# Passive Temporal Offset Estimation of Multichannel Recordings of an Ad-Hoc Microphone Array

Pasi Pertilä*, *Member, IEEE,* Matti S. Hämäläinen, *Member, IEEE,* and Mikael Mieskolainen

## Abstract

In recent years ad-hoc microphone arrays have become ubiquitous, and the capture hardware and quality is increasingly more sophisticated. Ad-hoc arrays hold a vast potential for audio applications, but they are inherently asynchronous, i.e., temporal offset exists in each channel, and furthermore the device locations are generally unknown. Therefore, the data is not directly suitable for traditional microphone array applications such as source localization and beamforming.

This work presents a least squares method for temporal offset estimation of a static ad-hoc microphone array. The method utilizes the captured audio content without the need to emit calibration signals, provided that during the recording a sufficient amount of sound sources surround the array. The Cramer-Rao lower bound of the estimator is given and the effect of limited number of surrounding sources on the solution accuracy is investigated. A practical implementation is then presented using non-linear filtering with automatic parameter adjustment. Simulations over a range of reverberation and noise levels demonstrate the algorithm's robustness. Using smartphones an average RMS error of 3.5 samples (at 48 kHz) was reached when the algorithm's assumptions were met.

Manuscript received March 26, 2013. Revised Aug 16, 2013. Accepted Oct 3, 2013. (online version: 22 Oct 2013)

P. Pertilä is with the Department of Signal Processing, Tampere University of Technology (TUT), P.O. Box 553, FI-33101 Tampere, Finland, e-mail: pasi.pertila@tut.fi, Tel.+358 40 849 0786, Fax. +358 3 364 1352.

M.S. Hämäläinen is with Media Technologies Laboratory at Nokia Research Center, Visiokatu 3, 33720 Tampere, Finland, e-mail: matti.s.hamalainen@nokia.com.

M. Mieskolainen is at Department of Physics, University of Helsinki, and Helsinki Institute of Physics, P.O. Box 64, FI-00014, Finland, email: mikael.mieskolainen@cern.ch. At the time of research he was with the Department of Signals Processing, TUT.

**Index Terms**

Synchronization, microphone arrays, ad hoc networks, acoustic measurements, calibration.

## I. Introduction

Today, mobile devices such as smartphones are used to capture audio and video in social events. Their location is generally unknown and there is no precise temporal synchronization between the devices. Such devices are here referred as an ad-hoc sensor network. Automatic compilation of video mixtures of user captured material from an ad-hoc network of smartphones has been proposed in e.g. [1], where users upload their footage of a concert and receive a compilation of the different streams. For this approach, the streams must be first time-aligned to avoid any mixing artifacts. Video based synchronization techniques [2], [3], [4] provide synchronization up to frame rate accuracy of $\pm 40$ ms (using 25 fps). In addition, the audio tracks can be used to synchronize the video captures using music information retrieval techniques [5] such as audio fingerprinting, as in [2]. The audio fingerprint can provide synchronization up to the feature window length, which typically ranges from few milliseconds to tens of milliseconds ($\pm 11.6$ ms in [2]).

Combining the audio of an ad-hoc microphone array using beamforming would offer improved sound quality. However, traditional beamforming relies on the knowledge of microphone positions and assumes sample synchronized audio channels [6], [7]. In addition, sound source localization is known to be highly sensitive to microphone synchronization and position errors [8], and temporal offset error in source separation using independent component analysis (ICA) lowers the separation quality [9]. Therefore, several spatial audio processing algorithms would become applicable on an ad-hoc microphone array if i) synchronization and ii) microphone locations were available.

In this article, we present a closed-form temporal offset estimation of ad-hoc microphone array recordings of sources surrounding the array without the need to use controlled emissions. Previously, we have proposed such method for microphone self-localization [10] and the current work extends this approach to temporal offset estimation. The Cramer-Rao lower bound (CRLB) for the estimator variance is presented. Simulations are used to test the effect of having a limited number of sources surround the array. Then, a practical estimator based on non-linear filtering with automatic parameter adjustment is presented to handle data containing outliers. The robustness of the algorithm is examined in various levels of reverberation and noise using simulations. Finally, we demonstrate the accuracy of the approach by considering actual recordings made in three different rooms with up to ten smartphones. The results are contrasted to data captured with a multichannel AD-converter.

The paper is organized as follows. Section II overviews solutions of related problems of wireless time-synchronization and microphone self-localization. In Section III, the problem of offset estimation is formulated and a solution is presented. The CRLB is derived and the effect of a limited number of sources is investigated in Section IV. Section V presents the practical estimator to handle outlier-corrupted data. The algorithm is evaluated in different reverberation and noise conditions in Section VI. Section VII describes the measurement setup, followed by the results Section VIII, and a discussion Section IX. Finally, conclusions are drawn in Section X.

## II. Time-Synchronization and Self-Localization

The time-synchronization of wireless network devices is a related problem to microphone synchronization, but requires a network interface to send and receive messages containing time-stamps and identification information. Various synchronization schemes have been devised and the interested reader can refer to e.g. [11], [12] for an overview. In [13] radio-frequency is used to synchronize capturing devices for audio signal separation. However, many devices such as portable audio recorders and video cameras lack a network interface and/or appropriate software. The audio tracks can even be obtained after the recording event, in which case the network synchronization algorithms are not applicable.

A related and partially overlapping field of research to microphone synchronization is microphone self-localization, i.e., the determination of the sensor geometry from either controlled emissions or from ambient sounds. In general, methods that use controlled emission are able to provide accurate microphone location information but require a loudspeaker. The methods [14], [15] utilize loudspeakers with known positions to emit synchronized calibration signals to locate the microphones. Methods presented in [16], [17], [18] utilize synchronized audio capturing but do not require known source positions or controlled emissions. The method presented in [19] assumes that the time of flight (TOF) from sources to sensors is known when locating sensors and sources. The active ranging of two asynchronous devices that have a microphone and a loudspeaker is researched in [20]. Specific self-localization methods for ad-hoc arrays of such devices have been proposed in [21], [22], [23], [24]. They are able to achieve high accuracy self-localization geometry, and [21], [24] are fundamentally closed-form estimators in contrast to iterative solutions. The method presented in [21] also includes acoustically determined orientation estimate of the device, assuming the device contains a microphone array. The method has been utilized in an asynchronous source localization framework [25].

Passive self-localization methods utilize only ambient sounds and include a method by McCowan et al. [26], which assumes a diffuse noise field. It estimates the microphone pairwise distance by fitting the measured noise coherence to the theoretical model of the field. Another approach is presented by Pertilä et al. [10], which determines the sensor pairwise distances by estimating the maximal sound travel times between the microphones (both negative and positive direction), i.e., minimum and maximum time difference of arrival (TDOA) values. The limit of the method [10] is that sources must reside in the endfire directions of microphone pairs. Such a conditions can be expected e.g. in a meeting room with participants seated in front of their mobile phones, which form the ad-hoc array. The methods [26], [10] use the estimated pairwise distances between devices to obtain Euclidean sensor coordinates using multidimensional scaling (MDS) [27]. The self-localization research by Raykar et al. [28] utilizes active emissions and formulates a maximum likelihood equation between the unknown parameters (time offsets, microphone positions) and the measurements (TDOA or TOF). The cost function is minimized iteratively to jointly solve the unknown parameters. Ono et al. [29], [30] present a TDOA based cost function for joint self-localization, source localization, and temporal offset estimation without the need for controlled calibration signals. The fundamental issues with the discussed cost function minimization approaches are high dimension of the search space, and providing a good initial guess for the iterative algorithm to avoid convergence into a local minimum. An online approach for the problem is presented by Miura et al. [31], which utilizes simultaneous localization and mapping (SLAM). More specifically, extended Kalman filtering and delay-and-sum beamforming are used to calibrate the stationary array by observing hand-clappings from directions surrounding the array.

## III. Temporal Offset estimation theory

Let $\mathbf{m}_i \in \mathbb{R}^3$ be the $i$th receiver position, and $i \in 1, \dots, M$, where $M$ is the number of microphones. In a noiseless and anechoic room the signal at microphone $i$ can be modeled as a time-shifted version of the source signal $s(t)$ as

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

where $*$ is convolution, $t$ is time, $\delta(\cdot)$ is the Dirac's delta function, and $\tau_i$ is time of arrival (TOA) in the $i$th microphone consisting of the sum of the physical propagation delay and the unknown time offset of the $i$th sensor denoted as $\Delta_i$

$$\tau_i = c^{-1}\|\mathbf{s} - \mathbf{m}_i\| + \Delta_i, \tag{2}$$

where $c$ is the speed of sound, and $\mathbf{s} \in \mathbb{R}^3$ denotes source position. The time difference of arrival (TDOA) between a source and microphones $i$ and $j$ is

$$\tau_{ij} \triangleq \tau_i - \tau_j = c^{-1}(\|\mathbf{s} - \mathbf{m}_i\| - \|\mathbf{s} - \mathbf{m}_j\|) + y_{ij}, \tag{3}$$

where $y_{ij} \triangleq \Delta_i - \Delta_j$ is the pairwise time offset. Based on *reverse triangle inequality* the minimum and maximum TDOA values are limited by the microphone separation distance

$$\Bigg| \|\mathbf{s} - \mathbf{m}_i\| - \|\mathbf{s} - \mathbf{m}_j\| \Bigg| \leq \|(\mathbf{m}_j - \mathbf{m}_i)\|, \tag{4}$$

therefore only a source located on the line that connects the microphones, outside of the microphone pair can cause the maximum TDOA observation. All source positions on the line can be defined as

$$\mathbf{s}(\alpha) \triangleq \mathbf{m}_i + \alpha(\mathbf{m}_j - \mathbf{m}_i), \tag{5}$$

where parameter $\alpha \in \mathbb{R}$ controls the position of the source on the line. The TDOA of such (near-field) source is obtained by substituting $\mathbf{s}(\alpha)$ into (3)

$$\tau_{ij}(\alpha) = c^{-1}(|\alpha| - |\alpha - 1|)\|(\mathbf{m}_j - \mathbf{m}_i)\| + y_{ij}. \tag{6}$$

This can be simplified as

$$\tau_{ij}(\alpha) = y_{ij} + c^{-1} \begin{cases} -\|(\mathbf{m}_j - \mathbf{m}_i)\|, & \text{if } \alpha \leq 0 \\ \|(\mathbf{m}_j - \mathbf{m}_i)\|, & \text{if } \alpha \geq 1 \\ (|\alpha| - |\alpha - 1|)\|(\mathbf{m}_j - \mathbf{m}_i)\|, & \text{otherwise} \end{cases} \tag{7}$$

The definition for maximum TDOA is $\tau_{ij}^{\text{max}} \triangleq \tau_{ij}(\alpha)|\alpha \geq 1$ and for the minimum TDOA $\tau_{ij}^{\text{min}} \triangleq \tau_{ij}(\alpha)|\alpha \leq 0$ is used, since the first two rows of solution (7) represent the lower and upper limits of a TDOA observation. Now, (7) can be written

$$\tau_{ij}(\alpha) = \begin{cases} \tau_{ij}^{\text{min}}, & \text{if } \alpha \leq 0 \\ \tau_{ij}^{\text{max}}, & \text{if } \alpha \geq 1 \\ y_{ij} + c^{-1}(|\alpha| - |\alpha - 1|)\|(\mathbf{m}_j - \mathbf{m}_i)\|, & \text{otherwise} \end{cases} \tag{8}$$

We point out two trivial cases, which are suitable for naïve estimation of the microphone pairwise offset. In the first case, the source and microphone geometry is known. Therefore, the pairwise offset values $y_{ij}$ can be directly estimated by subtracting the propagation related delays from the TDOA value: $y_{ij} = \tau_{ij} - c^{-1}(\|\mathbf{s} - \mathbf{m}_i\| - \|\mathbf{s} - \mathbf{m}_j\|)$ in (3). In the second case the array is collapsed, i.e., $\mathbf{m}_i = \mathbf{m}_j \ \forall \ \{i, j\}$, then $c^{-1}(\|\mathbf{s} - \mathbf{m}_i\| - \|\mathbf{s} - \mathbf{m}_j\|) = 0$ and $y_{ij} = \tau_{ij}$. Note that TDOA values $\tau_{ij}$ can be measured, e.g., using correlation.

**Theorem 1.** In the non-trivial case, the microphone pairwise offset $y_{ij}$ is

$$y_{ij} = \frac{1}{2}\left(\tau_{ij}^{\text{max}} + \tau_{ij}^{\text{min}}\right). \tag{9}$$

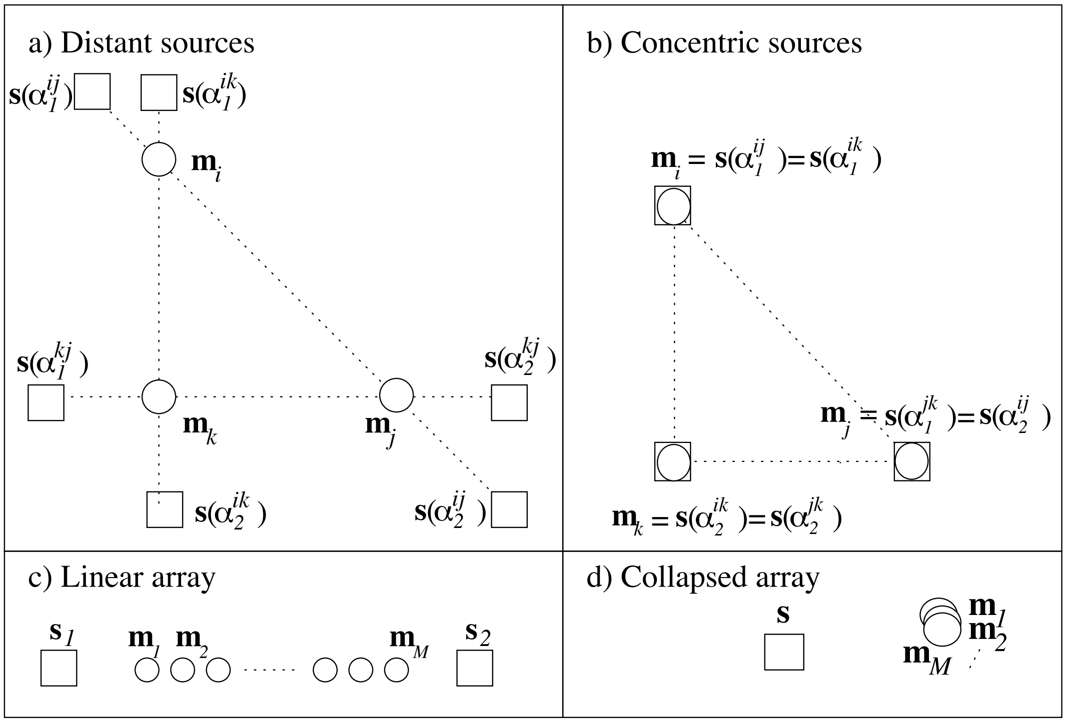


Fig. 1: The circles represent microphone locations $\mathbf{m}_i$, $i = 1, \ldots, M$ and squares represent source positions $\mathbf{s}$. Panels a)-d) illustrate geometries that fulfill the requirements (11) of Corollary 1, i.e., the proposed method's assumptions. In panel a) this is achieved with $P = M(M-1)/2$ sources. In panel b) equal amount of concentric transducers ($P = M$) is sufficient. Panel c) linear array requires only two sources ($P = 2$) in endfire directions. Panel d) A collapsed array requires only a single source in any point.

*Proof.* By Using (7)

$$\begin{aligned}\frac{1}{2}\left(\tau_{ij}^{\max} + \tau_{ij}^{\min}\right) &= \frac{1}{2}(c^{-1}\|\mathbf{m}_i - \mathbf{m}_j\| + y_{ij} + \\ -c^{-1}\|\mathbf{m}_i - \mathbf{m}_j\| + y_{ij}) &= \frac{1}{2}(y_{ij} + y_{ij}) = y_{ij} \quad \square \end{aligned} \tag{10}$$

The Theorem 1 is proven also for the far-field sound propagation model in Appendix A.

**Corollary 1.** The pairwise offset $y_{ij}$ depends on $\tau_{ij}^{\max}$ and $\tau_{ij}^{\min}$ which can only be observed when sources are exactly on the line defined by (5) at both sides of the microphone pair. Formally the requirement is

$$\forall\{i,j\}:\ \exists\,\{\mathbf{s}(\alpha_1^{ij}), \mathbf{s}(\alpha_2^{ij})\}, \alpha_1^{ij} \leq 0, \alpha_2^{ij} \geq 1. \tag{11}$$

Different microphone and source geometries that satisfy Corollary 1 requirements (11) are visualized in Fig. 1. Since there are $P = M(M-1)/2$ unique microphone pairs and each pair requires two sound sources as stated by (11), a total of $M(M-1)$ sources is sufficient to estimate the temporal offset of $M$ microphones in the general case, see panel a). However, a special case of the condition is reached when sources are concentric with the microphones, refer to panel b), where $P = M$ sources fulfill the requirements (11) of Corollary 1. A second special case is the linear array, where only $P = 2$ sources are required in both endfire directions for the determination of the offsets between $M$ channels, refer panel c). The trivial case is visualized by the collapsed array in panel d), which requires only a single source $P = 1$ to estimate the offset between $M$ microphones. The collapsed array can be approximated in practice by bringing the microphones close to each other.

The pairwise offset measurements $\mathbf{y}$ are generated from the offset values $\boldsymbol{\Delta}$ by the observation matrix $\mathbf{H}$

$$\mathbf{y} = \mathbf{H}\boldsymbol{\Delta} \tag{12}$$

where

$$\mathbf{y} = [y_{1,2}, y_{1,3}, \ldots, y_{M-1,M}]^{\mathrm{T}}, \ \mathbf{y} \in \mathbb{R}^{P\times 1}$$

$$\boldsymbol{\Delta} = [\Delta_1, \Delta_2, \Delta_3, \ldots, \Delta_M]^{\mathrm{T}}, \ \boldsymbol{\Delta} \in \mathbb{R}^{M\times 1}$$

$$\begin{aligned}\mathbf{H} = [\ &\mathbf{e}_1 - \mathbf{e}_2, \mathbf{e}_1 - \mathbf{e}_3, \ldots, \mathbf{e}_1 - \mathbf{e}_M, \mathbf{e}_2 - \mathbf{e}_3, \ldots, \\ &\mathbf{e}_2 - \mathbf{e}_M, \ldots, \mathbf{e}_{M-1} - \mathbf{e}_M]^{\mathrm{T}}, \ \mathbf{H} \in \mathbb{R}^{P\times M},\end{aligned} \tag{13}$$

where the $i^{\text{th}}$ unit vector $\mathbf{e}_i$ has a 1 in the $i^{\text{th}}$ position, and 0s in all other positions. E.g. in case of three sensors (13) is

$$\mathbf{H} = [\mathbf{e}_1 - \mathbf{e}_2, \mathbf{e}_1 - \mathbf{e}_3, \mathbf{e}_2 - \mathbf{e}_3]^{\mathrm{T}} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

To solve $\boldsymbol{\Delta}$ in (12) the cost function to be minimized is

$$\epsilon(\boldsymbol{\Delta}) = (\mathbf{H}\boldsymbol{\Delta} - \mathbf{y})^{\mathrm{T}}(\mathbf{H}\boldsymbol{\Delta} - \mathbf{y}) \tag{14}$$

and the Moore-Penrose matrix inverse is used to solve the offset values

$$\hat{\boldsymbol{\Delta}} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y}. \tag{15}$$

However, the rows of observation matrix $\mathbf{H}$ are linearly dependent. To eliminate this extra degree of freedom we arbitrarily set the first sensor's offset to zero and remove the first column of the observation matrix $\mathbf{H}$ when solving the rest $M-1$ offset values in Eq. (15). In practice this means that the global temporal offset cannot be determined.

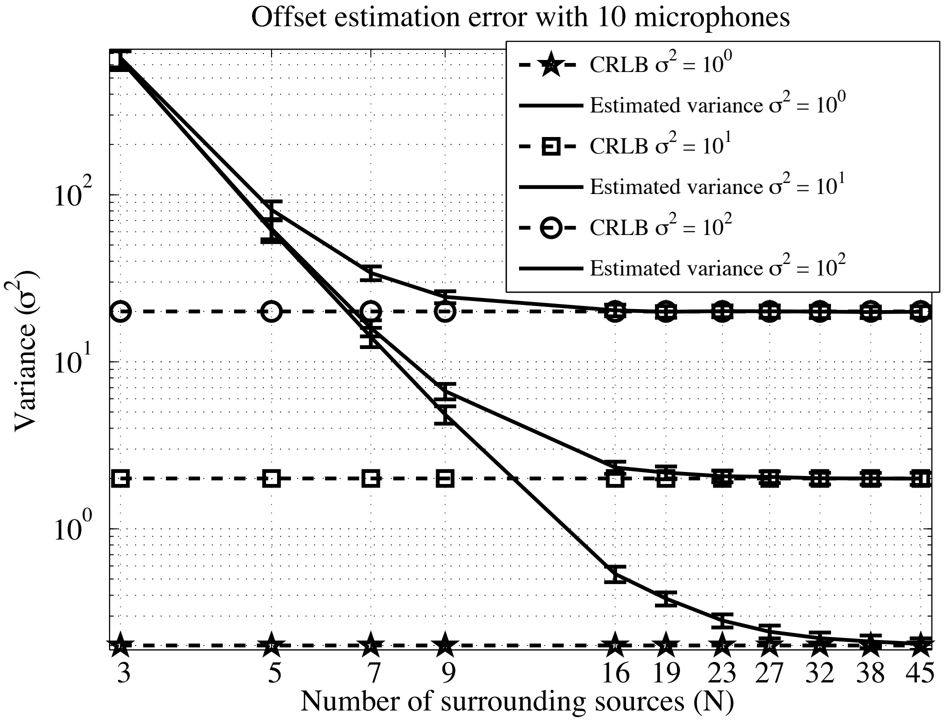


Fig. 2: The Cramer-Rao lower bound for the proposed offset estimator and the estimator's variance are plotted as a function of number of sources surrounding a 2D array. Three different levels of offset noise is used with 48 kHz sampling rate.

## IV. Cramer-Rao lower bound of offset estimation

The microphone pairwise offset observation can be modeled as

$$\mathbf{y} = \mathbf{H}\boldsymbol{\Delta} + \mathbf{w}, \tag{16}$$

where elements of Gaussian noise vector $\mathbf{w}$ are assumed independent between pairs and identically distributed with covariance $\sigma^2\mathbf{I}$. The CRLB gives the minimum achievable variance of any unbiased estimator. The CRLB of the estimator (16) is [32]

$$\mathbf{C}_{\boldsymbol{\Delta}} = \sigma^2(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}. \tag{17}$$

When the requirements (11) of Corollary 1 are exactly met, the CRLB is reached. However, the CRLB does not show the effect of not meeting the requirements.

### A. *Effect of the Number of Surrounding Sources on Estimation Accuracy Using an Arbitrary Array*

A simulation is made to test the behavior of the offset estimation method when the requirements (11) of Corollary 1 are not exactly met, i.e., both endfire directions of each pair only approximately contains a source. For this purpose, a random 2D array geometry with x and y microphone coordinates in meters is drawn from normal distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ with parameters $\boldsymbol{\mu} = [0, 0]^{\mathrm{T}}, \sigma^2 = 1$. Then, sound sources are placed on a circle with equal angular spacing to surround the microphones at four-meter radius, and the number of sound sources is in the range from 3 to 45. The pairwise offset measurements $\mathbf{y}$ are obtained from the near-field sound propagation model (3) using the generated microphone and sound source positions with speed of sound set to $c = 343$ m/s. The ground truth offset $\boldsymbol{\Delta}$ is drawn from a multivariate normal distribution $\boldsymbol{\Delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \times \mathbf{I})$, where $\sigma^2 = 100^2$.

The measurement noise $\mathbf{w}$ in (16) is zero mean normal distributed with different levels of variance $\sigma^2 = 1, 10$, and $100$ (the offset unit is samples at 48 kHz). Eq. (15) is used to estimate the offset. For each number of sources 50000 Monte Carlo trials are run. During each run, the offset values, measurement noise, and microphone and source coordinates are drawn from their corresponding distributions. Fig. 2 displays the resulting average variance of the offset estimator with the standard deviation as vertical bars $(\pm\sigma)$. The dashed lines indicate the mean value of the diagonal elements of the CRLB matrix.

The results indicate that the CRLB is asymptotically reached when the number of sources surrounding the array is increased. In practice, the requirements of Corollary 1 do not need to strictly hold to produce low variance offset estimates.

## V. Practical offset estimation

If the time difference between audio streams is larger than the processing frame length $L$, the streams are first frame-aligned before any processing. This can be done e.g. by aligning the signal energy envelope curves.

A voice activity detection (VAD) scheme is used as a pre-processing stage to remove frames that do not contain sound activity. Here, a quantile based voice activity detection (VAD) is applied [33], which has two parameters: the SNR threshold $\lambda_{\mathrm{SNR}}$ and amount of history frames $N_v$. Other VAD schemes could also be considered, refer to e.g. [34] for a review of different approaches. The algorithm is run on each channel separately, and if all channels are detected to contain voice, the corresponding time frame is marked as active.

For the remaining active frames $t = 1, \ldots, T$, the generalized cross-correlation (GCC) between sampled

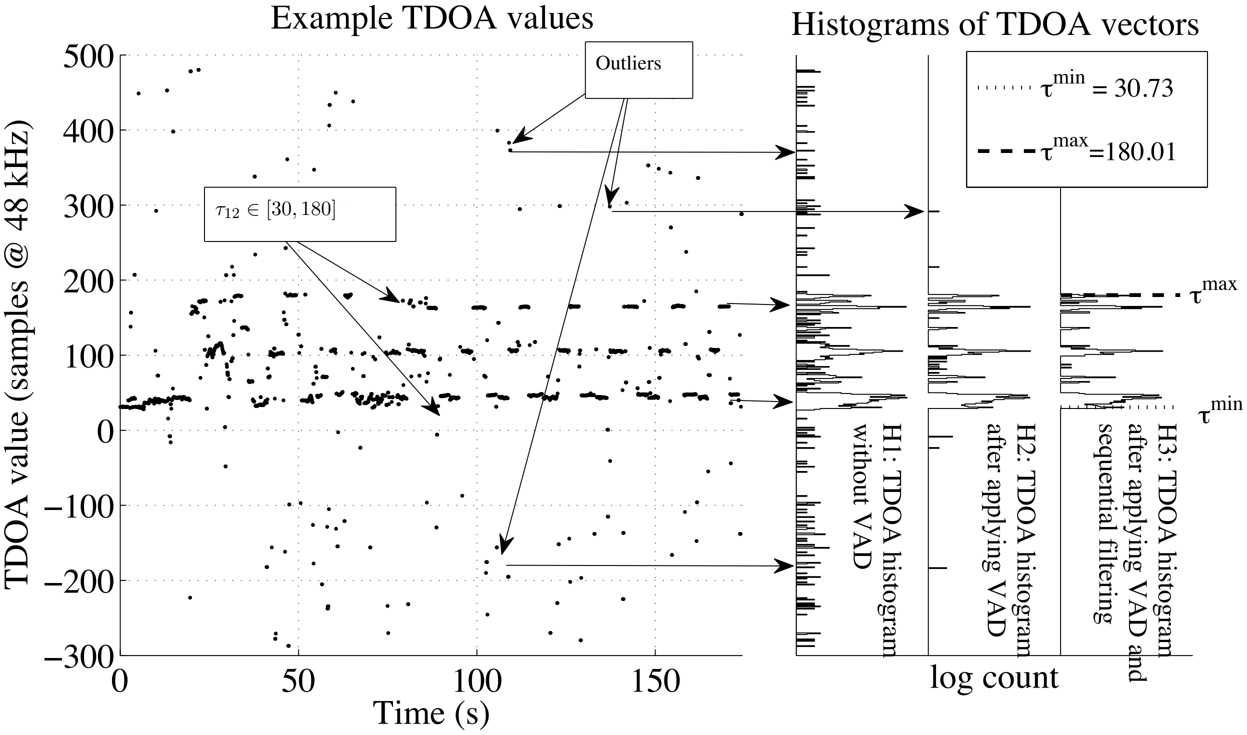


Fig. 3: An example of a microphone pairwise TDOA values ($\tau_{12}$) from a three-minute recording without using VAD or sequential filtering. Right panel displays three histograms H1–H3 of which the leftmost (H1) is obtained from the raw TDOA data. The effect of using VAD [33] is illustrated by the middle histogram (H2). The rightmost histogram (H3) displays the use of sequential filtering. Minimum and maximum TDOA values of the filtered set are displayed.

microphone signals $i, j$ with PHAT weight $\Psi(t, \omega) = |X_i(t, \omega)X_j^*(t, \omega)|^{-1}$ is obtained using [35]

$$r_{ij}(t, \tau) = \sum_{\omega} \Psi(t, \omega)X_i(t, \omega)X_j^*(t, \omega)\exp(j\omega\tau) \tag{18}$$

where $()^*$ is complex conjugate, $X_i(t, \omega)$ is Fourier transform of input frame at time $t$ at angular frequency $\omega$, and $\tau$ is time delay. A TDOA value is estimated by searching the correlation function peak index value

$$\hat{\tau}_{ij}(t) = \underset{\tau}{\operatorname{argmax}}\, r_{ij}(t, \tau). \tag{19}$$

A sub-sample TDOA estimate is then obtained by parabolic peak interpolation using the two adjacent values of the peak. The processing is performed in short time frames of length $L$ samples and $\hat{\boldsymbol{\tau}}_{ij} \in \mathbb{R}^{T\times 1}$ denotes a vector of TDOA values from all $T$ active frames, and $\boldsymbol{\tau} = [\hat{\boldsymbol{\tau}}_{1,2}, \ldots, \hat{\boldsymbol{\tau}}_{M-1,M}]$ denotes a matrix of all pairwise TDOA vectors.

The TDOA of speech sources is assumed stable between sequential frames, while uncorrelated background noise results in seemingly random TDOA values. Therefore, a heuristic gating procedure is applied to filter out TDOA values that differ sequentially more than $\lambda_G$ samples between any three sequential frames after [10].

### A. Naïve Offset Estimation

The naïve offset estimation algorithm assumes that all microphones are located in the same point in space, i.e., the array is collapsed. The pairwise offset value $y_{ij}$ is then the expected value of the corresponding TDOA: $y_{ij} = E[\tau_{ij}]$. The final microphone offset estimate is obtained by using the Moore-Penrose inverse (15). Figure 5 summarizes the method, where the discussed sequential filtering of TDOA values is presented on line 4.

### B. Offset Estimation using Minimum and Maximum TDOA Estimation

The array is no longer assumed as being collapsed. A small number of outliers may still exist after the VAD and sequential filtering. Using minimum and maximum operators to estimate $\tau_{ij}^{\max}$ and $\tau_{ij}^{\min}$ from the TDOA vector may result in an outlier pairwise offset value $y_{ij}$, which is likely to corrupt the offset estimator. Therefore, the $q$-quantile operator is used instead. The algorithm is listed in Fig. 6.

After the sequential filtering (line 4), the remaining TDOA values are sorted for each pair, i.e., $\tilde{\boldsymbol{\tau}}_{ij}(0) \leq \tilde{\boldsymbol{\tau}}_{ij}(1) \leq \ldots \leq \tilde{\boldsymbol{\tau}}_{ij}(T_{ij})$, where $T_{ij}+1$ is the amount of TDOA values after the sequential filtering operation for the microphone pair. The $q$-quantile TDOA is $\tilde{\boldsymbol{\tau}}_{ij}(\lfloor q \cdot T_{ij} \rceil)$, where $\lfloor \cdot \rceil$ denotes rounding to nearest integer, and $q \in [0,1]$. It is assumed that the errors are symmetrically distributed. Therefore a single quantile parameter is used for both minimum and maximum estimation as described on lines 5-8 in

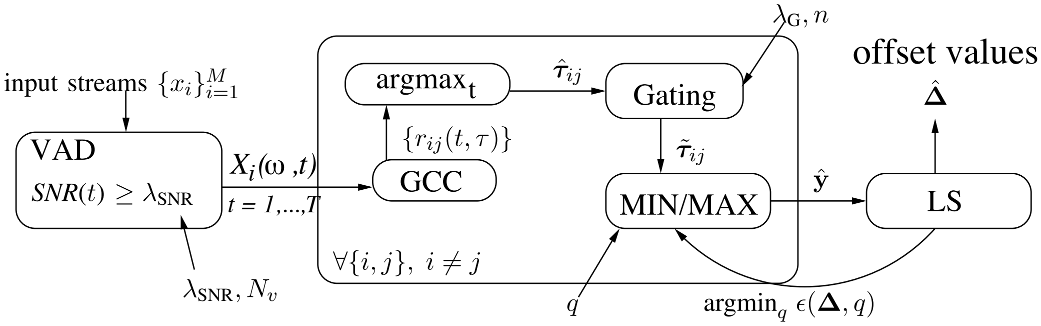

input streams $\{x_i\}_{i=1}^M$
VAD
$SNR(t) \geq \lambda_{\mathrm{SNR}}$
$\lambda_{\mathrm{SNR}}, N_v$
$X_i(\omega, t)$
$t = 1,\ldots,T$
argmax$_t$
$\{r_{ij}(t,\tau)\}$
GCC
$\hat{\boldsymbol{\tau}}_{ij}$
Gating
$\lambda_G, n$
$\tilde{\boldsymbol{\tau}}_{ij}$
MIN/MAX
$\forall\{i,j\},\ i \neq j$
$q$
$\hat{\mathbf{y}}$
LS
offset values
$\hat{\boldsymbol{\Delta}}$
$\mathrm{argmin}_q\ \epsilon(\boldsymbol{\Delta}, q)$

Fig. 4: Block diagram of the implementation of the proposed synchronization method.

```
1:  procedure $\boldsymbol{\Delta} = \textsc{NaiveOffset}(\boldsymbol{\tau}, \mathbf{H}, \lambda_G, n)$
2:      for $i = 1$ to $M - 1$ do
3:          for $j = i + 1$ to $M$ do
4:              $\tilde{\boldsymbol{\tau}}_{ij} = \{\tau_{ij}(t) \mid \lambda_G > \|\tau_{ij}(t) - \tau_{ij}(t-n)\|, \forall t\}$
5:              $\hat{y}_{ij} = \text{mean}(\tilde{\boldsymbol{\tau}}_{ij})$                    ▷ (9)
6:          end for
7:      end for
8:      $\mathbf{y} = [\hat{y}_{12}, \hat{y}_{13}, \dots, \hat{y}_{M-1,M}]^{\mathrm{T}}$
9:      $\boldsymbol{\Delta} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y}$                    ▷ (15)
10:     return $\boldsymbol{\Delta}$
11: end procedure
```

Fig. 5: A naïve algorithm for closed-form microphone temporal offset estimation.

Fig. 6. The pairwise offset is then obtained by combining the estimated minimum and maximum TDOA values using (9) on line 9. Finally, after the pairwise offset estimation is completed, the Moore-Penrose inverse is used to solve the channel offset values. Figure 4 illustrates the block diagram of the method.

### *C. Example*

The left axis in Fig. 3 illustrates microphone pairwise TDOA values from an actual three-minute recording of speech between four persons without using VAD or sequential filtering. The data is gathered at 48 kHz sampling rate. The x-axis represents time, y-axis represents the TDOA values between a pair of microphones with 56 cm separation, and 12 TDOA values per second are obtained. In the range ($\tau_{12} \in [30, 180]$), the TDOA values are caused by actual speakers. The outlier values are outside of this physically possible delay range, and are introduced by sensor noise, background noise and reverberated components. The right panel displays three histograms labeled as H1–H3, where the leftmost histogram H1 is obtained from the raw TDOA data. The horizontal arrows illustrate how TDOA values are mapped into histogram bins. The physically bounded TDOA region can be distinguished from the data due to its larger bin counts. Using VAD ($\lambda_{\text{SNR}} = 0$ dB, $N_v = 30$) to remove frames clears the majority of outliers, refer to the middle histogram H2. Applying the sequential filter ($\lambda_{\text{G}} = 6,\ n = 2$) to this TDOA data results in the rightmost histogram H3. Now, the maximum and minimum TDOA values can be extracted using the $q$-quantile where $q$ is close to one. In the example, the ground truth pairwise offset $y_{1,2}$ is

```
procedure $\{\mathbf{\Delta}, \epsilon\} = \textsc{EstOffset}(\boldsymbol{\tau}, \mathbf{H}, \lambda_G, q, n)$
    for $i = 1$ to $M - 1$ do
        for $j = i + 1$ to $M$ do
            $\tilde{\boldsymbol{\tau}}_{ij} = \{\hat{\tau}_{ij}(t) \mid \lambda_G > \|\tau_{ij}(t) - \tau_{ij}(t - n)\|, \forall t\}$
            $\tilde{\boldsymbol{\tau}}_{ij} \leftarrow \text{sort}\{\tilde{\boldsymbol{\tau}}_{ij}\}$
            $T_{ij} = \text{length}(\tilde{\boldsymbol{\tau}}_{ij}) - 1$
            $\hat{\tau}_{ij}^{\text{min}} = \tilde{\boldsymbol{\tau}}_{ij}(\lfloor(1 - q) \cdot T_{ij}\rceil)$
            $\hat{\tau}_{ij}^{\text{max}} = \tilde{\boldsymbol{\tau}}_{ij}(\lfloor q \cdot T_{ij}\rceil)$
            $\hat{y}_{ij} = \left(\hat{\tau}_{ij}^{\text{max}} + \hat{\tau}_{ij}^{\text{min}}\right)/2$                ▷ (9)
        end for
    end for
    $\mathbf{y} = [\hat{y}_{12}, \hat{y}_{13}, \ldots, \hat{y}_{M-1,M}]^{\text{T}}$
    $\mathbf{\Delta} = (\mathbf{H}^{\text{T}}\mathbf{H})^{-1}\mathbf{H}^{\text{T}}\mathbf{y}$                ▷ (15)
    $\epsilon = (\mathbf{H}\mathbf{\Delta} - \mathbf{y})^{\text{T}}(\mathbf{H}\mathbf{\Delta} - \mathbf{y})$                ▷ (14)
    return $\mathbf{\Delta}$, and $\epsilon$
end procedure
```

Fig. 6: An algorithm for closed-form microphone temporal offset estimation.

105.1 samples, and the pairwise offset estimate from the minimum and maximum TDOA values (9) $\hat{y}_{1,2}$ is 105.4 samples and thus the error is less than one sample (0.3 samples, corresponding to 5.3 $\mu s$).

### D. Selection of Filtering Parameters

The presented non-linear filtering algorithm contains five parameters that control the TDOA data outlier rejection: VAD parameters (for [33] $\lambda_{\text{SNR}}$ and $N_v$), sequential filter parameters (gate width $\lambda_{\text{G}}$, and order $n$), and the quantile value $q$. The VAD and sequential filter parameters can be selected with insight into the problem. VAD parameters $\lambda_{\text{SNR}} = 0$ dB, $N_v = 30$ are selected to filter out frames that are below the background level. With these values roughly $58$ % of the frames were labeled as inactive in the tested meeting recordings. Sequential TDOA values from the same source will have a low variance whereas the outlier TDOA values seem to be uniformly distributed having a very large variance. Therefore, it is not sensitive to the threshold. Here, $\lambda_{\text{G}} = 6, n = 2$ are selected as in [10]. These parameters are held fixed for all the data processed.

The remaining quantile parameter $q$ is equal to one if there are no outliers, and less than one when outliers are present. The use of a parameter value $q$ that minimizes the least squares error in (15) is proposed. A penalty term is used to multiply the LS cost function (14) to favor $q$ values close to one. The proposed restricted error criterion is

$$\epsilon(\mathbf{\Delta}, q) = \epsilon(\mathbf{\Delta})(1 + \mu \cdot (1 - q)), \tag{20}$$

where $\mu$ is a weight parameter enforcing $q$ values close to one. Figure 7 displays the scaled error criterion $\epsilon(\mathbf{\Delta}, q)$ as a function of the quantile value $q$ for two different recordings[1]. The quantile cost function parameter is empirically set to $\mu = 100$. In addition, the offset RMS error (21) using ground truth is displayed. The offset RMS error is defined here as

$$\text{RMSE}(\mathbf{\Delta}) = \sqrt{\frac{1}{M-1}\sum_{i=2}^{M}(\hat{\Delta}_i - \Delta_i)^2}, \tag{21}$$

where the offset of the first reference channel is aligned to zero, and the offset RMSE is estimated from channels $i = 2, 3, \dots, M$. The circle represents the value obtained by minimizing (20). The $q$ values that minimize $\epsilon(\mathbf{\Delta}, q)$ also have low RMS error in the graphs, and there is no single optimal value of $q$ to use for both recordings. An algorithm for offset estimation that minimizes the modified cost function is presented in Fig. 8. Refer to Fig. 4 for the block diagram.

## VI. Simulations

The GCC based TDOA estimation will fail after a threshold value of reverberation is reached [36]. Similarly, when the SNR of the received signal drops below a threshold value the estimator is known to degrade in steps [37]. Therefore, the performance of the proposed algorithm (Fig. 8) is estimated using the image method [38] to control the amount of reverberation in different SNR conditions. The room geometry is modeled after an existing meeting room of floor size $8.0 \times 5.2$ m. A ten-microphone letter ”U” shaped 2D array is simulated, refer to array labeled as ”Room 1” in Fig. 10d. A 13 second female voice speech sequence is played sequentially from 42 cm proximity of every microphone. The source is displaced 30 cm horizontally from the microphone, into the outward direction from array center. In addition, a 30 cm elevation displacement is added. The source placement with respect to the microphone aims to simulate a seated person behind every microphone. Note that the geometry is inspired by a natural use case, and does not strictly fulfill the requirement (11). Therefore, exact results are not expected.

[1] Data from recording 1 (room 1) in left panel, and recording 15 (laboratory) in right panel, refer to Section VII for description about data.

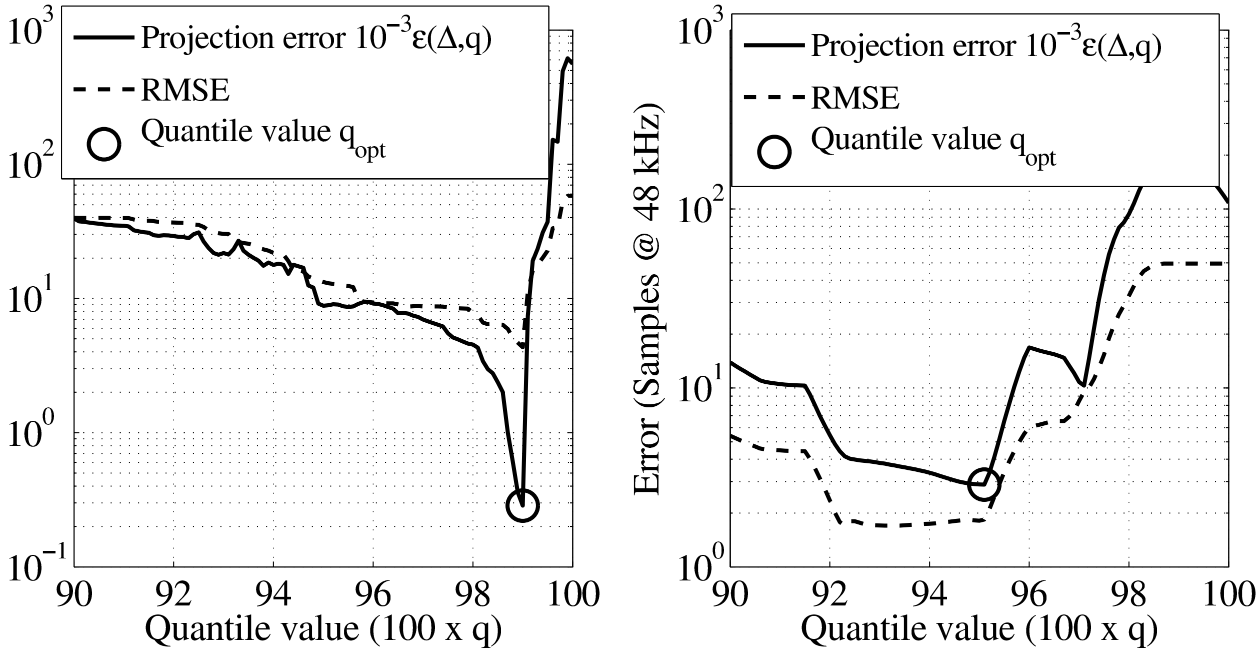


Fig. 7: The offset error criterion $\epsilon(\mathbf{\Delta}, q)$ plotted as function of the quantile value (solid line) for two recordings. The RMS error of the offset estimation is also plotted (dashed line).

```
procedure $\mathbf{\Delta} = \textsc{EstOffset2}(\boldsymbol{\tau}, \mathbf{H}, \lambda_G, n, \mu)$
    $q_{\text{opt}} = \underset{q}{\text{argmin}}\{\textsc{Cost}(\boldsymbol{\tau}, \mathbf{H}, \lambda_G, q, n, \mu)\}$
    $\mathbf{\Delta} = \textsc{EstOffset}(\boldsymbol{\tau}, \mathbf{H}, \lambda_G, q_{\text{opt}}, n)$
    return $\mathbf{\Delta}$
end procedure
procedure $\epsilon_2 = \textsc{Cost}(\boldsymbol{\tau}, \mathbf{H}, \lambda_G, q, n, \mu)$
    $\{\mathbf{\Delta}, \epsilon_1\} = \textsc{EstOffset}(\boldsymbol{\tau}, \mathbf{H}, \lambda_G, q, n)$
    $\epsilon_2 = \epsilon_1 \cdot (1 + \mu \cdot (1 - q))$                              ▷ (20)
    return $\epsilon_2$
end procedure
```

Fig. 8: Iterative offset estimation using modified cost function.

The sampling rate is set to 48 kHz and the utilized speech sequence from TIMIT database is sampled at 16 kHz. The SNR is varied in the range $[-10, 10]$ dB by adding white noise to the frequency band $[0, 8]$ kHz. The SNR is here defined as the ratio between the direct path sound amplitude $a_1$ at one-meter distance from the source and the amplitude of the noise $a_n$, i.e., SNR $= 20\log_{10}(a_1/a_n)$. The desired reverberation time T$_{60}$ is varied between $[0, 2.4]$ s and is obtained by controlling a common surface reflection coefficient $\beta$, which is estimated with Eyring's reverberation formula [39]. The results are averaged over ten rotations to lessen the effect of possible special reflections. In each simulation, the channel offset values were uniformly distributed over the range of 500 samples.

The data is processed in 8192 sample Hann weighted windows with parameter values discussed in Section V-D.

The performance of the proposed algorithm is plotted in Fig. 9. Each line indicates the offset estimation RMSE with constant reverberation time in different SNR conditions. The bottom dotted line is obtained by simulating a source 5 cm away from the microphone without reverberation to obtain practical performance limit of the method by fulfilling the requirement (11). In each reverberation level a 70 $\mu s$ offset estimation RMSE is reached when the SNR is sufficient. For the anechoic case $-5$ dB is a sufficient SNR where as in highly reverberant space of T$_{60}$ $= 2.4$ s reverberation time requires at least $+7.5$ dB SNR. Note that the direct path SNR drops 6 dB when the distance is doubled and therefore the SNR at each microphone is different for every source.

The non-monotonic decrease of the RMSE value as a function of SNR is due to bias caused by not having the sources exactly in the microphone connecting axis. In the case where the speaker placement fulfills the requirement (11) the RMSE is monotonously decreasing (dotted line). To summarize the simulation results, the algorithm can perform in highly reverberant conditions even in relatively low SNR values, e.g., when T$_{60}$ $= 1.4$ s and SNR only $+5$ dB the offset estimation RMSE of 2.2 samples is reached.

## VII. MEASUREMENTS

The actual measurement data is gathered in three different physical environments. The first two environments are meeting rooms, where ten smartphones are placed on a table. The first room's floor dimensions are $8.0 \times 5.2$ m, and reverberation time T$_{60}$ is approximately 320 ms, see Fig. 10a. The second room's floor dimensions are $5.1 \times 6.6$ m and the reverberation time T$_{60}$ is approximately 370 ms, see Fig. 10b. Figure 10d illustrates the centered microphone positions on the table surfaces in the two rooms. The table in room 1 is larger compared to room 2 which results in a larger array.

The third environment is a laboratory with floor size of $4.5 \times 4.0$ m and reverberation time T$_{60}$ approximately 260 ms, see Fig. 10c, where a fixed array of eight smartphones is illustrated. A second array consisted of seven handheld smartphones. Both array geometries are illustrated in Fig. 10e. The

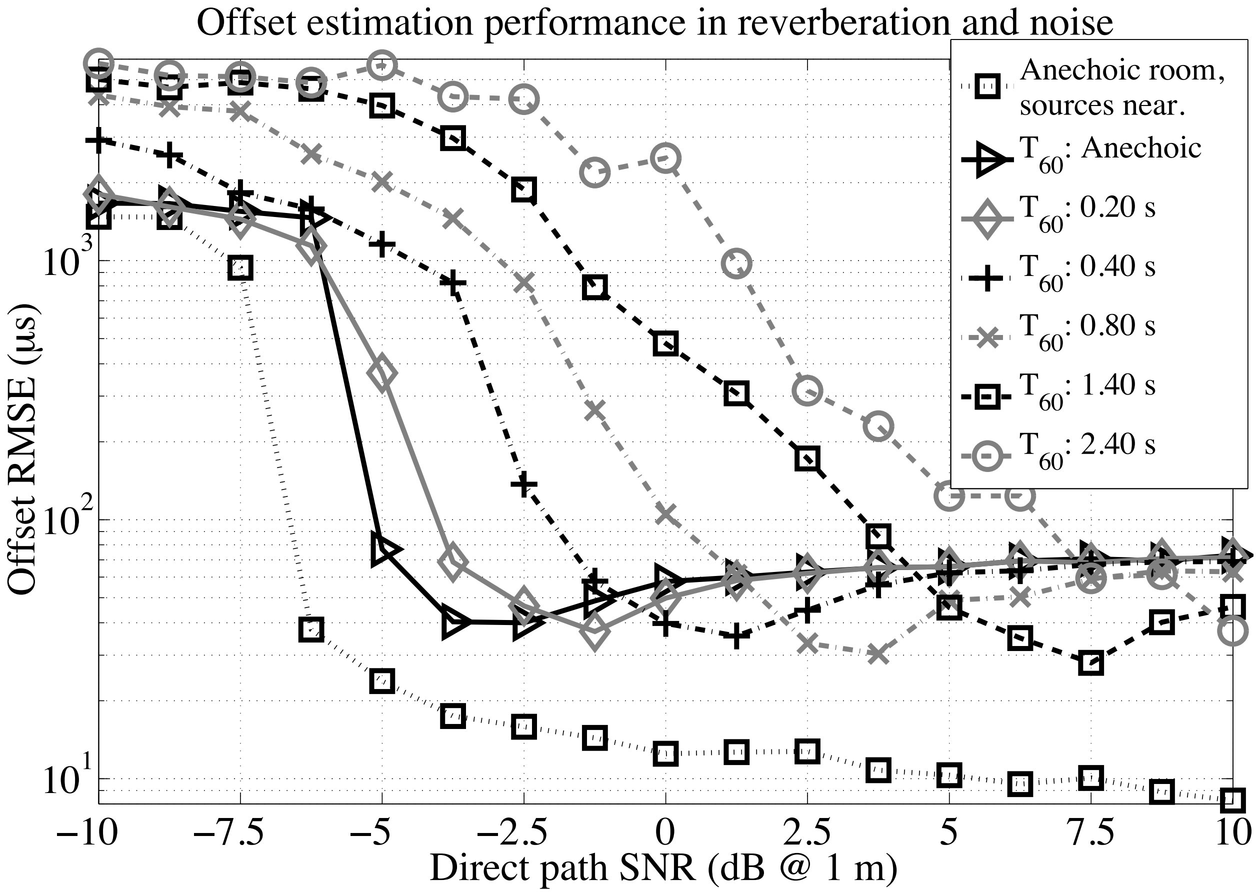


Fig. 9: Results of the proposed synchronization algorithm (refer to Fig. 8) using simulated reverberant speech.

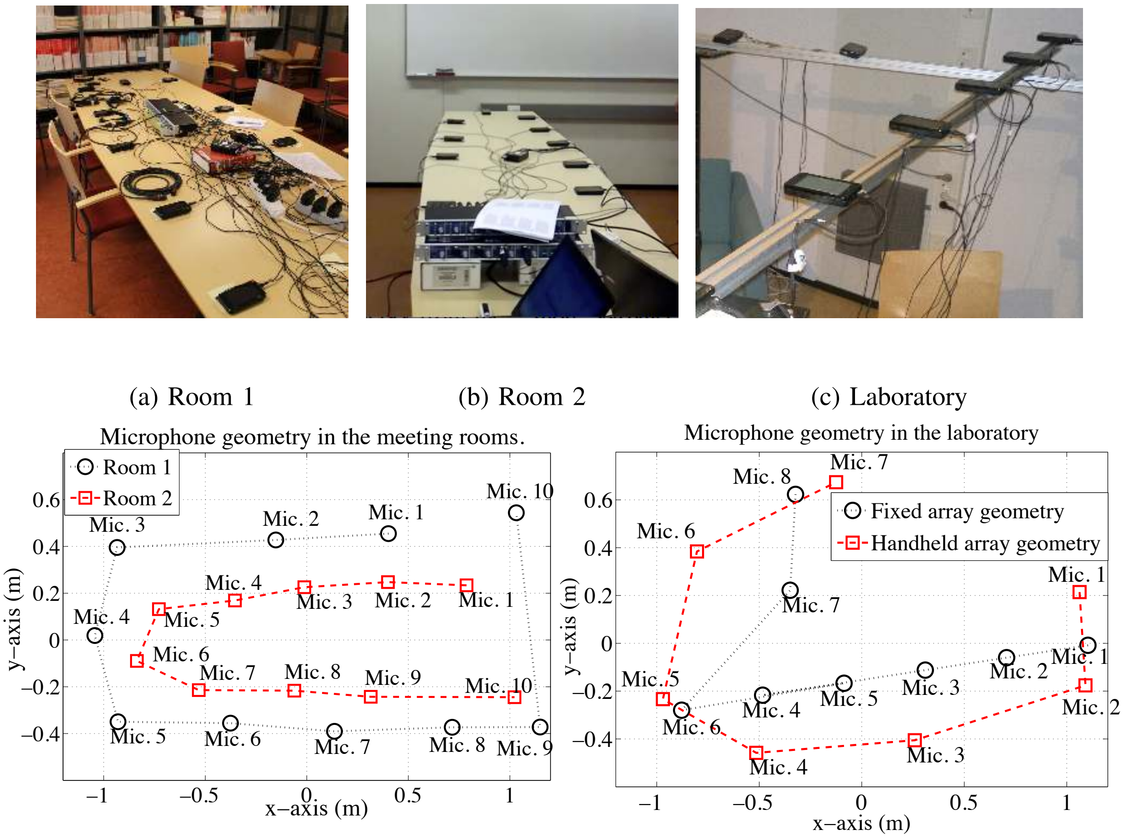
microphone positions are obtained by measuring the distances between microphones with a tape measure and then using multidimensional scaling to obtain a 2D geometry in Euclidean space [27][2].

### A. Use Cases

Three use cases are presented, based on the degree the geometry satisfies the requirement of the Corollary 1, i.e., are sources expected to be in the endfire directions.

*1) Recordings with geometries that satisfy requirements of Corollary 1:* Room 1 was used to capture recordings 1-7, Room 2 was used to capture recordings 8-10, and recordings 11-15 were captured in the laboratory with recording durations varying between 40 to 90 s. In recordings 1-10, four participants

[2] The distances between devices in the handheld case is estimated using [10].



(a) Room 1

(b) Room 2

(c) Laboratory

(d) The two meeting room arrays

(e) The laboratory geometry

Fig. 10: The recording setup is illustrated. Panels (a)–(b) depict the fixed arrays of the meeting rooms, and panel (c) depicts the fixed laboratory array. Panel (d) displays the corresponding centered microphone coordinates, and panel (e) displays the two microphone geometries in the laboratory.

surround the array and speak sentences towards the array from several different positions trying to excite endfire directions. In the laboratory recordings 11-14 one person walks behind every smartphone and reads a sentence. In recording 15, seven people hold the smartphones as if shooting a video and each person utters a sentence in turn. Endfire directions are assumed to be excited in these recordings, i.e., the geometry fulfills Corollary 1 requirements (11). This use case is expected to reveal the best achievable offset estimation performance in practical conditions.

*2) Recordings with geometries that only approximately satisfy requirements of Corollary 1:* Recordings 16-17 and 18-19 contain speech from moving talkers in the two meeting rooms, correspondingly. During the recordings participants walk around the table and read sentences. In this case, the sources are now more elevated from the table than in the previous case, where people were seated. Therefore, endfire directions are expected to be excited only approximately, and the geometry only partially meets the requirements (11) of Corollary 1. In recording 19 the four speakers are overlapping.

*3) Recordings with geometries that do not satisfy Corollary 1 requirements:* Recordings 20-29 contain speech from four static seated talkers in the two meeting rooms. This geometry does not meet the requirements (11) of Corollary 1. Even though the offset estimation is not expected to be accurate, this case is used to study if the proposed algorithm can provide improvement over the Naïve algorithm.

### *B. Measurement Equipment and Reference Offset Estimation*

Nokia N900 smartphones were used to capture audio at 48 kHz sampling rate and 16-bit accuracy. Additionally, electret condenser microphones (Sennheiser MKE 2 P-C) were attached to each smartphone as a reference microphone within 5 mm distance from the smartphone microphone inlet to approximate having concentric microphones. The reference microphone data was captured using a RME Fireface 800 and RME Octamic II AD-converters synchronized together with 24-bit floating point accuracy and 48 kHz sampling rate. All devices were connected to AC power.

The ground truth or reference offset between the audio streams captured by the reference microphone and the smartphone microphone is estimated in two stages. First, 128 sample length frames are used to calculate the signal energy from the smartphone and the corresponding reference microphone. The envelope curves are then aligned using GCC. In the second stage, the frame-aligned data is processed in frames of length $2^{14}$ samples, and GCC is again used to estimate the sample difference between the audio streams for each frame. The final sample offset was obtained by linear regression, where the estimated line slope indicates the amount of clock drift. The procedure is repeated for each channel separately. Finally, the offset of the first channel is subtracted from each channel to obtain relative offsets used in (21). The average absolute clock drift was estimated to be 0.26 ppm in the devices. This means that on average after 82 s of audio captured at 48 kHz one sample difference between tracks occurs. Due to its small size, the drift was omitted in the evaluation of the results.

## VIII. RESULTS

All recordings are processed with Hann windows of empirically set length 171 ms with 50 % overlap, using the algorithm presented in Fig. 8. The selection of VAD and sequential filtering parameters was discussed in Section V-D.

The proposed algorithm with quantile parameter $q$ estimation (algorithm of Fig. 8) is contrasted with the Naïve algorithm, that provides the baseline for offset estimation improvement (algorithm of Fig. 5).

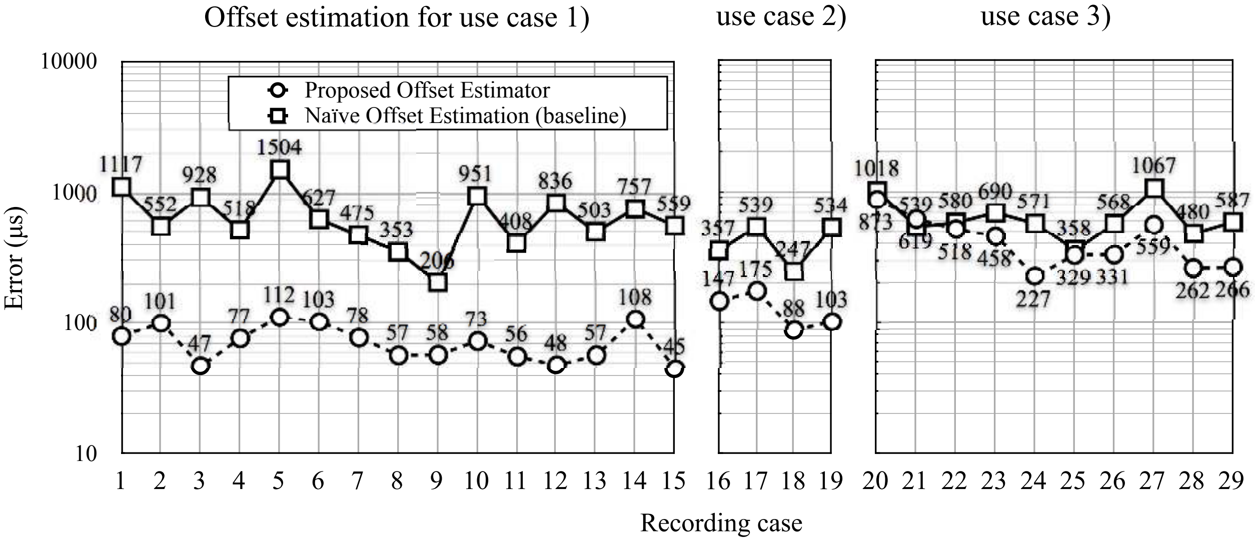


Fig. 11: Results for the examined three use cases of the offset estimation algorithms. The square markers represent baseline, and circles represent the proposed method (refer to Fig. 8). Refer to Table I for summary statistics of the use cases, and see Section VII-A for use case descriptions.

The offset estimation RMSE values ($\mu s$) for the three different use cases are presented in Fig. 11 in three separate panels. Table I summarizes the results using mean and standard deviation of the RMS errors of the three cases for both estimators.

The leftmost panel of Fig. 11 depicts the first use case, where the geometry agrees with Corollary 1 requirements (recordings 1–15). The proposed offset estimator has an average RMS error of 73 $\mu s$ corresponding to 3.5 samples at the used 48 kHz sampling rate, refer to Table I. The results are precise, and the baseline method shows ten times higher error. The null hypothesis that the baseline is as accurate as the proposed algorithm is tested with the paired Student's t-test. The resulting low $p$-value supports the rejection of this hypothesis.

In the uncontrolled geometry case with moving talkers the RMS error of the proposed estimator is roughly doubled, leading to average accuracy of 128 $\mu s$ corresponding to 6.2 samples. Since the endfire directions are excited by chance more variance and error are expected. Still the 6.2 sample average error is clearly more accurate than provided by the baseline (20.2 samples). The corresponding $p$-value of 0.020 also supports the significance of the difference. The overlapping of speakers in recording 19 did not affect the algorithm performance.

In the final use case, where the requirements of Corollary 1 are not met, in 9 out of 10 recordings the proposed algorithm has less RMS error than the baseline. Given a $p$-value of $0.005$ the difference can be considered statistically significant.

## IX. Discussion

If the microphone and source geometry are such that all endfire directions are excited, the proposed algorithm's offset estimation results are accurate in terms of RMS error. In the average offset error time of 73 $\mu s$ a sound wave travels roughly a 2.5 cm distance. In a more general case where endfire directions are not deliberately excited, but it is still reasonable to expect that sound will at some point arrive approximately from every endfire directions, the results are still significantly more accurate than the baseline. This is supported by the simulation results, which suggest that the CRLB can be reached when the number of randomly placed surrounding source positions increases. Finally, in the case where all endfire directions are not excited, the algorithm is still able to improve the baseline performance.

The error in sensor placement and synchronization is investigated for TDOA based closed-form localization in [8]. A 100 $\mu s$ error in a synchronization leads to approximately 12 cm source location error with a four microphone circular array with 1 m radius with the sources placed in the same location as the sensors. Considering the dimensions of typical sound sources such as mouth or loudspeaker, such an error is relatively small. Therefore, the proposed method is expected to be sufficiently accurate for localization applications.

Based on [9] a synchronization error of 100 samples at 8 kHz leads only to a drop of 0.5 dB in separation score (Signal-to-interference ratio) of an energy based ICA separation method. Our algorithm can reach an RMS error of 2 samples at 48 kHz that is clearly of sufficient accuracy for such energy based separation algorithms.

An issue not discussed in this article is clock drift of devices. It is known that blind source separation algorithms can deteriorate even with a 1–2 sample difference in sampling rates of the devices at 16 kHz [40]. This means that 3 to 6 sampling rate difference at 48 kHz lead to loss of separation quality during a sentence of speech. However, in contrast to [40] we used similar devices of the same manufacturer where on average one sample drift takes 82 s. Therefore, we did not face clock drift issues. By using more heterogeneous devices the effect of the clock drift may become more visible. However, this issue is left as a topic of future research.

## X. Conclusions

This research shows that it is relatively easy to passively estimate the temporal offset between different recordings made with devices that have small clock drifts. We used up to ten smartphones to demonstrate

TABLE I: The summary statistics of RMS offset estimation error using the Naïve baseline and the proposed approach, where $N$ is the number of data points. $p$-values of paired t-test of the algorithms are given. See Fig. 11 for detailed results.

| RMSE | Use case 1 $(N = 15)$ | | Use case 2 $(N = 4)$ | | Use case 3 $(N = 10)$ | |
|---|---|---|---|---|---|---|
| | $(\mu s)$ | (samples) | $(\mu s)$ | (samples) | $(\mu s)$ | (samples) |
| Naïve Alg. $(\mu \pm \sigma)$ | $686 \pm 335$ | $33.0 \pm 16.1$ | $419 \pm 143$ | $20.2 \pm 6.9$ | $646 \pm 226$ | $31.0 \pm 10.9$ |
| Proposed Alg. $(\mu \pm \sigma)$ | $73 \pm 23$ | $3.5 \pm 1.1$ | $128 \pm 40$ | $6.2 \pm 1.9$ | $444 \pm 203$ | $21.3 \pm 9.8$ |
| $p$-value (paired t-test) | $4.1 \cdot 10^{-6}$ | | 0.020 | | 0.0047 | |

that average temporal offset RMS error of 3.5 samples at 48 kHz sampling rate can be obtained in practical environments, where devices are surrounded by speakers in a geometric setting that fulfills the algorithm's requirements.

The method was also shown to achieve the Cramer-Rao lower bound in these conditions. The bound is achieved asymptotically as a function of number of surrounding randomly distributed sources. When the algorithm's assumptions are violated the synchronization cannot be expected to be accurate but it can still improve the baseline performance.

Simulation results show that the algorithm is able to provide accurate offset estimates even in highly reverberant environments with relatively low SNR.

Based on existing knowledge of the effect of synchronization error on sound source localization and blind source separation, these applications are expected to work in a satisfactory manner with the accuracy reached by the proposed method.

## APPENDIX A
## PROOF OF OFFSET ESTIMATION FOR FAR-FIELD SOURCES

Let two microphones $\mathbf{m}_i$ and $\mathbf{m}_j$ form a pair and let $\mathbf{r}$ be the pair's center point $\mathbf{r} = \frac{1}{2}(\mathbf{m}_i + \mathbf{m}_j)$. A sound emitted by a source is assumed to arrive as a plane wave[3] and its propagation direction is represented by vector $\mathbf{k} \in \mathbb{R}^3$. For convenience, let the length be $\|\mathbf{k}\| = c^{-1}$, where $c$ is speed of sound. The wavefront time of arrival at microphone $i$ with respect to center point $\mathbf{r}$ is [41, ch. 2]

$$\tau_i = \langle \mathbf{m}_i - \mathbf{r}, \mathbf{k} \rangle + \Delta_i \tag{22}$$

[3] The plane-wave assumption is made for each microphone pair separately.

where $\langle \cdot, \cdot \rangle$ is dot product, and $\Delta_i$ is the (stationary) sensor time-offset to global reference time. The TDOA is defined as

$$\tau_{ij} = \tau_i - \tau_j = \langle \mathbf{m}_i - \mathbf{m}_j, \mathbf{k} \rangle + y_{ij}, \tag{23}$$

where $y_{ij} = \Delta_i - \Delta_j$ is the pairwise offset. The propagation directions of wavefronts arriving from either of the two directions that are parallel to the microphone connecting axis $(\mathbf{m}_i - \mathbf{m}_j)$ can be written as [10]

$$\mathbf{k}(\beta) = \beta \frac{\mathbf{m}_i - \mathbf{m}_j}{\|\mathbf{m}_i - \mathbf{m}_j\|} c^{-1}, \beta = \{+1, -1\} \tag{24}$$

The TDOA for endfire directions is obtained by substituting the $\mathbf{k}(\beta)$ (24) as $\mathbf{k}$ in (23) [10]

$$\tau_{ij}(\beta) = \beta c^{-1} \|\mathbf{m}_i - \mathbf{m}_j\| + y_{ij}. \tag{25}$$

Note that since $\beta \in \{-1, +1\}$ for endfire sources the TDOA magnitude without the offset is the sound propagation time between the microphones and the sign corresponds to wavefront direction.

Using the definitions $\tau_{ij}^{\max} \triangleq \tau_{ij}(+1)$ and $\tau_{ij}^{\min} \triangleq \tau_{ij}(-1)$ the theorem (9) is proven by (10) for the far-field sources.

## Acknowledgments

This work was supported by the Finnish Academy project no. 138803 and Nokia Research Center. The first author wishes to thank M.Sc. Hanna Silen for her comments.

## References

[1] S. Vihavainen, S. Mate, L. Seppälä, F. Cricri, and I. D. Curcio, "We Want More: Human-Computer Collaboration in Mobile Social Video Remixing of Music Concerts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. ACM, 2011, pp. 287–296.

[2] P. Shrestha, M. Barbieri, H. Weda, and D. Sekulovski, "Synchronization of Multiple Camera Videos Using Audio-Visual Features," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 79 –92, jan. 2010.

[3] C. Lu and M. Mandal, "A Robust Technique for Motion-based Temporal Alignment of Video Sequences," *IEEE Trans. Multimedia*, vol. PP, no. 99, p. 1, 2012.

[4] F. Schweiger, G. Schroth, M. Eichhorn, A. Al-Nuaimi, B. Cizmeci, M. Fahrmair, and E. Steinbach, "Fully Automatic and Frame-accurate Video Synchronization using Bitrate Sequences," *IEEE Trans. Multimedia*, 2013.

[5] M. Goto and K. Hirata, "Recent Studies on Music Information Processing," *Acoustical Science and Technology*, vol. 25, no. 6, pp. 419–425, 2004.

[6] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*. John Wiley & Sons Ltd., 2009.

[7] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing, J. Benesty and W. Kellerman, Eds. Springer, 2008, vol. 1.

[8] W. van Herpen, S. Srinivasan, and P. Sommen, “Error Analysis on Source Localization in Ad-Hoc Wireless Microphone Networks,” in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2011.

[9] L. Zicheng, “Sound Source Separation with Distributed Microphone Arrays in the Presence of Clock Synchronization Errors,” in *Int. Workshop Acoustic Echo and Noise Control (IWAENC)*, 2008.

[10] P. Pertilä, M. Mieskolainen, and M. Hämäläinen, “Passive Self-Localization of Microphones Using Ambient Sounds,” in *EUSIPCO’12*, 2012.

[11] B. Sundararaman, U. Buy, and A. D. Kshemkalyani, “Clock Synchronization for Wireless Sensor Networks: A Survey,” *Ad Hoc Networks*, vol. 3, no. 3, pp. 281 – 323, 2005.

[12] I.-K. Rhee, J. Lee, J. Kim, E. Serpedin, and Y.-C. Wu, “Clock Synchronization in Wireless Sensor Networks: An Overview,” *Sensors*, vol. 9, no. 1, pp. 56–85, 2009.

[13] S. Wehr, I. Kozintsev, R. Lienhart, and W. Kellermann, “Synchronization of Acoustic Sensors for Distributed Ad-Hoc Audio Networks and its use for Blind Source Separation,” in *in Proc. IEEE Multimedia Software Engineering*. IEEE, 2004, pp. 18–25.

[14] J. Sachar, H. Silverman, and W. Patterson, “Position calibration of large-aperture microphone arrays,” in *ICASSP’02*, vol. 2. IEEE, 2002.

[15] S. Birchfield and A. Subramanya, “Microphone Array Position Calibration by Basis-Point Classical Multidimensional Scaling,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1025 – 1034, Sept. 2005.

[16] S. Thrun, “Affine Structure From Sound,” in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2005.

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

[18] T. Janson, C. Schindelhauer, and J. Wendeberg, “Self-localization Based on Ambient Signals,” in *In Algorithms for Sensor Systems*, ser. Lecture Notes in Computer Science. Springer, 2010, vol. 6451, pp. 176–188.

[19] 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.

[20] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, “BeepBeep: A High Accuracy Acoustic Ranging System using COTS Mobile Devices,” in *Sensys’07*, 2007.

[21] P. Pertilä, M. Mieskolainen, and M. Hämäläinen, “Closed-Form Self-Localization of Asynchronous Microphone Arrays,” in *HSCMA’11*, 2011, pp. 139 –144.

[22] M. Hennecke and G. Fink, “Towards Acoustic Self-Localization of Ad Hoc Smartphone Arrays,” in *HSCMA’11*, 2011, pp. 127 –132.

[23] H. Fan and C. Yan, “Asynchronous Differential TDOA for Sensor Self-Localization,” in *ICASSP*, 2007, pp. II–1109 –II–1112.

[24] B. C. Dalton and V. M. B. Jr., “Audio-Based Self-Localization for Ubiquitous Sensor Networks,” in *Proc. 118th AES Convention*, 2005.

[25] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, “Acoustic Source Localization with Distributed Asynchronous Microphone Networks,” *IEEE Trans. Audio, Speech, Language Process.*, vol. PP, no. 99, p. 1, 2012.

[26] 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, p. 666, 2008.

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

[28] V. Raykar, B. Yegnanarayana, S. Mahadeva Prasanna, and R. Duraiswami, “Speaker localization using excitation source information,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 751 – 761, 2005.

[29] N. Ono, H. Kohno, N. Ito, and S. Sagayama, “Blind Alignment of Asynchronously Recorded Signals for Distributed Microphone Array,” in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA ’09. IEEE Workshop on*, oct. 2009, pp. 161 –164.

[30] K. Hasegawa, N. Ono, S. Miyabe, and S. Sagayama, “Blind Estimation of Locations and Time Offsets for Distributed Recording Devices,” in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds. Springer Berlin Heidelberg, 2010, vol. 6365, pp. 57–64.

[31] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, “SLAM-based Online Calibration of Asynchronous Microphone Array for Robot Audition,” in *In Intelligent Robots and Systems (IROS)*, 2011, pp. 524–529.

[32] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, ser. Prentice Hall Signal Processing Series. Prentice Hall, 1998.

[33] V. Stahl, A. Fischer, and R. Bippus, “Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering,” in *In Proc. ICASSP’00*, vol. 3. IEEE, 2000, pp. 1875–1878.

[34] P. Loizou, *Speech Enhancement: Theory and Practice*, ser. Signal processing and communications. CRC Press, 2007.

[35] 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.

[36] B. Champagne, S. Bedard, and A. Stephenne, “Performance of Time-Delay Estimation in the Presence of Room Reverberation,” *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148 – 152, 1996.

[37] E. Weinstein and A. Weiss, “Fundamental Limitations in Passive Time Delay Estimation – Part II: Wide-Band Systems,” *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 31, no. 2, pp. 472 – 486, 1983.

[38] 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.

[39] H. Kuttruff, *Room Acoustics*, 5th ed. Spon Press, 2009.

[40] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, “On the Importance of Exact Synchronization for Distributed Audio Signal Processing,” in *ICASSP’03*, vol. 4, april 2003, pp. IV – 840–3 vol.4.

[41] L. J. Ziomek, *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*. CRC Press, 1995.

PLACE
PHOTO
HERE

**Pasi Pertilä** (M’10) received the M.Sc. and Ph.D. degrees from Tampere University of Technology (TUT), in 2003 and 2009, from the fields of information technology and signal processing, respectively. He works as a postdoctoral researcher in the Department of Signal Processing at TUT. His research interests include microphone array technologies for source tracking, speech enhancement and separation, and self-localization and synchronization of wireless microphone arrays. He serves as an officer in the IEEE Finland Section’s SP & CAS Chapter.

PLACE
PHOTO
HERE

**Matti S. Hämäläinen** (S'91, M'93) received the M.Sc. degree in information technology from Tampere University of Technology, in 1992. He joined Nokia Research Center in 1995 and has worked in different research and research management roles in the area of audio and multimedia technologies. Currently, he is research team leader at Nokia Research Center. His research interests include multimedia technologies, audio signal processing, microphone array technologies and spatial audio capture.

PLACE
PHOTO
HERE

**Mikael Mieskolainen** studied signal processing at Tampere University of Technology (TUT) and particle physics at ETH Zürich and University of Helsinki, and received the M.Sc. degree from TUT in 2013. He worked at Department of Signal Processing at TUT in between 2010 and 2012. Currently, he is a researcher at Helsinki Institute of Physics affiliated with CERN. His current research topics are diffractive hadron-hadron scattering processes at the LHC and mathematical methods for particle physics data analysis.