Passive online geometry calibration of acoustic sensor networks

Axel Plinge, *Member, IEEE,* Gernot A. Fink, *Senior Member, IEEE*, and Sharon Gannot *Senior Member, IEEE*

***Abstract*—As we are surrounded by an increased number of mobile devices equipped with wireless links and multiple microphones, e.g., smartphones, tablets, laptops and hearing aids, using them collaboratively for acoustic processing is a promising platform for emerging applications. These devices make up an acoustic sensor network comprised of nodes, i.e. distributed devices equipped with microphone arrays, communication unit and processing unit. Algorithms for speaker separation and localization using such a network require a precise knowledge of the nodes' locations and orientations. To acquire this knowledge, a recently introduced approach proposed a combined direction of arrival (DoA) and time difference of arrival (TDoA) target function for off-line calibration with dedicated recordings. This paper proposes an extension of this approach to a novel online method with two new features: First, by employing an evolutionary algorithm on incremental measurements, it is online and fast enough for real-time application. Second, by using the sparse spike representation computed in a cochlear model for TDoA estimation, the amount of information shared between the nodes by transmission is reduced while the accuracy is increased. The proposed approach is able to calibrate an acoustic senor network online during a meeting in a reverberant conference room.**

***Index Terms*—microphone array, geometry calibration, speech-based geometry calibration, acoustic sensor network**

I. Introduction

Nowadays, we are surrounded by a number of devices with microphones and wireless links. When these devices work together in a collaborative way, they act as the nodes of a wireless acoustic sensor network (WASN). For speaker tracking, knowledge of the spatial arrangement of the nodes is required [1]–[3]. Location information can also be beneficial in speech enhancement and speaker separation applications [4], [5].

As manual geometry calibration is a cumbersome task and hardly practical in applications with either many microphones or ad hoc configurations, automated methods of geometry calibration for acoustic nodes in WASNs are required.

Several offline approaches for localizing distributed microphones have been proposed in recent years (cf. [6] and references therein for a more detailed overview). Active approaches use a loudspeaker in each node to play sounds in a dedicated calibration step [7], [8]. Most passive approaches also require a calibration phase and additional equipment [9]–[11] while a few only require natural speech utterances [12], [13]. Many of these methods require long processing time and large amounts of data sharing and are therefore not well suited for real-time application in a WASN.

We recently presented a method for multiple speaker tracking in a WASN [3]. It is highly robust against reverberation through the application of neurobiologically inspired models. A cochlear model, based on insights from human hearing, computes a sparse spike representation of the microphone signals. By computing the cross-correlation of these signals, the time differences of arrival (TDoAs) and subsequently the directions of arrival (DoAs) of the speech signals can be derived. In order to compute speaker positions via triangulation, the geometry of the sensor network has to be known. This leads to the development of our off-line method for calibration [13]. Our calibration method provided a geometry estimate with sufficient accuracy for tracking. However, it is not applicable to online processing, since it is not adequately computationally efficient.

In this paper, we introduce an online version of our previous approach [13] with two main improvements: First, by using an evolutionary algorithm [14] and incremental measurements, faster and more accurate estimates can be achieved. Second, the amount of information shared between the nodes is reduced by exchanging only the sparse spike representation rather than the microphone signals themselves. The new method neither requires a dedicated calibration step nor additional hardware. It also works in parallel to the speaker tracking and incrementally calibrates the geometry with increasing accuracy.

The paper is organized as follows: First, the problem is stated. Then the novel method is described. Thereafter, an experimental study is presented to show that the accuracy of the proposed method exceeds that of the previous off-line version. The online capability is demonstrated in a meeting scenario, where sufficient accuracy for tracking is achieved.[1] Finally, a short conclusion will be given.

II. Problem statement

We address a conference room scenario where acoustic sensor nodes are deployed on a table and several speakers are talking from different positions. The task is the automated estimation of the two dimensional geometric arrangement of G nodes passively using only speech, cf. Fig 1. The tth speech utterance is uttered by a speaker at position $\boldsymbol{s}_t$. Each node g at position $\boldsymbol{r}_g$ captures this utterance by its microphones. Define the

A. Plinge and G. A. Fink are with the Department of Computer Science, TU Dortmund University, Dortmund, Germany. S. Gannot is with the Faculty of Engineering at Bar Ilan Univesity, Ramat Gan, Israel.
We would like to thank Shmulik Markovich-Golan for helpful discussions.
This work was supported by a fellowship within the FITweltweit program of the German Academic Exchange Service (DAAD).

[1] Video demonstration available at `https://vimeo.com/177715229`

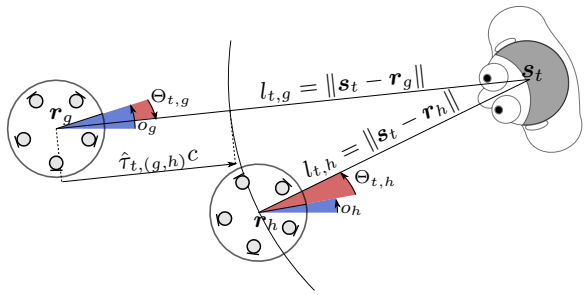

Figure 1. Geometric relations of two nodes at $\boldsymbol{r}_g$ and $\boldsymbol{r}_h$ and a source at $\boldsymbol{s}_t$. The intersection of the rays with the DoAs $\Theta_{t,g}$ and $\Theta_{t,h}$ is the source position. The distance difference corresponds to the TDoA $\hat{\tau}_{t,(g,h)}$.

captured signals as $\boldsymbol{y}_{g,1,t}, \boldsymbol{y}_{g,2,t} \dots \boldsymbol{y}_{g,I,t}$. For simplicity, it is assumed that I, the number of microphones per node, is equal for all nodes. The sampling rate and offset are synchronized across the nodes by a suitable method, e.g. [15]–[17].

Both the array's position $\boldsymbol{r}_g \in \mathbb{R}^2$ and orientation $o_g \in [-180, 180)$ have to be estimated, cf. Fig 1. As the calibration is purely acoustic, only a relative geometry with arbitrary rotation and translation of the entire network can be computed, cf. [6]. Therefore, the first node's parameters can be set to an arbitrary value, e.g. $r_1 = (0, 0)^T$ and $o_1 = 0$. Define the free parameter set as the location and orientation of the nodes $2, \dots, G$:

$$\boldsymbol{\gamma} = (r_{21}, r_{22} \dots r_{1G}, r_{2G}, o_2 \dots o_G) \tag{1}$$

The goal of the proposed method is to estimate $\boldsymbol{\gamma}$ online from the speech utterances. For that, two types of measurements are inferred from the microphone signals, namely the DoAs $\Theta_{t,g}$ at node g of the speaker at position $\boldsymbol{s}_t$ and the corresponding TDoA $\tau_{t,(g,h)}$ between the pair of nodes g, h, cf. Fig 1.

III. Passive online WASN calibration

A block diagram of the geometry calibration algorithm implemented on the WASN is shown in Fig. 2. Two nodes, g and h are depicted. We will describe the algorithm from the viewpoint of node g. The same description applies to node h, *mutatis mutandis*.

When a speech event occurs at position $\boldsymbol{s}_t$, node g computes its DoA $\hat{\Theta}_{g,t}$. This is broadcast along with spike data $\boldsymbol{z}_{t,g}$, inferred from the microphone signals $\boldsymbol{y}_{g,\cdot,t}$, to all other nodes. Using the spike data from the other nodes, the TDoAs $\hat{\tau}_{g,\cdot}$ are computed and again broadcast to all other nodes. Thereafter, each node continuously computes geometry estimates and broadcasts the result to all other nodes. Then all nodes update a weighted mean of the geometry estimate.

A. DoA and TDoA computation

In this paper we use the robust DoA estimation procedure as explained in [13] and references therein. Details are omitted due to space constraints. In this section, we focus on the TDoA estimation. While in [13] the generalized cross correlation with phase transform (GCC-PHAT) procedure [18] was used for the computation of the TDoA, here an alternative procedure is introduced. First, from the microphone signals $\boldsymbol{y}_{g,1,t}, \dots \boldsymbol{y}_{g,I,t}$ a sparse spike representation $\boldsymbol{q}_{g,1,1,t}, \dots \boldsymbol{q}_{g,I,B,t}$ in B frequency bands is computed by the cochlear model [19] (also used for the DoA). The spikes in all frequency bands are aggregated to compute

$$\boldsymbol{z}_{t,g} = \sum_{b=1}^{B} (\boldsymbol{q}_{g,1,b,t}, \boldsymbol{q}_{g,2,b,t}, \dots, \boldsymbol{q}_{g,I,b,t})^T. \tag{2}$$

This is broadcast to the other nodes along with the estimated DoA. As the spike representation is very sparse, the amount of shared data to be broadcast is reduced and less bandwidth is required for transmission. In each node, these signals are then cross-correlated with the nodes' own spike signals to estimate the TDoA. The estimation is done with the arrays' center as reference point. The resulting TDoAs are then broadcast to the other nodes.

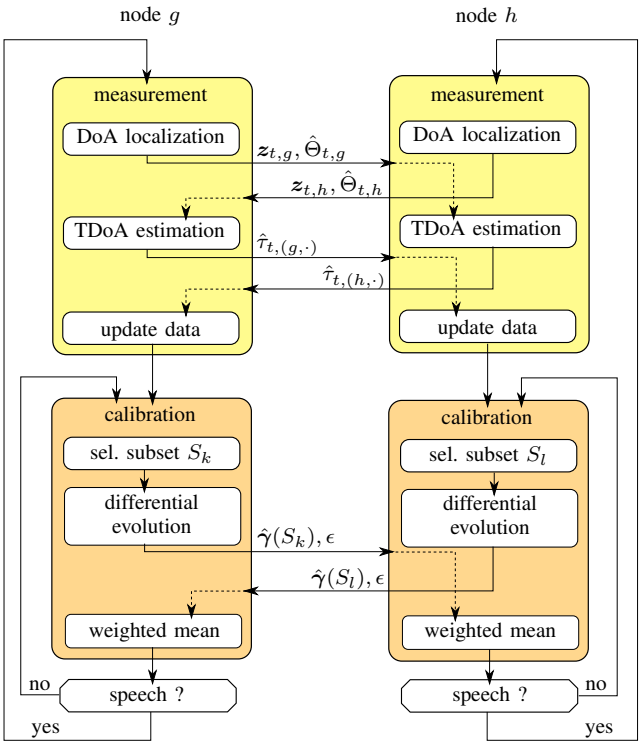

Figure 2. Distributed online computation of WASN geometry from speech.

B. Estimating γ: Target function

In order to blindly estimate the geometry, a target function is formulated that reaches a minimum value at the correct geometry $\boldsymbol{\gamma}$. The key idea is to use the mathematical relation of the DoAs and the TDoA for each pair of microphone arrays with respect to a speaker position, cf. Fig. 1. The following procedure is a variant of the one described in [13].

With a given geometry $\boldsymbol{\gamma}$, the source at event t can be localized by triangulation in each node using the broadcast DoAs and TDoAs. The source position estimate $\hat{\boldsymbol{s}}_{t,(g,h)}$, as viewed by the node pair g, h, is found as follows. Define $\boldsymbol{a}(\alpha) = (\cos\alpha, \sin\alpha)^T$ as the function computing the 2D unit vector directed towards the angle α. With this definition we can compute the distances $l_{t,g}$ and $l_{t,h}$ between the source and the nodes g, h by solving

$$\hat{\boldsymbol{r}}_g + l_{t,g}\boldsymbol{a}\left(\hat{o}_g + \hat{\Theta}_{t,g}\right) = \hat{\boldsymbol{r}}_h + l_{t,h}\boldsymbol{a}\left(\hat{o}_h + \hat{\Theta}_{t,h}\right) \tag{3}$$

for any given $\boldsymbol{\gamma}$. Now we can compute the intersection using either distance, e.g.

$$\hat{\boldsymbol{s}}_{t,(g,h)}(\boldsymbol{\gamma}) = \hat{\boldsymbol{r}}_h + l_{t,h}\boldsymbol{a}\left(\hat{o}_h + \Theta_{t,h}\right). \tag{4}$$

Then the mean of the pairwise location estimates is computed:

$$\hat{\boldsymbol{s}}_t(\boldsymbol{\gamma}) = \frac{2}{G(G-1)} \sum_{g<h} \hat{\boldsymbol{s}}_{t,(g,h)}(\boldsymbol{\gamma}) \ . \tag{5}$$

Using this location estimate, we can compute the error with respect to the TDoA $\hat{\tau}_{t,(g,h)}$ multiplied by the speed of sound c. A negative distance $l_{t,g} < 0$ or $l_{t,h} < 0$ implies no intersection for a specific configuration $\boldsymbol{\gamma}$. Such a configuration is penalized with a constant $\epsilon_{\max} = 10\,\mathrm{m}$.

$$\epsilon_{g,h}(\boldsymbol{s}_t, \boldsymbol{\gamma}) = \begin{cases} \epsilon_{\max}^2 & \text{no intersection} \\ \left(||\boldsymbol{s}_t - \hat{\boldsymbol{r}}_h|| - ||\boldsymbol{s}_t - \hat{\boldsymbol{r}}_g|| - \hat{\tau}_{t,(g,h)}c\right)^2 & \text{otherw.} \end{cases} \tag{6}$$

Assume we have collected t utterances and calculated their respective DoAs and TDoAs. In order to robustly estimate the network geometry $\boldsymbol{\gamma}$ a subset of t_0 out of the t utterances is randomly selected, where $t_0 \geq 3$ is a predefined number of utterances. Let S_k denote the kth choice of such a subset. For this subset the mean error over all utterances t' in S_k is computed as

$$\epsilon(S_k, \boldsymbol{\gamma}) = \frac{2}{G(G-1)|S_k|} \sum_{g<h} \sum_{t' \in S_k} \epsilon_{g,h}(\hat{\boldsymbol{s}}_{t'}(\boldsymbol{\gamma}), \boldsymbol{\gamma}) \tag{7}$$

The geometry $\hat{\boldsymbol{\gamma}}$ best explaining the measurements can be identified by finding the minimum value of this function.

$$\hat{\boldsymbol{\gamma}}(S_k) = \operatorname*{argmin}_{\boldsymbol{\gamma}} \epsilon(S_k, \boldsymbol{\gamma}) \tag{8}$$

New subsets are continuously chosen from the available utterances. The overall estimate is computed as an average weighted by the reciprocal of the target function values, i.e.,

$$\hat{\boldsymbol{\gamma}}^* = \left(\sum_k \frac{1}{\epsilon(S_k, \hat{\boldsymbol{\gamma}}(S_k))} \right)^{-1} \sum_k \frac{\hat{\boldsymbol{\gamma}}(S_k)}{\epsilon(S_k, \hat{\boldsymbol{\gamma}}(S_k))} \ . \tag{9}$$

This subset mean provides robustness against outliers. Subsets are computed incrementally as more speech events become available, starting when t_0 utterances have been accumulated. To allow for slow geometry changes, (9) can be recomputed after discarding old utterances.

C. Estimating γ: Differential evolution

The high-dimensional target function is non-continuous and non-differentiable, cf. Fig. 3. Therefore, gradient descent is not applicable to find the minimum as in, e.g., [20]. The previous method [13] used an exhaustive grid search to minimize the target function value, which led to long computation times. In order to speed up the process, different optimization strategies were investigated. While simulated annealing also showed good results, the best performing method was the differential evolutionary algorithm [14] with binomial recombination of the best member. It converged faster and more reliably.

A population of U candidate solutions for $\boldsymbol{\gamma}$ is used. They are initialized with random values and repeatedly mutated to form a new generation. Trial candidates are generated by mutating the member with the best fitness value ϵ using the difference between two random members of the current generation.

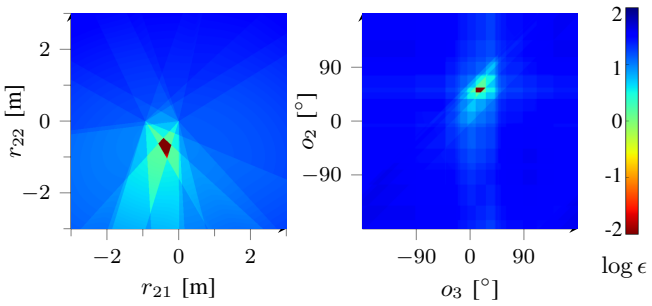

Figure 3. Cuts through the high-dimensional target function $\epsilon(S_k, \boldsymbol{\gamma})$ for an example with $G = 3$, i.e., $\boldsymbol{\gamma} = (r_{21}, r_{22}, r_{31}, r_{32}, o_2, o_3)^T$. The color represents the log of the target function, red=lowest, blue=highest. The minimum is located at $\hat{\gamma}^* = (-0.78, -0.30, 0.0, -0.93, 52, 18)$. On the left, the position of the second node $\boldsymbol{r}_2$ is varied while keeping the other values fixed to $\hat{\gamma}^*$. On the right, the two orientations are varied while keeping both positions fixed to the values of $\hat{\gamma}^*$. The visible plateaus are due to non-intersections penalized by $\epsilon_{\max}$.

The individual values of $\boldsymbol{\gamma}$ are replaced by the mutated ones according to a binomial random variable. When the fitness value is better for the trial candidate, it replaces the current one. Once the population converges to a set with low variance, the optimization terminates.

IV. Experimental Study

The proposed method was investigated in several aspects using real recordings. The influence of the parameters is evaluated. The online application in a meeting is demonstrated. Additionally, the influence of the two key improvements is tested in an comparative evaluation with the previous approach.

A. Setup

All experiments were undertaken in a smart conference room at TU Dortmund university. The reverberation time was measured as $T_{60} = 0.67\,\mathrm{s}$ [21]. Three circular arrays composed of five microphones with 10 cm diameter were embedded in a conference table. Each array was captured by a separate sound card at 48 kHz. The sound cards were synchronized with a remaining jitter of 22 μs between them.

For online estimation, a meeting was recorded. Four different speakers talked to each other, first standing around the table, then sitting down and later standing up again. An additional recording was done where a single speaker stands or sits in 19 different positions around the table.

As the geometry estimates exhibit an arbitrary rotation and translation to the reference geometry, the calibration result is aligned automatically before computing the errors (cf. [6]). The error e_r is computed as Euclidean distance of the aligned estimated positions to the true positions. The orientation error e_o is computed as mean over the nodes. From our previous experiments, we know that the minimum accuracy required for triangulation is about 10 cm and 5°, cf. [6], [13]. All computations were run on a standard PC (i7-3770 CPU, 3.3 GHz) in multiprocessing, using two cores for each node in the basic python implementation.[2]

[2]For additional material and source code visit
`http://patrec.cs.tu-dortmund.de/pubs`

Table I
COMPARISON OF DIFFERENT PARAMETERIZATIONS

t_0	U	t_ϵ[s]	e_r [cm]	e_o [°]
5	10	4.5±1.1	8.2±0.1	1.6±0.2
5	20	8.8±2.4	8.3±0.2	1.7±0.2
5	30	11.8±2.6	8.3±0.2	1.6±0.3
5	40	19.5±5.1	8.3±0.3	1.8±0.4
4	10	4.1±1.0	8.5±0.1	1.5±0.2
5	10	4.5±1.1	8.2±0.1	1.6±0.2
6	10	5.3±1.2	8.1±0.1	1.7±0.2
7	10	5.7±1.3	8.0±0.1	1.7±0.1
8	10	8.2±1.6	7.9±0.1	2.0±0.1

B. Parameter evaluation

Two parameters are critical for the performance of the algorithm: First, the number of utterances in the subsets t_0. It was varied between the minimal value of three and seven. Second, the population size U for the differential evolutionary optimization. In order to evaluate the performance of the algorithm, we have chosen three figures of merit: The computation time t_ϵ for optimizing one set of utterances S_k as well as the mean position error e_r, and mean orientation error e_o over the duration of the meeting. The results averaged over 100 Monte-Carlo runs are shown in Table I. The results are quite similar for the different configurations, showing the robustness of the approach. Using small population sizes U enables to compute much more subsets S_k within the real time constraint, as less time t_ϵ is required for each one. This improves the weighted mean by using more random subsets. Subset sizes t_0 of five to seven perform well. As the time before the initial estimate as well as the computation time is increased by larger t_0, five was chosen.

C. Example run

The averages above do not show the algorithm's behavior over time. In order to illustrate this, a typical run using $t_0 = 5$ and $U = 10$ is shown in Fig. 4. After one minute, seven speech events are present and the algorithm achieves around 8 cm and 2° error. The orientation error stabilizes after two minutes at about 1°.

D. Method comparison

The proposed method extends the previous one [13] by two new features. The optimization process of exhaustive search was replaced by the differential evolution in an online fashion and the GCC-PHAT was replaced by the new TDoA estimation procedure. The old 'off-line' and the new 'online' algorithm were applied with both 'PHAT' and 'spikes' options for the distance measurement, resulting in a total of four variants. Both the meeting and a recording of a single speaker taking up 19 static positions were used.

As the ground truth positions are known for the second sequence, it was possible to compute the accuracy of the TDoA measurements. The RMS error of the iter-array distances was 10.2 cm for the spike method and 12.2 cm for the GCC-PHAT. The error induced by the speaker elevation is around 5 cm. This is also the amount the calibration deteriorates from the 2D assumption.

Table II
COMPARISON OF CALIBRATION METHODS ON TWO RECORDINGS

recording	method	error (mean / end) e_r [cm]	e_o [°]	comp. time t_ϵ [s]	time[%]
single speaker	PHAT off-line	– /8.9±0.7	– /2.1±1.2	112.2	390±13
	spikes off-line	– /7.3±0.8	– /2.8±1.3	105.5	366±17
	PHAT online	8.8±0.2/8.8±0.2	1.0±0.3/1.2±0.4	4.9	50±03
	spikes online	6.8±0.2/6.8±0.2	1.2±0.2/1.6±0.2	4.5	52±04
meeting	PHAT off-line	– /9.0±1.7	– /2.1±1.3	111.1	199±09
	spikes off-line	– /9.7±1.4	– /1.9±1.0	103.0	184±08
	PHAT online	8.9±0.7/8.6±0.7	0.9±0.4/0.7±0.4	5.4	62±01
	spikes online	8.2±0.1/8.3±0.2	1.6±0.2/1.1±0.2	4.5	63±01

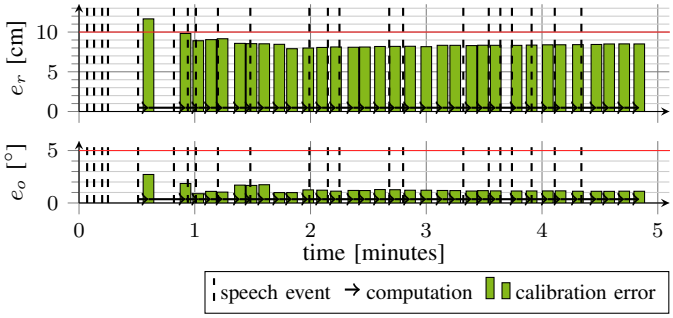

Figure 4. Online calibration. Realtime processing using the measurements (dashed lines) to update the geometry estimate over the course of a meeting. Position (top) and orientation (bottom) errors over time. The red line indicates the required minimum target accuracy for spatial processing.

With the off-line version [13] the weighted mean estimate was computed over random choices of 32 different subsets with a size of five. The new online approach was applied to an increasing number of utterances under realtime constraints with $t_0 = 5$ and $U = 10$. For each of the four method variants, Table II lists the calibration errors averaged over time and at the end of the sequence, as well as the computation times over 100 runs. The off-line method performed worse over all. The online method using differential evolution is more robust and achieves the best accuracy when applied with the spikes. It can be seen that the exhaustive search method takes much longer to compute an individual estimate than the differential evolution. The last column lists the computation time as a percentage of the total recording duration. It can be seen that the off-line method takes more than real-time to compute the 32 estimates.

V. CONCLUSION

A novel geometry calibration method was proposed. It is passive since it works with natural speech utterances only. It surpasses the earlier off-line version in three regards: By using speech events incrementally, the method is now online. By introducing differential evolution for TDoA-DoA optimization, the computation time could be sufficiently reduced to allow for real-time application. By using the sparse spike representation for TDoA estimation, the amount of shared information could be significantly reduced without affecting the calibration quality. The method can be used online in real-time, as shown by the application to a recording of a meeting. The orientation and position accuracy of around 2° and 8 cm are well within the requirements for practical spatial processing applications.

REFERENCES

[1] A. Griffin and A. Mouchtaris, "Localizing multiple audio sources from DoA estimates in a wireless acoustic sensor network," in *IEEE Workshop on Appl. of Signal Process. to Audio and Acoustics*, 2013.

[2] Y. Oualil and D. Klakow, "Multiple concurrent speaker short-term tracking using a Kalman filter bank," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Florence, Italy, May 2014, pp. 1458–1462.

[3] A. Plinge and G. A. Fink, "Multi-Speaker tracking using multiple distributed microphone arrays," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Florence, Italy, May 2014, pp. 614–618.

[4] Y. Dorfan, D. Cherkassky, and S. Gannot, "Speaker localization and separation using distributed expectation-maximization," in *European Signal Process. Conf.*, Nice, France, Aug. 2015.

[5] M. Taseska and E. A. P. Habets, "Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Trans. Acoust., Speech, Language Process.*, vol. 24, no. 7, pp. 1291–1304, 2016.

[6] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 14–29, July 2016.

[7] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 13, no. 1, pp. 70–83, 2005.

[8] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Closed-form self-localization of asynchronous microphone arrays," in *J. Works. on Hands-Free Speech Commun. and Microphone Arrays*, Edinburgh, UK, May 2011, pp. 139–144.

[9] S. D. Valente, M. Tagliasacchi, F. Antonacci, P. Bestagini, A. Sarti, S. Tubaro, and P. Milano, "Geometric calibration of distributed microphone arrays from acoustic source correspondences," in *IEEE Workshop on Multimedia Signal Process.*, 2010, pp. 13–18.

[10] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Calibration of distributed sound acquisition systems using ToA measurements from a moving source," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Florence, Italy, May 2014.

[11] D. B. Haddad, W. A. Martins, M. d. V. M. da Costa, L. W. P. Biscainho, L. O. Nunes, and B. Lee, "Robust acoustic self-localization of mobile devices," *IEEE Trans. Mobile Computing*, vol. 15, no. 4, pp. 982–995, 2016.

[12] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "DoA-based microphone array postion self-calibration using circular statistisc," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Vancover, Canada, May 2013.

[13] A. Plinge and G. A. Fink, "Geometry calibration of multiple microphone arrays in highly reverberant environments," in *Int. Works. on Acoustic Signal Enh.*, Antibes – Juan les Pins, France, Sept. 2014.

[14] R. Storn and K. Price, "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces," *Journ. Global Optimization*, vol. 11, pp. 341–359, 1997.

[15] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Int. Works. on Acoustic Signal Enh.*, Aachen, Germany, Sept. 2012.

[16] P. Pertilä, M. S. Hämäläinen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[17] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *Int. Works. on Acoustic Signal Enh.*, Antibes – Juan les Pins, France, Sept. 2014.

[18] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[19] A. Plinge and G. A. Fink, "Online multi-speaker tracking using multiple microphone arrays informed by auditory scene analysis," in *European Signal Process. Conf.*, Marrakesh, Morocco, 2013.

[20] A. Plinge and G. A. Fink, "Geometry calibration of distributed microphone arrays exploiting audio-visual correspondences," in *European Signal Process. Conf.*, Lisbon, Portugal, Sept. 2014.

[21] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Int. Works. on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.