TOWARDS ACOUSTIC SELF-LOCALIZATION OF AD HOC SMARTPHONE ARRAYS

Marius H. Hennecke *Gernot A. Fink*

Intelligent Systems Group, Robotics Research Institute, TU Dortmund University, Germany
{*marius.hennecke, gernot.fink*}@*tu-dortmund.de*

ABSTRACT

The advent of the smartphone in recent years opened new possibilities for the concept of ubiquitous computing. We propose to use multiple smartphones spontaneously assembled into an ad hoc microphone array as part of a teleconferencing system. The unknown spatial positions, the asynchronous sampling and the unknown time offsets between clocks of smartphones in the ad hoc array are the main problems for such an application as well as for almost all other acoustic signal processing algorithms. A maximum likelihood approach using time of arrival measurements of short calibration pulses is proposed to solve this self-localization problem. The global orientation of each phone, obtained by the means of nowadays common built-in geomagnetic compasses, in combination with the constant microphone-loudspeaker distance lead to a non-linear optimization problem with a reduced dimensionality in contrast to former methods. The applicability of the proposed self-localization is shown in simulation and via recordings in a typical reverberant and noisy conference room.

Index Terms— smartphone array, self-localization, maximum likelihood estimation

1. INTRODUCTION

The most well-known commercial application of microphone arrays is their use in teleconferencing systems. Typically, acoustic beamforming is applied on one central microphone array for obtaining enhanced versions of signals recorded from distant speakers. Acoustic source localization could be used additionally to control the visual capturing of the meeting participants, when multiple and potentially active cameras are present. Though a number of technical solutions for teleconferencing systems are available on the market, such systems have not found their way into the standard repertoire of modern conference rooms. This is attributed to the high costs of such specialized audio-visual systems and the practical problems concerning installation and maintenance. In contrast to that, today every participant of a business meeting will certainly carry a mobile phone able to provide hands-free speech communication with reasonable quality. Therefore, we propose

Fig. 1: An exemplary ad hoc array comprised of four commercial off-the-shelf smartphones.

to use an ad hoc network of smartphones for providing the acoustic communication capabilities required in a teleconferencing scenario. The main challenge in forming a microphone array from an ad hoc network of mobile phones lies in the fact that the spatial configuration of the sensors is not known and has to be estimated automatically, as this prior knowledge is mandatory for most acoustic signal processing algorithms.

In this contribution we present a method for the acoustic self-localization of nodes in an ad hoc array of commercial off-the-shelf smartphones placed on a conference room table (Fig. 1). Time of arrival (TOA) measurements of known acoustic calibration signals and the orientation of the phones, provided by built-in geomagnetic compasses, are used to estimate the spatial locations of the nodes. Additionally, as such an ad hoc array samples the sound field spatially in an asynchronous manner, i.e. no central word clock is available, the delays of the single platforms w.r.t. a reference time have to be jointly estimated.

2. RELATED WORK

Self-localization for wireless sensor networks in general is a well studied topic [1]. In such networks, sensor nodes communicate with neighbouring nodes in a peer-to-peer manner to exchange TOA, received signal strength or angle of ar-

This work was in part supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under contract number Fi 799/5-1.

rival measurements. The nodes are typically equipped with specialized RF sensing hardware for the respective modalities.

Ad hoc arrays are always asynchronously sampled and hence self-localization systems based on time measurements must jointly estimate the respective local time frames w.r.t. a reference channel. In order to avoid the joint estimation of positions and exact timing information Liu et al. [2] propose an energy based self-localization where only a rough time synchronization of the nodes is needed. The ad hoc array is formed by laptops with the assumption of one speaker in front of each laptop. The pairwise relative speech energy attenuation in combination with an energy decay model results in distance estimates which are the basis for a nonlinear optimization procedure. In a mildly reverberant room Liu et al. achieve an average position error of 21 cm for seven speakers and hence seven laptops.

Matrix factorization based approaches for finding suitable starting points for iterative nonlinear self-localization have been proposed. A closed-form rank-3 based factorization algorithm for the joint source and sensor localization is given by Thrun [3]. Under the assumption of far-field sources the self-localization problem is solved in the space of affine geometry. Finding a transformation from the affine solution to Euclidean space involves an iterative non-linear optimization of a low-dimensional problem. A probabilistic extension of this method exists which additionally takes the measurement uncertainty into account [4]. If the far-field assumption is not valid, a rank-5 factorization is necessary [5]. The drawback of the closed-form approaches is the amount of microphones and sources needed. For the rank-5 factorization method at least ten microphones and four sources or vice versa are necessary.

Our work is part of a larger research endeavour with the goal of an unsupervised joint calibration of an audio-visual sensor network. Prior work focused on the calibration of a distributed synchronously sampled microphone array [6]. The employed high-quality audio equipment allowed for a calibration using only ambient noise and naturally occurring sounds. The hierarchical approach first calibrates local arrays via ambient noise, which is assumed to be diffuse. The known distance dependent coherence of a diffuse noise field for omnidirectional microphones leads to pairwise distance estimates. Classical multidimensional scaling (CMDS) [7] reveals the spatial configuration of the local arrays. In the second stage acoustic sources are localized with every locally calibrated array. A matching procedure robust against localization outliers leads to the global spatial array configuration. Unlike [6] we focus in this contribution on the calibration of low-quality unsynchronized mobile phone audio hardware and resort to acoustic calibration signals to approach the self-localization task.

The proposed approach is inspired by the maximum likelihood (ML) method described by Raykar et al. [8]. The ML estimation of microphone coordinates, loudspeaker coordinates and capture start times of an ad hoc array of general purpose computers (laptops, etc.) is formulated as an unconstrained nonlinear optimization problem. The starting point for the iterative nonlinear optimization of the ML cost function is obtained via a closed form approximation. Microphones and loudspeaker are supposed to have the same coordinates which results in pairwise distance measurements. Employing CMDS [7] an approximate solution is found which serves as a starting point for an iterative solver. In contrast to [8] we propose to make use of the known and constant loudspeaker-microphone distance per smartphone and the global orientations of the phones obtained via their integrated geomagnetic compasses. This leads to a reduction of the number of free model parameters, as the loudspeaker coordinates are then solely dependent on the microphone coordinates.

3. SELF LOCALIZATION

The goal of our proposed self-localization procedure is to obtain the D-dimensional Euclidean positions

$$\boldsymbol{M} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_N) \tag{1}$$

of the microphones of all N smartphones contained in a specific ad hoc array. Due to the asynchronous sampling, the capture start times

$$\boldsymbol{T}_c = (t_{c_1}, t_{c_2}, \ldots, t_{c_N}) \tag{2}$$

w.r.t an arbitrary reference channel, e.g. $t_{c_1} = 0$ must be estimated as well. In the examined scenario every smartphone has additionally one loudspeaker and a built-in geomagnetic compass. The constant distance $d_i = \|\boldsymbol{m}_i - \boldsymbol{s}_i\|$ of smartphones i microphone to its loudspeaker $\boldsymbol{s}_i$ is assumed to be known in advance. This quantity is not going to change for a given type of smartphone. In combination with the global orientation φ_i measured by the i-th compass, the loudspeaker coordinates in the two-dimensional case can be expressed as

$$\boldsymbol{S} = \boldsymbol{M} + \begin{pmatrix} d_1 \cos(\varphi_1) & d_2 \cos(\varphi_2) & \cdots & d_N \cos(\varphi_N) \\ d_1 \sin(\varphi_1) & d_2 \sin(\varphi_2) & \cdots & d_N \sin(\varphi_N) \end{pmatrix}. \tag{3}$$

An extension of this formulation to the three-dimensional case is possible but unnecessary in the case of a table top ad hoc array considered here.

The TOA at microphone $\boldsymbol{m}_i$ of a sound emitted at time t_{s_k} from the k-th loudspeaker in the local time of channel i is defined by

$$t_i^{(k)} = t_{s_k} + \frac{\|\boldsymbol{s}_k - \boldsymbol{m}_i\|}{c} - t_{c_i}. \tag{4}$$

This equation incorporates the Euclidean distance $\|\cdot\|$ of the originating loudspeaker $\boldsymbol{s}_k$ to the receiver, and the speed of sound c. The speed of sound is assumed constant for the rest of this paper[1]. For a complete description of a TOA the capture

[1]The speed of sound depends on environmental effects [9], especially on the temperature. For our purposes, the assumption of a constant speed of sound $c(\theta = 20\,°\mathrm{C}) \approx 343\,\mathrm{m\,s^{-1}}$ is sufficient.

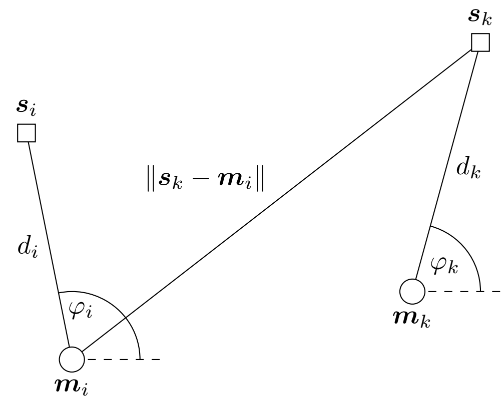

Fig. 2: Array setup depicting the given global orientations φ_*, distances d_* and the Euclidean distance of the loudspeaker $\boldsymbol{s}_k$ to the microphone $\boldsymbol{m}_i$.

start time t_{c_i} of channel i w.r.t an arbitrary reference channel must be known. Fig. 2 shows the distances and angles involved in an example ad hoc network with two phones.

Mobile phones are usually equipped with low-quality audio hardware and hence accurate timing information for capturing and playback is missing. Especially the exact signal emission time t_{s_k} is not known, due to the uncertainty about the audio pipelines latency. In order to circumvent the estimation of the signal emission times, only time differences of arrivals (TDOA)

$$\tau_{ij}^{(k)} = t_i^{(k)} - t_j^{(k)} \tag{5}$$

$$= \frac{\|\boldsymbol{s}_k - \boldsymbol{m}_i\|}{c} - \frac{\|\boldsymbol{s}_k - \boldsymbol{m}_j\|}{c} - t_{c_i} + t_{c_j} \tag{6}$$

between receiving channels i and j for a calibration signal originated at loudspeaker k are used for the self-localization task. The unknown signal emission time t_{s_k} gets cancelled out and hence is irrelevant for the TDOA case.

Actual TOA measurements are distorted by noise. Under a Gaussian uncorrelated noise assumption and taking into acount the fact that the difference of two normally distributed random variables is again Gaussian, the TDOA measurements are expressed as

$$\hat{\tau}_{ij}^{(k)} \sim \mathcal{N}\left(t_i^{(k)} - t_j^{(k)}, \left(\sigma_i^{(k)}\right)^2 + \left(\sigma_j^{(k)}\right)^2\right) \tag{7}$$

$$= \mathcal{N}\left(\tau_{ij}^{(k)}, \left(\sigma_{ij}^{(k)}\right)^2\right) \tag{8}$$

with a standard deviation $\sigma_i^{(k)}$ for the TOA case and $\sigma_{ij}^{(k)}$ for the respective TDOA case. The ML estimate [8] of the model parameters $\boldsymbol{M}$ and $\boldsymbol{T}_c$ incorporating measurements from pairs $(i, j) \subseteq \mathcal{P} = \{1, 2, \ldots, N\}^2$ leads to a weighted least squares minimization problem

$$(\hat{\boldsymbol{M}}, \hat{\boldsymbol{T}}_c) = \arg\min_{\boldsymbol{M}, \boldsymbol{T}_c} \sum_{k=1}^{N} \sum_{(i,j)\in\mathcal{P}} \left(\left(\hat{\tau}_{ij}^{(k)} - \tau_{ij}^{(k)}\right) / \sigma_{ij}^{(k)}\right)^2. \tag{9}$$

In the following, all $\sigma_{ij}^{(k)}$ are assumed to be equal. Consequently, they can be ignored in the minimization as they are independent of the model parameters. The ML estimate can be extended to incorporate repeated measurements to make it more robust in terms of noise. It is simply another sum over all repetitions and omitted for the sake of brevity in the above formulation.

Without prior knowledge about the array geometry an ML estimate of the nonlinear optimization problem is not invariant against rotation, translation and reflection. In order to ensure a unique solution and to further reduce the number of free parameters the first microphone is chosen arbitrarily as the spatial origin $\boldsymbol{m}_1 = (0, 0)^{\mathrm{T}}$ as well as the temporal origin $t_{c_1} = 0$. Furthermore, the second sensor is confined to lie on the first positive axis $\boldsymbol{m}_2 = (0, m_{2,y})^{\mathrm{T}}$, $m_{2,y} \geq 0$. In order to eliminate the reflection ambiguity, the third microphone is constrained to lie in the first two quadrants of the two-dimensional Euclidean space formed by the first two sensors, i.e. $m_{3,x} \geq 0$ and $m_{3,y} \geq 0$. Due to these constraints the number of free model parameters is reduced by a total of four with the additional benefit of a well-defined coordinate system.

For solving the above bound-constrained nonlinear optimization problem we employ the iterative trust-region-reflective algorithm [10]. The key parts of the partial derivatives needed for iteratively minimizing the least squares problem (9) are

$$\frac{\partial \tau_{ij}^{(k)}}{\partial \boldsymbol{m}_a} = \begin{cases} -\frac{\boldsymbol{s}_k - \boldsymbol{m}_i}{c\|\boldsymbol{s}_k - \boldsymbol{m}_i\|} & a = i, a \neq j, a \neq k \\ \frac{\boldsymbol{s}_k - \boldsymbol{m}_j}{c\|\boldsymbol{s}_k - \boldsymbol{m}_j\|} & a \neq i, a = j, a \neq k \\ \frac{\boldsymbol{s}_k - \boldsymbol{m}_i}{c\|\boldsymbol{s}_k - \boldsymbol{m}_i\|} - \frac{\boldsymbol{s}_k - \boldsymbol{m}_j}{c\|\boldsymbol{s}_k - \boldsymbol{m}_j\|} & a \neq i, a \neq j, a = k \\ \frac{\boldsymbol{s}_k - \boldsymbol{m}_i}{c\|\boldsymbol{s}_k - \boldsymbol{m}_i\|} & a \neq i, a = j, a = k \\ -\frac{\boldsymbol{s}_k - \boldsymbol{m}_j}{c\|\boldsymbol{s}_k - \boldsymbol{m}_j\|} & a = i, a \neq j, a = k \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

$$\frac{\partial \tau_{ij}^{(k)}}{\partial t_{c_a}} = \begin{cases} -1 & a = i \\ 1 & a = j \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Note that the loudspeaker coordinates $\boldsymbol{s}_k$ are a function of the microphone coordinates $\boldsymbol{m}_k$, which leads to the partial derivative parts where $a = k$.

Iterative gradient descent based optimization algorithms tend to get stuck in local minima. Hence, the starting point plays a crucial role in such a method. A good approximation of the model parameters can increase the probability to find a global optimum. In the following we will shortly revisit the distance-based self-localization approximation given by [8]. Taking the difference of TDOAs between a pair of phones

$$\tau_{ij}^{(i)} - \tau_{ij}^{(j)} = \frac{\|\boldsymbol{s}_i - \boldsymbol{m}_i\|}{c} - \frac{\|\boldsymbol{s}_i - \boldsymbol{m}_j\|}{c} - \left(\frac{\|\boldsymbol{s}_j - \boldsymbol{m}_i\|}{c} - \frac{\|\boldsymbol{s}_j - \boldsymbol{m}_j\|}{c}\right), \quad (12)$$

the corresponding capture start times are eliminated. In order to obtain an approximate distance estimate d_{ij} between phone i and j the influence of the per phone loudspeaker-microphone distance is neglected, i.e. $\|\boldsymbol{s}_i - \boldsymbol{m}_i\| = 0$ and $\|\boldsymbol{s}_j - \boldsymbol{m}_j\| = 0$. This results in equal inter loudspeaker-microphone distances $\|\boldsymbol{s}_i - \boldsymbol{m}_j\| = \|\boldsymbol{s}_j - \boldsymbol{m}_i\|$. The TDOA difference (12) simplifies with these assumptions to an approximate distance estimate

$$d_{ij} = \|\boldsymbol{m}_i - \boldsymbol{m}_j\| \approx \frac{c}{2}\left(t_j^{(i)} - t_i^{(i)} + t_i^{(j)} - t_j^{(j)}\right) \quad (13)$$

for all phone pairs. Employing CMDS [7], a spatial representation $\tilde{\boldsymbol{M}}_{(0)}$ of the microphone coordinates is found. CMDS finds the best embedding of the distance measurements in a least squares sense into a lower-dimensional subspace. In the examined self-localization task where the phones are supposed to lie flat on a table, this subspace has two principal components. Transforming the CMDS approximation such that it adheres to the aforementioned constraints leads to the starting point for the microphone coordinates $\boldsymbol{M}_{(0)}$ of the iterative optimization algorithm.

A starting point for the capture start times $\boldsymbol{T}_{c,(0)}$ could be calculated in a similar manner. But a simpler recursive method can be used due to the one-dimensionality of the capture start times. Taking the sum of TDOAs between a pair of phones opposed to the difference in Eq. (12) but using the same assumptions, an estimate for all capture start times is found recursively

$$t_{c_i,(0)} = \frac{t_{i-1}^{(i-1)} - t_i^{(i-1)} + t_{i-1}^{(i)} - t_i^{(i)}}{2} + t_{c_{i-1},(0)}. \quad (14)$$

The arbitrarily chosen temporal origin $t_{c_1,(0)} = 0$ ends the recursion. To avoid local minima during multiple invocations of the iterative optimization procedure, a small amount of uncorrelated white Gaussian noise is added to the initial parameter approximations for the microphone coordinates and the capture start times.

The known loudspeaker-microphone distance per phone and the global orientation of each phone given by the integrated compasses reduce the number of free model parameters in comparison to [8]. Due to the aforementioned constraints on the microphone coordinates and the capture start times, which guarantee a unique solution, $P = (D+1)N - D - 1$ parameters need to be estimated with dimensionality D. Whereas a total of $(2D+1)N - D - 1$ model parameters exist if the loudspeaker coordinates have to be estimated as well. In combination with a good initial guess even an under determined system can be solved in practice. Hence, self-localization of even two phones is possible.

4. EVALUATION

In the following, the performance of our proposed self-localization procedure for ad hoc smartphone arrays is evaluated. First, a Monte Carlo simulation evaluating different noise levels and array sizes is examined. The nowadays widely available Android-based smartphones are utilized for demonstrating the applicability of our self-localization method in a real-world scenario, a reverberant meeting room.

The accuracy of the self-localization is measured in terms of the root mean square error (RMSE)

$$\varepsilon(\boldsymbol{M}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\|\boldsymbol{m}_i - \boldsymbol{m}_i^o\|)^2}. \quad (15)$$

The ground truth positions are marked by the superscript $(\cdot)^o$. The RMSE is analogously defined for the capture start times $\boldsymbol{T}_c$. For $\varepsilon(\boldsymbol{T}_c)$ however, the Euclidean norm $\|\cdot\|$ reduces to an identity. Note that the error of the capture start time estimates can not be evaluated in the real-world scenario as there is no ground truth due to the lack of accurate timing information on the Android-based smartphones.

4.1. Calibration Signals

Measuring the exact TOA of an arbitrary acoustic signal is in general a hard task. The unknown microphone directivity patterns of the different smartphones and the unfavorable direction of the downwards facing loudspeakers in the employed smartphones makes a TOA estimation error-prone. If, however, the excitation signal is known in advance, a convolution of the captured signal with the inverse of the excitation signal yields the impulse response of the transmitter-receiver channel [11]. In our case this channel consists of the loudspeaker, the receiving microphone and a reverberant room. The direct path component of the resulting impulse response is an estimate for the TOA.

The first step in the design of the aforementioned self-localization method is the choice of an adequate excitation signal for measuring the impulse responses' direct path components and hence the TOA. Due to its flexibility in frequency resolution and length we used a logarithmically swept sine chirp [11, 12] throughout our experiments. Since we are only interested in the direct path component and not in an accurate estimate of the impulse response, the signal properties can be chosen arbitrarily. A single chirp should be short in order to reduce the measurement time and its audibility in the deployment environment. Informal tests showed a good performance for 100 ms long sweeps with a frequency range from

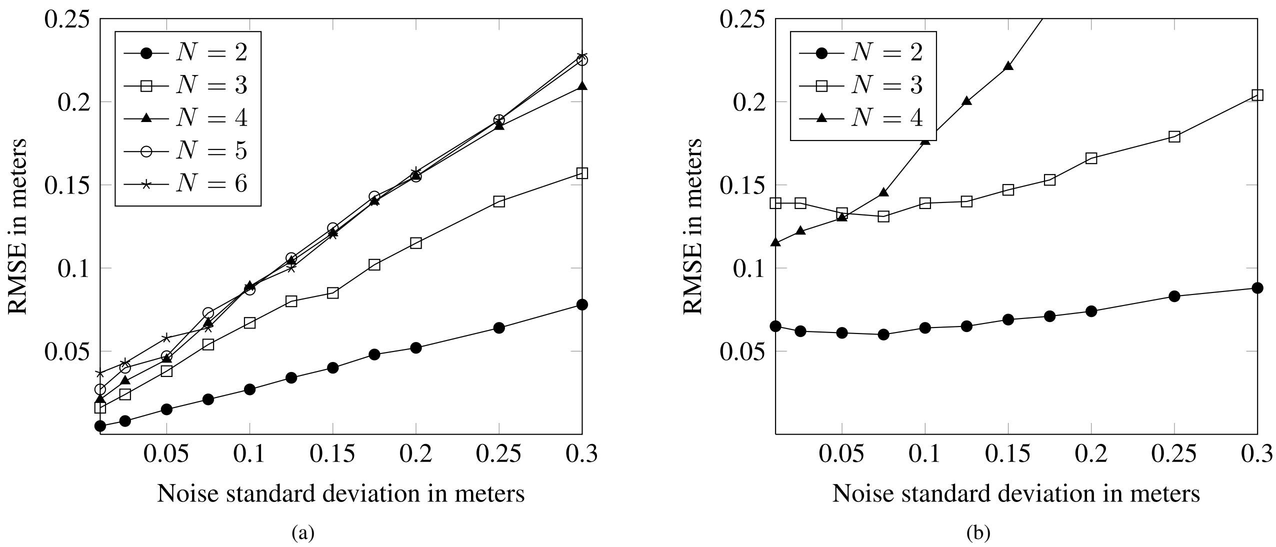

Fig. 3: Simulation results for an increasing number of sensors and increasing amount of noise. The results of the proposed approach (a) are shown in direct comparison to the results of the same optimization procedure not using the prior knowledge (b).

5 kHz to 16 kHz. All recordings use a sampling frequency of $f_s = 48$ kHz.

It is common for consumer-oriented acoustic hardware to have an automatic gain control (AGN) built-in. This is also the case for the employed smartphones without a software-side option to turn it off. Therefore, care has to be taken in the experiment design concerning the playback gain of the calibration signals, as an active AGN leads to non-linear distortions of the impulse response estimates which have an undesirable effect on the final TOA estimates.

4.2. Monte Carlo Simulation

The proposed self-localization method is evaluated for different measurement noise levels via Monte Carlo simulation. For each noise level and number of phones, 1000 random ad hoc array configurations are generated. Five independent TOA measurements afflicted by additive white Gaussian noise are used as a basis to self-localize every configuration. The generated arrays are constrained to lie inside a square of 1 m side length. Gaussian white noise with a standard deviation of $10°$ is added to the ground truth compass data.

Figure 3a depicts the average positional RMSE for increasing measurement noise σ in meters and different array sizes. The estimation error for the capture start times is negligible over all configurations. For $\sigma = 0.3$ m and $N = 6$ the error is 1.5 ms and far below a millisecond for all the other configurations. The resulting RMSE is proportional to the measurement noise and for more than three phones, no further increase in the RMSE due to more phones is visible. Exemplarily, an optimization result for $N = 6$ and $\sigma = 0.1$ m is shown in Fig. 4. The non-filled circles mark the microphone ground truth and the squares the loudspeaker ground truth. The filled objects mark the respective optimization results. Each gray cross is an intermediate microphone coordinate result of a specific iteration forming a trajectory. The first cross is the starting point of the optimization obtained via the aforementioned approximation. The depicted example took ten iterations to converge with a final RMSE of 6.2 cm.

The same simulation was carried out ignoring the proposed prior knowledge, i.e. the microphone loudspeaker distances and the global orientations of the phones. The results are shown in Fig. 3b. Due to the increased number of model parameters the optimization fails to reveal adequate solutions for $N = 4$ and $\sigma > 0.2$ m. For more than four phones the optimizations did not converge at all. This direct comparison shows the superior performance of the proposed self-localization approach.

4.3. Real-world test

The RMSE performance for the distance estimation task is given in Fig. 5 for two phones with distances ranging from 10 cm to 40 cm. Five independent measurements per distance were recorded. The individual distances from the pairwise TOA measurements are estimated according to the approximate distance (13). For small distances (less than 20 cm) the RMSE translates to a deviation of less than two samples at the used sampling rate and speed of sound.

Recordings were made in a reverberant room with a reverberation time of 600 ms for three ad hoc array configurations consisting of four smartphones. The evaluated arrays have an approximate diameter of 40 cm and the phones are placed with arbitrary orientations. For the three configurations with ten independent recordings per configuration, an average positional RMSE of 6.8 cm, 6.8 cm and 8.5 cm is achieved respectively.

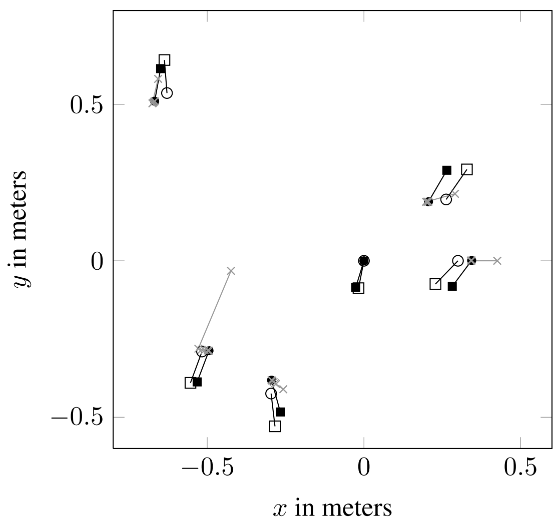

Fig. 4: Exemplary self-localization result for $N = 6$ smartphones and a measurement noise with standard deviation $\sigma = 0.1$ m converged after 10 iterations.

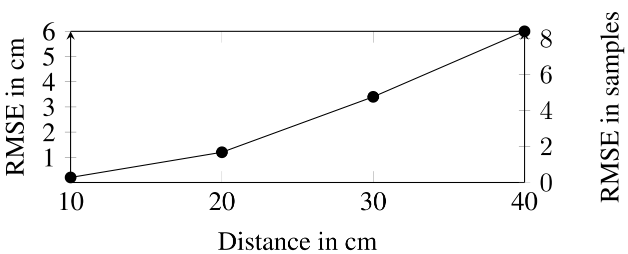

Fig. 5: RMSE for increasing distance of two phones in meters (left axis) and corresponding samples (right axis) according to Eq. (13) for a sampling rate of $f_s = 48$ kHz.

5. CONCLUSION

In this contribution, we proposed a self-localization method for ad hoc smartphone arrays. Such a sensor network could be used as part of a teleconferencing system. The described ML approach for estimating the unknown model parameters, i.e. the microphone coordinates, uses short calibration signals in order to obtain TOA measurements. With the phones global orientation, provided by built-in magnetic compasses, and the fixed and hardware-dependent microphone-loudspeaker distances, a reduction of the number of model parameters is achieved in comparison to a formulation without this prior knowledge. We showed the applicability and superior performance of the proposed approach in simulation side by side to a model which does not take the prior knowledge into account. Finally, we demonstrated that our self-localization method works also in a real-world scenario – a reverberant conference room – using off-the-shelf smartphones.

6. REFERENCES

[1] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero III, R. L. Moses, and N. S. Correal, "Locating the nodes: cooperative localization in wireless sensor networks," vol. 22, no. 4, pp. 54–69, 2005.

[2] Z. Liu, Z. Zhang, L.-W. He, and P. Chou, "Energy-Based Sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. II–761–II–764.

[3] S. Thrun, "Affine structure from sound," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2005, pp. 1353–1360.

[4] C.-H. Lin and C.-C. Wang, "Probabilistic structure from sound and probabilistic sound source localization," in *Advanced robotics and Its Social Impacts (ARSO), IEEE Workshop on*, Aug. 2008, pp. 1–6.

[5] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 2445–2448.

[6] M. Hennecke, T. Plötz, G. A. Fink, J. Schmalenströer, and R. Häb-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *IEEE/SP 15th Workshop on Statistical Signal Processing*, Cardiff, UK, 2009, pp. 257–260.

[7] T. F. Cox and M. A. Cox, *Multidimensional Scaling*, ser. Monographs on statistics and applied probability. CRC Press, 2001, no. 88.

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

[9] D. A. Bohn, "Environmental effects on the speed of sound," *Journal of the Audio Engineering Society*, vol. 36, no. 4, pp. 223–231, 1988.

[10] T. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *Mathematica*, vol. 6, no. 2, pp. 418–445, 1996.

[11] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, 2001.

[12] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of Different Impulse Response Measurement Techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, Apr. 2002.