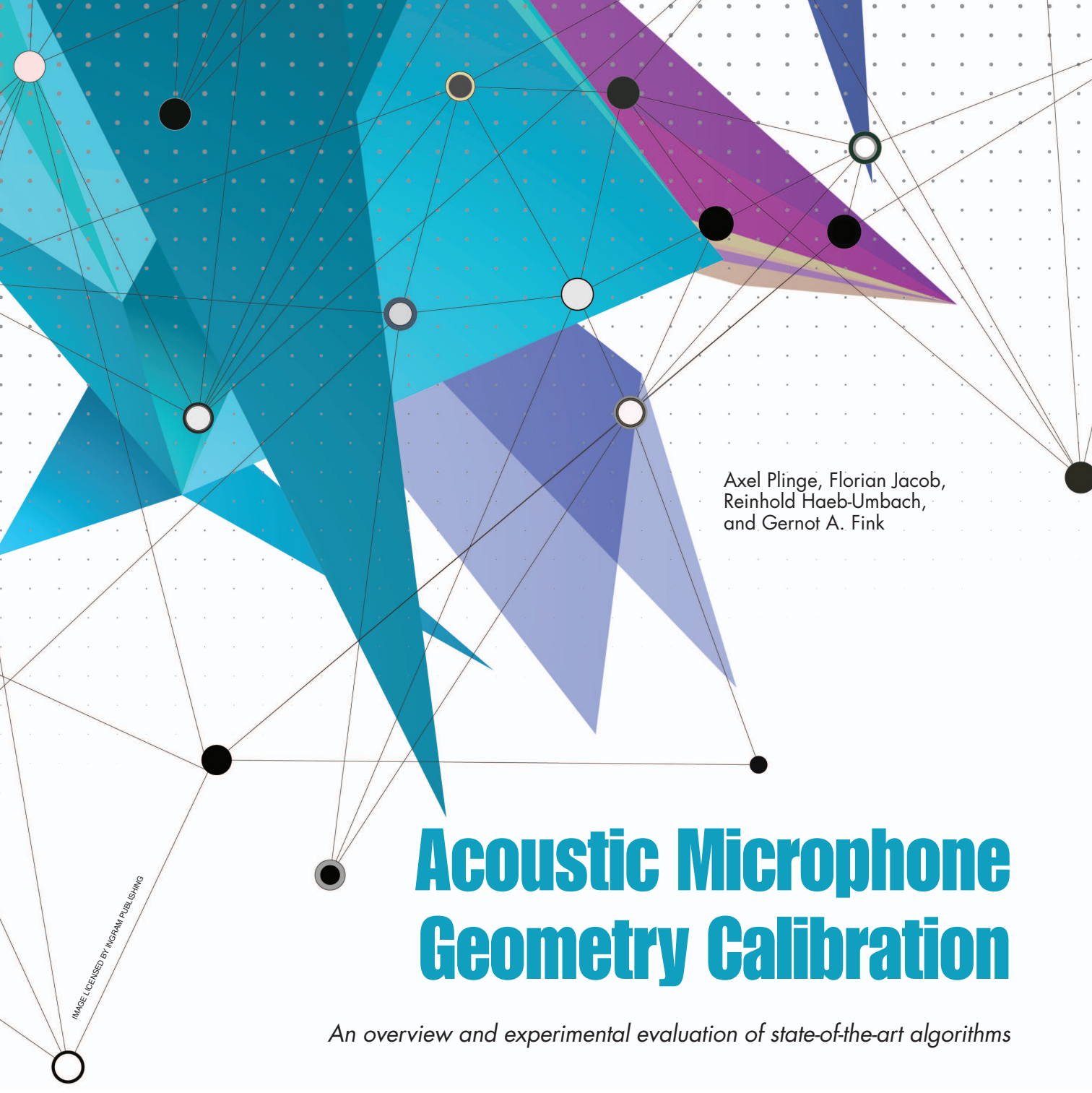

Axel Plinge, Florian Jacob,
Reinhold Haeb-Umbach,
and Gernot A. Fink

Acoustic Microphone Geometry Calibration

An overview and experimental evaluation of state-of-the-art algorithms

Today, we are often surrounded by devices with one or more microphones, such as smartphones, laptops, and wireless microphones. If they are part of an acoustic sensor network, their distribution in the environment can be beneficially exploited for various speech processing tasks. However, applications like speaker localization, speaker tracking, and speech enhancement by beamforming avail themselves of the geometrical configuration of the sensors. Therefore, acoustic microphone geometry calibration has recently become a very active field of research. This article provides an application-oriented, comprehensive survey of existing methods for microphone position self-calibration, which will be categorized by the measurements they use and the scenarios they can calibrate. Selected methods will be evaluated comparatively with real-world recordings.

Digital Object Identifier 10.1109/MSP.2016.2555198
Date of publication: 1 July 2016

Introduction

Wireless acoustic sensor networks (WASNs) are a promising approach for sound capturing and processing systems [2].

 1053-5888/16©2016IEEE

With the low cost of acoustic sensors and wireless communication devices, such networks are more common. They are composed of devices such as smartphones, tablet computers, wireless microphones, or hearing aids that are equipped with a single or multiple microphones. Due to their distribution in an environment, it is likely that at least one device is close to every relevant sound source. Thus, WASNs deliver a signal with improved quality compared to traditional microphone arrays, which sample a sound field only locally. As a consequence of the ad hoc nature of many WASNs, the position of the sensor nodes is often unknown and may even vary over time. However, a number of important speech processing tasks rely on the estimation of the location of sound sources, which in turn requires the location of the recording devices to be known.

A popular application of distributed microphones or microphone arrays is source localization and tracking. The localization results are used to enable subsequent applications, such as camera control and speech enhancement. But only if the positions and orientations of the microphones are known can the source position or direction be estimated by triangulation, trilateration, or other approaches. Errors in the assumed geometric arrangement of the microphones have a significant effect on the localization accuracy, as is well demonstrated by the experiment described in "Impact of Geometry Errors on Source Localization."

While signal extraction from distributed microphone arrays can be achieved without explicit estimation of the position of the sources [22], the speech enhancement performance can be improved by incorporating source location information. Faster adaptation to changing acoustic environments can be obtained by so-called informed spatial filtering approaches, where the adaptation of the source extraction filters is supported by information on the source location [42]. Moreover, in parametric spatial processing, parameters describing the sound field, such as the source location, are employed for spatial audio coding and reproduction, source enhancement, or acoustic scene analysis.

Given the importance of the aforementioned speech processing tasks, which require the geometric configuration of the acoustic sensors to be known, research in microphone geometry calibration has substantially increased in recent years. While, in early approaches, the positions of the microphones were determined by hand or computed from manual measurements of pairwise microphone distances, this clearly becomes impractical in the ad hoc scenarios typical of WASNs.

The goal of recent research efforts is to devise methods to infer the position of the sensors solely from the acoustic signals they capture, a strategy termed *acoustic geometry calibration*. Research articles also refer to the task as *microphone self- or auto-localization*, or *position self-calibration*. The basic idea is to extract a quantity from the microphone signals that is related to their geometric arrangement—for example, the time difference of an arriving sound at two nodes, or the direction under which an acoustic event is observed. The extracted information is used in an objective function that essentially scores the deviation of the actual measurement from the measurement as predicted by the assumed geometry. Since the extraction of location information from the microphone signals is often achieved by measurement of time or time difference, there exists a close dependence of localization on clock synchronization, as we will see in the following.

We also consider node localization in nonacoustic sensor networks. However, there are some fundamental differences. Popular localization systems, such as satellite, cellular, or Wi-Fi-based systems, rely on knowledge of the location of anchor nodes, i.e., the satellites or base stations that transmit a radio signal, to infer the position of user terminals. Such anchor nodes are, in general, not available in acoustic geometry calibration. Furthermore, the nodes in wireless sensor networks are typically assumed to consist of a transceiver, i.e., a radio transmitter and a receiver at the same location. Thus, active localization can be performed by exchanging time stamps or other signaling information, from which position-related information is estimated.

In an acoustic sensor network, we wish to perform localization via acoustic signals only. There are also active approaches, which can be used for sensor nodes such as smartphones or laptop computers equipped with both loudspeakers and microphones. In the general case of passive calibration, however, the sensor is unable to produce a sound by itself, and the process has to rely on external acoustic events such as speech or ambient noise. Then there is no time synchronization between transmitter and receiver, and time or time difference estimation is further complicated by the unfavorable correlation properties of the signals.

Only recently has this unconstrained scenario been tackled by acoustic geometry calibration algorithms. Earlier approaches targeted laboratory installations and required a known arrangement of loudspeakers, dedicated calibration signals, and strict time synchronization. By removing the earlier constraints, modern approaches attempt to calibrate ad hoc arrangements using ambient sounds such as speech. Furthermore, the constraints of time synchronization or colocation of microphones and loudspeakers are relaxed.

This ongoing research is being performed in a number of directions, focusing on different scenarios and employing different measurements and optimization strategies. The purpose of this article is to categorize the individual approaches with respect to the scenario addressed and the signals and objective functions employed. This should give the reader a basic understanding of the algorithms used, their applicability to different scenarios, and the expected localization performance.

Application scenarios

Acoustic geometry calibration approaches have been developed for different application scenarios. In the following survey, we distinguish three fundamental types of microphone arrangements: small compact arrays, distributed individual microphones, and distributed microphone arrays

Impact of Geometry Errors on Source Localization

To investigate the dependence of acoustic source localization accuracy on the microphone position calibration performance, a recent speaker tracking method [32] was applied to simulated data. Five nodes with circular microphone arrays composed of five microphones each were used, located in the middle of the room. A speaker was localized at 18 positions around the arrays in a reverberant room ($T_{60} = 0.5$ seconds) of size 6.5 m $\times$ 3.5 m $\times$ 2.5 m. An erroneous geometry calibration was simulated as follows: the calibration error was drawn from a zero mean normal distribution with increasing standard deviation. This way, a fixed mean calibration error from 0.1 m and 2° to 0.5 m and 10° over all arrays was simulated. The plots in Figure S1 show the mean speaker localization error $\varepsilon_l(m)$ as a function of mean position [Figure S1(a)] and orientation [Figure S1(b)] geometry calibration error. Furthermore, the error bars indicate the standard deviation over 100 experiments. The localization error should be no larger than the size of a human head, i.e., below 30 cm (indicated by the orange line), for practical applications such as camera control. Due to reverberation deteriorating the measurements, the resulting localization error is already close to that level even without any calibration error. When the calibration error exceeds 10 cm or 2°, it increases further. Beyond this, the performance of source localization will start to suffer.

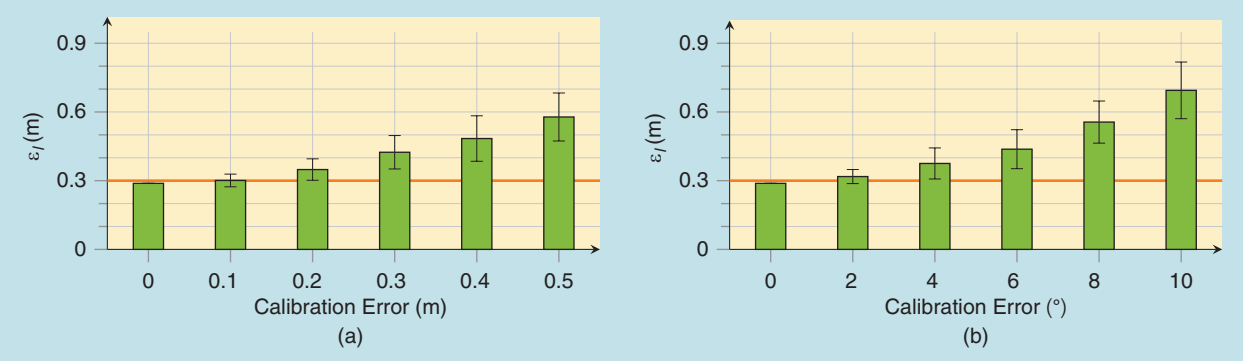

FIGURE S1. The mean speaker localization error $\varepsilon_l(m)$ as a function of (a) mean position and (b) orientation geometry calibration error.

(Figure 1). These three application scenarios, among which only the last two are relevant for WASNs, can be characterized as follows:

1) The first scenario (S1) addresses the calibration of individual microphones that are arranged in a microphone array of small geometric dimension. This configuration is characterized by the fact that the microphones are so close to each other that there exists some acoustic coherence between the captured signals. Furthermore, one can expect that all microphones share the same time base. The calibration of the positions of the individual microphones within the array is termed *array shape calibration*.
2) If each sensor node consists of a single microphone and the sensor nodes are distributed in an environment such that the microphones no longer form a compact array, the calibration process is called *microphone configuration calibration* (S2). Because of the distribution of the microphones, time synchronization between the sensor nodes cannot, in general, be expected.
3) The third scenario (S3) addresses the calibration of sensor nodes that are distributed in the environment, and where each sensor node consists of a small microphone array. This task is termed *array configuration calibration*. The relative geometric arrangement of the microphones within each array is assumed to be known. Since each node is composed of more than one microphone, the arrangement of the microphones within a node is given in terms of the array position and its orientation in space. Furthermore, the absence of time synchronization between the sensor nodes has to be expected, while the microphones within each node usually share the same time base.

Approaches to geometry calibration

The common goal of all geometry calibration algorithms is the estimation of the geometric sensor arrangement. In the following, the sensor positions will be denoted by $\mathbf{m}_m$, $m = 1, \ldots, M$, where M indicates the number of sensor nodes or microphones. The microphones are assumed to be omnidirectional, so the orientation of the individual microphones does not need to be estimated. In the case of array configuration calibration, $\mathbf{m}_m$ refers to the center of the mth array. To be able to infer the microphone positions within the array, the orientation of the sensor node also needs to be determined. It consists of the azimuth γ_m

for two-dimensional localization and of the azimuth and elevation (γ_m, φ_m) for three-dimensional localization. The positions are gathered in a $P \times M$ matrix $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_M]$. The dimensionality of the geometric space P can be either two-dimensional $(P = 2)$ or three-dimensional $(P = 3)$.

Microphone position self-calibration algorithms extract measurements from the received microphone signals, which depend on the geometric arrangement of the microphones in relation to each other or to a sound source. Four basic types of acoustic measurements can be distinguished, as illustrated in Figure 2:

1) Pairwise distance (PD) measurements (M1) $\tilde{d}_{m,n}$ can be derived from measuring the noise coherence between microphones m and n.
2) Time of arrival (ToA) measurements (M2) are obtained by receiving sound from a number of positions that will, in general, be unknown. The ToA at microphone m of a sound emitted at source k at time t_k is denoted as $t_{k,m}$. The time difference $t_{k,m} - t_k$, which is also referred to as *time of flight* (*ToF*), is proportional to the source-to-microphone distance in the case of a direct sound propagation from source to sensor.
3) Time difference of arrival (TDoA) measurements (M3) also use a number of external source positions. However, the source signals' emission times are unknown. We denote the time delay between nodes m and n of a sound emitted at position $\mathbf{e}_k$ by $\tau_{k,(m,n)}$.
4) Direction of arrival (DoA) measurements (M4), sometimes referred to as *AoA*, are measured only in the third scenario (S3), where distributed microphone arrays are considered. We denote the direction at which sound k impinges on node m with a unit norm vector $\mathbf{u}_{k,m}$.

In addition to the four acoustic measurements shown in Figure 2, briefly discussed are bimodal arrangements where video cameras are employed in addition to acoustic sensors (M5). We have not, however, discussed the visual localization problem but have concentrated on acoustic localization instead, assuming that the location of the visual sensor nodes was known.

These measurements are related to the geometric arrangement: the PDs derived from noise coherence allow for a direct computation of the geometry (M1). The ToA (M2) yields the distance between source and sensor. From TDoA measurements (M3), the difference of the propagation path from a source to two receiving microphones can be inferred, and information on a microphone array's orientation is gleaned from DoAs (M4). Furthermore, cameras on known positions are employed to resolve ambiguities and to anchor an estimated geometry in a coordinate system (M5).

Geometry calibration approaches cannot recover absolute positions. Thus, the coordinate origin is placed, without loss of generality, at the location of the first sensor node, and the second sensor is used to align the orientation. Since the calibration usually recovers only relative positions, the estimated locations exhibit an arbitrary translation and rotation with respect to the ground truth that cannot be fixed from the acoustic measurements alone. Sometimes an arbitrary reflection can occur as well. In the case of DoA-only measurements (M4), there arises an additional scale indeterminacy.

An objective function judges the deviation of the measurements from what would be expected under an assumed geometry. The form of the objective function naturally depends on the measurements used. The sought-after geometry is the one that best predicts the actual measurements. In general, the optimization problem is nonlinear and nonconvex, exhibiting multiple local minima. To avoid unfavorable local minima, researchers introduced additional information or constraints, such as low-rank matrix approximation, the assumption that the acoustic sources are in the far field, or the colocation of at least one acoustic source and sensor.

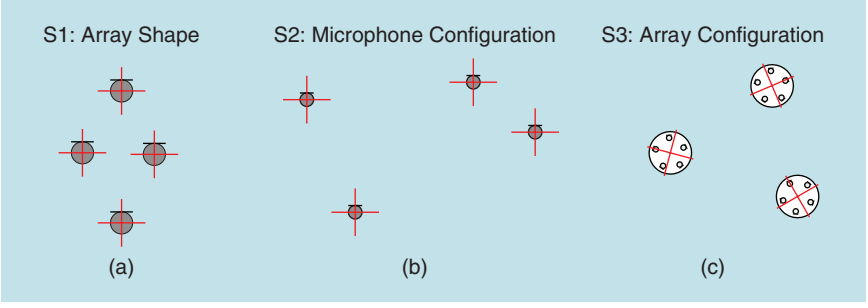

FIGURE 1. Three calibration scenarios: (a) The array shape calibration addresses the task of determining the positions of the microphones forming a compact array. (b) Microphone configuration calibration determines the position of individual microphones distributed in the room. (c) Microphone array configuration calibration seeks to determine both the position and the orientation of distributed microphone arrays whose intra-array shape is known.

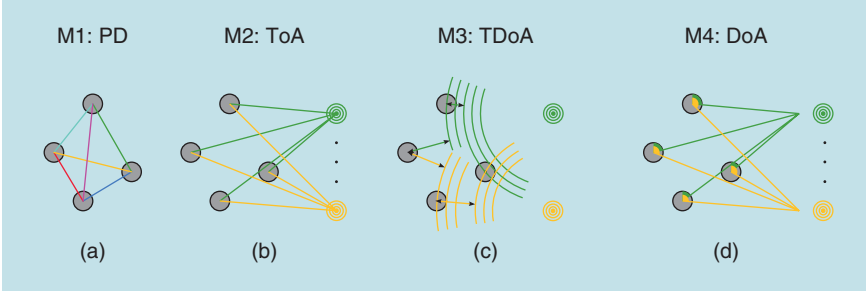

FIGURE 2. The four different types of measurements used in acoustic geometry calibration. (a) Noise coherence leads to PD measurements. (b) ToA provides distance information between a sound source and microphones. (c) The TDoA measurement yields distance-difference information between microphones in the direction of the sound source. (d) The DoA measurement provides the angle at which a sound source is observed by a sensor node.

Noise and reverberation will inevitably degrade any of the aforementioned measurements. In particular, reflections from nearby walls or objects cause errors of the measured quantities. Reverberation or the absence of a direct propagation path can especially lead to serious measurement outliers. Due to such outliers, the suitability of least squares (LS) type of objective functions, which are prevalent in geometry calibration, is somewhat questionable. However, in many cases the number of available measurements is larger than required, resulting in an overconstrained problem. Consequently, it may be attempted to remove outliers from the set of measurements. One of the most popular methods to eliminate outliers is the random sample consensus (RANSAC) [10]. Using data that provides more measurements than required, the RANSAC requires that random subsets be chosen; those that agree on a solution are used. This method is effective for removing outlier measurements and outlier geometry estimates [14], [39], [44].

To obtain the previously introduced measurements (M1–M5), three basic types of signals are used: spatially diffuse noise, dedicated point sources emitting calibration sounds such as sweeps or chirps, and natural sounds such as speech. The use of dedicated calibration signals yields the highest accuracy. However, results using speech are precise enough for most applications.

Having external sources and distributed sensor nodes raises the issue of time synchronization. On the one hand, this concerns the synchronization of the acoustic sources with the microphones and, on the other hand, the synchronization among the microphones themselves. While, in the first case, emission times will be unknown, in the second case, the measurements (M1–M4) will be affected. Microphones in an array of small geometric dimensions are usually connected to the same sampling device, resulting in a synchronized capture of the acoustic signals. However, distributed sensor nodes have different sampling devices with different frequencies and phases. Without synchronization, the TDoA estimation of a nonmoving source would indicate a movement, since the audio signals will diverge [40]. Therefore, the question of synchronization is tightly linked to the particular formulation of the geometry calibration problem. Different assumptions on synchronization will be used to group the methods discussed. The methods involve fully synchronized setups [38], two-step approaches that first solve the synchronization and then the localization problem [11], and joint localization and synchronization methods that estimate sensor and source locations together with the time synchronization parameters [26].

In the following, an overview of approaches to geometry calibration is given, which is ordered according to the measurements (M1–M5) used. For each measurement, the optimization criterion will be formulated, and relevant publications reviewed. The notation that will be used in the following is summarized in "Notation of the Most Common Quantities." A summary where the approaches are ordered according to the application scenario addressed concludes the survey.

Noise coherence (M1)

This class of approaches infers the microphone positions from distance measurements between pairs of microphones, which are obtained by evaluating the coherence function of a diffuse noise field. Since the assumption of a diffuse noise

Notation of the Most Common Quantities

Estimates are marked as $\hat{\cdot}$, measurements as $\tilde{\cdot}$.

To be estimated	$\mathbf{m}_m$	Position of sensor node m	$\mathbf{M}$	$P \times M$ matrix
	$\mathbf{e}_k$	Position of acoustic event k	$\mathbf{E}$	$P \times K$ matrix
	γ_m	Azimuth angle of microphone array m	γ	$1 \times M$ vector
	φ_m	Elevation angle of microphone array m	φ	$1 \times M$ vector
	$\mathbf{u}_{k,m}$	Unit length orientation vector pointing from node m toward source k in global coordinates		
Intermediate	δ_m	Absolute time offset		
	$\delta_{m,n}$	Pairwise time offset		
	t_k	Emission time of acoustic event k		
To be measured	$d_{m,n}$	Distance between nodes m and n	$\mathbf{D}$	$M \times M$ matrix
	$t_{k,m}$	Time of arrival of event k at node n	$\mathbf{T}$	$M \times K$ matrix
	$\tau_{k,(m,n)}$	Time difference of arrival of acoustic event k between nodes m and n		
	$\mathbf{v}_{k,m}$	Unit length orientation vector pointing from node m toward source k, measured in array's local coordinate system		

field is valid only for small microphone distances, this class of approaches is relevant only for the first scenario (S1). The coherence is defined as the normalized cross-power spectrum of two microphone signals. In the case of a diffuse noise field, it exhibits a sinc shape. The nulls of the sinc are proportional to the distance between the microphones. By fitting the theoretical diffuse noise coherence (DNC) function to its measurement, an estimate of the distance between the microphones can be obtained [24]. Note that the estimation of the coherence function requires the two microphones to be time synchronized.

Now, the microphone position matrix $\hat{\mathbf{M}}$ is determined, such that the estimated PDs $\|\hat{\mathbf{m}}_m - \hat{\mathbf{m}}_n\|$ are closest to the measured distances $\tilde{d}_{m,n}$:

$$\hat{\mathbf{M}} = \underset{\mathbf{M}}{\operatorname{argmin}} \sum_{m=1}^{M} \sum_{n=m+1}^{M} (\|\mathbf{m}_m - \mathbf{m}_n\| - \tilde{d}_{m,n})^2. \tag{1}$$

This optimization problem can be solved in a closed form using multidimensional scaling (MDS) [3]. Given all squared PD measurements $\tilde{d}^2_{m,n}$ that are arranged in distance matrix $\tilde{\mathbf{D}}$, MDS will find the spatial configuration of microphones. The fundamental insight is that the microphone configuration can be derived by eigenvalue decomposition from the scalar product matrix $\mathbf{B} = \mathbf{M}^{\mathrm{T}}\mathbf{M}$. This matrix $\mathbf{B}$ can be computed from the squared distance matrix, as shown in the following. Since the largest variance in this matrix is caused by the geometric displacement, the P-dimensional subspace wherein the microphones reside is spanned by the eigenvectors corresponding to the P largest eigenvalues of $\mathbf{B}$. Here we give only a brief overview and refer to [9] for further details.

Using the matrix of squared distance measurements $\tilde{\mathbf{D}}$, an estimate of the scalar product matrix $\hat{\mathbf{B}}$ is computed as

$$\hat{\mathbf{B}} = -\frac{1}{2}\mathbf{Q}\tilde{\mathbf{D}}\mathbf{Q}, \quad \text{where} \quad \mathbf{Q} = \mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^{\mathrm{T}} \tag{2}$$

is the row and column-wise centering matrix. Here $\mathbf{I}$ is a $M \times M$ identity matrix and $\mathbf{1}$ an m-dimensional column vector of ones. Since $\hat{\mathbf{B}}$ is symmetric and positive semidefinite, it can be decomposed into

$$\hat{\mathbf{B}} = \mathbf{V}\Lambda\mathbf{V}^{\mathrm{T}}. \tag{3}$$

The diagonal matrix Λ contains the eigenvalues, while $\mathbf{V}$ is the matrix composed of the corresponding unit length eigenvectors. Given that $\mathbf{B} = \mathbf{M}^{\mathrm{T}}\mathbf{M}$, we see the relation to the singular value decomposition (SVD) as $\mathbf{V}\Sigma\mathbf{W}^{\mathrm{T}} = \mathrm{svd}(\mathbf{M})$, with $\Sigma = \Lambda^{(1/2)}$. The P largest eigenvalues are used to compute the geometry estimate as

$$\hat{\mathbf{M}} = \mathbf{V}_P\Sigma_P, \tag{4}$$

where Σ_P is obtained from Σ by removing all except the P largest eigenvalues. Similarly, $\mathbf{V}_P$ is the correspondingly truncated matrix of eigenvectors.

In their experiments, McCowan et al. [23] reported a microphone positioning error of around 1.5 cm. Hennecke et al. [14] performed their experiment in a reverberant conference room and achieved an accuracy of around 1 cm, with the array not too close to a wall. Otherwise, reflections from nearby walls introduce a directional bias. This shows that the assumption of diffuseness has to hold to obtain precise results.

Instead of directly matching the theoretical DNC function with the measured one, Velasco et al. [46] recently derived a model for the expected generalized cross-correlation with phase transform (GCC-PHAT) output in a diffuse noise field. With this approach, they were able to decrease the error of the PD estimates to around 0.5 cm in moderately reverberant environments.

The PDs constitute a matrix with much higher dimension than its rank. Therefore, it is possible to derive the geometry with a subset of distance measurements. Taghizadeh et al. [41] developed a low-rank matrix completion strategy to deal with incomplete measurements and to handle configurations where some microphones are not inside a compact array but half a meter away.

As an alternative to MDS, Asaei et al. [1] used nonnegative matrix factorization (NMF) to derive the microphone positions from PD measurements. The low-rank property of the distance matrix was exploited to estimate the missing values by NMF.

ToA (M2)

ToA approaches measure the arrival times of acoustic events. Assuming a point source and a direct propagation path, the ToA is proportional to the distance between the source and the microphone. Let $\mathbf{e}_k, k = 1, \ldots, K$ denote the location of an acoustic event k, which will, in general, be unknown. All event locations are gathered in a $P \times K$ matrix $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_K]$. The ToA of the event k at the microphone at position $\mathbf{m}_m$ is given by [11] as

$$t_{k,m} = \frac{\|\mathbf{m}_m - \mathbf{e}_k\|}{c} + t_k - \delta_m, \tag{5}$$

where, t_k denotes the onset time of the event, δ_m is the internal recording delay, while c refers to the speed of sound. Knowledge of t_k and δ_m assumes that source and sensor share the same time base and that the signal at the source is available. While the former requires time synchronization, the latter assumes an artificial sound source such as a loudspeaker. Relevant work is discussed where this is the case, before the more challenging case of joint localization and synchronization is examined.

Artificial sound sources and common time base

If t_k and δ_m are known, we can immediately compute the distance between $\mathbf{m}_m$ and $\mathbf{e}_k$ from the measured ToA $\tilde{t}_{k,m}$:

$$\tilde{d}_{k,m} = (\tilde{t}_{k,m} - t_k + \delta_m) \cdot c, \tag{6}$$

where the term in parentheses corresponds to the ToF. The ToA $\tilde{t}_{k,m}$ can be obtained by cross-correlating the microphone

signal with the calibration sound. The corresponding objective function, which jointly estimates all source and sensor positions, is

$$(\hat{\mathbf{M}}, \hat{\mathbf{E}}) = \underset{\mathbf{M},\mathbf{E}}{\operatorname{argmin}} \sum_{m=1}^{M} \sum_{k=1}^{K} (\| \mathbf{m}_m - \mathbf{e}_k \| - \tilde{d}_{k,m})^2. \quad (7)$$

Note the similarity with (1). Thus, a direct solution of (7) is possible by employing a variant of the MDS algorithm called base point MDS (BMDS). It infers the microphone positions from PDs between the microphones and the source positions [4]. First, it computes the coordinates of a P-dimensional basis from the PDs of $P + 1$ nodes. Second, the distance measurements to the base points are used to infer the microphone positions relative to the base points. Third, a full distance matrix is constructed to run the conventional MDS algorithm.

Sachar et al. [38] used a fixed loudspeaker construction composed of four speakers on the edges of a pyramid that emitted calibration pulses. They calibrated a small 16-element microphone array with an accuracy of 0.8 cm and a large aperture array consisting of 448 microphones with an accuracy of 3 cm. Contini et al. [6] used a synchronized loudspeaker, which was moved to 25×5 positions of a rectangular grid, and white noise as a calibration signal. They were able to calibrate a linear array with an accuracy of 1 cm in both an anechoic and a strongly reverberant room.

Crocco et al. [8] used chirp signals with known emission times to obtain precise ToA estimates. An elaborate formulation of the objective function allows exploiting the constraints that the sensor and event locations are rank P matrices. By using SVD and a rank approximation technique, they were able to derive a closed-form solution in the affine space. The transformation of this solution into the Euclidean space requires an estimation of a matrix with P^2 unknown parameters. Thus, the number of unknowns became independent of the number of sensors and events. The estimation of this matrix still leads to a nonlinear optimization problem. However, this is much simpler to solve than an optimization of (7). In an experiment with eight microphones and 21 source positions, they reported an accuracy of 1 cm [7].

FIGURE 3. An ad hoc array composed of smartphones. (Photo used courtesy of TU Dortmund.)

Active devices

Several methods use active devices such as smartphones or laptops. Here, microphones and loudspeakers are colocated. PDs can be obtained from the correlation of a calibration pulse played back by the loudspeaker at one node with the signal received by the microphone at other nodes. No time synchronization is necessary, since the emission offset is canceled out by using a pair of devices as both sender and receiver. Typically, an initial estimate is computed by MDS. Thereafter, a maximum likelihood (ML) estimation is performed, incorporating the known distance between loudspeaker and microphone in an active device [13], [35].

Raykar et al. [36] derived a joint clustering-based ML estimation. In an experiment with laptops on a table, they achieved an accuracy of around 7 cm without time synchronization and 3 cm with it. Hennecke et al. [13] used smartphones lying close to one another on a table as shown in Figure 3. Using the known speaker-microphone distance and the known smartphone orientation in the second step improved the estimate and allowed resolution of the invariance to mirrored solutions. They achieved an accuracy of 7 cm for four smartphones, which were approximately 40 cm apart from each other.

Joint localization and synchronization

If the signals' onset times and recording delays of the ToA measurements are unknown, the localization problem is considerably more difficult. Gaubitch et al. [11] suggested a two-step approach, where first the timing information and then the locations were estimated. For both estimation procedures, the low-rank structure of the sensor and source location matrices was exploited.

First (5) is rewritten as

$$\frac{\mathbf{m}_m^{\mathrm{T}}\mathbf{m}_m + \mathbf{e}_k^{\mathrm{T}}\mathbf{e}_k - 2\mathbf{m}_m^{\mathrm{T}}\mathbf{e}_k}{c^2} = t_{k,m}^2 + t_k^2 + \delta_m^2 - 2(t_{k,m}t_k - t_{k,m}\delta_m + t_k\delta_m). \quad (8)$$

Next, the equations for $m = 1$ and for $k = 1$ are subtracted from (8). If this is done for $k = 2, \ldots, K$ and $m = 2, \ldots, M$, and the ToA measurements $\tilde{t}_{k,m}$ are used, the resulting system of equations can be expressed in matrix form as

$$\frac{-2\bar{\mathbf{M}}^{\mathrm{T}}\bar{\mathbf{E}}}{c^2} = \tilde{\mathbf{T}} + \boldsymbol{\Gamma}(\boldsymbol{\theta}), \quad (9)$$

where $\bar{\mathbf{M}}$ is the $P \times (M-1)$ dimensional matrix of the microphone locations relative to the first microphone, with entry $(\mathbf{m}_m - \mathbf{m}_1)$ in the $(\mathbf{m}-1)$st column, and where $\bar{\mathbf{E}}$ is the $P \times (K-1)$-dimensional location matrix of the acoustic events relative to the first event, with entry $(\mathbf{e}_k - \mathbf{e}_1)$ in the $(k-1)$st column. Furthermore, $\tilde{\mathbf{T}}$ contains the squares of the measured ToA values, and the matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ gathers the terms that depend on the unknown timing parameters $\boldsymbol{\theta} = [t_2, \ldots, t_k, \delta_1, \ldots, \delta_M]$.

Since the rank of both $\bar{\mathbf{M}}$ and $\bar{\mathbf{E}}$ is P, the rank of the matrix on the left-hand side of (9) is equal to P, which is usually much smaller than both M and K. The matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ can be viewed as correcting $\tilde{\mathbf{T}}$, such that $\tilde{\mathbf{T}} + \boldsymbol{\Gamma}(\boldsymbol{\theta})$ also has rank P. This observation opens the way to determining the timing parameters $\boldsymbol{\theta}$ [11]: first, determine the best rank-P approximation of the right-hand side of (9). This can be achieved via SVD. Then determine the parameters $\boldsymbol{\theta}$ such that the distance between the rank-P approximation and $\tilde{\mathbf{T}} + \boldsymbol{\Gamma}(\boldsymbol{\theta})$ is as small as possible. These two steps are alternated until convergence. In a subsequent work, the authors employed an alternative low-rank approximation method, the structured total LS algorithm, which was much faster [15]. However, this was achieved by playing back a sequence of chips with known timing.

Once the timing parameters $\boldsymbol{\theta}$ are estimated, the location matrices $\bar{\mathbf{M}}$ and $\bar{\mathbf{E}}$ are determined, again exploiting their rank-P property. An SVD gives

$$\bar{\mathbf{M}}^{\mathrm{T}}\bar{\mathbf{E}} = \mathbf{U}\Sigma\mathbf{W}^{\mathrm{T}}, \tag{10}$$

from which $\bar{\mathbf{M}}^{\mathrm{T}}$ and $\bar{\mathbf{E}}$ can be recovered as

$$\bar{\mathbf{M}}^{\mathrm{T}} = \mathbf{U}_P\mathbf{C} \text{ and } \bar{\mathbf{E}} = \mathbf{C}^{-1}\boldsymbol{\Sigma}_P\mathbf{W}_P^{\mathrm{T}}, \tag{11}$$

where $\boldsymbol{\Sigma}_P = \boldsymbol{\Sigma}$, all except for the P largest eigenvalues are truncated, and $\mathbf{U}_P$ and $\mathbf{W}_P$ consist of the corresponding left and right singular vectors, respectively.

The remaining problem is the estimation of the $P \times P$ matrix $\mathbf{C}$, which is much easier than the estimation of $\bar{\mathbf{M}}$ and $\bar{\mathbf{E}}$, because it is of much lower dimension. In [11], it was formulated as a nonlinear LS problem, while [8] showed that a closed-form solution can be found if one sensor is colocated with a source. If the timing parameters are known, only the second part of the algorithm needs to be carried out, as described in [8].

Using microphones randomly distributed on a table, an accuracy of 2 cm was achieved with this method [11]. The ToA measurements were obtained from recordings of handclaps. However, the authors needed to label the largest peak in each handclap signal manually to obtain precise estimates.

TDoA (M3)

The TDoA is proportional to the distance difference of a pair of microphones to the source, when the direct path from the source to the microphones exists. The TDoA itself may be measured by the maximum in the correlation of the two microphone signals, or by onset detection. To derive the geometry, this measurement is related to the positioning as follows: The TDoA from the kth source to the mth and nth sensor is given by

$$\tau_{k,(m,n)} = t_{k,m} - t_{k,n} = \frac{\|\mathbf{m}_m - \mathbf{e}_k\| - \|\mathbf{m}_n - \mathbf{e}_k\|}{c} - \delta_m + \delta_n. \tag{12}$$

Note that the onset time t_k cancels out. Here, $-\delta_m + \delta_n$ is the time offset between the recording devices.

Relevant work is discussed first, which assumes time synchronization and thus absence or, equivalently, knowledge of the delays, before we turn to the case where the delays have to be estimated as ancillary parameters.

Synchronized microphones

In reverberant environments, the steered response power with phase transform is often employed for TDoA-based localization. This is equivalent to a filter-and-sum beamformer, when steering the beamformer to all possible locations and selecting the position where the output energy is maximized [5]. The individual source positions obtained from several distributed microphone arrays located at the ceiling of a highly reverberant conference room were used by Hennencke et al. to perform a coordinate mapping [14]. In their experiments, speech and noise emitted from random positions was used together with a RANSAC scheme. They achieved an accuracy of 10 cm with speech and white noise in most cases.

If a sensor node consists of a microphone array, an acoustic camera can be formed by a delay-and-sum beamformer applied to the received signals of each array. Redondi et al. [37] used pure sinusoids as source signals to obtain acoustic images. This enables the application of computer vision techniques. They used camera models to extract positions in Cartesian coordinates, which were used as input for a subsequent coordinate mapping approach. The coordinate mapping approach recovered the rotation and translation for each microphone array to a selected reference array.

Thrun [43] showed that the localization problem can be significantly simplified if the sources are in the far field of the microphones. Then a source signal impinges on all sensors from the same angle, and the actual position of the source is immaterial. Thus, we can write $\|\mathbf{m}_m - \mathbf{e}_k\| = \mathbf{u}_k^{\mathrm{T}}(\mathbf{m}_m - \mathbf{e}_k)$, where the unit-length direction vector $\mathbf{u}_k$ only depends on the source and is independent of which microphone is considered. Then the right-hand side of (12) simplifies to $\mathbf{u}_k^{\mathrm{T}}(\mathbf{m}_m - \mathbf{m}_1)/c$, which has to be compared to the measured TDoA $\tilde{\tau}_{k,(m,1)}$, leading to the overdetermined system of equations

$$\mathbf{U}^{\mathrm{T}}\bar{\mathbf{M}}/c = \tau, \tag{13}$$

where $\mathbf{U}$ is the $P \times K$ matrix, with $\mathbf{u}_k$ on the kth column, and $\boldsymbol{\tau}$ is the $K \times (M-1)$ matrix of measured TDoAs. Note that Thrun assumes that the δ terms are known and thus can be set to zero.

Again, the rank argument can be invoked. The rank of the left-hand side of (13) is P, and so must be the rank of the right-hand side. Thus, we can apply the same rank approximation by SVD as explained in (10) and (11) to infer the microphone positions. This approach of Thrun and others has become known as *affine structure from sound* (*ASfS*).

Unsynchronized microphones

In the general case of unsynchronized microphones, the recording delays δ_m, $m = 1, \ldots, M$ are different and unknown.

A joint estimation with the locations leads to the nonlinear LS problem [25]

$$\hat{\mathbf{M}}, \hat{\mathbf{E}}, \hat{\boldsymbol{\delta}} = \underset{\mathbf{M},\mathbf{E},\boldsymbol{\delta}}{\operatorname{argmin}} \sum_{m=2}^{M} \sum_{k=1}^{K} \left(\frac{\|\mathbf{m}_m - \mathbf{e}_k\| - \|\mathbf{m}_1 - \mathbf{e}_k\|}{c} - \delta_m + \delta_1 - \tilde{\tau}_{k,(m,1)} \right)^2. \tag{14}$$

In [25], the optimization problem was solved using an auxiliary function-based algorithm, which is an extension of the expectation-maximization (EM) algorithm. The alternating optimization of the positions and timing parameters was shown to have better convergence properties than gradient descent.

As an alternative to this one-stage scheme that estimates the unknown positions and timing parameters simultaneously, Wang et al. [47] recently proposed a two-stage method. This method again uses a rank approximation technique. While the aforementioned rank approximation-based algorithms [8], [11], [12] are suitable for ToA and require at least the onset times or the internal delays to be known, Thrun [43] was able to work with TDoA; however, both onset time and internal delays needed to be known. Wang et al. [47] coped with these limitations and derived an approach that uses TDoA measurements and estimates the unknown onset times and internal delays. Furthermore, Wang et al. showed that the combination of their algorithm and the EM algorithm-based calibration from [25] can outperform the individual methods.

An interesting insight is that the maximum TDoA (mTDoA) can be used to estimate PDs [28]. The maximum of the TDoA values is obtained if the source is aligned with the microphone-to-microphone direction (the so-called endfire position). To achieve this condition, sound from a variety of distributed source positions is used and the maximum is computed over the whole period over all of them. Unlike the noise coherence method, this approach does not require a time synchronization between the microphones and is able to handle larger microphone distances. On the other hand, the intermicrophone distance will be underestimated by the mTDoA approach if no event is observed in the endfire position. Let

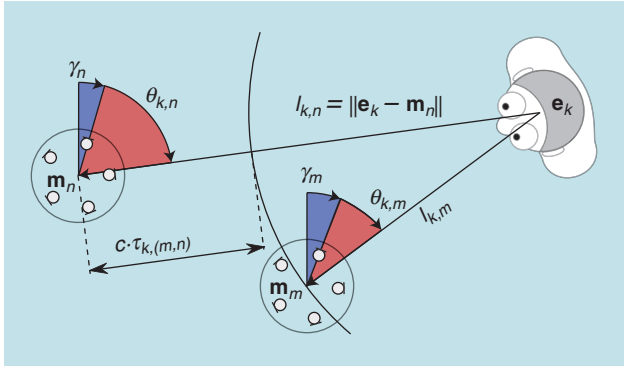

FIGURE 4. Two sensor nodes consisting of five-element circular arrays at $\mathbf{m}_m$ and $\mathbf{m}_n$ and a source at $\mathbf{e}_k$. The DoA θ is offset by the angle γ with respect to the global coordinate system. The TDoA $\tau_{k,(m,n)}$ corresponds to the difference between the distances of each array to the speaker. The illustration is in 2-D, so $\mathbf{v}_{k,m} = (\cos\theta_{k,m}, \sin\theta_{k,m})^T$.

$$\tilde{\tau}_{n,m}^{\max} = \max_k \{\tilde{\tau}_{k,(m,n)}\} = \|\boldsymbol{m}_n - \boldsymbol{m}_m\|/c + \delta_m - \delta_n \tag{15}$$

and

$$\tilde{\tau}_{n,m}^{\min} = \min_k \{\tilde{\tau}_{k,(m,n)}\} = -\|\boldsymbol{m}_n - \boldsymbol{m}_m\|/c + \delta_m - \delta_n, \tag{16}$$

where we assume that an acoustic event exists in either endfire position. Thus, it follows that the time offset can be computed as in [27], i.e., that

$$\hat{\delta}_{m,n} = \delta_m - \delta_n = \frac{1}{2}(\tilde{\tau}_{m,n}^{\max} + \tilde{\tau}_{m,n}^{\min}), \tag{17}$$

and that the distance is

$$\hat{d}_{m,n} = (c/2)(\tilde{\tau}_{m,n}^{\max} - \tilde{\tau}_{m,n}^{\min}). \tag{18}$$

Now, MDS can be invoked again to derive an estimate of the microphone geometry [28]. The performance of the result depends on whether the mTDoA is observed. However, reasonable results are obtained even if this is not true for all pairs. Both Parviainen et al. [26] and Pertilä et al. [28] performed experiments with smartphones in a meeting room, where they achieved around a 12-cm accuracy.

DoA (M4)

DoA-based calibration can be conducted only for the third scenario (S3) described in the "Application Scenarios" section, since it requires a microphone array per sensor node rather than a single microphone to acquire DoA estimates.

An objective function is formed by comparing the direction of the kth acoustic event, as measured by array m, $\tilde{\mathbf{u}}_{k,m}$, with the direction predicted by the assumed geometry: $\mathbf{u}_{k,m} = (\mathbf{m}_m - \mathbf{e}_k)/\|\mathbf{m}_m - \mathbf{e}_k\|$. Since we are working with directions, a Euclidean distance measure is inappropriate. A cosine distance measure is used instead

$$1 - \cos(\sphericalangle(\tilde{\mathbf{u}}_{k,m}, \hat{\mathbf{u}}_{k,m})) = 1 - \tilde{\mathbf{u}}_{k,m}^{\mathrm{T}} \mathbf{u}_{k,m}. \tag{19}$$

The vector $\tilde{\mathbf{u}}_{k,m}$ describes the measurement of the impinging angle with respect to a global coordinate system. However, a measurement $\tilde{\mathbf{v}}_{k,m}$, which can be obtained from the microphone signals, is located in the local coordinate system of the sensor node. For an illustration in two dimensions, see Figure 4. Since the array exhibits an unknown azimuth γ_m and elevation φ_m with respect to the global coordinate system, the rotation has to be compensated. This is achieved using a rotation matrix $\mathbf{R}(\gamma_m, \varphi_m)$

$$\tilde{\mathbf{u}}_{k,m} = \mathbf{R}^{-1}(\gamma_m, \varphi_m)\tilde{\mathbf{v}}_{k,m}. \tag{20}$$

This transformation allows the combination of all measurements in a common objective function, if (20) is plugged in to (19) and we eventually arrive at

$$\hat{\mathbf{M}}, \hat{\mathbf{E}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}} = \underset{\mathbf{M},\mathbf{E},\boldsymbol{\gamma},\boldsymbol{\varphi}}{\operatorname{argmin}} \sum_{m=1}^{M} \sum_{k=1}^{K} (1 - \tilde{\mathbf{v}}_{k,m}^{\mathrm{T}} \mathbf{R}(\gamma_m, \varphi_m)\mathbf{u}_{k,m}). \tag{21}$$

The optimization of (21) was carried out in [19] by using the Newton algorithm. Since [19] was solely working on DoA estimates, neither a time synchronization between the sensor nodes nor a synchronization between the source and sensors was required. The only requirement was that the microphones within an array were synchronized, since the DoA is eventually measured by TDoAs. To handle noisy measurements, the RANSAC framework was employed.

A DoA-based localization is obviously unable to determine the scale of the geometry. Scale ambiguity can be resolved by employing an additional TDoA measurement, as proposed in [39]. These measurements provide distance differences, which are used to scale the DoA-only calibration result, such that the geometry matches the distance differences. The overall system is shown in Figure 5. However, the interarray TDoA estimation requires a time synchronization between sensor nodes. Jacob et al. [16] showed that the knowledge of the intra-array geometry of a circular array is sufficient to solve the scale ambiguity problem. Thus, the TDoA can be omitted.

The method introduced in [31] combined DoA measurements from the individual microphone arrays and TDoA measurements between the arrays. For each pair m, n of arrays, the position and orientation have to fulfill the geometric relations with respect to the measured TDoA and DoA, as illustrated in Figure 4. Given an estimate of the azimuthal $\hat{\gamma}_m, \hat{\gamma}_n$ and elevation displacements $\hat{\varphi}_m, \hat{\varphi}_n$, and estimates of the positions $\hat{\mathbf{m}}_m, \hat{\mathbf{m}}_n$, the measured directions $\tilde{\mathbf{u}}_{k,m}$ and $\tilde{\mathbf{u}}_{k,n}$ can be used to compute the source position by triangulation [31]

$$\begin{aligned}\hat{\mathbf{e}}_k(m,n) &= \hat{\mathbf{m}}_m + \hat{l}_{k,m}\tilde{\mathbf{v}}_{k,m}\mathbf{R}^{-1}(\hat{\gamma}_m, \hat{\varphi}_m) \\ &= \hat{\mathbf{m}}_n + \hat{l}_{k,n}\tilde{\mathbf{v}}_{k,n}\mathbf{R}^{-1}(\hat{\gamma}_n, \hat{\varphi}_n).\end{aligned} \tag{22}$$

Here, the notation $\hat{\mathbf{e}}_k(m,n)$ indicates that the source position estimate is obtained from measurements of arrays m and n. The line intersection provides estimates $\hat{l}_{k,m} = \|\mathbf{e_k} - \mathbf{m}_m\|$ and $\hat{l}_{k,n}$ of the distances to the source. If either of these distances is negative, there is no intersection and the solution points to a mirrored position.

Next, the measured TDoA is compared with the value predicted by the current estimates of source and sensor positions. As can be seen in Figure 4, they are related as $c\cdot\tau_{k,(m,n)} = l_{k,n} - l_{k,m}$. This comparison allows the formation of an objective function that is determined for three or more source positions. It is used to estimate the geometry hierarchically (Figure 6). First, the geometry of all pairs $(1, m), m = 2, \ldots, M$ is estimated individually. Then these estimates are used as a starting point to estimate the geometry of all microphone arrays jointly. The procedure is repeated to improve the estimation and remove a bias that can be the result of an individual DoA error. Random subsets of three or more positions are used, and the final geometry estimation result is obtained as the average of several such subsets.

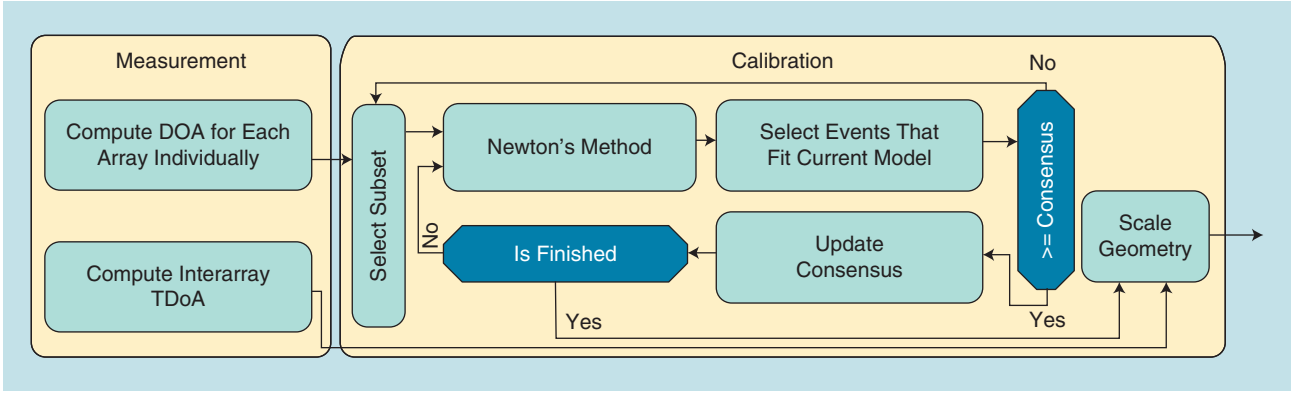

FIGURE 5. A DoA-based calibration embedded into a RANSAC framework [19] and combined with scale factor estimation from [39].

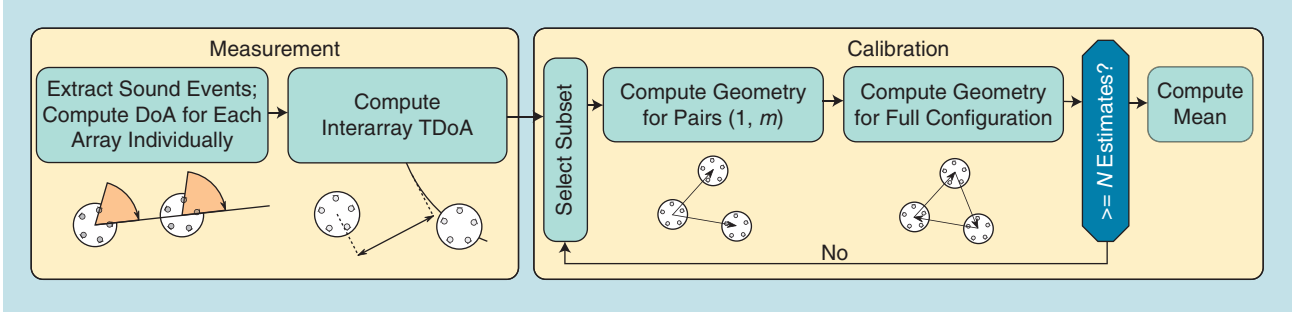

FIGURE 6. A DoA–TDoA method for the calibration of distributed microphone arrays [31].

The combined DoA–TDoA method for microphone array configuration calibration achieved a position accuracy of 10 cm and an angular accuracy of 2° using speech, compared to 1 cm and 1° accuracy when using noise [31]. These values were measured in a 2-D scenario with tabletop microphones.

Visual support (M5)

Acoustic sensors are often accompanied by visual ones. This allows joint calibration of the audiovisual sensor network employing both the acoustic and visual modality. However, the discussion of the geometry estimation of a camera network and thus the joint calibration of audiovisual sensor networks is beyond the scope of this survey. We will only briefly discuss how known positions of cameras can be used to resolve the translation and rotation indeterminacy and, if required, also the scale indeterminacy of a purely acoustic geometry calibration.

If a speaker is tracked separately, both by a camera network with known camera positions and by a microphone array whose positions have been estimated by one of the aforementioned geometry calibration methods, the microphone positions have to be embedded into the coordinate system provided by the camera network. This is achieved by matching the acoustic and visual speaker trajectory [17], thus fixing the translation, rotation, and scale of the acoustic sensor network. An alternative to this postmatching of trajectories is "online" joint audiovisual localization, as described in [18].

In [30], Plinge and Fink assumed that the visual modality provides the absolute positions of the speakers, from which the microphone geometry is calibrated using DoA measurements. This approach can be seen as a reverse localization problem, where the source positions are known and the sensor positions need to be estimated. As the relation of the DoA and the source position is straightforward (Figure 4), each node's position and orientation can be estimated along with the distances $\boldsymbol{l}_m = (l_{1,m} \ldots l_{K,m})^{\mathrm{T}}$. For any set of three or more speaker positions e_k, the corresponding equations can be combined into an overdetermined system of equations. The geometric arrangement of the microphone arrays can be determined by solving the resulting objective function

$$\hat{\gamma}_m, \hat{\mathbf{m}}_m, \hat{\boldsymbol{l}}_m = \underset{\boldsymbol{m}_m, \gamma_m, \boldsymbol{l}_m}{\operatorname{argmin}} \sum_k \| \mathbf{e}_k - \mathbf{m}_m - l_{k,m} \tilde{\boldsymbol{v}}_{k,m} \boldsymbol{R}^{-1}(\hat{\gamma}_m, \hat{\varphi}_m) \|. \tag{23}$$

Given reasonably accurate visual localizations, this objective function is approximately convex, since it increases monotonously with both position and orientation errors. Therefore, gradient descent can be used, and a solution is found regardless of the initialization.

By embedding the microphone positions into the coordinate system of the camera system, the cross-modality speaker tracking and localization outperformed a solely acoustic and a solely visual localization. Thus, it is beneficial to embed the microphone positions into the coordinate system of the camera system.

Summary

Table 1 gives an overview of the calibration approaches. They are grouped by the application scenarios (S1–S3) introduced in the "Application Scenarios" section and the measurements employed (M1–M5). The references given are accompanied by a few remarks concerning the methodology and types of experiments conducted. In the following, we summarize some design considerations.

Objective functions

MDS or BMDS is the preferred approach if PDs are given. It is used in all three application scenarios. If solely compact arrays are considered, the distance can be obtained by the diffuse noise approach. For distributed active devices, the required distance measurements can be obtained by correlating transmitted and received signals. However, the mTDoA approach is able to work with ambient sounds or speech signals. ToA, TDoA, and DoA lead to nonlinear LS problems, where care has to be taken to avoid unfavorable local minima.

Number of source positions

The required number of source positions for a successful calibration varies significantly. Only the diffuse noise approach does not require any active sources—except for the presence of diffuse noise. The number of different source positions used in the other experiments ranges from about ten to more than 100, depending on the method and objective function used. In the past, several approaches used loudspeakers mounted at fixed positions [20], [38]. With the advent of more elaborate methods, a single moving source could be utilized. A particularly convenient method is to employ a handheld smartphone playing noise or chirps [12], [31].

Reverberation will degrade any measurement. Apart from increasing the number of source positions, additional steps to increase the robustness have to be added. These include robust estimation [29] or RANSAC [39], [45].

Signal types

We can distinguish three classes of sounds used: diffuse noise, dedicated calibration signals, and natural sounds such as speech. Spatially diffuse noise can be used only for the calibration of small microphone arrays [3], [41].

Many calibration methods use calibration signals with good correlation properties, such as white noise or chirps, that allow for exact TDoA measurements. The use of a calibration signal provides more accurate calibration results than the use of natural sounds.

For online and ad hoc scenarios, speech is preferred—but the use of speech is challenging, since it will provide less accurate measurements compared to white noise due to its less sharp autocorrelation function. Additionally, when speech is used, it is necessary to preprocess the microphone signal by voice activity detection or signal classification to exclude badly localized sounds produced, for example, by furniture, chairs, footsteps, or doors [33]. The accuracy of speech-based

Table 1. An overview of the calibration methods discussed, ordered by scenario and measurement.

	Array Shape (S1)	Microphone Configuration (S2)	Array Configuration (S3)
PD (M1)	• MDS — Manual measurements [3] — Diffuse noise [24], [46] • Rank approximation — Diffuse noise + far-field microphones [41]	• NMF — Incomplete + Noisy distance [1]	
ToA (M2)	• BMDS — Known distances [4] • Direct minimization — Known loudspeaker configuration [6], [38]	• Direct minimization — Known loudspeaker configuration [38] • Rank approximation via SVD — Known emission time [7], [8] — Impulse train [12] — Manually labeled handclaps [11] • Active devices — One microphone and event are colocated [35] — Unsynchronized + MDS [13], [36]	
TDoA (M3)		• Direct minimization — No initialization [34] — Auxiliary function [25] • Rank approximation via SVD — Far-field sources [43] — Unsynchronized [47] • mTDoA — MDS + speech [26]–[28]	• Coordinate mapping — Direct minimization [45] — Random sampling [14] — Acoustic camera [37]
DoA (M4)			• RANSAC — Random walk + speech [16], [19] • Intra-array TDoA — Fixed speaker positions [32] — Random walk [39]
Visual support (M5)			• Audiovisual — Trajectory mapping [17] — Visual speaker localization [30] — Joint calibration [18]

methods tends to be lower, but fortunately high enough for practical applications.

Synchronization

If the clocks of transmitter and receiver are synchronized, PDs can be obtained from ToA measurements. TDoA measurements require only a synchronization among the microphones. A further relaxation is possible if only DoA measurements are incorporated. Some algorithms couple the estimation of the sampling deviation and the calibration process itself [11], [25], [47], while the mTDoA method elegantly removes a potential unknown delay [26], [27] (cf. the treatment of timing difference by the mTDOA approach in the "Unsynchronized Microphones" section).

Experimental evaluation of selected methods

The section "Approaches to Geometry Calibration" provided an overview about a broad range of geometry calibration algorithms. The authors of the corresponding publications evaluate their algorithms usually on their proprietary data sets, which makes a comparison among different approaches difficult. This section tries to fill this gap and provides a comparison under a common evaluation framework. We conducted experiments for all three application scenarios, and for each scenario we evaluated a selection of algorithms in a two-dimensional calibration experiment in a reverberant laboratory environment. We also did our best to correctly implement and optimize those algorithms that we have not proposed. We do not claim, however, that we achieved their best possible performance. Table 2 provides an overview of the algorithms selected and the scenarios where they have been applied. In the following, we first describe the test environment and the performance evaluation metrics used. Afterward, we present the results for each of the three scenarios. The evaluation is concluded by a short summary.

Table 2. An overview of the methods used in the evaluation.

Array Shape (S1)	Microphone Configuration (S2)	Array Configuration (S3)
DNC + MDS [24]	ToA rank [11]	DoA + TDoA scaling [19], [39]
mTDoA + MDS [28]	mTDoA + MDS [28]	DoA–TDoA [31]
ASfS [43]	ASfS [43]	DoA + Video [30]

Evaluation setups and metrics

The location was a highly reverberant 3.7 m × 6.8 m × 2.6 m conference room of a smart house installation at TU Dortmund University. Signals from three circular microphone arrays that were arranged in an irregular triangle of an approximate edge size of 1 m were recorded at 48 kHz (Figure 7). Each array was embedded in the table and consisted of five microphones arranged equidistantly on a circle of radius 5 cm. The signals were captured synchronized. A reverberation time (T_{60}) of 0.67 seconds was calculated using a blind estimation algorithm [21]. Five cameras mounted at the ceiling captured the scene at 10 frames/second and 384 × 288 pixel resolution. They have a field of view of 48° × 36°. Acoustic events were produced from ten locations around the table. For the first recording, a smartphone was held at the same height as the microphones, and a white noise signal was played back. In the second case, a speaker was either sitting (four positions) or standing (six positions) at the table. Consequently, his mouth was approximately 0.4 m or 0.7 m above the microphones, respectively.

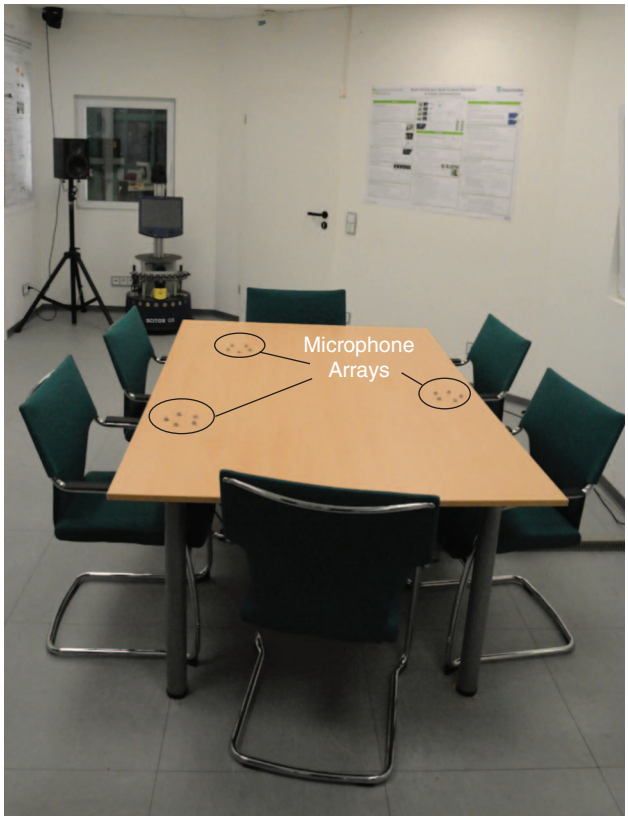

FIGURE 7. The recording setup. The conference table is embedded with three circular microphone arrays.

As mentioned previously, the estimated geometry $\hat{\mathbf{M}}$ exhibits an arbitrary translation and rotation with respect to the true geometry that needs to be removed before an error can be measured. To this end, a translation vector $\mathbf{t}$ and a rotation matrix $\mathbf{R}$ are determined by SVD, such that the mean location error between the estimated microphone positions after correction and the true microphone positions is minimized. Therefore the estimated positions after correction are given by $\hat{\mathbf{m}}_i' = \mathbf{R}\hat{\mathbf{m}}_i + \mathbf{t}$, where $\hat{\mathbf{m}}_i$ represents the original estimates. The performance measure for the positions is the root-mean-square (RMS) error

$$\epsilon_p = \sqrt{\frac{1}{M}\sum_{i=1}^{M} \| \boldsymbol{m}_i - \boldsymbol{m}_i' \|^2}. \tag{24}$$

In the case of array configuration calibration, the orientation of the arrays is also an important parameter to estimate. The estimate has an arbitrary rotation relative to the ground truth. To compensate for this, an angle d_γ is determined such that the deviation of the ground truth from the estimated angles $\hat{\gamma}_j$ after rotation by d_γ is minimal. Then the average orientation error is computed as

$$\epsilon_\gamma = \frac{1}{M}\sum_{m=1}^{M} |\gamma_m - d_\gamma - \hat{\gamma}_m|. \tag{25}$$

Since several algorithms solve nonlinear problems and use random initializations, each experiment is repeated ten times. The average and standard deviation computed over these runs is reported.

Microphone array shape calibration (S1)

We compare the DNC approach [24] and the mTDoA [28] algorithm, which both estimate PDs, from which the overall geometry is inferred by MDS. The comparison includes the ASfS approach [43], which originally was developed for microphone configuration calibration but can also be employed for microphone array shape calibration.

The methods achieved a positioning accuracy of 0.3–1.5 cm for the microphones of the circular array described above. The overall results for the DNC and mTDoA method were rather similar, while ASfS, which was developed for microphone configuration calibration problems, performed slightly worse (Figure 8). We observed, though, that the mTDoA + MDS method does not degrade as quickly as the DNC approach when the interelement distance is increased. The DNC approach relies on the presence of ambient diffuse noise, while the mTDoA algorithm relies in the presence of sound sources being in the endfire position of microphone pairs. Either requirement may not always be fulfilled.

Microphone configuration calibration (S2)

The microphone configuration calibration performance was evaluated on all 15 microphones of the three circular arrays. The mTDoA approach combined with MDS [28] worked for pairs of microphones from different arrays with a distance of up to about 1 m. We therefore included it in our evaluation and compared it to the ASfS approach [43] and the ToA rank approximation scheme [11].

Figure 9 compares the calibration error of the methods using either white noise or speech as calibration signals. The best localization is achieved with mTDoA, resulting in an RMS error of 4.5 cm ± 1.8 cm using a speech signal, while 1.3 cm ± 0.7 cm was achieved with noise excitation. The ASfS approach performed slightly worse with 6.0 cm ± 2.7 cm and 1.8 cm ± 0.8 cm, respectively. Our implementation of the ToA rank scheme did not perform well. This might be a consequence of the arrangement, since experiments with uniformly distributed microphones performed significantly better.

Array configuration calibration (S3)

For the array configuration calibration, the DoA–TDoA method [31], the DoA + Video method [30], and the DoA + TDoA scaling approach [19], [39] were used.

The required DoAs were estimated by a neurobiologically inspired method [29], since it is robust to noise and excludes nonspeech sounds. The event locations were automatically identified as segments with a small angular variance. The error in the DoA angle estimation was around 3°. The TDoA information was extracted by computation of the GCC-PHAT.

The TDoAs over all microphone pairs had an error of around 6 cm. For the multimodal method, visual localization by background subtraction and an upper-body detector was used [30]. Seven localizations with an accuracy of 20 cm were derived for the ten detected speech segments. For the noise sequence, the ground truth positions marked on the floor where the sounds were produced were used.

The calibration results are shown in Figure 10. All methods achieved an average position error ϵ_p of less than 7 cm. The maximum position error for the DoA–TDoA method was 0.6 cm for noise and 6.0 cm for speech. For the DoA + Video approach, it was 5 cm for noise and 7 cm for speech. For the DoA + TDoA scaling method, it was 4 cm for noise and 2.5 cm for speech. The angular error is close to 1° for all methods, except for the DoA–TDoA method using speech, where only 3° was achieved.

Summary

In our experiments, array calibration could be performed quite accurately using diffuse noise or mTDoA from multiple distributed speech events. Below 1-cm precision is close to the requirements for beamforming. While the diffuse noise approach is limited to small array sizes, the latter method also allowed the calibration of distributed microphones on a table using speech or noise with 10-cm and 3-cm precision,

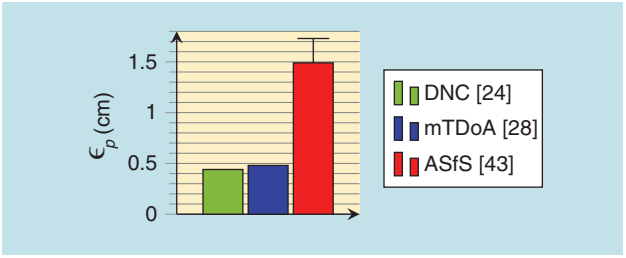

FIGURE 8. Array shape calibration. The mean position error for tabletop microphones in a reverberant smart room.

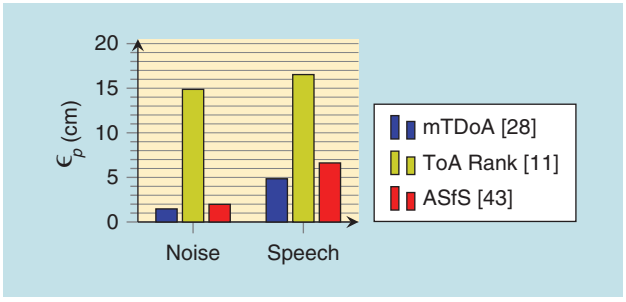

FIGURE 9. Microphone configuration calibration. The mean position error using either noise or speech as input signals.

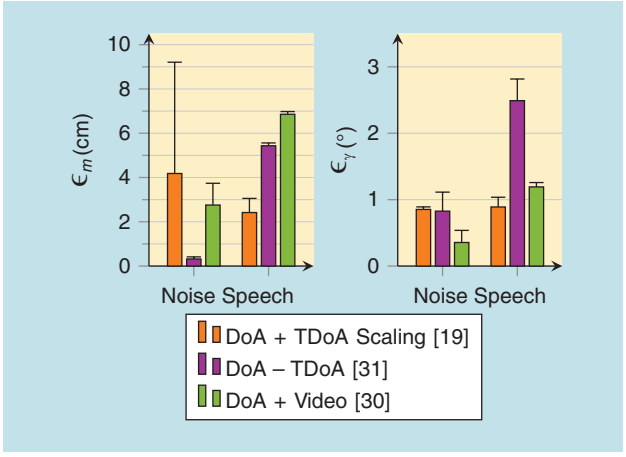

FIGURE 10. Array configuration calibration. The mean position and orientation error for three DOA-based algorithms.

respectively. This can also be achieved using the ASfS method [43], with a slightly higher error. We were able to show that state-of-the-art methods [19], [31] are capable of calibrating array configurations with an orientation error of well below 5° and a position error well below 10 cm. This provides sufficient accuracy for triangulation-based processing algorithms.

Conclusions and outlook

This article provided a survey of acoustic geometry calibration algorithms, which attempt to reveal the position of microphones solely from the acoustic signals received by them. The algorithms can be categorized on the basis of such things as the kind of acoustic signals used, the kind of position-related measurements employed, the kind of objective function used, and the necessity of synchronization.

We have chosen to organize the presentation according to two criteria. The first is based on the scenario addressed: the position estimation of individual microphones within a microphone array (S1), the localization of distributed microphones (S2), and the positioning of distributed microphone arrays (S3). The second criterion is the measurements used, from which position-related information is extracted: the noise coherence function (M1), which is related to the distance between two microphones; the ToA (M2), which is related to the distance between a sound source and a sensor; the TDoA (M3) between the signals at two microphones, which is related to the distance between the microphones and the angle at which the sound is observed; and finally the DoA (M4), from which the relative geometry of sources and sensors can be revealed. These two categorizations have been chosen to provide, on the one hand, an application-oriented point of view and, on the other hand, a technology-oriented perspective. While the first allows a practitioner to quickly identify which approach is suitable for which application, the second may help researchers to sort the different approaches according to which input data are used.

The survey and the experimental evaluation showed that the highest accuracy is achieved if dedicated calibration signals are used, such as chirps from known positions, and if transmitter and receiver are synchronized. A positioning accuracy on the order of 1 cm has been reported in all scenarios. For usability reasons, however, a localization from natural sounds such as speech is preferable. But even with speech, accuracies in the range of 5 cm are achievable, which is high enough for applications like speaker tracking.

Further improvements in accuracy and usability can be obtained along different lines of research. First, the quality of the measurements should be improved. The aforementioned measurements are already the result of some signal processing. They are often obtained from cross-correlations of acoustic signals, and these correlations are heavily affected by noise and reverberation. More robust correlation results would immediately lead to improved input data for the calibration process and thus improved calibration results.

Second, the objective functions are often nonlinear LS-type functions, which exhibit multiple local minima and which are often optimized iteratively. However, due to reverberation and reflections, the measurement error is not normally distributed, putting the suitability of LS into question. Today, mostly ad hoc countermeasures such as RANSAC are used, while a more principled approach to handle outliers has yet to come.

Furthermore, the optimization should be guided by a priori knowledge or sensible assumptions to avoid unfavorable local minima. Examples that have proven effective are the exploitation of the low-rank property of position and distance matrices or the far-field assumption. Such constraints are the more important the more complex the optimization problem becomes. A particularly complex case is the calibration of unsynchronized microphones from ToA or TDoA measurements using speech input in noisy reverberant environments. This requires the synchronization problem to be solved before or along with the localization problem. We expect more work in this field, both because of the difficulty, and thus attractiveness, of the problem, and because of its practical importance.

Acknowledgment

This work was supported by the German Research Foundation under contracts Fi 799/5-1 and Ha3455/7-2.

Authors

Axel Plinge (axel.plinge@tu-dortmund.de) received his diploma degree with distinction in computer science with a minor in philosophy from the Technical University of Dortmund, Germany, where he worked at the Leibniz Research Centre for Working Environment and Human Factors in different areas of psychophysical research from hearing to color vision and depth perception. In two European research projects, he developed innovative speech enhancement and replacement methods for persons with severely sensory-impaired hearing. He has published several papers on a variety of topics, including human hearing and vision, speech enhancement, pattern recognition, speaker tracking, and geometry calibration.

Florian Jacob (jacob@nt.uni-paderborn.de) received his Dipl.-Ing. degree in computer engineering with a specialization in electrical engineering from Paderborn University, Germany, in 2011. He has been a research staff member in the Department of Communications Engineering since 2011. His research interests include acoustic signal processing and unsupervised geometry calibration of distributed sensor networks. He is currently working toward his Ph.D. degree.

Reinhold Haeb-Umbach (haeb@nt.uni-paderborn.de) received his Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from Rheinisch-Westfälische Technische Hochschule Aachen, Germany, in 1983 and 1988, respectively. From 1988 to 1989, he was a postdoctoral fellow at the IBM Almaden Research Center, San Jose, California, and conducted research on coding and signal processing for recording channels. From 1990 to 2001, he was with Philips Research and worked on various aspects of automatic speech recognition. Since 2001, he has been a professor of communications engineering at Paderborn University, Germany. His main research interests are statistical speech signal processing and recognition. He has published more than 150 papers in peer-reviewed journals and conferences.

Gernot A. Fink (Gernot.Fink@udo.edu) received his diploma in computer science from the University of Erlangen-Nuremberg, Germany, in 1991. From 1991 to 2005, he was with the Applied Computer Science Group at Bielefeld University, Germany, where he received his Ph.D. degree (Dr.-Ing.) in 1995 and his venia legendi (Habilitation) in 2002. Since 2005, he has been a professor at the Technical University of Dortmund, Germany, where he heads the Pattern Recognition in Embedded Systems Group. His research interests are machine perception, statistical pattern recognition, and document analysis. He has published more than 150 papers and a textbook on Markov models for pattern recognition.

References

[1] A. Asaei, N. Mohammadiha, M. J. Taghizadeh, S. Doclo, and H. Bourlard, "On application of non-negative matrix factorization for ad hoc microphone array calibration from incomplete noisy distances," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Brisbane, Queensland, Australia, Apr. 2015, pp. 2694–2698.

[2] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symp. Commun. and Vehicular Technol. in the Benelux*, Ghent, Belgium, Nov. 2011, pp. 1–6.

[3] S. T. Birchfield, "Geometric microphone array calibration by multidimensional scaling," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, Hong Kong, Apr. 2003, pp. 157–160.

[4] S. T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 1025–1034, Sept. 2005.

[5] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Springer Science and Business Media, Berlin, Germany: Springer-Verlag, 2001.

[6] A. Contini, A. Canclini, F. Antonacci, M. Compagnoni, A. Sarti, and S. Tubaro, "Self-calibration of microphone arrays from measurement of times of arrival of acoustic signals," in *Int. Symp. Commun. Control and Signal Processing*, May 2012, pp. 1–6.

[7] M. Crocco, A. Del Bue, M. Bustreo, and V. Murino, "A closed form solution to the microphone position self-calibration problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 2597–2600.

[8] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, Feb. 2012.

[9] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: Essential theory, algorithms, and applications," *IEEE Signal Processing Mag.*, vol. 32, no. 6, pp. 12–30, Nov. 2015.

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

[11] N. Gaubitch, W. Kleijn, and R. Heusdens, "Auto-localization in adhoc microphone arrays," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 106–110.

[12] N. Gaubitch, W. Kleijn, and R. Heusdens, "Calibration of distributed sound acquisition systems using TOA measurements from a moving acoustic source," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 7455–7459.

[13] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Proc. Workshop Hands-Free Speech Commun. and Microphone Arrays*, Edinburgh, UK, May 2011, pp. 127–132.

[14] M. H. Hennecke, T. Plötz, G. A. Fink, J. Schmalenstroer, and R. Haeb-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *Proc. IEEE Workshop Statistical Signal Processing*, Cardiff, UK, Aug. 2009, pp. 257–260.

[15] R. Heusdens and N. Gaubitch, "Time-delay estimation for TOA-based localization of multiple sensors," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 609–613.

[16] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "DOA-based microphone array position self-calibration using circular statistics," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vancouver, British Columbia, Canada, May 2013, pp. 116–120.

[17] F. Jacob and R. Haeb-Umbach, "Coordinate mapping between an acoustic and visual sensor network in the shape domain for a joint self-calibrating speaker tracking," in *Proc. 11. ITG Fachtagung Sprachkommunikation*, Erlangen, Germany, Sept. 2014, pp. 1–4.

[18] F. Jacob and R. Haeb-Umbach, (2015). Absolute geometry calibration of distributed microphone arrays in an audio-visual sensor network. *CoRR* [Online]. vol. abs/1504.03128. Available: http://arxiv.org/abs/1504.03128

[19] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "Microphone array position self-calibration from reverberant speech input," in *Proc. Int. Workshop Acoustic Signal Enhancement*, Aachen, Germany, Sept. 2012, pp. 1–4.

[20] A. Lauterbach, K. Ehrenfried, L. Koop, and S. Loose, "Procedure for the accurate phase calibration of a microphone array," presented at *Proc. 15th AIAA/CEAS Aeroacoustics Conf.*, Miami, FL, May 2009.

[21] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Int. Workshop Acoustic Echo and Noise Control*, Tel Aviv, Israel, Sept. 2010, pp. 1–4.

[22] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 21, no. 2, pp. 343–356, Feb. 2013.

[23] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[24] I. McCowan and M. Lincoln, "Microphone array shape calibration in diffuse noise fields," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 16, no. 3, pp. 666–670, Mar. 2008.

[25] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. IEEE Workshop Applicat. Signal Processing Audio, Acoustics*, New Paltz, NY, Oct. 2009, pp. 161–164.

[26] M. Parviainen, P. Pertilä, and M. S. Hämäläinen, "Self-localization of wireless acoustic sensors in meeting rooms," in *Proc. Joint Workshop Hands-Free Speech Commun. and Microphone Arrays*, Nancy, France, May 2014, pp. 152–156.

[27] P. Pertila, M. S. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an adhoc microphone array," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[28] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Passive self-localization of microphones using ambient sounds," in *Proc. Euro. Signal Processing Conf.*, Bucharest, Romania, Aug. 2012, pp. 1314–1318.

[29] A. Plinge and G. A. Fink, "Online multi-speaker tracking using multiple microphone arrays informed by auditory scene analysis," in *Proc. Euro. Signal Processing Conf.*, Marrakesh, Morroco, Sept. 2013, pp. 1–5.

[30] A. Plinge and G. A. Fink, "Geometry calibration of distributed microphone arrays exploiting audio-visual correspondences," in *Proc. Euro. Signal Processing Conf.*, Lisbon, Portugal, Sept. 2014, pp. 116–120.

[31] A. Plinge and G. A. Fink, "Geometry calibration of multiple microphone arrays in highly reverberant environments," in *Proc. Int. Workshop Acoustic Signal Enhancement*, Juan les Pins, France, Sept. 2014, pp. 243–247.

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

[33] A. Plinge, R. Grzeszick, and G. A. Fink, "A Bag-of-Features approach to acoustic event detection," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014 , pp. 3704–3708.

[34] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, NV, Mar. 2008, pp. 2445–2448.

[35] V. C. Raykar and R. Duraiswami, "Automatic position calibration of multiple microphones," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, May 2004, pp. iv–69–72.

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

[37] A. Redondi, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Geometric calibration of distributed microphone arrays," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, Rio de Janeiro, Brazil, Oct. 2009, pp. 1–5.

[38] J. Sachar, H. Silverman, and W. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 1, pp. 42–52, Jan. 2005.

[39] J. Schmalenstroeer, F. Jacob, R. Haeb-Umbach, M. H. Hennecke, and G. A. Fink, "Unsupervised geometry calibration of acoustic sensor networks using source correspondences," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 597–600.

[40] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A gossiping approach to sampling clock synchronization in wireless acoustic sensor networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 7575–7579.

[41] M. J. Taghizadeh, R. Parhizkar, P. N. Garner, H. Bourlard, and A. Asaei, "Ad hoc microphone array calibration: Euclidean distance matrix completion algorithm and theoretical guarantees," *Signal Processing*, vol. 107, pp. 123–140, Feb. 2015.

[42] M. Taseska and E. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 7, pp. 1195–1207, July 2014.

[43] S. Thrun, "Affine structure from sound," in *Proc. Conf. Neural Informat. Process, Systems*, Vancouver, Canada, 2005, pp. 1353–1360.

[44] S. D. Valente, F. Antonacci, M. Tagliasacchi, A. Sarti, and S. Tubaro, "Self-calibration of two microphone arrays from volumetric acoustic maps in non-reverberant rooms," in *Proc. Int. Symp. Commun., Control and Signal Processing*, Limassol, Cyprus, Mar. 2010.

[45] 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 *Proc. IEEE Workshop Multimedia Signal Processing*, Saint Malo, France, Oct. 2010, pp. 13–18.

[46] J. Velasco, M. J. Taghizadeh, A. Asaei, H. Bourlard, C. J. Martín-Arguedas, J. Macias-Guarasa, and D. Pizarro, "Novel GCC-PHAT model in diffuse sound field for microphone array pairwise distance based calibration," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 2669–2673.

[47] L. Wang, T. K. Hon, J. Reiss, and A. Cavallaro, "Self-localization of adhoc arrays using time difference of arrivals," *IEEE Trans. Signal Processing*, vol. 64, no. 4, pp. 1018–1033, Feb. 2016.

SP