The LOCATA Challenge Data Corpus for Acoustic Source Localization and Tracking

Heinrich W. Löllmann$^{1)}$, Christine Evers$^{2)}$, Alexander Schmidt$^{1)}$, Heinrich Mellmann$^{3)}$,
Hendrik Barfuss$^{1)}$, Patrick A. Naylor$^{2)}$, and Walter Kellermann$^{1)}$

$^{1)}$Friedrich-Alexander University Erlangen-Nürnberg, $^{2)}$Imperial College London,
$^{3)}$Humboldt-Universität zu Berlin

***Abstract*—Algorithms for acoustic source localization and tracking are essential for a wide range of applications such as personal assistants, smart homes, tele-conferencing systems, hearing aids, or autonomous systems. Numerous algorithms have been proposed for this purpose which, however, are not evaluated and compared against each other by using a common database so far. The IEEE-AASP Challenge on sound source localization and tracking (LOCATA) provides a novel, comprehensive data corpus for the objective benchmarking of state-of-the-art algorithms on sound source localization and tracking. The data corpus comprises six tasks ranging from the localization of a single static sound source with a static microphone array to the tracking of multiple moving speakers with a moving microphone array. It contains real-world multichannel audio recordings, obtained by hearing aids, microphones integrated in a robot head, a planar and a spherical microphone array in an enclosed acoustic environment as well as positional information about the involved arrays and sound sources represented by moving human talkers or static loudspeakers.**

I. Introduction

Acoustic source localization and tracking equip machines with positional information about nearby sound sources required for applications such as tele-conferencing systems, smart environments, hearing aids, or humanoid robots (see e.g., [1–5]). Instantaneous estimates of the source Direction Of Arrival (DOA), independent of information acquired in the past, can be obtained with at least two microphones using, e.g., the Generalized Cross-Correlation (GCC) Phase Transform (PHAT) [6], Steered Response Power (SRP) PHAT [2, 7], subspace-based approaches and beamsteering [8–10], adaptive filtering [11], Independent Component Analysis (ICA)-based approaches [12, 13] or localization in the Spherical Harmonics (SH)-domain [14, 15]. Smoothed trajectories of the source positional information can be obtained from the instantaneous DOA estimates using acoustic source tracking approaches. Kalman filter variants and particle filters are applied in, e.g., [1, 16] for tracking of a single moving sound source. Multiple moving sources are tracked from Time Delay of Arrival (TDOA) estimates using Probability Hypothesis Density (PHD) filters in [17]. Using a moving microphone array, the 3D source positions are probabilistically triangulated from 2D DOA estimates in [18, 19], and are tracked directly from the acoustic signals without the need of DOA or TDOA extraction in [20]. Moreover, acoustic Simultaneous Localization And Mapping (SLAM) [19, 21] equips autonomous machines, such as robots, with the ability to localize the machine's position and orientation within the environment whilst jointly tracking the 3D positions of nearby sound sources.

The evaluation of localization and tracking approaches is mostly conducted with simulated data where reverberant enclosures are commonly simulated by means of the image-method [22] or its variants [23]. An additional evaluation of such algorithms with real-world data seems appropriate to demonstrate their practicality. Such an evaluation of localization algorithms for a fixed array and speaker position can be found in, e.g., [2, 24, 25]. In [16, 26], tracking algorithms are evaluated by measured data for a single moving speaker. However, such evaluation results can hardly be compared with those for other algorithms since no common publicly available database is used. Moreover, information on the accuracy of the ground-truth position data is often not provided or lies in a range of several centimeters, e.g., [16].

More recently, the single- and multichannel audio recordings database (SMARD) was published [27]. The recordings were conducted in a low-reverberant room ($T_{60} = 0.15\,\mathrm{s}$) using different microphone arrays and loudspeakers which played back either artificial sounds, music or speech signals. However, this database considers only a single source scenario and microphone arrays and loudspeakers at fixed positions.

This paper presents a novel, open-access data corpus for acoustic source localization and tracking that i) provides audio recordings in a real acoustic environment using four different microphone arrays for a variety of scenarios encountered in practice, ii) involves static loudspeakers, moving human talkers, and microphone arrays installed on a static as well as a moving platform, and iii) includes ground-truth positional data of all microphones and sources with an accuracy of less than $1\,\mathrm{cm}$. The data corpus is released as part of the IEEE Audio and Acoustic Signal Processing (AASP) Challenge on acoustic source *LOCalization And TrAcking* (LOCATA).

II. The LOCATA Challenge

The scope of the LOCATA Challenge is to objectively benchmark state-of-the-art localization and tracking algorithms using one common, open-access data corpus of scenarios typically encountered in speech and acoustic signal processing applications. The offered challenge tasks are the localization and/or tracking of:

- **Task 1**: A single, static loudspeaker using a static microphone array
- **Task 2**: Multiple static loudspeakers using a static microphone array
- **Task 3**: A single, moving talker using a static microphone array
- **Task 4**: Multiple moving talkers using a static microphone array
- **Task 5**: A single, moving talker using a moving microphone array
- **Task 6**: Multiple moving talkers using a moving microphone array.

Similar to previous IEEE-AASP challenges, such as CHIME [28] or ACE [29], the data corpus is divided into a development and evaluation database. The development database contains three recordings for each of the tasks and each of the four microphone arrays described later, i.e., 72 recordings in total. The development database should enable participants of the challenge to develop and tune their algorithms. Ground-truth data of the position and orientation for all microphone arrays and sound sources is therefore provided. The evaluation database contains the ground-truth positional information for all microphone arrays, but not the sound sources. For Task 1 and 2, it comprises 13 recordings for each microphone configuration and task and 5 recordings per task and array otherwise, i.e., 184 recordings in total.

Upon completion of the LOCATA Challenge, the full data corpus containing the ground-truth positional information for all scenarios will be released. Further information about the challenge can be found in [30].

III. Data Corpus

The recordings for the LOCATA data corpus were conducted in the computing laboratory of the Department of Computer Science at the Humboldt University Berlin. This room with dimensions of about $7.1\,\mathrm{m}\times 9.8\,\mathrm{m}\times 3\,\mathrm{m}$ is equipped with the optical tracking system OptiTrack [31], which is typically used to track the positions of robots deployed for the soccer competition RoboCup.

A. Microphone Arrays

Four different microphone arrays (see Fig. 1) were used for the recordings to emulate scenarios typically encountered in speech signal processing applications, such as smart environments, hearing aids or robot audition.

- *DICIT array*: A planar array with 15 microphones which includes four nested linear uniform sub-arrays with microphone spacings of 4, 8, 16 and 32 cm. The array has a length of $2.24\,\mathrm{m}$ and a height of $0.32\,\mathrm{m}$, and has been developed as part of the EU-funded project "Distant talking Interfaces for Control of Interactive TV (DICIT)", cf., [32].
- *Eigenmike*: The em32 Eigenmike$^{\circledR}$of the manufacturer mh acoustics is a spherical microphone array with 32 microphones and a diameter of $84\,\mathrm{mm}$ [33].

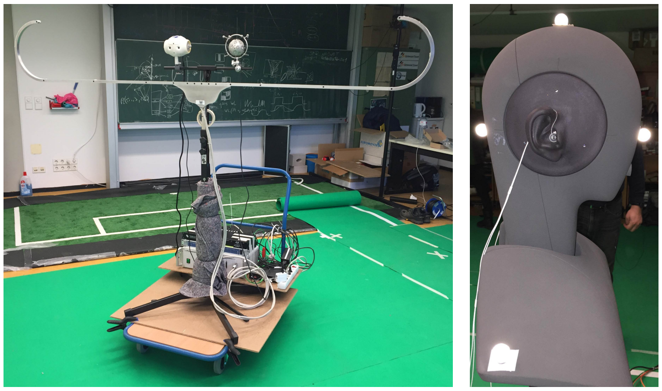


Figure 1. Recording environment and used microphone arrays with markers.

- *Robot head*: A pseudo-spherical array with 12 microphones integrated in a prototype head for the humanoid robot NAO. This prototype head was developed as part of the EU-funded project "Embodied Audition for Robots (EARS)", cf., [34, 35].
- *Hearing aids*: A pair of hearing aid dummies (Siemens Signia, type Pure 7mi) mounted on a dummy head (HMS II of HeadAcoustics). Each hearing aid dummy is equipped with two microphones (Sonion, type 50GC30-MP2) at a distance of $9\,\mathrm{mm}$, and the spacing of both hearing aid dummies amounts to $157\,\mathrm{mm}$.

The multichannel recordings ($f_\mathrm{s} = 48\,\mathrm{kHz}$) were synchronized with the ground-truth positional data acquired by an optical tracking system (see Sec. III-C). The recordings were conducted in a real acoustic environment and were hence subject to room reverberation ($T_{60} = 0.55\,\mathrm{s}$) and noise, including measurement and ambient noise. A detailed description of the array configurations and recording conditions is given by [36].

B. Speech Material

For the scenarios involving static sound sources, sentences of the CSTR VCTK1 database [37], downsampled to $48\,\mathrm{kHz}$, were played back by loudspeakers. For the scenarios involving moving sound sources, randomly selected sentences of the CSTR VCTK1 database were read live by 5 non-native moving human talkers, equipped with microphones near their mouths to record the close-talking speech signals. The source signals are provided as part of the development dataset, but not the evaluation dataset.

C. Ground-Truth Position Data

The positions and orientations of the arrays and sound sources were determined by the optical tracking system OptiTrac [31], equipped with 10 synchronized infra-red cameras (type Flex 13) and positioned along the perimeter of a $4\,\mathrm{m}\times 6\,\mathrm{m}$ recording area within the acoustic enclosure. The optical tracking system provides position estimates at a frame rate of $120\,\mathrm{Hz}$ and an error of less than $1\,\mathrm{mm}$ as per manufacturer specification [31]. The optical tracking system uses reflective markers for localizing objects, i.e., the microphone arrays and sound sources for LOCATA (see

Fig. 1), by optical cameras. Multiple markers were attached to each object, forming marker groups – or trackables – used to determine the orientation and position of the objects over time. The camera system determines the marker positions by triangulation. The position estimates were labeled with time stamps to synchronize it with the audio recordings with an accuracy of approximately ±1 ms.

The microphone positions were obtained from the individual marker positions of each trackable based on models derived from caliper measurements and technical drawings of the microphone configuration. Each model contains the marker positions of each trackable and the microphone positions w.r.t. the local coordinate system (local reference frame) of the object (trackable). The origin and orientation of the local coordinate system for the arrays, for example, are given, by their physical center and 'look direction', respectively. An exact specification for all microphone arrays and sound sources is provided by [36].

For convenient transformations of coordinates between the global and local reference frames, the data corpus provides the positions, translation vectors and rotation matrices for all sound sources and arrays for each time stamp of the ground-truth data. Moreover, the microphone positions are provided relative to the global reference frame for each array.

Reflections of the infra-red light emitted by the OptiTrack system on the surfaces of the objects could cause the detection of 'ghost markers' or missing detections. In addition, some markers were occasionally occluded during the recordings with moving objects. These effects led in isolated instances to outliers for the position and orientation estimates which were replaced by reconstructed and interpolated values. The calculation of the Mean-Square Error (MSE) between the unprocessed and processed marker positions led to values of less than 1 cm.

IV. Baseline Results

Baseline results obtained with the development database are presented to illustrate the character of the challenge.

A. Algorithms

For all algorithms, the microphone signals are processed in the Short-Time Fourier Transform domain at 48 kHz sampling rate, for 1024 Discrete Fourier Transform points, and a frame duration of 0.03 ms. The source DOAs are estimated only during periods of voice activity which are estimated by applying the Voice Activity Detector (VAD) of [38] for a window length of 10 ms to one arbitrarily selected channel of each microphone array. The following algorithms serve as baseline approaches for the challenge and, hence, are not adapted to the specific array geometries (e.g., by performing SH-domain processing for the Eigenmike).

1) Multiple Signal Classification (MUSIC): The instantaneous source DOAs are estimated by evaluating the MUSIC [9, 10] pseudo-spectrum for each frequency bin and block size of 100 frames. The step-size between consecutive blocks is 10 frames. The MUSIC resolution is 5° in azimuth and inclination, respectively. To obtain a single pseudo-spectrum per block, the spectra are summed over all frequency bins [39]. A single DOA estimate per block corresponds to the peak direction in the summed spectrum. Due to different rates of the blocks and ground-truth position data, the MUSIC estimates are interpolated to the sampling rate of the ground-truth data.

2) Single-source Kalman filter: For the single-source scenarios in Task 1, 3, and 5, smoothed trajectories of the source azimuth are estimated using the Kalman filter [40] from the uninterpolated MUSIC estimates of the source azimuth only. The Kalman filter avoids interpolation to the ground-truth data rate by 1) predicting the source tracks at the ground-truth data rate, and 2) updating the predictions using the MUSIC estimates at the block rate. The Kalman filter uses a constant-velocity source motion model [41] with process noise standard deviation of 5° in azimuth and 0.1° per second in speed. The measurement noise standard deviation is 20°.

3) Multi-source Kalman filter: A one-to-one mapping between each MUSIC estimate and a predicted source track is established by means of the association algorithm in [42], using the azimuth error as cost function. If the nearest track corresponds to an angular distance of over 20°, a new, temporary track is initialized. To avoid false track initializations due to MUSIC estimates directed away from the sound sources, e.g., due to early reflections, the following track confirmation scheme is used: A 'full' track is confirmed if the track is associated with a DOA estimate in 3 consecutive time-frames. To avoid an exponential explosion in the number of tracks, any temporary and confirmed tracks that are unassociated in 5 consecutive time-frames are terminated.

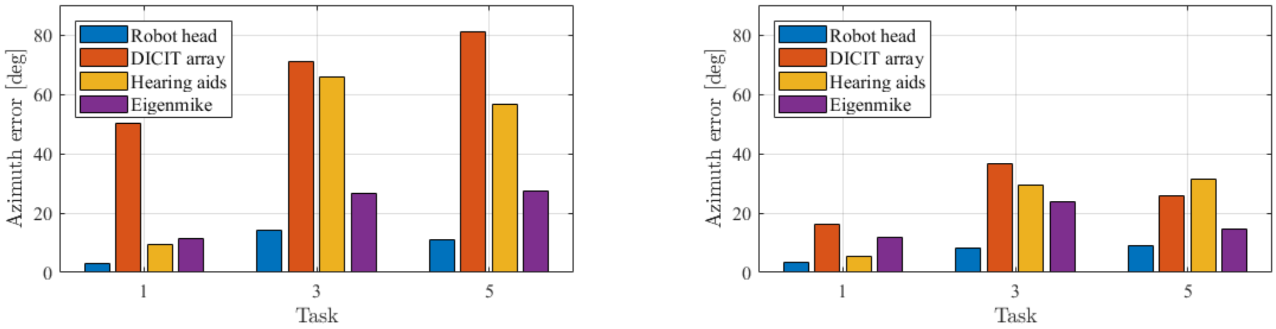
B. Metrics

The performance of the baseline algorithms is evaluated in this paper based on the azimuth accuracy of the DOA estimates. In the case of MUSIC, the error between the ground-truth source azimuth and the interpolated azimuth estimates is evaluated. For the multi-source scenarios in Tasks 2, 4 and 6, the minimum azimuth error between the interpolated MUSIC estimates and any of the ground-truth DOAs is used.

In contrast to MUSIC, the Kalman filter implementation may estimate multiple source tracks for each time step. Therefore, the average azimuth error is evaluated between all ground-truth source trajectories and estimated tracks. The resulting cost matrix is used for the association algorithm in [42] to establish a one-to-one assignment between the ground-truth trajectories and track estimates. The overall azimuth error per recording is given by the azimuth error averaged over all pairs of tracks and their associated ground-truth trajectories.

C. Results

The results in Fig. 2 show the azimuth error, averaged over each recording and all voice activity periods, for Task 1, 3 and 5. Fig. 2a shows that the pseudo-spherical robot head achieves the highest azimuth accuracy, with DOA estimation errors of 2.9° for Task 1 and 14.2° for Task 3. The less challenging Task 1 to localize a static sources with a static microphone array leads to the lowest error for all configurations. The errors increase for Task 3, involving a single, moving source; e.g., the



(a) DOA Estimation

(b) Tracking

Figure 2. Azimuth accuracy for Tasks 1, 3, 5 involving single sources for (a) baseline DOA estimator and (b) baseline tracker.

Table I
AZIMUTH ERROR FOR BASELINE LOCALIZATION ALGORITHMS.

Task	Robot head		DICIT array		Hearing aids		Eigenmike	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	2.9	0.0	50.0	0.6	9.2	0.1	11.4	0.0
2	6.4	0.0	52.4	0.6	16.5	0.1	8.0	0.0
3	14.2	0.2	70.9	0.9	65.8	0.8	26.8	0.2
4	9.5	0.0	64.4	0.8	72.6	0.7	12.1	0.0
5	11.1	0.2	81.0	1.0	56.5	0.8	27.5	0.4
6	10.2	0.1	42.5	0.4	51.3	0.5	22.9	0.1

Table II
AZIMUTH ERROR FOR BASELINE TRACKING ALGORITHMS.

Task	Robot head		DICIT array		Hearing aids		Eigenmike	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	3.3	0.0	16.0	0.1	5.5	0.0	11.9	0.0
2	8.6	0.0	48.0	0.8	15.7	0.3	17.0	0.3
3	8.4	0.0	36.6	0.6	29.5	0.4	23.8	0.2
4	14.4	0.1	59.0	1.2	59.7	1.0	16.8	0.1
5	9.2	0.2	25.7	0.8	31.3	0.4	14.6	0.1
6	32.0	0.6	51.3	0.7	61.4	0.7	38.5	0.5

azimuth accuracy reduces by 56.8% for the Eigenmike from 11.4° for Task 1 to 26.8° for Task 3. The performance for Task 5, compared to Task 3, remains approximately constant for the Eigenmike. The robot head and hearing aids indicate small performance improvements relative to Task 3 of 14% and 21% respectively. Reflective of human-machine interaction applications, Task 5 involves microphone arrays that frequently approach the moving talker. Reductions in source-sensor range due to an approaching microphone array therefore lead to improvements in azimuth estimation accuracy.

As summarized in Table I, the results highlight that the DICIT array results in azimuth errors between 50° and 81°. To reduce the severe effects of spatial aliasing due to the large spacings of some microphones for the DICIT array and in order to use the same algorithms (which do not account for nested sub-arrays) for all four arrays, a linear, uniform subarray of the DICIT array with only 3 microphone and a spacing of 4cm has been used, which necessarily leads to front-back ambiguities.

DOA estimation using the signals recorded by the hearing aids result in an azimuth error of 9.2° for Task 1. The azimuth errors for the hearing aids is degraded to 65.8° for Task 3 and 56.5° for Task 5. The microphone configuration of the hearing aids mounted on the dummy head leads to ambiguities in the elevation, and hence azimuth angle, of the MUSIC pseudo-spectra. These ambiguities are particularly severe for the tasks involving moving sources as the motion of a walking human leads to elevation variations in and between blocks.

The performance results for the tracking algorithm are shown in Fig. 2b and summarized in Table II. The results highlight that extrapolation of the source trajectories using temporal models of the source dynamics, rather than interpolation, lead to performance improvements for all arrays in Task 3 and 5. For example, the azimuth estimates obtained from the DICIT array recordings in Task 3 are improved by 55.3°, i.e., 68%, compared to the MUSIC estimates. However, the performance results in Table II indicate that the tracking accuracy is mostly degraded for the multi-source scenarios of Task 2, 4, and 6, compared to the single-source scenarios of Task 1, 3, and 5. This performance degradation is caused by the association uncertainty between the MUSIC estimates and tracks, and ambiguities due to overlapping speech segments from multiple sound sources.

V. SUMMARY

This paper presents a novel, open-access data corpus of multichannel audio recordings for the objective evaluation of sound source localization and tracking algorithms as part of the LOCATA Challenge. The recordings were conducted using a planar array, a spherical and a pseudo-spherical array, as well as a pair of hearing aids. Scenarios include static loudspeakers, moving human talkers, as well as static and moving arrays. Baseline results are presented using the development dataset of the LOCATA Challenge for broadband MUSIC DOA estimation and Kalman filter-based source tracking.

Acknowledgment

The authors would like to thank Claas-Norman Ritter and Ilse Sofía Ramírez Buensuceso Conde for their contributions as well as the hearing aid manufacturer Sivantos for providing the hearing aids.

REFERENCES

[1] N. Strobel, S. Spors, and R. Rabenstein, “Joint Audio-Video Signal Processing for Object Localization and Tracking,” in *Mircophone Arrays*, M. S. Brandstein and H. F. Silvermann, Eds., chapter 10, pp. 203–225. Springer, Berlin, 2001.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust Localization in Reverberant Rooms,” in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., Digital Signal Processing, pp. 157–180. Springer, Berlin, Germany, 2001.

[3] J. C. Chen, L. Yip, J. Elson, H. Wang, D. Maniezzo, R. E. Hudson, K. Yao, and D. Estrin, “Coherent Acoustic Array Processing and Localization on Wireless Sensor Networks,” *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1154–1162, Aug. 2003.

[4] W. Noble and D. Byrne, “A Comparison of Different Binaural Hearing Aid Systems for Sound Localization in the Horizontal and Vertical Planes,” *British Journal of Audiology*, vol. 24, no. 5, pp. 335–346, 1990.

[5] V. Tourbabin and B. Rafaely, “Speaker Localization by Humanoid Robots in Reverberant Environments,” in *Proc. of IEEE Conv. of Electrical and Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, Dec. 2014, pp. 1–5.

[6] C. Knapp and G. Carter, “The Generalized Correlation Method for Estimation of Time Delay,” *IEEE Trans. on Acoustics, Speech, and Signal Processsing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[7] H. Do and H. F. Silverman, “SRP-PHAT Methods of Locating Simultaneous Multiple Talkers Using a Frame of Microphone Array Data,” in *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas (Texas), USA, Mar. 2010, pp. 125–128.

[8] E. D. D. Claudio and R. Parisi, “Multi-Source Localization Strategies,” in *Mircophone Arrays*, M. S. Brandstein and H. F. Silvermann, Eds., chapter 9, pp. 181–201. Springer, Berlin, 2001.

[9] H. L. van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, Wiley, New York, 2002.

[10] J. P. Dmochowski, J. Benesty, and S. Affes, “Broadband MUSIC: Opportunities and Challenges for Multiple Source Localization,” in *Proc. of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz (New York), USA, Oct. 2007, pp. 18–21.

[11] G. Doblinger, “Localization and Tracking of Acoustical Sources,” in *Topics in Acoustic Echo and Noise Control*, E. Hänsler and G. Schmidt, Eds., chapter 4, pp. 91–124. Springer, Berlin, 2006.

[12] F. Nesta and M. Omologo, “Cooperative Wiener-ICA for Source Localization and Separation by Distributed Microphone Arrays,” in *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas (Texas), USA, Mar. 2010, pp. 1–4.

[13] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, “TDOA Estimation for Multiple Sound Sources in Noisy and Reverberant Environments Using Broadband Independent Component Analysis,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[14] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, “Robust Localization of Multiple Sources in Reverberant Environments Using EB-ESPRIT with Spherical Microphone Arrays,” in *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 117–120.

[15] A. H. Moore, C. Evers, and P. A. Naylor, “Direction of Arrival Estimation in the Spherical Harmonic Domain Using Subspace Pseudointensity Vectors,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 178–192, Jan. 2017.

[16] D. B. Ward, E. A. Lehmann, and R. C. Williamson, “Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[17] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, “Tracking an Unknown Time-Varying Number of Speakers Using TDOA Measurements: A Random Finite Set Approach,” *IEEE Trans. on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, Sept. 2006.

[18] C. Evers, J. Sheaffer, A. H. Moore, B. Rafaely, and P. A. Naylor, “Bearing-Only Acoustic Tracking of Moving Speakers for Robot Audition,” in *Proc. of IEEE Intl. Conf. on Digital Signal Processing (DSP)*, Singapore, July 2015.

[19] C. Evers and P. A. Naylor, “Acoustic SLAM,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1484–1498, Sept. 2018.

[20] C. Evers, Y. Dorfan, S. Gannot, and P. A. Naylor, “Source Tracking Using Moving Microphone Arrays for Robot Audition,” in *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans (Louisiana), USA, Mar. 2017.

[21] C. Evers and P. A. Naylor, “Optimized Self-Localization for SLAM in Dynamic Scenes Using Probability Hypothesis Density Filters,” *IEEE Trans. on Signal Processing*, vol. 66, no. 4, pp. 863–878, Feb. 2018.

[22] J. B. Allen and D. A. Berkley, “Image Method for Efficiently Simulating Small-Room Acoustics,” *Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 943–950, Apr. 1979.

[23] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, “Simulating Room Impulse Responses for Spherical Microphone Arrays,” in *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 129–132.

[24] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, “Performance of Real-Time Source-Location Estimators for a Large-Aperture Microphone Array,” *IEEE Trans. on Acoustics, Speech, and Signal Processsing*, vol. 13, no. 4, pp. 593–606, July 2005.

[25] A. Brutti, M. Omologo, and P. Svaizer, “Comparison Between Different Sound Source Localization Techniques Based on a Real Data Collection,” in *Proc. of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, May 2008.

[26] M. Omologo, P. Svaizer, A. Brutti, and L. Cristoforetti, “Speaker Localization in CHIL Lectures: Evaluation Criteria and Results,” in *Machine Learning for Multimodal Interaction. MLMI 2005. Lecture Notes in Computer Science*, vol. 3869. Springer, Berlin, 2006.

[27] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, “The Single- and Multichannel Audio Recordings Database (SMARD),” in *Proc. of Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes, France, Sept. 2014.

[28] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The Third ‘CHiME’ Speech Separation and Recognition Challenge: Dataset, Task and Baselines,” in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale (Arizona), USA, Dec. 2015, pp. 504–511.

[29] J. Eaton, A. H. Moore, N. D. Gaubitch, and P. A. Naylor, “The ACE Challenge - Corpus Description and Performance Evaluation,” in *Proc. of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz (New York), USA, Oct. 2015.

[30] “LOCATA website,” www.locata-challenge.org, Feb. 2018.

[31] OptiTrack, *Product Information about OptiTrack Flex13*, [Online], http://optitrack.com/products/flex-13/, Feb. 2018.

[32] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo, “WOZ Acoustic Data Collection for Interactive TV,” *Language Resources and Evaluation*, vol. 44, no. 3, pp. 205–219, Sept. 2010.

[33] mh acoustics, *EM32 Eigenmike microphone array release notes (v17.0)*, Oct. 2013, www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf.

[34] V. Tourbabin and B. Rafaely, “Theoretical Framework for the Optimization of Microphone Array Configuration for Humanoid Robot Audition,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, Dec. 2014.

[35] V. Tourbabin and B. Rafaely, “Optimal Design of Microphone Array for Humanoid-Robot Audition,” in *Proc. of Israeli Conf. on Robotics (ICR)*, Herzliya, Israel, Mar. 2016, (abstract).

[36] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, *IEEE-AASP Challenge on Source Localization and Tracking: Documentation for Participants*, Apr. 2018, [Online], www.locata-challenge.org.

[37] C. Veaux, J. Yamagishi, and K. MacDonald, “English Multi-speaker Corpus for CSTR Voice Cloning Toolkit,” [Online] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html, 2018.

[38] J. Sohn, N. S. Kim, and W. Sung, “A Statistical Model-Based Voice Activity Detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[39] O. Nadiri and B. Rafaely, “Localization of Multiple Speakers under High Reverberation Using a Spherical Microphone Array and the Direct-Path Dominance Test,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 10, Oct. 2014.

[40] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle Filters for Tracking Applications*, Artech House, Boston, 2004.

[41] X.-R. Li and V. P. Jilkov, “Survey of Maneuvering Target Tracking. Part I: Dynamic Models,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[42] H. W. Kuhn, “The Hungarian Method for the Assignment Problem,” *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, Mar. 1955.