SELF-LOCALIZATION OF WIRELESS ACOUSTIC SENSORS IN MEETING ROOMS

Mikko Parviainen, Pasi Pertilä
Department of Signal Processing, Tampere, Finland
{mikko.p.parviainen, pasi.pertila}@tut.fi

Matti S. Hämäläinen
Media Technologies Laboratory
Nokia Research Center
Tampere, Finland
matti.s.hamalainen@nokia.com

ABSTRACT

This paper presents a passive acoustic self-localization and synchronization system, which estimates the positions of wireless acoustic sensors utilizing the signals emitted by the persons present in the same room. The system is designed to utilize common off-the-shelf devices such as mobile phones. Once devices are self-localized and synchronized, the system could be utilized by traditional array processing methods. The proposed calibration system is evaluated with real recordings from meeting scenarios. The proposed system builds on earlier work with the added contribution of this work is i) increasing the accuracy of positioning, and ii) introduction data-driven data association. The results show that improvement over the existing methods in all tested recordings with 10 smartphones.

1. INTRODUCTION

Traditional microphone array methods such as beamforming and source localization [1] require that microphone positions are known and that there is no temporal offsets between the captured signals, i.e., the signals are synchronized. The advancement of modern communication devices such as smartphones, tablets, and more recently wearable devices have created ubiquitous microphone arrays. Unfortunately, such microphone locations and their temporal offsets are generally not available in accurate enough form, which would allow the direct utilization of the traditional array processing methods.

The problem of simultaneously locating devices, estimating the temporal offsets, and locating external sources using only passive listening, can be principally solved by minimizing a global cost function that incorporates all the unknowns and the corresponding measurements [3]. However, this approach requires a good initial guess to avoid converging to local minima. In [5] a self-localization solution that estimates distances between devices from diffuse sound field is proposed. The positions of microphones are estimated using Multidimensional Scaling (MDS) [5].

Recent advances in self-localization [6] and temporal offset estimation [7] provide accurate initial guesses when the assumptions of the methods are met. Once the initial positions and temporal offsets are available, traditional source localization techniques [1] can be used to obtain initial locations. Since an error in sensor location can in some cases lead to double the error in source localization [8], the estimates of microphone positions and offsets should be further refined. Fortunately, the minimization of the global cost-function can be performed, once the captured data is assigned to its corresponding source. The assignment in itself is a separate research problem referred as data-association, which also can deal with detecting measurement errors caused by clutter and noise (see e.g. [9, Ch. 16]).

In this work, we propose a "divide and conquer" approach for the problem of microphone self-localization and in meeting room scenario, where devices are static on a table, and the speakers are seated at the table. The novelty of this work is the combination of the initial guess methods with a provided data-association technique. The performance of self-localization using actual smartphone recordings is contrasted to the initial estimates to demonstrate accuracy improvement.

2. FORMULATION

Let \( m_i \in \mathbb{R}^3 \) be the \( i \)th receiver position \( i \in \{1, \ldots, N\} \), with \( N \) microphones. In an anechoic room the signal \( m_i(t) \) can be modeled as a delayed source signal \( s_k(t) \) as

\[
m_i(t) = s_k(t - \tau_{ik}^s) + n_i(t),
\]

where \( t \) is time, \( k \in \{1, \ldots, K\} \) denotes source index with \( K \) sources, and \( \tau_{ik}^s \) is time of arrival (TOA) from source \( k \) to the \( i \)th microphone

\[
\tau_{ik}^s = c^{-1}||s_k - m_i|| + \delta_i,
\]

where \( \delta_i \) is unknown time offset, \( c \) is speed of sound, and \( s_k, m_i \in \mathbb{R}^3 \) are source and microphone positions. The time difference of arrival (TDOA) between microphone pair \( \{i,j\} \) for source \( k \) is

\[
\tau_{ik}^s \pm \tau_{jk}^s = c^{-1}||s_k - m_i|| - ||s_k - m_j|| + \delta_{ij},
\]

where pairwise time offset is \( \delta_{ij} \equiv \delta_i - \delta_j \). The vector of all TDOA values is denoted \( \tau = [\tau_{11}^s, \tau_{12}^s, \ldots, \tau_{K(K-1)/2}^s]^T \), and \( \tau \in \mathbb{R}^{K(N-1)} \). TDOA estimates can be obtained e.g. using correlation ([10]) between all \( P = N(N - 1)/2 \) unique microphone pairs for each source. The General Self-Localization Problem (GSLP) is solved by the following minimization problem [3]

\[
J(\hat{S}, \hat{M}, \hat{\delta}) = \min_{S, M, \delta} \sum_{i,j,k} \left(c^{-1}||s_k - m_i|| - ||s_k - m_j|| + \delta_{ij} - \tau_{ij,k}\right)^2,
\]

where the sum is over all \( K \) sources and \( P \) microphone pairs, \( S = [s_1, s_2, \ldots, s_K]^T \), \( M = [m_1, m_2, \ldots, m_N]^T \), \( \delta = [\delta_1, \delta_2, \ldots, \delta_N]^T \). Note that the result of the minimization is subject to arbitrary rotation, reflection, translation, and arbitrary common time offset. This leads to \( N_{s} = 3(N + K) + N - 7 \) number of unknown variables, which can not exceed the \( N_{s} = S(N - 1) \) independent measurements [3]. The degrees of freedom (DOF) is defined here as

*This work is funded by Nokia Research Center and Finnish Academy project no. 138803.
DOF = \( N_m - N_u \), where DOF \( \geq 0 \) should be forced. The redundant DOFs are removed by fixing a coordinate system \( m_1 = [0, 0, 0]^T \) (translation), \( m_2 = [m_{2x}, 0, 0]^T \) and \( m_3 = [m_{3x}, m_{3y}, 0]^T \). Furthermore, the clock offset of the first microphone is set to zero i.e. \( \delta_1 = 0 \).

3. THE SELF-LOCALIZATION SYSTEM

The proposed system (Figure 1) incorporates previously developed techniques for processing speech signal in meeting room scenario. These include TDOA estimation (A), temporal offset estimation (B), initial microphone position estimation (C), and sound source localization (D). These techniques combined with the proposed data association scheme (E) enable the iterative self-localization (F). The fundamental goal is to enhance initial microphone position and offset estimates provided by (B), and (C).

3.1. Time Difference of Arrival (TDOA) Estimation (A)

Time-delay estimation is one possible key to self-localization, since it is a function of all the unknown variables (2), and it can be directly measured from the sensors. TDOA estimates are utilized by all the subsystems, which is computationally efficient. The peak index value of the generalized cross-correlation with a weighting function is used to calculate the TDOA \( \tau_{ij} \) between the signals of microphone pair \( i, j \), refer to [10]. The PHAT weight function removes the amplitude information and in practice shapes the correlation function to make the peak of cross correlation function more prominent and robustness [11].

The TDOA estimates contain a significant amount of instances which do not originate from a speech source. Therefore a sequential filter is applied as follows. If a TDOA estimate in the current frame has changed over a threshold compared to the previous frame in any microphone pair, then the current frame is labeled as an outlier and is not processed by the system.

3.2. Offset Estimation (B)

The proposed method is designed for an ad hoc device network. To utilize the data collected by devices, a common time base needs to be established. This is done by offset estimation subsystem. The offset estimation is conducted using the method presented in [7]. The pairwise offsets \( \delta_{ij} \) are obtained as follows.

\[
\delta_{ij} = \frac{1}{2} (\tau_{ij}^{\text{max}} + \tau_{ij}^{\text{min}}),
\]

where \( \tau_{ij}^{\text{min}} \) and \( \tau_{ij}^{\text{max}} \) are the minimum and the maximum TDOA values for microphone pair \( i, j \). The proof of (5) is presented in [7]. The observation of the minimum and maximum time delays requires that during the recording signals are emitted from the line that connects each microphone pair \( (i, j) \) (see Figure 2). Due to one missing degrees of freedom, the method [7] produces microphone offset estimates \( \delta_i, i = 2, \ldots, N \) from the pairwise measurements (5), where the offset values are relative to the first microphone, which can be (arbitrarily) set to zero \( \delta_1 = 0 \).

3.3. Initial Microphone Position Estimation (C)

The purpose of the initial microphone position is to roughly estimate minimum and the maximum TDOA values for microphone pairs. The method presented in [6] is used. The fundamental idea is to is to estimate the pairwise distance using speed of sound \( c \)

\[
d_{ij} = \frac{c}{2} (\tau_{ij}^{\text{max}} - \tau_{ij}^{\text{min}}),
\]

From the distance matrix, consisting of all pairwise distances, the positions of the microphones in a relative coordinate system are obtained using MDS [12].

3.4. Source Localization (D)

Closed-form source localization techniques such as [13] and [14] are attracting due to computational efficiency, but involve a linearization of quadratic equations. The accuracy of closed-form may be sufficient enough for providing initialization for source variables in general self-localization problem (4) [8]. In testing the closed-form source localization methods presented [13] and [14], they turn out
to be suffering from inaccuracies in microphone positioning and offset estimation provided by subsystem (C) with the data used in this work, which is a similar finding to [8].

Source localization is thus done via iterative optimization by solving (4) for each source individually. Microphone position estimates \( \hat{m}_1, \ldots, \hat{m}_N \) are obtained from subsystem (B) and pairwise temporal offset estimates \( \delta_{12}, \ldots, \delta_{N-1,N} \) from subsystem (C). Pairwise offset values are subtracted from TDOA values to enable traditional source localization methods that assume perfectly synchronized microphones with known locations.

In optimization of the cost function (4) Matlab function \texttt{lsqnonlin} [15] is used with trust-region reflective algorithm. The number of iterations allowed is set to 5000. Termination tolerances for variables change and objective function change are set to \( 1 \times 10^{-12} \).

3.5. Data Association (E)

The purpose of data association is to determine to which sound source TDOA measurements belong. The data association (DA) scheme in the proposed system is based on assumption that sources are non-moving and only one source is active at a time. However, this kind of assumption is reasonable for instance in a meeting, in which it is common that people are sitting at a table and talking one at a time. A more advanced DA technique is required if the conditions are not met [9, Ch. 16]. The fundamental idea is to detect from TDOA measurements changes that result from the fact that active sources has changed.

There are a total of \( P = \binom{N}{2} / 2 \) TDOA measurements per time frame. Thus the input signal \( \tau \) to the DA subsystem is a \( P \times T \) matrix, where \( T \) is amount of frames used for TDOA estimation. Since multiple source positions can be mapped into the same TDOA value, all \( P \) pairs of TDOA values should be considered when trying to identify a sound source. We propose to use Principal Component Analysis (PCA) to obtain a reduced set of TDOA features. PCA seeks a projection in least squares sense that best presents the data to enable simple detection methods such as clustering or peak-picking [16].

PCA results in principal components \( P \) and \( S \) scores. The original data \( \tau \) is reconstructed as follows.

\[
\hat{\tau} = SP^T
\]

The operation of DA subsystem is as follows. (I) Perform PCA for the data. (II) Estimate probability density of the reduced data (i.e. all data up to current frame). (III) Detect peaks from density estimate. (IV) Find time indices that correspond to the density maxima. (V) Use the indices to find appropriate TDOA values.

Figure 3 illustrates the input data, reduced data and its use for identifying sources. Figure 3a presents input data \( \tau \) i.e. the original 45 dimensional TDOA data over a recording. Each dimension is presented by a different color. Figure 3b presents the data mapped from 45 to 2 over the recording. The blue line corresponds to the first principal component and the green line corresponds to the second principal component. Approximately constant segments e.g. from frame 125 to 175 represent time periods when a particular source is active. Transitions to another static segment corresponds to source change. Figure 3c presents probability density estimation of reduced data corresponding to the first two PCAs, where peaks correspond to sound sources. Comparing the panels 3c and 3b one can observe that peaks occur in probability density estimate at the same locations where there are a lot of data points in Figure 3b.

By detecting peaks from probability density estimate Figure 3c and searching for corresponding time instants from TDOA, one can label frames by a unique source ID number. The resulting TDOA labeling using the proposed DA is illustrated in Figure 3d.

3.5.1. Iterative Self-localization (F)

The iterative self-localization solves (4) via optimization. In the used data the number of microphones is \( N = 10 \) and the number of sources is \( K = 4 \). Taking into account the redundant degrees of freedom, the number of unknown variables is 45. Using the information provided by subsystems (B)-(E), it is possible to initialize the optimization problem (4) with good values and thus increase the probability of convergence to a reasonable solution.

The subsystem (F) outputs microphone position estimates \( \hat{m}_1, \ldots, \hat{m}_N \) as well as source position estimates \( \hat{s}_1, \ldots, \hat{s}_K \) of which the former is in the primary interest of this system.

4. DATA DESCRIPTION

Data is recorded using ten Nokia N900 mobile handsets running Maemo operating system. Each device records the data using its microphone with an added Sennheiser MKE 2P-C microphone attached near the microphone inlet for reference purposes. Sampling rate is 48000 Hz and bit depth is 16 bits. The analysis window in TDOA estimation (A) is 8192 samples (approximately 171 ms) with 50 % overlap.
The recordings were made in meeting rooms. The scenarios imitate a meeting of four participants. The participants are seated around a table and N900s are placed on the table in front of meeting participants. In the actual recording, each participant utters a sentence after which another participant utters another sentence etc. The length of each recording is approximately 1 minute 30 s. In all recordings except Recording ID 3 participants are placed approximately at the same positions with respect to devices. However, they are slightly altering their posture from one recording to another. Figure 4a illustrates the scenario. In Recording ID 3 one participant is sitting at the end of the table as illustrated in Figure 4b. Position of devices in each room are same.

The reference microphone positions for evaluation were obtained with a tape measure.

5. RESULTS

The presentation of results focuses on positioning accuracy. We stress that the proposed data association method is novel and has an essential role in the operation of the system.

The results obtained with the proposed system are presented as error to the ground truth coordinates. The error is defined as

$$e = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{D} \sum_{d=1}^{D} (m_i - \hat{m}_i)^2},$$

where $N$ is the number of microphones, $D$ is the dimension (here $D = 3$), $m_i$ is the location estimate and $\hat{m}_i$ is the ground truth position of microphone $i$. Table 1 presents the results obtained in six real recordings. The $e$ is the error after initial microphone positioning (B) and $e$ is the error after iterative self-localization (F). The coordinate estimates are rotated, translated and reflected to match before analyzing the distance in (8).

Comparing the error using paired t-test between initial microphone positioning (B) and iterative self-localization, it can be seen that the results were improved in a statistically significant way ($p<0.05$).

<table>
<thead>
<tr>
<th>Recording ID</th>
<th>$e_{initial}$ [m]</th>
<th>$e_{enhanced}$ [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.167</td>
<td>0.131</td>
</tr>
<tr>
<td>2</td>
<td>0.142</td>
<td>0.128</td>
</tr>
<tr>
<td>3</td>
<td>0.112</td>
<td>0.072</td>
</tr>
<tr>
<td>4</td>
<td>0.161</td>
<td>0.147</td>
</tr>
<tr>
<td>5</td>
<td>0.132</td>
<td>0.127</td>
</tr>
<tr>
<td>6</td>
<td>0.129</td>
<td>0.123</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

A self-localization system utilizing an ad hoc device network was presented. The system is able to estimate positions of devices from acoustic measurement. For instance, in a meeting scenario, the participants can use their mobile devices to establish a network. Using the proposed system the network can be self-localized and thus be used e.g. speaker localization, and annotation. All of the mentioned applications of the can be used for enhanced teleconferencing experience. The system was tested with data recorded in two meeting rooms. The recordings imitate a meeting scenario; participants are seated at a table and one person is speaking at a time. The proposed system achieves root-mean-square device position error of 7 – 15 cm.

7. REFERENCES


