Distributed combined acoustic echo cancellation and noise reduction using GEVD-based distributed adaptive node specific signal estimation with prior knowledge

Santiago Ruiz, Toon van Waterschoot and Marc Moonen
Dept. of Electrical Engineering, ESAT-STADIUS
KU Leuven
Leuven, Belgium
Email: {santiago.ruiz, toon.vanwaterschoot,marc.moonen}@esat.kuleuven.be

***Abstract*—Distributed combined acoustic echo cancellation (AEC) and noise reduction (NR) in a wireless acoustic sensor network (WASN) is tackled by using a specific version of the PK-GEVD-DANSE algorithm (cfr. [1]). Although this algorithm was initially developed for distributed NR with partial prior knowledge of the desired speech steering vector, it is shown that it can also be used for AEC combined with NR. Simulations have been carried out using centralized and distributed batch-mode implementations to verify the performance of the algorithm in terms of AEC quantified with the echo return loss enhancement (ERLE), as well as in terms of the NR quantified with the signal-to-noise ratio (SNR).**

***Index Terms*—Distributed signal processing, wireless acoustic sensor networks, acoustic echo cancellation, noise reduction.**

I. Introduction

Many speech and audio signal processing applications, such as teleconferencing/telepresence, in-car and full-duplex communication, voice recognition and ambient intelligence, suffer from acoustic echoes and background noise which corrupt the desired speech signal. Acoustic echo cancellation (AEC) and noise reduction (NR) techniques can be used to enhance the desired speech signal while reducing undesired components [2]–[5].

Centralized approaches are usually prohibitive in a wireless acoustic sensor network (WASN) in terms of complexity and communication cost [6], hence distributed signal processing techniques have been developed such as, e.g., the distributed delay-and-sum beamformer for NR based on randomized gossiping presented in [7], which was then extended to a distributed MVDR beamformer based on message passing in [8]. Both algorithms do not have a topology constraint and provide good performance at the expense of a high communication cost [7].

The distributed adaptive node-specific signal estimation (DANSE) algorithm as developed in [9], performs distributed NR, i.e., optimally enhances the local microphone signals in each WASN node, as if all signals in the WASN were available to each and every node, while each node is still sharing only a fused version of its microphone signals with the other nodes. None of the algorithms mentioned so far considered the fact that there may be signals available in the WASN that do not contribute to the desired speech steering vector.

Here, distributed combined AEC and NR is considered in a WASN where a node has microphones as well as loudspeakers. The loudspeakers play given (far-end) signals, and generate echo signals in the microphones (also in other nodes). The prior knowledge (PK) generalized eigenvalue decomposition (GEVD)-based DANSE algorithm (PK-GEVD-DANSE), although initially developed for NR with partial prior knowledge of the desired speech steering vector, will be used here for AEC combined with NR. In this case, the PK expresses the fact that the loudspeaker signals in an AEC scenario do not contribute to the desired speech steering vector. The PK-GEVD-DANSE algorithm then performs AEC and NR at each node based on sharing not only fused microphone and loudspeaker signals, which act as desired signal references, but also fused loudspeaker signals, which act as noise references. Each node then effectively implements a cascaded approach to perform first AEC and then NR, using fused signals from the other nodes. Nevertheless, the PK-GEVD-DANSE achieves a performance as if all signals in the WASN were available to each and every node. A centralized PK multichannel Wiener filter (PK-MWF) algorithm is also described and used for comparison.

The paper is organized as follows. The data model is presented in Section II. The formulations for the centralized and distributed algorithm are provided in Sections III, IV and V. Simulations are shown in Section VI and Section VII concludes the paper.

This work was carried out at the ESAT Laboratory of KU Leuven in the frame of KU Leuven Internal Fund Projects VES/19/004 and C2-16-00449 ”Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking“, FWO/FNRS EOS Project nr. 30452698 ”MUSE-WINET - Multi-Service Wireless Network” and European Research Council under the European Union’s Horizon 2020 Research and Innovation Program / ERC Consolidator Grant: SONORA (no. 773268) . The scientific responsibility is assumed by its authors.

II. Data model

A fully connected WASN with K nodes is considered as shown in Fig. 1 in which a node $k \in \mathcal{K} = \{1, \ldots, K\}$ has access to the short-time Fourier transform (STFT) domain $n_k \times 1$ signal vector $\mathbf{y}_k(\kappa, l) = \begin{bmatrix} \mathbf{x}_k(\kappa, l) \\ \mathbf{u}_k(\kappa, l) \end{bmatrix}$, where κ is the frequency bin index, l the time frame index (for brevity κ and l will be omitted in the following, except for a few cases where l has to be included explicitly), $n_k = m_k + P \cdot l_k$, $\mathbf{u}_k$ contains l_k local loudspeaker signals sampled at the current and previous $P-1$ frames, i.e.,

$$\mathbf{u}_k(l) = \begin{bmatrix} u_1(l) \\ \vdots \\ u_1(l-P+1) \\ \vdots \\ u_{l_k}(l) \\ \vdots \\ u_{l_k}(l-P+1) \end{bmatrix} \tag{1}$$

whereas $\mathbf{x}_k$ contains m_k local microphone signals sampled only at the current frame and is modeled as

$$\mathbf{x}_k = \mathbf{s}_k + \mathbf{n}_k = \mathbf{a}_k s + \mathbf{n}_k. \tag{2}$$

Here, s is the desired speech signal (also known as the dry signal), $\mathbf{a}_k$ contains the acoustic transfer functions from the speech source position to the local microphones, $\mathbf{s}_k$ the desired speech component and $\mathbf{n}_k$ the noise component, specified as

$$\mathbf{n}_k = \mathbf{G}_k \mathbf{u}_k + \mathbf{G}_{-k} \mathbf{u}_{-k} + \mathbf{b}_k \tag{3}$$

where $\mathbf{G}_k$ is an $m_k \times Pl_k$ matrix representing the local echo paths from the local loudspeakers to the local microphones, $\mathbf{G}_{-k}$ is an $m_k \times P(L - l_k)$ matrix representing the echo paths from the loudspeakers in the other nodes to the local microphones, with $L = \sum_{k=1}^{K} l_k$, and similarly $M = \sum_{k=1}^{K} m_k$, $N = \sum_{k=1}^{K} n_k = M + P \cdot L$, $\mathbf{u}_{-k}$ contains the loudspeaker signals from the other nodes[1] and $\mathbf{b}_k$ is the background noise. Node k has immediate access to $\mathbf{y}_k$ only. The following vectors are also defined $\tilde{\mathbf{s}}_k = \left[\mathbf{s}_k^H \ \mathbf{0}_{1 \times Pl_k}\right]^H$, $\tilde{\mathbf{n}}_k = \left[\mathbf{n}_k^H \ \mathbf{u}_k^H\right]^H$ and $\tilde{\mathbf{a}}_k = \left[\mathbf{a}_k^H \ \mathbf{0}_{1 \times Pl_k}\right]^H$, where $\mathbf{0}_{1 \times Pl_k}$ is a Pl_k-dimensional all-zero vector and $(\cdot)^H$ denotes the conjugate transpose operator, so that $\mathbf{y}_k = \tilde{\mathbf{s}}_k + \tilde{\mathbf{n}}_k = \tilde{\mathbf{a}}_k s + \tilde{\mathbf{n}}_k$. The N-dimensional vectors, $\mathbf{y}, \mathbf{s}, \mathbf{n}$ and $\mathbf{a}$ are the stacked versions of $\mathbf{y}_k, \tilde{\mathbf{s}}_k, \tilde{\mathbf{n}}_k$ and $\tilde{\mathbf{a}}_k$ respectively, such that the signal vector $\mathbf{y}_k$ can be generalized as follows

$$\mathbf{y} = \mathbf{s} + \mathbf{n} = \mathbf{a}s + \mathbf{n}. \tag{4}$$

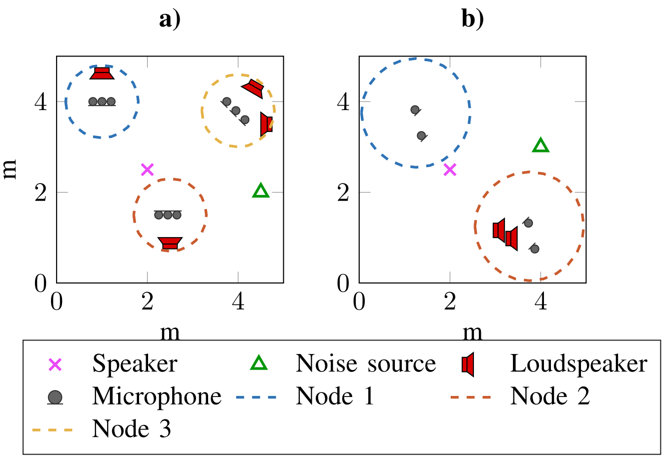

Fig. 1: Scenarios used composed of: **a**). Three nodes each with 3 microphones and 1 or 2 loudspeakers. The loudspeaker signals are three speech signals and one music signal. **b**). Two nodes each with 2 microphones. One node with a stereo loudspeaker signal.

III. Centralized combined AEC and NR without prior knowledge (MWF)

The node-specific task for node k is to estimate the desired signal d_k, defined here as the desired speech component in the first local microphone i.e. $d_k = [1\ \mathbf{0}]\,\mathbf{s}_k = \mathbf{e}_{d_k}^H \mathbf{s}$, where $\mathbf{0}$ is an all-zero vector with matching dimensions and $\mathbf{e}_{d_k}^H$ is a vector that selects this desired speech component in $\mathbf{s}$. The mean squared error (MSE) between the desired signal and the filtered microphone and loudspeaker signals is minimized defining an optimal filter

$$\hat{\mathbf{w}}_k = \arg\min_{\mathbf{w}_k} E\{||d_k - \mathbf{w}_k^H \mathbf{y}||^2\} \tag{5}$$

where $E\{\cdot\}$ is the expected value operator. The node-specific signal estimate is then obtained as $\hat{d}_k = \hat{\mathbf{w}}_k^H \mathbf{y}$. The solution to this is the well-known MWF [11] given by

$$\hat{\mathbf{w}}_k = \mathbf{R}_{\mathbf{yy}}^{-1} \mathbf{R}_{\mathbf{y}d_k} = \mathbf{R}_{\mathbf{yy}}^{-1} \mathbf{R}_{\mathbf{ys}} \mathbf{e}_{d_k} = \mathbf{R}_{\mathbf{yy}}^{-1} \mathbf{R}_{\mathbf{ss}} \mathbf{e}_{d_k} \tag{6}$$

where $\mathbf{R}_{\mathbf{yy}} = E\{\mathbf{y}\mathbf{y}^H\}$, $\mathbf{R}_{\mathbf{y}d_k} = E\{\mathbf{y}d_k^H\}$, $\mathbf{R}_{\mathbf{ys}} = E\{\mathbf{y}\mathbf{s}^H\}$ and $\mathbf{R}_{\mathbf{ss}} = E\{\mathbf{s}\mathbf{s}^H\}$ are signal correlation matrices. The final expression in (6) is obtained based on the assumption that s and $\mathbf{n}$ are uncorrelated. $\mathbf{R}_{\mathbf{ss}}$ is not directly observable and must be estimated. In practice, $\mathbf{R}_{\mathbf{yy}}$ and $\mathbf{R}_{\mathbf{nn}} = E\left\{\mathbf{n}\mathbf{n}^H\right\}$ are estimated, by using a voice activity detector (VAD), during ”*speech plus noise*” periods where the desired speech signal, loudspeaker signals and background noise are active, and ”*noise-only*” periods where there is no activity of the desired speech signal and the other components are active, respectively [12], i.e.,

$$\text{if VAD=1: } \hat{\mathbf{R}}_{\mathbf{yy}}(l) = \beta \hat{\mathbf{R}}_{\mathbf{yy}}(l-1) + (1-\beta)\mathbf{y}(l)\mathbf{y}^H(l) \tag{7}$$
$$\text{if VAD=0: } \hat{\mathbf{R}}_{\mathbf{nn}}(l) = \beta \hat{\mathbf{R}}_{\mathbf{nn}}(l-1) + (1-\beta)\mathbf{y}(l)\mathbf{y}^H(l)$$

where $\hat{\mathbf{R}}_{\mathbf{yy}}(l)$, $\hat{\mathbf{R}}_{\mathbf{nn}}(l)$, $\mathbf{y}(l)$ represent $\hat{\mathbf{R}}_{\mathbf{yy}}$, $\hat{\mathbf{R}}_{\mathbf{nn}}$, $\mathbf{y}$ at frame l, respectively. The forgetting factor $0 < \beta < 1$ can be chosen depending on the time-variation of the signal statistics i.e. if

[1] The model for the echo signals in (3) is an approximate model, the approximation being better when the STFT uses frequency selective analysis filters, and when the number of frames P matches with the true length of the echo paths [10]

the statistics change slowly β should be chosen close to 1 to obtain long-term estimates that mainly capture the spatial coherence between the microphone signals. For the time being, it is assumed that the loudspeaker signals are always active and sufficiently uncorrelated, hence that $\mathbf{R_{nn}}$ is full rank, and that the loudspeaker signals and background noise are stationary. Therefore the VAD should be able to detect the activity of the desired speech signal in the presence of loudspeaker signals which may contain speech signals, and other background noise signals. The following procedure will then be used to estimate $\mathbf{R_{ss}}$ based on the criterion [1], [11]

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \underset{\substack{\mathrm{rank}(\mathbf{R_{ss}})=1 \\ \mathbf{R_{ss}}\succeq 0}}{\arg\min} ||\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} - \mathbf{R_{ss}}\right)\hat{\mathbf{R}}_{\mathbf{nn}}^{-\mathrm{H}/2}||_F^2 \tag{8}$$

where $||\cdot||_F$ denotes the Frobenius norm. Spatial pre-whitening is applied by pre- and post-multiplying by $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}^{-\mathrm{H}/2}$, respectively. The solution to (8) is based on a generalized eigenvalue decomposition (GEVD) of the matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}}\}$ [11], [13]

$$\begin{aligned} \hat{\mathbf{R}}_{\mathbf{yy}} &= \mathbf{Q}\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{Q}^{\mathrm{H}} \\ \hat{\mathbf{R}}_{\mathbf{nn}} &= \mathbf{Q}\mathbf{\Sigma}_{\mathbf{nn}}\mathbf{Q}^{\mathrm{H}} \end{aligned} \tag{9}$$

where $\mathbf{\Sigma_{yy}}$ and $\mathbf{\Sigma_{nn}}$ are diagonal matrices and $\mathbf{Q}$ is an invertible matrix. The speech correlation matrix estimate $\hat{\mathbf{R}}_{\mathbf{ss}}$ is then given as [11]

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \mathbf{Q}\mathrm{diag}\{\sigma_{y_1} - \sigma_{n_1}, 0, \ldots, 0\}\mathbf{Q}^{\mathrm{H}} \tag{10}$$

where σ_{y_1} and σ_{n_1} are the first diagonal elements of $\mathbf{\Sigma_{yy}}$ and $\mathbf{\Sigma_{nn}}$, respectively, corresponding to the largest ratio $\sigma_{y_i}/\sigma_{n_i}$. Using (10) and $\hat{\mathbf{R}}_{\mathbf{yy}}$ (cfr. (9)) in (6), $\hat{\mathbf{w}}_k$ can be expressed as

$$\hat{\mathbf{w}}_k = \mathbf{Q}^{-H}\mathrm{diag}\left\{1 - \frac{\sigma_{n_1}}{\sigma_{y_1}}, 0, \ldots, 0\right\}\mathbf{Q}^H\mathbf{e}_{d_k}. \tag{11}$$

In this approach, the loudspeaker signals are included in the formulation, however, it fundamentally consists of applying NR without considering that there is no desired speech component in these loudspeaker signals.

IV. Centralized combined AEC and NR with prior knowledge (PK-MWF)

Exploiting the prior knowledge that $\mathbf{R_{ss}}$ has a specific zero structure (cfr. definition of $\mathbf{s}_k$ ans $\tilde{\mathbf{s}}_k$) , (8) can be redefined as

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \underset{\substack{\mathrm{rank}(\mathbf{R_{ss}})=1 \\ \mathbf{B}^{\mathrm{H}}\mathbf{R_{ss}}\mathbf{B}=0 \\ \mathbf{R_{ss}}\succeq 0}}{\arg\min} ||\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} - \mathbf{R_{ss}}\right)\hat{\mathbf{R}}_{\mathbf{nn}}^{-\mathrm{H}/2}||_F^2 \tag{12}$$

where $\mathbf{B}$ is a $N \times PL$ block diagonal matrix with k^{th} diagonal block $\mathbf{B}_k$ equal to

$$\mathbf{B}_k = \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{Pl_k} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{B}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{B}_K \end{bmatrix}, \tag{13}$$

with $\mathbf{I}_{Pl_k}$ a $Pl_k \times Pl_k$ identity matrix. In this particular case $\mathbf{B}$ is a selection matrix that selects the loudspeaker signals. In [1] it is shown that this leads to the reduced dimensional matrix pencil $\{\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}, \hat{\mathbf{R}}_{\hat{\mathbf{n}}\hat{\mathbf{n}}}\}$ with GEVD

$$\begin{aligned} \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} &= \hat{\mathbf{Q}}\mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Q}}^{\mathrm{H}} \\ \hat{\mathbf{R}}_{\hat{\mathbf{n}}\hat{\mathbf{n}}} &= \hat{\mathbf{Q}}\mathbf{\Sigma}_{\hat{\mathbf{n}}\hat{\mathbf{n}}}\hat{\mathbf{Q}}^{\mathrm{H}} \end{aligned} \tag{14}$$

where $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \mathbf{C}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{yy}}\mathbf{C}$, $\hat{\mathbf{R}}_{\hat{\mathbf{n}}\hat{\mathbf{n}}} = \mathbf{C}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{C}$, $\hat{\mathbf{y}} = \mathbf{C}^{\mathrm{H}}\mathbf{y}$, and with $\mathbf{C}$ a $N \times M$ matrix obtained from the linearly-constrained minimum variance (LCMV) beamformer optimization criterion

$$\mathbf{C} = \underset{\mathrm{s.t.}\quad \mathbf{H}^H\mathbf{C}=\mathbf{I}_M}{\arg\min} \mathrm{trace}\{\mathbf{C}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{C}\} \tag{15}$$

where $\mathbf{H}$ is a $N \times M$ block diagonal matrix with k^{th} diagonal block equal to

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{I}_{m_k} \\ \mathbf{0} \end{bmatrix}, \mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{H}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{H}_K \end{bmatrix}, \tag{16}$$

such that $\mathbf{H}^{\mathrm{H}}\mathbf{H} = \mathbf{I}_M$ and $\mathbf{B}^H\mathbf{H} = \mathbf{0}$. Hence $\mathbf{C}$ can be defined based on a generalised sidelobe canceller (GSC) implementation as [1], [14]

$$\mathbf{C} = \mathbf{H} - \mathbf{BF} \tag{17}$$

$$\mathbf{F} = (\mathbf{B}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{H}. \tag{18}$$

where the filter $\mathbf{F}$ operates on the loudspeaker signals and, in this particular case effectively serves as an AEC filter cancelling the echo components in the so-called fixed beamformer outputs corresponding to $\mathbf{H}$ i.e. the microphone signals. In practice, $\mathbf{F}$ can also be implemented adaptively via a normalized least mean squares (NLMS) algorithm, as in [2], [14].

The prior knowledge speech correlation matrix $\hat{\mathbf{R}}_{\mathbf{ss}}$, i.e., the solution to (12), is then given as [1], [11]

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \mathbf{H}\hat{\mathbf{Q}}\mathrm{diag}\{\sigma_{\hat{y}_1} - \sigma_{\hat{n}_1}, 0, \ldots, 0\}\hat{\mathbf{Q}}^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}, \tag{19}$$

where $\sigma_{\hat{y}_1}$ and $\sigma_{\hat{n}_1}$ are the first diagonal elements of $\mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{\Sigma}_{\hat{\mathbf{n}}\hat{\mathbf{n}}}$, respectively, corresponding to the largest ratio $\sigma_{\hat{y}_i}/\sigma_{\hat{n}_i}$. Using this expression and the reduced dimensional $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ (cfr. (14)), the filter $\check{\mathbf{w}}_k$ can finally be expressed as [1]

$$\hat{\mathbf{w}}_k = \mathbf{C}\hat{\mathbf{W}}_{GEVD}\mathbf{H}^{\mathrm{H}}\mathbf{e}_{d_k} = \mathbf{C}\hat{\mathbf{W}}_{GEVD}\hat{\mathbf{e}}_{d_k} \tag{20}$$

$$\hat{\mathbf{W}}_{GEVD} = \hat{\mathbf{Q}}^{-H}\mathrm{diag}\left\{1 - \frac{\sigma_{\hat{n}_1}}{\sigma_{\hat{y}_1}}, 0, \ldots, 0\right\}\hat{\mathbf{Q}}^{\mathrm{H}}. \tag{21}$$

where the node-specific signal estimate is finally obtained as $\hat{d}_k = \hat{\mathbf{w}}_k^H y$. The centralized solution then indeed corresponds to merely a cascade of AEC (cfr. (17)) and NR (crf. (21)).

V. Distributed combined AEC and NR with prior knowledge (PK-GEVD-DANSE)

In the distributed processing approach, the PK-GEVD-DANSE algorithm [1] is used where each node instead of broadcasting n_k microphone and loudspeaker signals, broadcasts only 2 fused signals, i.e., a desired signal reference and a noise reference. In the context of combined AEC and NR, the second fused signal is a fused loudspeaker signal. Each node then performs local operations, effectively corresponding to a reduced dimensional version ($n_k + 2(K-1)$ in node k) of the procedure of section IV (dimension N) but now based on its local microphone and loudspeaker signals and the received fused signals from the other nodes. The fused signals broadcast by node k are

$$z_k = \mathbf{p}_k^H \mathbf{y}_k \tag{22}$$

$$\underline{z}_k = \boldsymbol{\lambda}_k^H \mathbf{B}_k^H \mathbf{y}_k = \boldsymbol{\lambda}_k^H \mathbf{u}_k \tag{23}$$

where $\mathbf{p}_k$ is an n_k-dimensional fusion vector and $\boldsymbol{\lambda}_k$ is a Pl_k-dimensional fusion vector. Then each node has access to a signal vector $\check{\mathbf{y}}_k = \left[\mathbf{y}_k^H \; \mathbf{z}_{-k}^H \; \underline{\mathbf{z}}_{-k}^H\right]^H$, where the subscript $_{-k}$ refers to the concatenation of the fused signals of nodes other than k, so that $\mathbf{z}_{-k} = [z_1^H \ldots z_{k-1}^H \, z_{k+1}^H \ldots z_K^H]^H$ and similarly for $\underline{\mathbf{z}}_{-k}$. A modification must be introduced in $\mathbf{H}_k$ and $\mathbf{B}_k$ to account for the extra signals broadcast from the other nodes, hence

$$\mathbf{H}_k = \begin{bmatrix} \begin{bmatrix} \mathbf{I}_{m_k} \\ \mathbf{0} \end{bmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} \mathbf{I}_{K-1} \\ \mathbf{0} \end{bmatrix} \end{bmatrix}, \; \mathbf{B}_k = \begin{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{Pl_k} \end{bmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{K-1} \end{bmatrix} \end{bmatrix} \tag{24}$$

where $\mathbf{H}_k$ is an $(n_k+2K-2)\times(m_k+K-1)$ matrix and $\mathbf{B}_k$ is an $(n_k+2K-2)\times(Pl_k+K-1)$ matrix. Then equations (17) and (18) become respectively

$$\mathbf{C}_k = \mathbf{H}_k - \mathbf{B}_k\mathbf{F}_k \tag{25}$$

$$\mathbf{F}_k = (\mathbf{B}_k^H \hat{\mathbf{R}}_{\check{\mathbf{n}}_k\check{\mathbf{n}}_k} \mathbf{B}_k)^{-1} \mathbf{B}_k^H \hat{\mathbf{R}}_{\check{\mathbf{n}}_k\check{\mathbf{n}}_k} \mathbf{H}_k \tag{26}$$

where $\check{\mathbf{n}}_k$ corresponds to $\check{\mathbf{y}}_k$ in *"noise-only"* periods, $\hat{\mathbf{R}}_{\check{\mathbf{y}}_k\check{\mathbf{y}}_k} = E\{\check{\mathbf{y}}_k\check{\mathbf{y}}_k^H\}$, $\hat{\mathbf{y}}_k = \mathbf{C}_k^H\check{\mathbf{y}}_k$, $\hat{\mathbf{R}}_{\check{\mathbf{n}}_k\check{\mathbf{n}}_k} = E\{\check{\mathbf{n}}_k\check{\mathbf{n}}_k^H\}$, $\hat{\mathbf{R}}_{\hat{\mathbf{n}}_k\hat{\mathbf{n}}_k} = \mathbf{C}_k^H\hat{\mathbf{R}}_{\check{\mathbf{n}}_k\check{\mathbf{n}}_k}\mathbf{C}_k$ and $\hat{\mathbf{R}}_{\hat{\mathbf{y}}_k\hat{\mathbf{y}}_k} = \mathbf{C}_k^H\hat{\mathbf{R}}_{\check{\mathbf{y}}_k\check{\mathbf{y}}_k}\mathbf{C}_k$.The fusion vectors are defined as [1]

$$\boldsymbol{\lambda}_k = [\mathbf{I}_{Pl_k} \; \mathbf{0}](\mathbf{B}_k^H \hat{\mathbf{R}}_{\check{\mathbf{n}}_k\check{\mathbf{n}}_k} \mathbf{B}_k)^{-1} \mathbf{B}_k^H \hat{\mathbf{R}}_{\check{\mathbf{y}}_k\check{\mathbf{y}}_k} \hat{\mathbf{w}}_k. \tag{27}$$

$$\mathbf{p}_k = [\mathbf{I}_{n_k} \; \mathbf{0}]\hat{\mathbf{w}}_k \tag{28}$$

where the local filter $\hat{\mathbf{w}}_k$ is defined as

$$\hat{\mathbf{w}}_k = \mathbf{C}_k \hat{\mathbf{W}}_{GEVD,k} \mathbf{H}_k^H [1 \; \mathbf{0}]^H. \tag{29}$$

with $\hat{\mathbf{W}}_{GEVD,k}$ defined similar to (21) but using the GEVD of $\{\hat{\mathbf{R}}_{\hat{\mathbf{y}}_k\hat{\mathbf{y}}_k}, \hat{\mathbf{R}}_{\hat{\mathbf{n}}_k\hat{\mathbf{n}}_k}\}$. In each time frame the nodes broadcast fused signals (22) and (23) using their current fusion vectors. One node then updates its fusion vectors by means of (24)-(29). When the nodes update sequentially in a round-robin fashion (e.g. one node updates per time frame) the local signal estimates $\hat{d}_k = \hat{\mathbf{w}}_k^H \check{\mathbf{y}}_k$ have been shown to converge in each node to the centralized signal estimates obtained with (20) [1].

VI. Simulations

This section outlines the simulations carried out using the sections III and IV algorithms, MWF and PK-MWF, respectively and the PK-GEVD-DANSE algorithm described in section V. The scenario depicted in Fig. 1a was first used with $P = 1$. The performance of the algorithm was measured in terms of the echo return loss enhancement (ERLE) and the signal-to-noise ratio (SNR). The simulations were set up as follows. Firstly, the microphone and loudspeaker signals were simulated at each node using room impulse responses of 500 samples long with the randomized image method described in [15] and a sampling frequency of 16kHz. The reflection coefficient of all surfaces in the room was set to 0.15 (for a reverberation time $T_{60} = 0.1116$s), and the random displacement of the image sources to 0.13m. The inter-microphone distance of the arrays was set to 20cm for all the nodes. The microphone signals have been created such that the signal-to-echo-and-noise ratio at microphone 1 in node 2 was -5 dB. Then, the corresponding vector $\mathbf{y}_k$ for each node was transformed to the STFT domain using a square-root hann window of 512 samples using 50% overlap. The correlation matrices in (9), (14) and $\{\hat{\mathbf{R}}_{\check{\mathbf{y}}_k\check{\mathbf{y}}_k}, \hat{\mathbf{R}}_{\check{\mathbf{n}}_k\check{\mathbf{n}}_k}\}$ in section V where computed by selecting the time frames where the desired speech signal was active and not active, respectively, based on an ideal VAD. An ideal VAD was used to isolate the influence of VAD errors. In practice the VAD may not perform well because the echo signals may also be speech signals. This may be tackled by sharing VAD information among the nodes [16] and/or using a speaker-selective VAD [17]. All nodes in Fig. 1a had a loudspeaker reproducing a speech signal, which were simultaneously active only when the desired speech signal was not active. The second loudspeaker in node 3 was reproducing a music signal which was continuously active. The desired speech signal came from a speaker located around the centre of the room. A continuously active localized noise source was also included, producing babble noise.

PK-GEVD-DANSE was run with simultaneous node updating with a relaxation factor $\alpha_{rS} = 0.9$, to guarantee convergence as suggested in [12]. The ERLE was computed with non-overlapping windows of 1024 samples. The average ERLE (over the time frames) and SNR are shown in Fig. 2. Both metrics were computed for the first microphone in each node. The SNR was computed by filtering the noise component at each microphone signal with the filter obtained for each implementation. PK-GEVD-DANSE is abbreviated to PK-DANSE in the legends for brevity. The MSE for the three algorithms at the first microphone of each node is shown in Fig. 3.

It can be seen in all nodes that including the PK reduces the error in the estimation of the desired speech signal. It can also be seen that PK-GEVD-DANSE converges to the same results as PK-MWF. In node 3 PK-GEVD-DANSE and PK-MWF outperform MWF in terms of ERLE and SNR. Notice that node 3 is the furthest away from the desired speech source location, it is very close to the noise source and has two different loudspeaker signals.

The scenario depicted in Fig. 1b was simulated with echo paths of 4096 samples long and a $T_{60} \approx 1.1$s. A per-frame processing approach now is used for PK-GEVD-DANSE, where the correlation matrices are updated based on (7), and its MSE is shown in Fig. 4 and compared to batch results for MWF and PK-MWF. The significant MSE values before frame index 50 are due to the initial updating of the correlation matrices. The use of previous frames is investigated for $P =$

2, 4, 8 and 16 in the same scenario. Fig. 5 shows the MSE for PK-GEVD-DANSE with different values of P and the PK-MWF, with frame size of 512 samples. It is observed that for $P > 1$ PK-GEVD-DANSE outperforms PK-MWF ($P = 1$) at node 2, which is not the case at node 1. For $P > 16$ the MSE was not reduced further, which could be related to the presence of the babble noise and the increasing number of filter coefficients to be estimated.

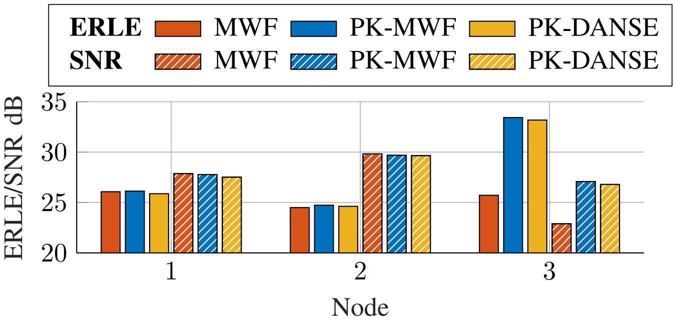

Fig. 2: Average ERLE and SNR computed at the first microphone of each node in Fig. 1a.

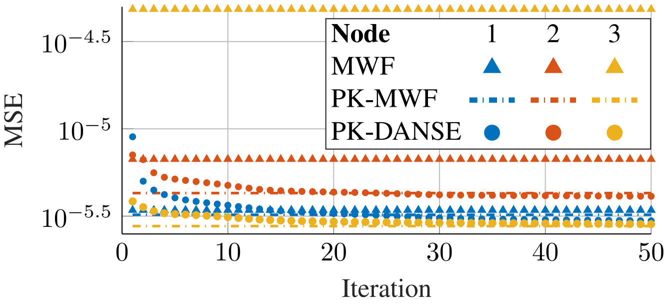

Fig. 3: MSE at each of the nodes in Fig. 1a.

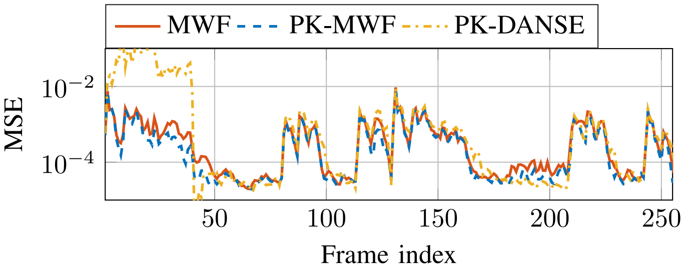

Fig. 4: MSE at node 2 in Fig. 1b.

VII. Conclusions

It has been shown that PK-GEVD-DANSE can be adopted for distributed combined AEC and NR in a WASN. It has also been shown that PK-GEVD-DANSE algorithm performing simultaneous updates in the nodes converges to the PK-MWF algorithm. The performance of the algorithms has been verified in terms of AEC quantified with the ERLE, as well as in terms of NR quantified with the SNR

References

[1] R. Van Rompaey and M. Moonen, "Distributed adaptive node-specific signal estimation in a wireless sensor network with partial prior knowledge of the desired source steering vector," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, 2019.

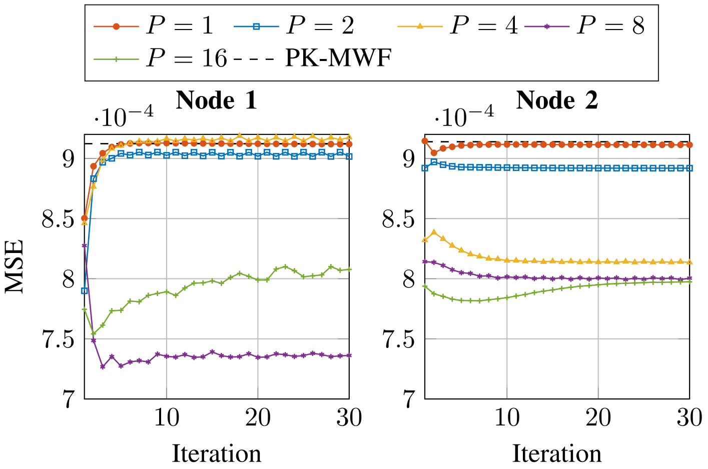

Fig. 5: MSE for PK-GEVD-DANSE at each of the nodes in Fig. 1b with $P = 1, 2, 4, 8, 16$.

[2] W. Herbordt, W. Kellermann, and S. Nakamura, "Joint optimization of acoustic echo cancellation and adaptive beamforming," *Topics in acoustic echo and noise control*, pp. 19–50, 2006.

[3] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective*, Elsevier, 2014.

[4] E. Böhmler, J. Freudenberger, and S. Stenzel, "Combined echo and noise reduction for distributed microphones," in *Proc. 2011 Joint Workshop Hands-Free Speech Comm. Microphone Arrays*, 2011, pp. 98–103.

[5] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal processing*, vol. 64, no. 1, pp. 21–32, 1998.

[6] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Computer networks*, vol. 51, no. 4, pp. 921–960, 2007.

[7] Y. Zeng and R. C. Hendriks, "Distributed delay and sum beamformer for speech enhancement via randomized gossip," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, 2013.

[8] R. Heusdens, G. Zhang, R. C Hendriks, Y. Zeng, and W. B. Kleijn, "Distributed MVDR beamforming for (wireless) microphone networks using message passing," in *2012 Proc. Int. Workshop Acoustic Signal Enhancement IWAENC*. VDE, 2012.

[9] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks—part i: Sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, 2010.

[10] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[11] 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/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[12] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP J. Adv. Signal Process*, vol. 2009, 2009.

[13] F. Jabloun and B. Champagne, "Signal subspace techniques for speech enhancement," in *Speech Enhancement*, pp. 135–159. Springer, 2005.

[14] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, *in Handbook on array processing and sensor networks "Acoustic beamforming for hearing aid applications"*, pp. 269–302, Wiley Online Library, 2008.

[15] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 23, no. 4, pp. 774–786, 2015.

[16] V. Berisha, H. Kwon, and A. Spanias, "Real-time implementation of a distributed voice activity detector," in *Proc. 4th IEEE Workshop Sensor Array Multichannel Process.*, pp. 659–662.

[17] S. Maraboina, D. Kolossa, P.K. Bora, and R. Orglmeister, "Multi-speaker voice activity detection using ica and beampattern analysis," in *Proc. 2006 14th Eur. Signal Process. Conf.*, 2006.