PASSIVE SELF-LOCALIZATION OF MICROPHONES USING AMBIENT SOUNDS

Pasi Pertilä[⋆] *Mikael Mieskolainen*[⋆] *Matti S. Hämäläinen*[•]

[⋆]Tampere University of Technology, Department of Signal Processing,
P.O.Box 553, Tampere, FI-33101, Finland, {pasi.pertila, mikael.mieskolainen}@tut.fi
[•]Nokia Research Center, Tampere, Finland, matti.s.hamalainen@nokia.com

ABSTRACT

This work presents a method to localize a set of microphones using recorded signals from surrounding continuous sounds such as speech. When a sound wave travels through a microphone array a time difference of arrival (TDOA) can be extracted between each microphone pair. A sound wave impinging towards a microphone pair from the end-fire direction results in the extreme TDOA value, leading to information about microphone distance. In indoors the reverberation may cause TDOA outliers, and a set of non-linear techniques for estimating the distance is proposed. The multidimensional scaling (MDS) is used to map the microphone pairwise distances into Cartesian microphone locations. The accuracy of the method and the effect of the number of sources is evaluated using speech signals in simulated environment. A self-localization RMS error of 7 cm was reached using ten asynchronous smartphones in a meeting room from a recorded conversation with a maximum of 3.7 m device separation.

Index Terms— Microphone arrays, Array Shape Calibration, Self-Localization, TDOA estimation, Multidimensional Scaling

1. INTRODUCTION

Automatic calibration of microphone arrays is essential in distributed microphone signal processing applications. Spatial signal processing methods such as beamforming and sound source localization are dependent on microphone positions. Multichannel AD-converters can offset sample synchronized multichannel audio, whereas synchronizing the signals from AD-converters of different mobile devices is more challenging. The ability to estimate the microphone positions from a set of asynchronous recordings without performing any active calibration, i.e., signal emissions, would bring the processing of distributed microphones a step closer to practical applications.

In [1] microphone calibration in a diffuse noise field is proposed. The analytic form of the coherence function is dependent on microphone separation. The separation can be solved by minimizing a distance between measured coherence and it's theoretical shape. However, diffuse noise field can not be always assumed. In [2, 3] discrete sound sources are located in the near-field by using the time difference of arrival (TDOA) values calculated from received signals. Self-localization is then performed (on a linear array in [3]) by minimizing a set of equations of source and microphone locations. Such iterative techniques require a good initial guess to enable convergence, and adding degrees of freedom to the microphone locations by allowing 2D and 3D array configurations leads to high dimensional search problems. In [4] a method for solving the source and sensor positions in a linear approach is proposed.

In [5] a method for using TDOA values observed from transient sound events between time synchronized smartphones is investigated. In addition, a two receiver case is treated by studying the theoretical shape of TDOA distribution for equally spread sources around the array. However, if the sources are not equally spread, e.g., in a typical meeting with static talkers, the TDOA distribution can contain multiple peaks corresponding to angles of the participants and reflected signals (such data is illustrated in Fig. 3). Fitting a theoretical model to such data may result in biased locations.

This work uses multidimensional scaling (MDS) algorithm [6] for localizing microphone coordinates based on pairwise distances between microphones. The distances are derived from the minimum and maximum observed TDOA values, and the proposed estimator cancels out the unknown sensor time-offsets. This enables the self-localization of asynchronous devices, such as smartphones. Two non-linear filtering techniques are then proposed for the minimum and maximum TDOA estimation. First, a sequential filter passes the TDOA values related to spatially consistent sources. Secondly, a histogram based thresholding operation filters remaining TDOA outliers.

The performance of the proposed method is characterized with simulations in different noise and reverberation levels. To verify the performance of the proposed method, recorded data from a meeting room environments is analyzed. The method is shown in practice to be suitable for the recovery of the array geometry based on the obtained asynchronous microphone signals. In a second simulation, the number of sound sources in a meeting room is varied to see how it affects self-localization error of the proposed method.

This work is funded by the Finnish Academy project no. 138803 and Nokia Research Center.

The advantages of the proposed method include that it does not require the knowledge of sound source positions, does not need synchronized receivers, and can operate using two or more microphones. The algorithm assumes that sound signals from both directions parallel to each microphone pair's axis are observed.

The paper is organized as follows. In Section 2, the pairwise distance estimator is derived from the signal model. Section 3 presents a non-linear implementation of the proposed estimator. Self-localization based on pairwise distances is briefly reviewed in Section 4. Section 5 describes the error metrics, and Section 6 investigates the algorithm's performance in different noise and reverberation levels with simulations as well as the performance using varying amount of sources. Measurement setup and the obtained results are detailed in Section 7. Section 8 concludes the discussion.

2. PAIRWISE DISTANCE ESTIMATION

Let $\mathbf{m}_i \in \mathbb{R}^3$ be the ith receiver position, where $i \in [1, M]$. The signal at microphone i can be modeled as a delayed version of the source signal $s(t)$ as

$$x_i(t) = s(t) * \delta(t - \tau_i), \tag{1}$$

where t is time, $\delta(\cdot)$ is the Dirac's delta function, and τ_i is propagation delay.

Assume that two microphones $\mathbf{m}_i$ and $\mathbf{m}_j$ form a pair and that a source $\mathbf{s}$ resides in the far field, i.e., $\|\mathbf{m}_i - \mathbf{m}_j\| \ll \|\mathbf{r} - \mathbf{s}\|$, where $\mathbf{r}$ is pair's center point $\mathbf{r} = \frac{1}{2}(\mathbf{m}_i + \mathbf{m}_j)$. Therefore, the sound arrives as a plane wave with propagation direction represented by vector $\mathbf{k} \in \mathbb{R}^3$, with length $\|\mathbf{k}\| = c^{-1}$, where c is speed of sound. The wavefront time of arrival at microphone i with respect to center point $\mathbf{r}$ is [7, ch. 2]

$$\tau_i = \langle \mathbf{m}_i - \mathbf{r}, \mathbf{k} \rangle + \Delta_i \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is dot product, and Δ_i is the sensor time-offset to reference time. If the sensors are synchronized, then $\Delta_i = 0$, but unfortunately this is not generally the case in ad-hoc networks with sensor specific clocks. The TDOA is defined as:

$$\tau_{ij} = \tau_i - \tau_j = \langle \mathbf{m}_i - \mathbf{m}_j, \mathbf{k} \rangle + \Delta_{ij}, \tag{3}$$

where $\Delta_{ij} = \Delta_i - \Delta_j$. The propagation vectors of wavefronts arriving from either of the two directions that are parallel to the microphone connecting axis, i.e., endfire directions, can be written as

$$\mathbf{k}(\beta) = \beta \frac{\mathbf{m}_j - \mathbf{m}_i}{\|\mathbf{m}_j - \mathbf{m}_i\|} c^{-1}, \beta \in \{-1, 1\} \tag{4}$$

Refer to Fig. 1, where two waves arrive from the endfire directions ($\beta = 1$, and $\beta = -1$). The TDOA for endfire source directions is obtained by substituting (4) into (3):

$$\tau_{ij}(\beta) = \beta c^{-1} \|\mathbf{m}_i - \mathbf{m}_j\| + \Delta_{ij}. \tag{5}$$

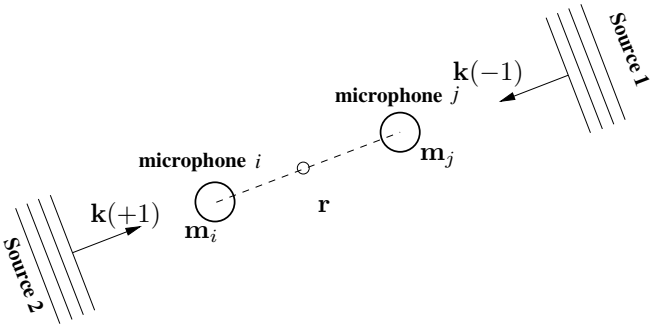


Figure 1: Two wavefronts impinge a microphone pair from directions parallel to the microphone pair's axis (marked as dotted line). The wavefronts are emitted by separate sources.

Note that since $\beta \in \{-1, +1\}$ the TDOA magnitude without the offset is the sound propagation time between the microphones and the sign corresponds to source direction. Since the magnitudes of both TDOA values represent the physical lower and upper limits of the observation we use terms $\tau_{ij}^{\max} \triangleq \tau_{ij}(+1)$ and $\tau_{ij}^{\min} \triangleq \tau_{ij}(-1)$.

Theorem 1. The microphone inter-distance d_{ij} is

$$d_{ij} = \frac{c}{2} \left(\tau_{ij}^{\max} - \tau_{ij}^{\min} \right). \tag{6}$$

Proof. By Using (5)

$$\frac{c}{2} \left(\tau_{ij}(+1) - \tau_{ij}(-1) \right) = \frac{1}{2} \Big(\|\mathbf{m}_i - \mathbf{m}_j\| + c\Delta_{ij} - (-\|\mathbf{m}_i - \mathbf{m}_j\| + c\Delta_{ij}) \Big) = \|\mathbf{m}_i - \mathbf{m}_j\| \triangleq d_{ij} \quad \square$$

In the distance estimation (6) the unknown offsets Δ_{ij} are canceled out. Note that (6) requires that i) maximum and minimum TDOA values $\tau_{ij}^{\max}$ and $\tau_{ij}^{\min}$ are measured from sources in the end-fire directions not located between the microphones, and ii) speed of sound c is known. In this work, we assume knowledge of c and present a novel threshold based method for estimating $\tau_{ij}^{\max}$ and $\tau_{ij}^{\min}$ in the following section.

3. MEASUREMENT OF PAIRWISE DISTANCES

First, a simplified signal energy based voice activity detection (VAD) is performed for the input data to remove frames that contain less energy than λ_{E} times the average frame energy.

Then, the generalized cross-correlation (GCC) between sampled microphone signals i, j with weight $\Psi(\omega)$ is obtained using [8]

$$r_{ij}(\tau) = \sum_{\omega} \Psi(\omega) X_i(\omega) X_j^*(\omega) \exp(j\omega\tau), \tag{7}$$

where $X_i(\omega)$ is frequency domain input signal, ω is angular frequency, $()^*$ is complex conjugate, and τ is time delay. A TDOA value is estimated by searching the correlation function peak index value

$$\hat{\tau}_{ij} = \underset{t}{\operatorname{argmax}}\, r_{ij}(t). \tag{8}$$

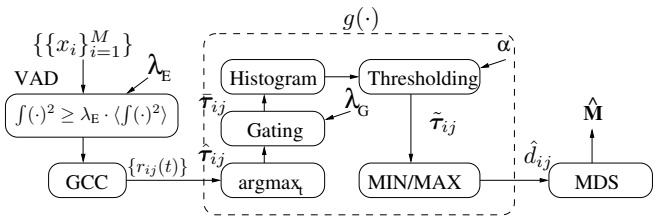


Figure 2: Block diagram of the proposed self-localization method.

A sub-sample TDOA estimate is then obtained by interpolation. The processing is performed in short time frames of length L and $\hat{\boldsymbol{\tau}}_{ij} \in \mathbb{R}^T$ denotes a vector of TDOA values from all T input frames.

A microphone pair (i, j) interdistance estimator can be described as a mapping $g : \{r_{ij}(\tau)\} \mapsto \hat{d}_{ij}$, where $\{r_{ij}(\tau)\}$ is a set of time cross-correlation vectors between a microphone pair i, j calculated over input frames. In this work, the distance mapping $g(\cdot)$ is a set of non-linear operations on the TDOA vector $\hat{\boldsymbol{\tau}}_{ij}$ obtained from (8). Figure 2 illustrates the block diagram of the method.

3.1. Sequential TDOA Gating

Since the TDOA information is based on natural sound source which are often continuous between sequential frames, a gating procedure is implemented to filter out TDOA values that differ sequentially more than λ_G samples. Let $\tau_{ij}(t)$ represent a TDOA value at time frame $t \in [1, T]$ between two channels i and j. The nth order filter is described as

$$\bar{\boldsymbol{\tau}}_{ij} = \{\hat{\tau}_{ij}(t) \mid \lambda_G > |\hat{\tau}_{ij}(t) - \hat{\tau}_{ij}(t-n)|, \forall t\}. \quad (9)$$

Here, the TDOA values are kept if they are passed by the first or second order filter, i.e., $n \in [1, 2]$.

3.2. TDOA Histogram Filtering

Next, a histogram of the filtered TDOA vector $\bar{\boldsymbol{\tau}}_{ij}$ is taken. The histogram bin count n_{ij}^k represent the number of TDOA values in the vector $\bar{\boldsymbol{\tau}}_{ij}$ that are closest to the value k, where $k \in [-K, K]$ and K is TDOA upper histogram limit in samples. A histogram threshold operation is then performed to select delay values with high enough occurrences

$$\tilde{\boldsymbol{\tau}}_{ij} = \{\bar{\tau}_{ij}^k | n_{ij}^k > \alpha \cdot \max(n_{ij}^{-K}, \ldots, n_{ij}^K), \forall k\}, \quad (10)$$

where $\alpha \in [0, 1]$ is a threshold parameter. Setting $\alpha = 0$ would keep all TDOA values, and $\alpha = 1$ would keep only the most frequent TDOAs. The proposed estimators for maximum and minimum TDOA values are

$$\hat{\tau}_{ij}^{\max} = \max(\tilde{\boldsymbol{\tau}}_{ij}) \quad (11)$$

$$\hat{\tau}_{ij}^{\min} = \min(\tilde{\boldsymbol{\tau}}_{ij}). \quad (12)$$

Figure 3 details an example of a microphone pairwise TDOA histogram from recorded speech data before any filtering (top), after sequential filtering (9) (center), and after sequential and histogram thresholding operations (10) (bottom). The x-axis is the sample delay value k, and y-axis is the

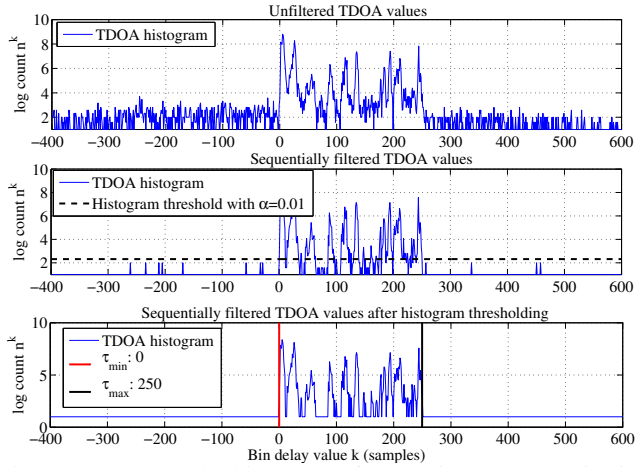


Figure 3: Example histogram from microphone pairwise TDOA vector $\hat{\boldsymbol{\tau}}$. The x-axis is histogram bin TDOA value and y-axis is the corresponding count of TDOA values ($\alpha = 0.01, \lambda_G = 6$ samples).

logarithmic transform of the bin counts n^k. The ground truth microphone distance is measured with tape to be 91 cm which corresponds to 254 sample difference between maximum and minimum TDOA with 48 kHz sampling rate, and $c = 344$ m/s. The difference from the TDOA data is 250 samples (see lower panel in Fig. 3). This indicates a 4 sample error between minimum and maximum TDOA values, which corresponds to 1.4 cm error in the distance (6). Note that the sequential filter removes almost all outlier TDOA values, and therefore α can remain relatively small.

4. MICROPHONE ARRAY SELF-LOCALIZATION

Let $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M] \in \mathbb{R}^{D \times M}$ be the microphone coordinate matrix to be determined in D dimensional space, $\delta_{ij} \triangleq \|\mathbf{m}_i - \mathbf{m}_j\|$ is the theoretical distance between microphones i and j, and $\hat{d}_{ij}$ is the measured distance. MDS [6] finds $\mathbf{M}$ that minimizes the cost function

$$\sigma_r(\mathbf{M}) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} (\hat{d}_{ij} - \delta_{ij})^2, \quad (13)$$

where $\mathbf{M}$ is subject to global isometries (distance preserving mappings) on Euclidean space, i.e. global rotations, translations and reflections.

5. PERFORMANCE METRICS

The RMSE in pairwise distance estimation is

$$\text{RMSE}(\hat{d}_{ij}) = \sqrt{\frac{1}{P} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \left(\hat{d}_{ij} - d_{ij}\right)^2}, \quad (14)$$

where the summation is over all $P = M(M-1)/2$ unique microphone pairs due to symmetry ($d_{ij} = d_{ji}$) and ($d_{ii} = 0$). The relative RMSE is here written $\text{RRMSE}(\hat{d}_{ij}) = 100\% \cdot \text{RMSE}(\hat{d}_{ij})/\bar{d}$, where $\bar{d}$ is the average pairwise distance $\bar{d} = \frac{1}{P} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} d_{ij}$. The RMS error of microphone coordinates is

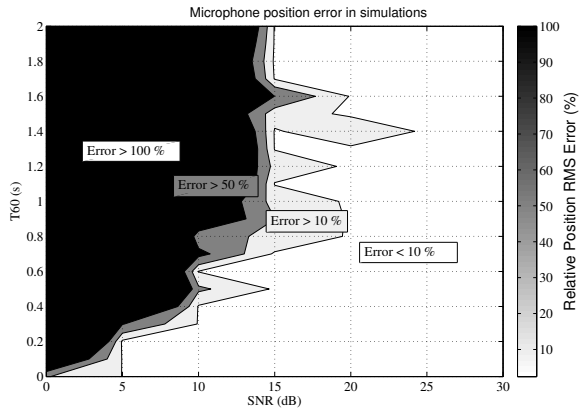


Figure 4: Relative position RMS error of microphones as a function of reverberation time (T_{60}) and SNR (dB).

$$\mathrm{RMSE}(\hat{\mathbf{M}}) = \sqrt{\frac{1}{M}\sum_{i=1}^{M} \|\hat{\mathbf{m}}_i - \mathbf{m}_i\|^2}, \qquad (15)$$

and the relative RMSE of the microphone positions is here written $\mathrm{RRMSE}(\hat{\mathbf{M}}) = 100\% \cdot \mathrm{RMSE}(\hat{\mathbf{M}}) / \sum_{i=1}^{M} \|\mathbf{m}_i - \bar{\mathbf{m}}\|^2$, where $\bar{\mathbf{m}}$ is the average microphone position $\bar{\mathbf{m}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{m}$.

6. SIMULATION RESULTS

A simulation is used to evaluate the performance of the proposed self-localization algorithm in different types of reverberation and noise conditions. A rectangular cuboid shape room is set to contain two sound sources at 1.1 m distance from a six microphone linear array with 10 cm element spacing. The sources are located on the same line as the array, and are on both sides of the array. The image method [9] is used in a $2.4 \times 5.9 \times 2.8$ size office space. The reflection coefficients of the surface are set identical and varied to result in a reverberation time $T_{60} = [0, 0.1, \ldots, 2.0]$ s using the Eyering's equation [10]. In addition, white Gaussian noise is used to corrupt the signals to result in SNR values between +30 dB and 0 dB. A 13 s female speech signal was used as the source signal. The data was sampled at 48 kHz. Table 1 details the empirically selected processing parameters. The locations are estimated in 3D.

The microphone position relative RMSE as a function of SNR and T_{60} is displayed in Fig. 4. The self-localization error increases when SNR decreases and reverberation time increases. It can be concluded that there is a threshold SNR value between 0 to 15 dB, below which the location error sharply rises. The algorithm is not so sensitive to increased reverberation when the SNR high.

Table 1: Processing parameter values.

Window length L, overlap, and type	4096, 50 %, Hanning
GCC weighting, $\Psi(\omega)$	$\lvert X_i(\omega) X_j^*(\omega)\rvert^{-1}$
Delay value parameter, K	1000 samples
VAD threshold, λ_{E}	0.2
Gating threshold, λ_{textG}	6 samples
Histogram threshold, α	0.1

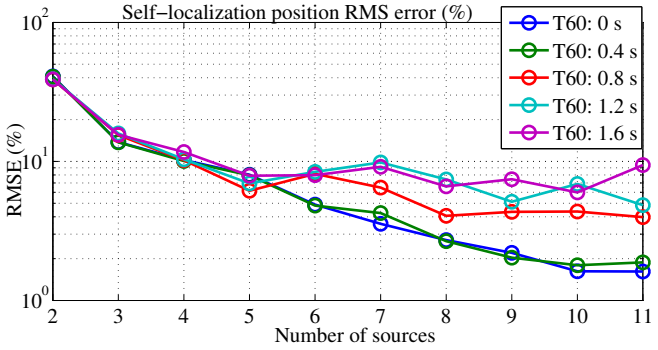
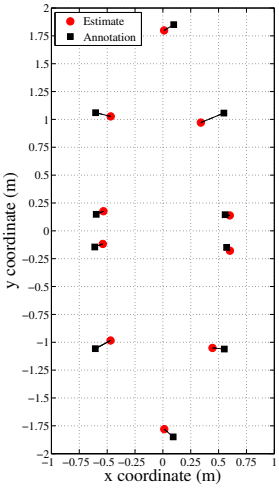


Figure 5: Relative RMS error of microphone positions $\mathrm{RRMSE}(\hat{\mathbf{M}})$ as a function of number of sources surrounding array of Fig. 6 in different reverberation times (T_{60}).

In the second simulation, the objective is to evaluate the amount of error produced by not having sources exactly at the end-fire directions. For this purpose, a meeting room of size $7 \times 7.4 \times 3$ m is used to place ten microphones at locations depicted in Fig. 6 at 1.5 m height. Speech sources are placed around the array center with radius of 3 m, with sources in equally spaced angles apart. The same source signal is used as in the previous simulation. The sources are rotated in 22.5° intervals over half a circle around the microphones. The 2D self-localization is evaluated separately for each rotated source geometry. The results are then averaged over the rotations to dampen the effect of special geometries. The number of sources is varied $S = [2, 3, \ldots, 11]$. Reverberation time is varied between 0 s and 1.6 s while SNR is fixed to 20 dB. Figure 5 displays the relative position RMSE (y-axis) averaged over the rotations for different number of sources (x-axis) in different reverberation (different curves). The results show that the RRMS error decreases approximately logarithmically as a function of number of sources in low reverberation $T_{60} \leq 0.4$ s. The high error with few sources is due to not having sources at all end-fire directions. In higher reverberation $T_{60} \geq 0.8$ s, the error does not decrease after a sufficient number of sources are present, i.e., the reflections cause more error into the distance estimates than the distance error caused by non end-fire sources. The minimum reached error level depends on the amount of reverberation.

7. MEASURED DATA RESULTS

Ten Nokia N900 smartphones were placed face up on a wooden table to capture audio at 48 kHz and 16 bit integer accuracy. The meeting room walls are wooden and one wall contains a large window partially covered with curtains. The floor consists of stone tiles and the ceiling is covered with coated fiberglass boards. The reverberation time T_{60} is measured to be 440 ms, and the room floor dimensions are 6×4 m and ceiling rises from 2.9 m to 3.5 m in the middle of the room. During the recording, three seated people talk in turns. The speakers switch chairs until speech has been emitted behind every phone. The ten minute recordings were



(a) Ground truth and estimates with real data.

(b) 10 Device array on a 4 m long table.

Figure 6: Measurement setup is illustrated.

automatically aligned between devices at one tenth of a frame level using the energy envelopes of the signals before any processing. A tape measure was used to obtain ground truth inter-distances of the devices d_{ij}, and MDS was used to obtain ground truth coordinates $\mathbf{M}$. Refer to Fig. 6 for a picture of the setup (right) and the ground truth positions (Fig. 6a, "□" -markers). The table also contained a laptop and other electronic devices.

The same processing parameters as in the simulations (Table 1) are used. The microphone signal SNR is estimated to be roughly 20 dB, and $[100, 13000]$ Hz band was used. The self-localization was performed in 2D. Figure 7 details the self-localization and distance errors as a function of time. Both absolute and relative values are illustrated (refer to Sec. 5) with two different scales. The solid lines represent the position error, and the dashed lines are distance errors. Both errors decrease after 140 s, and slowly decrease during the rest of the recording. The absolute position error reaches 6.9 cm and the relative error is 6.5 % after 10 minutes. The absolute distance RMSE is 13.1 cm and the relative error is 8.1 %. The final self-localization geometry is visualized in Fig. 6a ("○" -markers) along with the annotated geometry. It is noted that the estimated geometry is smaller than the true geometry. This can be explained by the participants not talking at the table height, but in a slightly elevated angle. Therefore, the maximal TDOA values are not exactly observed, since sound did not arrive directly from the end-fire directions. In addition, the reverberation is expected to degrade the performance, as demonstrated by simulations.

8. CONCLUSIONS

This work presents a novel microphone self-localization procedure based on observing the distances between a microphone pairs using time difference of arrival (TDOA) data and non-linear filtering.

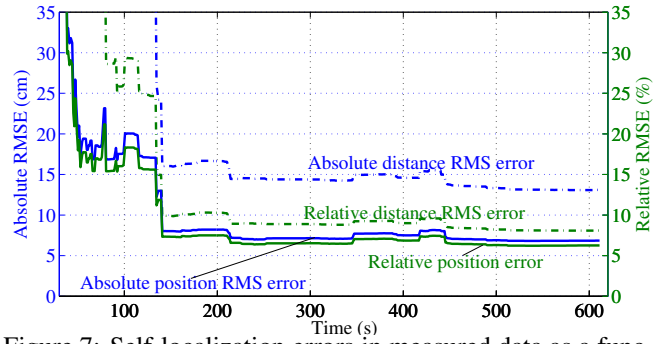


Figure 7: Self-localization errors in measured data as a function of time, refer to Sec 5 for error metrics.

The method does not require synchronous microphone signals or active calibration procedures. In contrast, the only requirement is that continuous audible sounds, such as speech, are observed from near end-fire directions of all microphone pairs. Simulations show that the proposed method is robust against reverberation, and that there is a threshold SNR below which the localization error sharply increases. Simulations showed that the algorithm works even if the sources are not strictly in the end-fire direction, which increases the practical value of the proposed method. Measurements with actual devices in a meeting room achieved relative RMS self-localization error less than 7 %.

9. REFERENCES

[1] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Trans. Audio Speech and Language Proc.*, vol. 16, no. 3, pp. 666, 2008.

[2] V.C. Raykar, I. Kozintsev, and R. Lienhart, "Self localization of acoustic sensors and actuators on distributed platforms," in *WOMTEC*, 2003.

[3] P.D. Jager, M. Trinkle, and A. Hashemi-Sakhtsari, "Automatic microphone array position calibration using an acoustic sounding source," in *ICIEA'09*, 2009, pp. 2110 –2113.

[4] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *ICASSP*, 2008, pp. 2445–2448.

[5] T. Janson, C. Schindelhauer, and J. Wendeberg, "Self-localization application for iphone using only ambient sound signals," in *IPIN'10*, 2010, pp. 1 –10.

[6] I. Borg and P.J.F. Groenen, *Modern Multidimensional Scaling Theory and Applications*, Springer Verlag, 2005.

[7] Lawrence J. Ziomek, *Fundamentals of acoustic field theory and space-time signal processing*, CRC Press, 1995.

[8] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 24, no. 4, pp. 320 – 327, Aug 1976.

[9] J. Allen and D. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943 – 950, 1979.

[10] H. Kuttruff, *Room Acoustics*, Spon Press, 5 edition, 2009.