REAL-TIME DISTRIBUTED SPEECH ENHANCEMENT WITH TWO COLLABORATING MICROPHONE ARRAYS

Amin Hassani, Alexander Bertrand, Marc Moonen

KU Leuven, Dept. of Electrical Engineering-ESAT,
STADIUS center for dynamical systems, signal processing and data analytics
E-mail: amin.hassani (*corresponding*), alexander.bertrand, marc.moonen@esat.kuleuven.be

1. INTRODUCTION

In this demonstration, we aim at presenting our recent implementation results and provide an evaluation testbed through which users can experiment and compare the outputs of the distributed speech enhancement algorithms in [1–3]. The system allows a user to assess the merits of these algorithms in any acoustic setup.

The multi-channel Wiener filter (MWF) is a well-known noise reduction algorithm for multi-microphone speech processing applications. In general, the noise reduction improves as the number of available microphones increases, since a better spatial sampling or diversity can be exploited. Motivated by this, wireless acoustic sensor networks (WASNs), consisting of a multitude of collaborating nodes with an embedded signal processing unit and microphone array, have been proposed to increase the spatial diversity of multi-microphone systems. However, due to the limited per-node computational power and communication bandwidth, *reduced-bandwidth distributed* processing is more favorable than a centralized processing where all the microphone signals are transmitted to a fusion center. In this demo, we evaluate the so-called distributed adaptive node-specific signal estimation (DANSE) algorithm [1] which is essentially a distributed realization of the MWFs of the individual nodes of a WASN and allows the nodes to cooperate by exchanging pre-filtered and compressed signals, while eventually converging to the same centralized MWF solutions as if each node would have access to all the microphone signals in the WASN [1,2]. In the original version of DANSE in [1], the required speech correlation matrices are estimated using a straightforward subtraction-based method. This method, however, has been shown to deliver an unsatisfying performance in the presence of second-order statistics error (e.g., due to low-SNR conditions, highly non-stationary noise or erroneous voice activity detections (VADs)) [4]. An alternative version of DANSE, called generalized eigenvalue decomposition (GEVD)-based DANSE, has been developed in [3] in which each node incorporates a GEVD-based low-rank approximation of the speech correlation matrix in its local MWF. An in-depth theoretical study of the underlying principals of the GEVD-based DANSE algorithm has been presented in [3]. In order to also evaluate the merits of the GEVD-based DANSE algortihm in a practical realistic environment, a real-time experimental setup has been developed which will be explained in the next section.

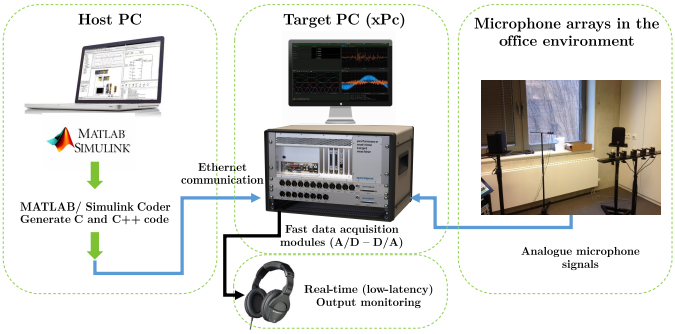

Fig. 1. block diagram of the real-time setup

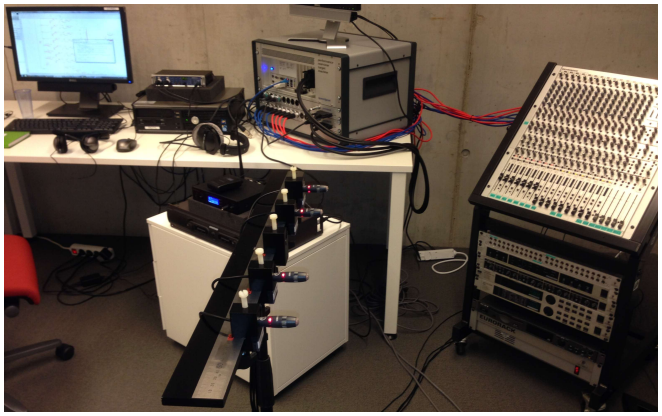

Fig. 2. Experimental setup in the lab

2. REAL-TIME SETUP DESCRIPTION

A block diagram of the real-time setup is depicted in Figure 1. Two microphone arrays, without any prior gain or phase calibration, are placed in random locations in a standard office of approximately 12m^2 (see also Figure 2). In particular, the first and the second array consist of 4 omni-directional AKG CK32 and 3 omni-directional AKG CK92 microphones, respectively. Both arrays have a uniform linear configuration with an inter-microphone spacing of 10 cm. The audio signals (both speech and noise) are generated by an RME Fireface

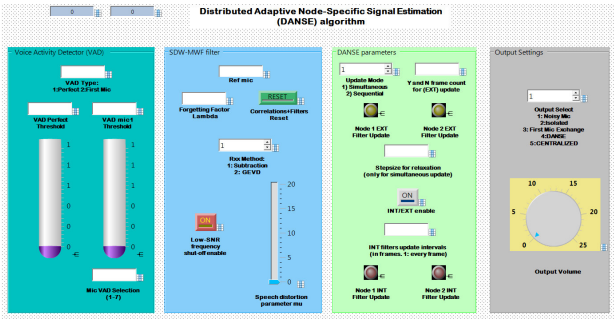

Fig. 3. screenshot of the GUI

soundcard which is controlled via a PC running Adobe Audition. The speech signal is then fed into the speech source loudspeaker (JBL control 1), while the noise signals are separately played back through other loudspeakers present in the scenario. As for the real-time signal processing unit, a so-called high performance multi-core xPC target machine (Speedgoat) is employed to execute the C-coded algorithms in a low latency regime. For this, the test algorithms are first developed on the Host PC in Simulink and Matlab environments, and then the Simulink Real-time toolbox compiles the complete signal processing scheme into C-code and transmits it through an Ethernet connection to the xPC (see Figure 1). The xPC target machine has a total of 8 input audio signals: the 4 signals of array 1, the 3 signals of array 2 and a copy of the clean speech signal which is used for constructing a perfect VAD for the algorithm (the user can switch between a real and a perfect VAD in real-time). The MWF and GEVD-based DANSE are implemented with a weighted overlap-add procedure with 50% overlap, where a sampling frequency of f_s=16kHz and a discrete Fourier transform (DFT) size of $L = 512$ were used. Note that this results in the frame size and overall I/O latency of 16ms. The user can then interact with the system and switch between the following cases in real-time:

- Isolated GEVD-based MWF: the *non-cooperative* case where each node has only access to its own microphone signals
- Centralized GEVD-based MWF: the *full-bandwidth cooperative* case where each node has access to all microphone signals in the WASN
- GEVD-based DANSE: the *cooperative distributed* case where nodes only exchange optimally-compressed single-channel signals
- Basic cooperation with single microphone exchange: this requires the same computational power and communication bandwidth as DANSE, but without the optimal compression of DANSE.

In addition, a graphical user interface (GUI), running on the Host PC, is provided such that the user can easily select the desired algorithm and change the corresponding parameters in real-time (depicted in Figure 3). In particular this GUI has 4 main control panels: 1) VAD, through which the user can choose between perfect or real VAD and input the corresponding VAD threshold. 2) MWF, through which the user can select the reference microphone and tune the speech distortion parameter [4]. Moreover, subtraction-based or GEVD-based estimation can be selected for the estimation of the speech correlation matrix. 3) DANSE, through which the user can choose the update mode (sequential vs. simultaneous [2]), tune the node update frequency (as a function of the number of collected speech+noise and noise-only frames) and step size for relaxation (when the simultaneous update mode is selected). 4) output, where the user can switch between the four algorithms and control the volume of the monitor headphone.

A video footage providing the output signals of the different cases while the setup is running in real-time has been presented in [5]. This indeed reveals how the GEVD-based DANSE algorithm remarkably enhances the speech signal in the presence of highly non-stationary construction and multi-talker noise and obtains the same performance as the centralized MWF.

3. ON-SITE DEMO DESCRIPTION

The main objective of this demo will be to provide an evaluation testbed such that the user can compare the different cases described in Section 2 and further evaluate the merits of the GEVD-based DANSE algorithm. To achieve this, output signals of the different cases will be recorded and taken to the conference, where together with the extra materials such as the video footage and the illustrative posters, the benefits of using the GEVD-based DANSE algorithm will be explained and demonstrated and its performance will be compared to the more expensive centralized MWF algorithm.

The first set of recordings will let the user compare the GEVD-based MWF with the subtraction-based MWF in conditions with low-SNR, highly non-stationary noise, or with an erroneous VAD (with the aim of assessing the merits of the former case). The second set of recordings provides the outputs of the following cases: 1) raw (unprocessed) reference microphone 2) isolated GEVD-based MWF 3) centralized GEVD-based MWF 4) GEVD-based DANSE and 5) GEVD-based single channel transmission case. The user will be expected to verify the superior performance achieved by the GEVD-based DANSE algorithm (after convergence to the centralized case).

4. REFERENCES

[1] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks part I: sequential node updating," in *IEEE Trans. Signal Process.*, 2010, vol. 58, pp. 5277–5291.

[2] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks part II: simultaneous and asynchronous node updating," in *IEEE Trans. Signal Process.*, 2010, vol. 58, pp. 5292–5306.

[3] A. Hassani, A. Bertrand, and M. Moonen, "GEVD-based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks," *IEEE Trans. on Signal Process.*, vol. 64, no. 10, pp. 2557–2572, May 2016.

[4] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, April 2014.

[5] A. Hassani, A. Bertrand, and M. Moonen, "Real-time distributed speech enhancement," available online at: https://youtu.be/dF4IfJU9xhE, 2016.