INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

Advanced semi-blind speaker extraction and tracking implemented in experimental device with revolving dense microphone array

J. Čmejla, T. Kounovský, J. Janský, J. Málek, M. Rozkovec, and Z. Koldovský

Faculty of Mechatronics, Informatics, and Interdisciplinary Studies
Technical University of Liberec, Czech Republic
jaroslav.cmejla@tul.cz

Abstract

We present a new device for speaker extraction and physical tracking and demonstrate its use in real conditions. The device is equipped with a dense planar array consisting of 64 microphones mounted on a rotating platform. State-of-the-art blind source extraction algorithms controlled by x-vector piloting are used to extract the desired speaker, which is being tracked by the rotating microphone array. The audience will experience the functionality of the device and the potential of the blind algorithms to extract the speaker from multi-source noisy recordings in a live situation.

Index Terms: acoustic tracking, blind source extraction, piloting

1. Description

The subject of our presentation is an experimental device shown in Fig. 1 when it is deployed on-site in a speaker extraction problem. The device consists of a revolving microphone array, a step motor, and electronic circuits for control, signal sampling, and transmission. It is connected to a computer running MATLAB, where advanced methods for semi-blind speaker extraction are processing signals online. The device output yields an extracted signal of the desired speaker.

The methods are interesting from a scientific perspective as they are performing speaker extraction based on a state-of-the-art blind source extraction (BSE) mixing model. The user will thus experience the potential of BSE in the speaker extraction problem.

The convergence of BSE to the desired speaker is partially controlled based on speaker identification using x-vectors. The speaker ID can be trained on-site. The method can then be switched to extract and track any speaker for whom a training utterance is available. The extraction performance is aided by steering the array physically towards the speaker in order to maximize the signal gain.

The motivation behind this device and its demonstration at Interspeech is threefold.

1. The dense microphone array provides a way to deal with omnidirectional noise sources that cannot be subtracted by spatial filtering; they can only be attenuated, a task in which a large number of microphones is helpful. At the same time, the array preserves its small size and ability to subtract directional sources such as other interfering speakers.
2. Turning the array towards the desired speaker can improve the signal gain and, at the same time, support the suppression of the other sources, especially when the microphones and the array construction are directional.
3. The transmission of signals using a standard communication interface to MATLAB provides programmers with a powerful tool for developing new algorithms. In particular, we focus on implementing state-of-the-art methods for blind source (speaker) extraction to demonstrate their potential in unknown situations without prior learning. However, other speaker enhancement methods can be implemented.

2. Device

The tracking device can be separated into two parts: the microphone array board and the rotary platform. The microphone array board is equipped with 64 MEMS microphones and an FPGA controller. The microphones are arranged into an 8×8 vertical planar grid with 1 cm equidistant spacing. The FPGA controller deals with the synchronized sampling of signals. The samples are batched, serialized, and sent through Ethernet using the standard Transmission Control Protocol (TCP).

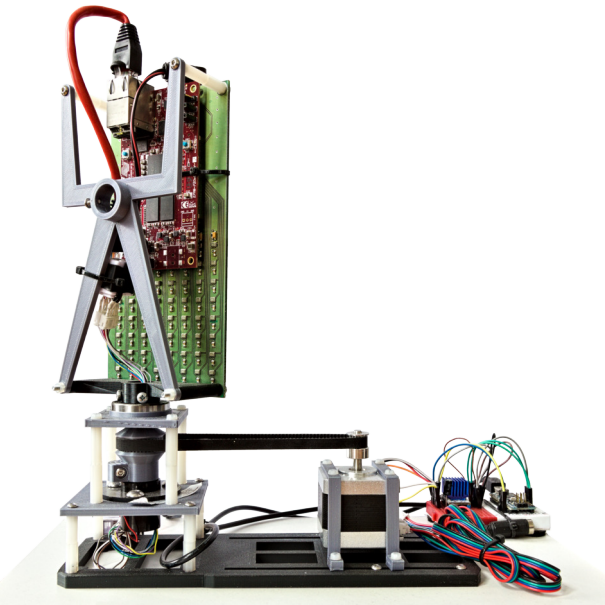

Figure 1: *64 microphone array board with a rotary platform.*

The rotary platform has been designed for unlimited rotation along the vertical axis. The construction parts of the platform were designed for and printed on a regular FDM 3D printer. The electronics parts of the rotary system consist of a stepper motor and a timing belt that drives the rotating platform. Unlimited smooth rotation of the platform is achieved by the use

of a slip-ring and bearings. The stepper motor is connected to a dedicated Arduino Nano microcontroller by a standalone Trinamic driver. The actual rotation commands are sent to Arduino using a serial interface.

3. Algorithm

The implemented algorithm is designed to extract one desired speaker of interest (SOI) and, additionally, to track the speaker by rotating the platform in the estimated direction. The source extraction itself is based on recent BSE algorithms, where the algorithm can be selected by the user: CSV-AuxIVE [1], QuickIVE [2], or FastDIVA [3]. These algorithms employ an advanced mixing model (CSV - Constant Separating Vector), which is a semi-time-variant linear mixing model that allows for movements of the SOI within the current data context [3, 4]. The algorithms are deployed in a standard batch-online processing regime (with an overlap between the batches).

An inherent problem of BSE is the uncertainty of the source (speaker) that is being extracted. To ensure that the algorithm converges to the desired speaker, the piloted version of CSV-AuxIVE is used in cooperation with an X-vector speaker ID system. This system labels frames where the desired speaker is dominant. It is composed of a time-delayed neural network for extracting speaker embeddings and a PLDA classifier that compares the embeddings to a previously obtained enrollment set [5, 6]. The labeling is used for generation of a pilot signal that influences the convergence of the algorithms.

4. Software

Device control software (screenshot shown in Fig. 2) is implemented in MATLAB. Currently, the software handles the following tasks:

- communication with the microphone board controller,
- audio data acquisition from the microphone board,
- online extraction processing and piloting,
- sending of commands to the motor controller.

The interface enables the user to select the speaker that is being targeted. This functionality requires a prepared enrollment set of speaker's utterances. For easy comparison, the user can switch between the noisy mixture as it is recorded by the 1st microphone and the output of the extraction algorithm.

The tracking switch enables commands to be sent to the motor controller in order to rotate the platform towards the SOI.

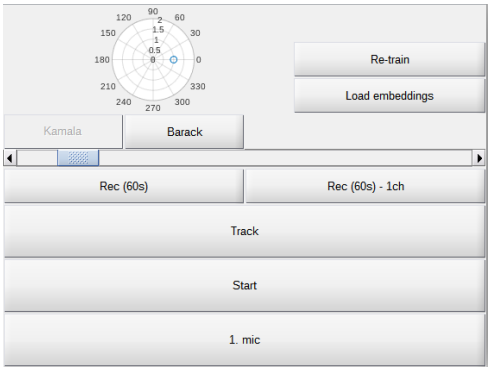

Figure 2: *Device control software interface (Matlab GUI).*

The current estimated speaker position relative to the front of the microphone array is shown in the polar plot.

5. Video

The video that we are providing as the part of our show&tell proposal shows a situation with two speakers situated in an office room at an approximate distance of 1 m (from the device). A female speaker is simulated by a loudspeaker attached to a rotating arm. Another loudspeaker simulates a male speaker in a static position.

In the video, the scene is shown along with a screencap of the software. In addition to control elements, a source direction estimation is included in a polar graph showing the estimated angle of the targeted speaker.

The video shows the scene while the audio contains the output from the software. During the footage, the operator switches between the 1st microphone and the output of the extraction algorithm. The operator also demonstrates the tracking ability of the device by turning on the tracking switch. When the tracking is enabled, the device turns in the direction of the targeted source. The extraction performance gets slightly worse just after the device rotates the array, which is caused by the fact that, at that moment, the extraction algorithm is not yet fully adapted to the new relative positions (a challenging problem to be solved). Nevertheless, it recovers quickly. The online tracking ability is demonstrated when the operator moves the loudspeaker slowly around. The device continues correctly tracking the source. After, the target source is switched to the male speaker. The algorithm converges to the new speaker in a manner of seconds, and the device turns to the new target.

6. References

[1] J. Janský, Z. Koldovský, J. Málek, T. Kounovský, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *arXiv 2002.12619*, 2020.

[2] Z. Koldovský, V. Kautský, T. Kounovský, and J. Čmejla, "Algorithm for independent vector extraction based on semi-time-variant mixing model," *arXiv 1910.10242*, 2021.

[3] Z. Koldovský, V. Kautský, P. Tichavský, J. Čmejla, and J. Málek, "Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2158–2173, 2021.

[4] N. Amor, J. Čmejla, V. Kautský, Z. Koldovský, and T. Kounovský, "Blind extraction of moving sources via independent component and vector analysis: Examples," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3725–3729.

[5] J. Janský, J. Málek, J. Čmejla, T. Kounovský, Z. Koldovský, and J. Žďánský, "Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 676–680.

[6] J. Málek, J. Janský, T. Kounovský, Z. Koldovský, and J. Žďánský, "Blind extraction of moving audio source in a challenging environment supported by speaker identification via x-vectors," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 226–230.