Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Exposé for Master's Thesis

Sound Source Localization (and more TBD) using the Azure Kinect's Microphonearray on a Robot

Departmet of Informatics
MIN Faculty
Universität Hamburg

Hamburg, January 20, 2023

Roland Fredenhagen

dev@modprog.de
M.Sc. Informatics
Matriculation number: 7031533

First Reviewer:
Second Reviewer:
Supervisor:

1 Research Question

TAMS has a robot equipped with an Azure Kinect providing not only a visual data, but also multichannel audio from a 7 microphone array [Mic23]. This spacial audio data should be used to perform sound source localization (SSL) and separation as well as reduction of ego and environmental noise.

2 Existing Work

2.1 Sound Source Localization (SSL)

A very extensive collection of algorithms and their performance in different scenarios is provided by Evers et al. [Eve+20] with the LOCATA challenge comparing different algorithms in situations very similar to the proposed usage in the research question (chapter 1). Containing scenarios with moving sensors and sources on top of background noise and distortions such as reverberations. Their work also contains tests with multiple audio sources.

The microphone setups matching ours most closely is the Robot head though theirs consists of a total of 12 microphones that are placed spherical opposed to the planar circle of the Kinect and the DICIT which has a two-dimensional configuration, but whose microphones are placed a lot further apart with a total range of almost 2 meters.

The best performing algorithm was by Madmoni et al. [Mad+18] using the Robot head microphone array was a combination of the Direct-Path Dominance Test [NR14] and the MUSIC-Algorithm [Sch86], though it was only tested with the static sound sources, and they assumed a priori knowledge of the number of sound sources.

The only submitted algorithm applied to all scenarios using the Robot Head was by Li et al. [Li+18]. They used three modules to provide both localization and tracking.

- A recursive direct-path relative transfer function (DP-RTF) estimation module based on the estimation of the convolutive transfer function proposed by Li et al. [Li+16].
- An online multiple-speaker localization module assigning DP-RTF features to sources adopting a complex Gaussian mixture model.
- A multiple-speaker tracking module using a variational expectation maximization algorithm.

This allows their algorithm to not only locate static sound sources but also allows scenarios with either the sound sources, the microphone array or both moving.

Another comprehensive set of SSL methods was collected by Grumiaux et al. [Gru+22] though this one was focused solely on deep learning approaches. Two of the most capable methods were proposed by Pinto, Bauerheim, and Parisot-Dupuis [PBP21] and Hammer et al. [Ham+21], with both providing support for multiple audio sources, but only the latter being able to track moving sources.

2.2 Noise Reduction

The specific use case of suppressing both ego-noise and environmental noise is investigated in [Fan+21] and [Fan+] the latter even using a 4 microphone array similar to our setup. They show that combining a pretrained algorithm for ego-noise suppression and an adaptive part able to better reduce environmental noise to a partially adaptive scheme, perform better than both fixed and adaptive systems on their own.

3 Thesis Content

The main focus of the thesis is the localization of audio sources, building on top of that dynamic tracking and source separation will be applied as well as the suppression of background noises, depending on the complexity and difficulties of implementing the different algorithms. On top of developing the systems to do localization and noise suppression the produced data should be provided for usage in the robot-operating-system (ROS).

As the proposed algorithms in chapter 2 are not completely available as executable code, they need to be implemented first to be able to run and test them. The final sound source localization software for running on ROS should be provided as Open Source software after thesis completion.

3.1 Localization

As three-dimensional localization is difficult to achieve using a static microphone array [Eve+20] the localization efforts will focus on the direction of arrival (DoA). For this the most promising algorithms proposed in section 2.1 will be compared when applied to the real world usage with the Kinect input and our robots noise. The data produced this way e.g. an angle pair then can be provided as a ROS-topic. For tracking these could be extended with a unique id to identify different moving sources.

3.2 Noise Reduction

For noise reduction the proposed scheme by Fang et al. [Fan+] will be applied to suppress both ego and environmental noise and provide clean audio for later consumption e.g. by voice recognition software.

4 | Time Plan

January	Registration of Master's Thesis
February	Implementation of the proposed algorithms
March	Developing test setup for algorithms on robot and comparing performance
April - May	Investigate feasibility of implementing noise reduction on top of localization
June	Finalizing the thesis
July	Submission

Bibliography

[Eve+20] Christine Evers et al. The LOCATA challenge: Acoustic source localization and tracking. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020), pp. 1620–1643.

[Fan+] Huajian Fang et al. Partially Adaptive Multichannel Joint Reduction of Ego-noise and Environmental Noise. In.

[Fan+21] Huajian Fang et al. Joint Reduction of Ego-noise and Environmental Noise with a Partially-adaptive Dictionary. In: Speech Communication; 14th ITG Conference. VDE. 2021, pp. 1–5.

[Gru+22] Pierre-Amaury Grumiaux et al. A survey of sound source localization with deep learning methods. In: The Journal of the Acoustical Society of America 152.1 (2022), pp. 107–151.

[Ham+21] Hodaya Hammer et al. Dynamically localizing multiple speakers based on the time-frequency domain. In: EURASIP Journal on Audio, Speech, and Music Processing 2021.1 (2021), pp. 1–10.

[Li+16] Xiaofei Li et al. Estimation of the direct-path relative transfer function for supervised sound-source localization. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.11 (2016), pp. 2171–2186.

[Li+18] Xiaofei Li et al. A cascaded multiple-speaker localization and tracking system. In: arXiv preprint arXiv:1812.04417 (2018).

[Mad+18] Lior Madmoni et al. Description of algorithms for Ben-Gurion University submission to the LOCATA challenge. In: arXiv preprint arXiv:1812.04942 (2018).

[Mic23] Microsoft. Azure Kinect DK. 2023. url: https://azure.microsoft.com/en-us/products/kinect-dk/ (visited on 01/05/2023).

[NR14] Or Nadiri and Boaz Rafaely. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 22.10 (2014), pp. 1494–1505.

[PBP21] Wagner Gonçalves Pinto, Michaël Bauerheim, and Hélène Parisot-Dupuis. Deconvoluting acoustic beamforming maps with a deep neural network. In: (2021).

[Sch86] Ralph Schmidt. Multiple emitter location and signal parameter estimation. In: IEEE transactions on antennas and propagation 34.3 (1986), pp. 276–280.