the median of the set $\{s(i)\ \forall\ k-v \leq i \leq k+v\}$ for a voting window of size v. The function operates on each dimension separately.

IV. EXPERIMENTS AND RESULTS

In Table I we show the results of the experimental validation for the selected trajectories (see Section II-A) from the LuViRA dataset. Each method is denoted as a string of individual letters, corresponding to the pipeline steps used in that order. For each method, we evaluate the result as the absolute error (in cm) of the estimated 3D positions. The metrics reported are mean (MAE) and median, as well as the standard deviation (SD), of the absolute error distribution for each trajectory. We also report the percentage of frames that fail to give valid results, mostly a problem for method A without any additional smoothing. The discussion will focus on the mean, but it is worth noting that the median is often far from the mean, indicating that a Gaussian distribution is not a good fit and the larger SD suffer from outliers.

As expected, the Table I results of A improve significantly with smoothing. The proposed ABCD method outperforms even ABC, with an improvement of about 7% in total. The median error is now mostly within 15 cm, but the improvement in mean and SD is limited on the trajectories that ABC already struggles with. When there are many outliers, i.e. few estimates near the target, for a length of time, the smooth motion assumption may instead lock on to the outliers.

Two alternative methods are shown in Table II, relying on median smoothing (O) without consideration to the original TDOA estimates. Adding this step at the end, method ABCDO

TABLE III

RESULTS ON RM3 FOR DIFFERENT METHOD SEQUENCES. THE RESULTS ARE BASED ON THE 3D DISTANCE (IN CM) FROM GT TO THE ESTIMATE.

Method	Mean	SD	Median	<5 cm	<15cm
A	73.3^a	70.3^a	49.9^a	1.8%	17.5%
AB	12.8	15.7	8.20	16.4%	81.3%
ABC	10.2	8.24	7.64	24.9%	81.6%
ABCD	8.06	4.37	7.12	22.6%	95.0%
AO	26.6	19.6	20.5	1.8%	33.0%
ABO	10.5	10.8	7.07	19.9%	85.4%
ABCO	10.2	8.24	7.57	**25.2%**	81.7%
ABCDO	7.95	4.22	7.12	22.9%	**95.3%**
AOB	24.1	19.7	16.8	4.3%	43.9%
ABOC	9.44	6.13	7.36	18.4%	85.6%
ABCOD	8.06	4.36	7.12	22.7%	95.0%
ABOCD	8.02	4.07	**6.99**	22.3%	93.2%
ABOCDO	**7.88**	**3.76**	7.03	23.0%	94.2%

Invalid estimates excluded at the following rate: ^{a}2%.

results are very similar to ABCD, showing as expected that the estimates are smooth among themselves. The ABOCDO method, on the other hand, provides yet another 15% improvement on the total.

By adding the O step already after B, the smoothness of the trajectory is enforced earlier, disrupting the misdirection of the estimated TDOA. A successful example is the RC1 trajectory, where the MAE is halved and the SD considerably reduced. Since the median remains largely unchanged, it is evident that the improvement comes from outliers joining the majority. In this case, the following steps (C and D) were able to benefit from the better initial estimate and find the appropriate context in the TDOA estimates. This pattern also follows, to a lesser extent, for other trajectories with weak results for ABCD. The RM5 trajectory still shows a high SD, but it should be noted that ABOCDO overcomes the 6-7% invalid estimates discarded in the other methods. There are no invalid estimates with this method. Table II also shows that the ABOCDO method performs worse on certain trajectories. In particular, RD1 and RL1 deteriorate in both MAE and SD. Further analysis is needed to find a pipeline that can build on the TDOA-agnostic smoothing without this problem.

Ablation experiments (see Table III) outline the effect of O in different positions, for one of the trajectories. The first group of rows confirms that each step improves MAE, as well as SD, without O. The next group evaluates standard filtering, and although O is promising, it cannot replace any of the other steps, e.g. AB is still better than AO. The possibility of using O as an intermediate step is evaluated in group three. AOB is an improvement over AO, but not better than AB. ABCOD is discarded for a similar reason. The exception is ABOC, which improves on both ABC and ABCO. The last group shows that the ABOCD is not better than the equivalent ABCDO, while ABOCDO gives a very promising result on this trajectory. As discussed above, Table II indicates that the same is not true for all trajectories.

V. CONCLUSIONS

In this paper, we have presented several components using motion assumptions for 3D localization from TDOA estimates. We have investigated how these components can be combined in different sequences to form robust systems and evaluated how they perform on real data with ground truth estimates. The experiments show that there are several methods that improve over the benchmark. The idea of using RANSAC on larger time windows and allowing TDOA to influence the smoothing step seems to be fruitful. Future work to configure the pipeline without reference to ground truth may need an objective smoothness metric. We aim to extend these methods for broader use in self-calibration and motion estimation.

REFERENCES

[1] K. Åström, M. Larsson, G. Flood, and M. Oskarsson, "Extension of time-difference-of-arrival self calibration solutions using robust multilateration," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 870–874.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320 – 327, aug 1976.

[3] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[4] J. Scheuing and Y. Bin, "Disambiguation of TDOA estimates in multi-path multi-source environments (DATEMM)," in *International Conference on Acoustics Speech and Signal Processing*, vol. 4. IEEE, 2006, pp. 837–840.

[5] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, 2011.

[6] J. Gulin and K. Åström, "GCC-PHAT re-imagined - a U-Net filter for audio TDOA peak-selection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 8806–8810.

[7] A. Berg *et al.*, "wav2pos: Sound source localization using masked autoencoders," in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, Oct. 2024, p. 1–8.

[8] G. Jekateryńczuk and Z. Piotrowski, "A survey of sound source localization and detection methods and their applications," *Sensors*, vol. 24, p. 68, 12 2023.

[9] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 07 2022.

[10] I. Yaman *et al.*, "LuViRA dataset validation and discussion: Comparing vision, radio, and audio sensors for indoor localization," *IEEE Journal of Indoor and Seamless Positioning and Navigation*, pp. 1–11, 2024.

[11] S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, 2004.

[12] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," in *Defense, Security, and Sensing*, 1997.

[13] C. Hue, J.-P. Le Cadre, and P. Perez, "Tracking multiple objects with particle filtering," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 791–812, 2002.

[14] Y. Oualil, F. Faubel, and D. Klakow, "A multiple hypothesis gaussian mixture filter for acoustic source localization and tracking," in *International Workshop on Acoustic Echo and Noise Control*, 09 2012.

[15] I. Yaman *et al.*, "The LuViRA dataset: Synchronized vision, radio, and audio sensors for indoor localization," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 920–11 926.

[16] K. Batstone *et al.*, "Robust self-calibration of constant offset time-difference-of-arrival," in *International conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[17] M. Larsson, G. Flood, M. Oskarsson, and K. Åström, "Fast and robust stratified self-calibration using time-difference-of-arrival measurements," in *International conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[18] S. Thrun, "Affine structure from sound," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2005.