Applied Acoustics 117 (2017) 76–85

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Self-localization of dynamic user-worn microphones from observed speech

Mikko Parviainen *, Pasi Pertilä [1]

Department of Signal Processing, Tampere University of Technology (TUT), FI-33101 Tampere, Finland

ARTICLE INFO

Article history:
Received 21 April 2016
Received in revised form 11 October 2016
Accepted 19 October 2016
Available online 9 November 2016

Keywords:
Self-localization
Ad hoc networks
Microphone arrays
Acoustic measurements
Kalman filtering
Data association

ABSTRACT

The increase of mobile devices and most recently wearables has raised the interest to utilize their sensors for various applications such as indoor localization. We present the first acoustic self-localization scheme that is passive, and is capable of operating when sensors are moving, and possibly unsynchronized. As a result, the relative microphone positions are obtained and therefore an ad hoc microphone array has been established. The proposed system takes advantage of the knowledge that a device is worn by its user e.g. attached to his/her clothing. A user here acts as a sound source and the sensor is the user-worn microphone. Such an entity is referred to as a node. Node-related spatial information is obtained from Time Difference of Arrival (TDOA) estimated from audio captured by the nodes. Kalman filtering is used for node tracking and prediction of spatial information during periods of node silence. Finally, the node positions are recovered using multidimensional scaling (MDS). The only information required by the proposed system is observations of sounds produced by the nodes such as speech to localize the moving nodes. The general framework for acoustic self-localization is presented followed by an implementation to demonstrate the concept. Real data collected by off-the-shelf equipment is used to evaluate the positioning accuracy of nodes in contrast to image based method. The presented system achieves an accuracy of approximately 10 cm in an acoustic laboratory.

1. Introduction

Self-localization is one of the enabling technologies in acoustic sensor networks. The self-localization means that the physical locations of the nodes are determined automatically. This enables fast deployment of such a network for spatial applications e.g. sound source localization via TDOA [1–4] and audio enhancement via spatial filtering techniques, such as beamforming [5], which traditionally rely on node geometry and temporal synchronization of microphones. Higher-level applications that can utilize the self-localization as underlying technology include automatic meeting transcriptions [6] and providing aid for hearing impaired persons by signal enhancement [7,8].

The increase of smart technology embedded into mobile phones, tablets, wrist watches, fitness bands, apparel, and jewellery has created a need for self-localization of the networks created by the sensors of the devices. Once the sensors are self-localized, they can be used for various tasks including the ones mentioned above. There are many challenges in taking such ad hoc sensor networks to use. The devices in general are different, and the quality of microphones analog-to-digital converters may vary significantly.[2] Furthermore, there may be unpredictable processing delays in audio path.

To be transparent and easy to adapt for many applications, and to be equipment agnostic, the self-localization must take place without extra hardware components or the use of intrusive signals. Furthermore, the most versatile form of self-localization and synchronization is applicable even after the capture event. This basically means using the environmental sounds in self-localization from unsynchronized audio streams. So far, this level of performance has been achieved in a scenario with static device [9,10]. Furthermore, in the user-worn microphone scenario, the self-localization must take the node motion into account, and therefore continuous self-localization is needed.

In this article we present an acoustic self-localization method for the dynamic sensor scenario in 3D space where the nodes of the acoustic sensor network are continuously changing their places. The node positions are calculated from the speech signal

* Corresponding author.
E-mail addresses: mikko.p.parviainen@tut.fi (M. Parviainen), pasi.pertila@tut.fi (P. Pertilä).
[1] Co-author.

[2] In this work homogenous hardware is used and it is acknowledged that future work should include research with heterogeneous hardware.

http://dx.doi.org/10.1016/j.apacoust.2016.10.019
0003-682X/

produced by the nodes themselves. Each node contains a microphone m and a source s (see Fig. 2).

The proposed method extends the previous work of [10,11] by allowing the nodes to be in motion while estimating their position from the audio produced by the nodes themselves. The data streams recorded by the nodes are unsynchronized, which is completely different from using wireless microphones that can utilize radio frequencies for side channel synchronization. The real data recordings using off-the-shelf hardware are used to evaluate the proposed system. The evaluation is made by comparing the estimated node paths to the reference node paths obtained from an implemented multiview camera setup.

This article is organized as follows. Section 2 reviews background of passive acoustic self-localization. Section 3 presents the theory of the proposed system. Section 4 describes how the theoretical presentation is implemented. In Section 5 the measurement procedure is described. Section 6 presents the performance evaluation procedure. In Section 7 clock drift analysis of the used equipment is presented. Section 8 presents real data performance of the proposed system. Section 9 discussion about the results and further development of the system is provided. Section 10 concludes the article.

2. Background

The general self-localization problem of acoustic sensor networks is stated as solving the positions of the nodes of the network. The network in general consists of sensors (microphones) and sound sources. Usually, the sensor positions are of interest in self-localization, but once they are obtained, sound sources can be localized if desired using e.g. multilateration. The general self-localization problem is solved by the following minimization problem (see [12])

$$J(\widehat{\mathbf{S}},\widehat{\mathbf{M}},\hat{\boldsymbol{\alpha}}) = \underset{\mathbf{S},\mathbf{M},\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{\forall\{i,j,k\}} \Big(c^{-1}(\|\mathbf{s}_k - \mathbf{m}_i\| - \|\mathbf{s}_k - \mathbf{m}_j\|) + \alpha_i - \alpha_j - \tau_{i,j}^k \Big)^2, \quad (1)$$

where the sum is over all $k = 1,\ldots,K$ sound sources and $\frac{N(N-1)}{2}$ unique microphone pairs (i,j). $\mathbf{S} = [\mathbf{s}_1,\ldots,\mathbf{s}_K]^{\mathrm{T}}$ and $\mathbf{M} = [\mathbf{m}_1,\ldots,\mathbf{m}_N]^{\mathrm{T}}$ are matrices whose columns are the Cartesian coordinates of sensor and microphone positions, respectively. $\boldsymbol{\alpha} = [\alpha_1,\alpha_2,\ldots,\alpha_N]^{\mathrm{T}}$ are temporal offsets of the sensors, and c is the speed of sound. $\tau_{i,j}^k$ is TDOA measured by sensor pair i,j from sound source k. The global minimum of (1) can be interpreted as correct positions of sound source positions, microphone positions, and temporal offsets of the captured audio.

Especially, with increasing number of sensors and sound sources (1) becomes an optimization problem plagued by local minima. Furthermore, any solution of (1) is subject to transformations that preserve distances between two points, such as translation, rotation, and reflection. In 3D space, the set of such transformations is referred to as the Euclidean group 3, E(3) [13]. This means that even the global optimum of (1) may differ from physical ground truth node coordinates.

In [14] a self-localization method based on time-of-flight and time-difference-of-flight is presented. Multidimensional Scaling (MDS) [15] is used to initialize the optimization problem (similar to (1)). The method also estimates temporal offsets for each node. The method assumes that nodes have capability to emit and receive sounds unique to each node and estimates pairwise distance matrix (corresponding to τ above).

In [16] a Time of Arrival (TOA) based self-localization method is presented. TOA is estimated from sounds naturally occurring in the environment. The synchronization of the receivers is crucial for the method and therefore it has a side-channel and infrastructure for it.

In [17] is a matrix factorization method and it attacks the general self-localization problem by dividing it into a sequence of simpler problems to avoid getting stuck to local minima. The general self-localization is eased by assuming that sound events occur in far field with respect to receivers, which enables to the simplification of the optimization problem. The resulting constraint of the simplified optimization problem is used to obtain the positions of the receivers. An extension of [17] is presented in [18] which takes measurement uncertainty into account. In near field, a rank-5 factorization method is needed [19] which requires at least ten microphones and four sources or vice versa. The methods [17–19] expect synchronous audio streams.

A method presented in [20] is directed to microphone array self-localization or *calibration* in diffuse soundfield. An extension to [20] is presented in [21], which uses multiple arrays and sound source localization to estimate relative rotation and translation of an array pair. Both methods are designed for relatively small intra-sensor distances of approximately 20 cm or smaller.

In [22] a method for ad hoc sensors is presented. The method is able to estimate the relative smartphone positions from measured TDOA. Pairwise sensor distances are estimated and MDS is performed to obtain the initial positions to an optimization problem similar to (1). Furthermore, four of all the variables in the optimization problem are fixed to establish a coordinate system (2D case). The method requires known calibration signals, which are audible and in frequency range from 5 kHz to 16 kHz. Another system using active calibrations signals is presented in [9]. The system performs direction of arrival estimation and distance estimation for self-localization.

In [23] a self-localization method of moving receiver based on TDOA estimation from ultrasound is presented. Using frequencies outside human hearing is attractive due to unobtrusiveness. The drawback of the system is requirement of an infrastructure of ultrasonic transmitters, their careful placement in a room, and side channel for data association.

3. Theory

The general idea of an acoustic self-localization system in user-worn devices scenario is presented in Fig. 1. Like in [10,11] the fundamental idea is that the system estimates pairwise distances between all the nodes in the network. From pairwise distance matrix relative coordinates can be estimated by finding node geometry in the Euclidean space that fulfills the restrictions of the distance matrix. The novelty is in proposing a tracking of the distance matrix, which allows self-localization of moving nodes continuously in contrast to [10,11]. The theoretical presentation of each subsystem illustrated in Fig. 1 is given in this section.

3.1. Signal model

Let $\mathbf{m}_i \in \mathbb{R}^3$ be the ith node position and $i \in 1,\ldots,N$. In an anechoic room the signal $m_i(t)$ can be modeled as a delayed source signal $s_k(t)$ as

$$m_i(t) = s_k(t - \Delta_i^k) + n_i(t), \quad (2)$$

where t is time, $k \in [1,\ldots,N]$ denotes active node index with N nodes, $n_i(t)$ is noise component, and Δ_i^k is TOA from active node k to the ith node

$$\Delta_i^k = c^{-1}\|\mathbf{s}_k - \mathbf{m}_i\| + \alpha_i, \quad (3)$$