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*[⋆]

[⋆] *KU Leuven*, ESAT-STADIUS, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[†] *KU Leuven*, ESAT-ETC, e-Media Research Lab, Andreas Vesaliusstraat 13, Leuven, Belgium

ABSTRACT

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

1. INTRODUCTION

Many speech and audio signal processing applications, such as teleconferencing/telepresence, in-car and full-duplex communication, voice recognition and intelligent ambience, suffer from acoustic echoes and interfering 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 signal while reducing undesired components [2, 3, 4, 5].

Here, distributed combined AEC and NR is considered in the context of wireless acoustic sensor networks (WASN) [6]. The distributed adaptive node-specific signal estimation (DANSE) algorithm as developed in [7], performs distributed NR i.e. optimally enhances the local microphone signals of each WASN node, as if all signals in the WASN were available to each and every node, but by sharing only a fused version of the microphone signals at each node with the 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, is used here for AEC combined with NR. A centralized PK multi-channel Wiener filter (PK-MWF) is also used for comparison. The PK exploits the fact that the loudspeaker reference signals do not contribute to the desired speech steering vector in an AEC scenario, hence may be cancelled out before the signals are fed into the MWF.

The paper is organized as follows. The data model is presented in Section 2. The formulations for the centralized and distributed algorithms are provided in Sections 3, 4 and 5. Simulations using speech signals are shown in Section 6 and Section 7 concludes the paper.

This work was carried out at the ESAT Laboratory of KU Leuven in the frame of KU Leuven internal funding 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.

2. DATA MODEL

A fully connected WASN with K nodes is considered as shown in Figure 1 in which a node $k \in \mathcal{K} = \{1, \ldots, K\}$ has access to the short-time Fourier transform (STFT) domain $M_k \times 1$ 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, l the time frame (for brevity κ and l will be omitted in the following), $M_k = m_k + l_k$, $\mathbf{u}_k$ contains l_k loudspeaker signals and $\mathbf{x}_k$ contains m_k microphone signals defined as

$$\mathbf{x}_k = \mathbf{s}_k + \mathbf{n}_k = \mathbf{a}_k \breve{s} + \mathbf{n}_k. \tag{1}$$

Where $\breve{s}$ is the desired speech signal (also known as the dry signal), $\mathbf{a}_k$ are the acoustic transfer functions from $\breve{s}$ to the node's microphones, $\mathbf{s}_k$ is the desired speech component and $\mathbf{n}_k$ is the noise component at node k, specified as

$$\mathbf{n}_k = \mathbf{G}_k \mathbf{u}_k + \mathbf{G}_{-k} \mathbf{u}_{-k} + \mathbf{b}_k \tag{2}$$

where $\mathbf{G}_k$ represents the local echo paths from the local loudspeakers to the local microphones, $\mathbf{G}_{-k}$ represents the

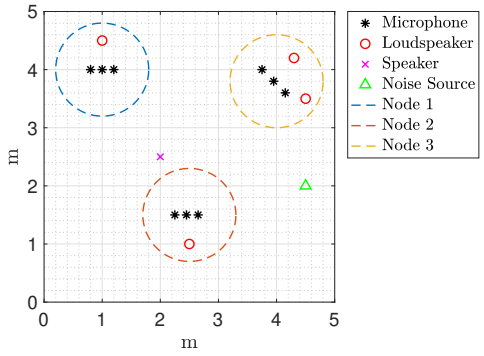

Fig. 1. Scenario used composed of three nodes each with a 3-microphone linear array. Three speech signals and one voiceless music signal were used as loudspeaker signals.

echo paths from the loudspeakers in other nodes to the local microphones and $\mathbf{u}_{-k}$ are the loudspeaker signals from the other nodes. $\mathbf{b}_k$ is the background noise. A node k has access to $\mathbf{y}_k$ only, while the remaining signals are unknown. The following vectors are also defined $\tilde{\mathbf{s}}_k = \left[\mathbf{s}_k^H \ \mathbf{0}_{1\times l_k}\right]^H$, $\check{\mathbf{n}}_k = \left[\mathbf{n}_k^H \ \mathbf{u}_k^H\right]^H$ and $\tilde{\mathbf{a}}_k = \left[\mathbf{a}_k^H \ \mathbf{0}_{1\times l_k}\right]^H$, where $\mathbf{0}_{1\times l_k}$ is an l_k-dimensional all-zero vector and $(\cdot)^H$ denotes the conjugate transpose operator. The M-dimensional vectors, where $M = \sum_{k=1}^{K} M_k$, $\mathbf{y}, \mathbf{s}, \mathbf{n}$ and $\mathbf{a}$ are the stacked versions of $\mathbf{y}_k, \tilde{\mathbf{s}}_k, \check{\mathbf{n}}_k$ and $\tilde{\mathbf{a}}_k$ respectively, such that the signal vector $\mathbf{y}_k$ can be generalized as follows

$$\mathbf{y} = \mathbf{s} + \mathbf{n} = \mathbf{a}\check{s} + \mathbf{n}. \tag{3}$$

3. CENTRALIZED COMBINED AEC AND NR WITHOUT PRIOR KNOWLEDGE

The node-specific task for node k is to estimate the desired signal d_k, defined as the desired speech component in the node's first 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 the correct speech component in $\mathbf{s}$. The mean squared error (MSE) between the desired signal and the filtered microphone and loudspeaker signals is minimized to obtain an optimal filter

$$\check{\mathbf{w}}_k = \arg\min_{\mathbf{w}_k} E\{||d_k - \mathbf{w}_k^H\mathbf{y}||^2\} \tag{4}$$

where $E\{\cdot\}$ is the expected value operator. The solution to this is the well-known MWF [8, 9]. given by

$$\check{\mathbf{w}}_k = \mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{\mathbf{y}d_k} = \mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{\mathbf{ys}}\mathbf{e}_{d_k} = \mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{\mathbf{ss}}\mathbf{e}_{d_k} \tag{5}$$

where $\mathbf{R}_{\mathbf{yy}} = E\{\mathbf{y}\mathbf{y}^H\}$, $\mathbf{R}_{\mathbf{y}d_k} = E\{\mathbf{y}d_k^H\}$, $\mathbf{R}_{\mathbf{ss}} = E\{\mathbf{s}\mathbf{s}^H\}$ and $\mathbf{R}_{\mathbf{ys}} = E\{\mathbf{y}\mathbf{s}^H\}$ are signal correlation matrices. The final expression in (5) is obtained based on the assumption that $\check{s}$ and $\mathbf{n}$ are uncorrelated. $\mathbf{R}_{\mathbf{ss}}$ is not directly observable and must be estimated in a robust way. 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, [10] i.e.

$$\begin{aligned}\hat{\mathbf{R}}_{\mathbf{yy}}(l) &= \beta\hat{\mathbf{R}}_{\mathbf{yy}}(l-1) + (1-\beta)\mathbf{y}(l)\mathbf{y}^H(l) \\ \hat{\mathbf{R}}_{\mathbf{nn}}(l) &= \beta\hat{\mathbf{R}}_{\mathbf{nn}}(l-1) + (1-\beta)\mathbf{y}(l)\mathbf{y}^H(l)\end{aligned} \tag{6}$$

where $\mathbf{y}(l)$ represents $\mathbf{y}$ at frame l. The forgetting factor $0 < \beta < 1$ can be chosen depending on the variation of the statistics of the signals i.e. if 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. It is assumed that $\mathbf{R}_{\mathbf{nn}}$ is full rank which implies that the loudspeaker signals are assumed to be different 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 $\mathbf{R}_{\mathbf{ss}}$ based on the criterion [1, 8]

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \underset{\substack{\text{rank}(\mathbf{R}_{\mathbf{ss}})=1\\ \mathbf{R}_{\mathbf{ss}}\succeq 0}}{\arg\min} \left|\left|\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} - \mathbf{R}_{\mathbf{ss}}\right)\hat{\mathbf{R}}_{\mathbf{nn}}^{-\mathrm{H}/2}\right|\right|_F^2 \tag{7}$$

where $||\cdot||_F$ denotes the Frobenius norm. Spatial prewhitening, also called orthogonalisation, 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 (7) is based on a generalized eigenvalue decomposition (GEVD) of the matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}}\}$ [8, 11]

$$\begin{aligned}\hat{\mathbf{R}}_{\mathbf{yy}} &= \mathbf{Q}\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{Q}^{\mathrm{H}} \\ \hat{\mathbf{R}}_{\mathbf{nn}} &= \mathbf{Q}\mathbf{\Sigma}_{\mathbf{nn}}\mathbf{Q}^{\mathrm{H}}\end{aligned} \tag{8}$$

where $\mathbf{\Sigma}_{\mathbf{yy}}$ and $\mathbf{\Sigma}_{\mathbf{nn}}$ are diagonal matrices and $\mathbf{Q}$ is an invertible matrix. The speech correlation matrix estimate $\hat{\mathbf{R}}_{\mathbf{ss}}$ is then [8]

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \mathbf{Q}\text{diag}\{\sigma_{y_1} - \sigma_{n_1}, 0, \ldots, 0\}\mathbf{Q}^{\mathrm{H}} \tag{9}$$

where σ_{y_1} and σ_{n_1} are the diagonal elements of $\mathbf{\Sigma}_{\mathbf{yy}}$ and $\mathbf{\Sigma}_{\mathbf{nn}}$, respectively, corresponding to the largest ratio $\sigma_{y_i}/\sigma_{n_i}$. 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.

4. CENTRALIZED COMBINED AEC AND NR WITH PRIOR KNOWLEDGE

Exploiting the prior knowledge that $\mathbf{R}_{\mathbf{ss}}$ has a specific zero structure (cfr. definition of $\mathbf{s}$ ans $\tilde{\mathbf{s}}_k$) , (7) can be redefined as

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \underset{\substack{\text{rank}(\mathbf{R}_{\mathbf{ss}})=1\\ \mathbf{B}^{\mathrm{H}}\mathbf{R}_{\mathbf{ss}}\mathbf{B}=0\\ \mathbf{R}_{\mathbf{ss}}\succeq 0}}{\arg\min} \left|\left|\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} - \mathbf{R}_{\mathbf{ss}}\right)\hat{\mathbf{R}}_{\mathbf{nn}}^{-\mathrm{H}/2}\right|\right|_F^2 \tag{10}$$

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

$$\mathbf{B}_k = \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{l_k} \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & 0 & \dots & 0 \\ 0 & \mathbf{B}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{B}_K \end{bmatrix}, \tag{11}$$

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

$$\begin{aligned} \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} &= \hat{\mathbf{Q}}\mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\hat{\mathbf{Q}}^{\mathrm{H}} \\ \hat{\mathbf{R}}_{\hat{\mathbf{n}}\hat{\mathbf{n}}} &= \hat{\mathbf{Q}}\mathbf{\Sigma}_{\hat{\mathbf{n}}\hat{\mathbf{n}}}\hat{\mathbf{Q}}^{\mathrm{H}} \end{aligned} \tag{12}$$

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

$$\mathbf{C} = \underset{\text{s.t.} \quad \mathbf{H}^H\mathbf{C}=\mathbf{I}_L}{\arg\min} \ \mathrm{trace}\{\mathbf{C}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{C}\} \tag{13}$$

with $L = \sum_{k=1}^{K} l_k$ and $\mathbf{H}$ is a block diagonal matrix with k^{th} diagonal block

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{I}_{m_k} \\ \mathbf{0} \end{bmatrix}, \ \mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & 0 & \dots & 0 \\ 0 & \mathbf{H}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{H}_K \end{bmatrix}, \tag{14}$$

such that $\mathbf{H}^{\mathrm{H}}\mathbf{H} = \mathbf{I}$ and $\mathbf{B}^H\mathbf{H} = \mathbf{0}$. Hence $\mathbf{C}$ is defined based on a generalised sidelobe canceller (GSC) implementation as [1, 12]

$$\mathbf{C} = \mathbf{H} - \mathbf{B}\mathbf{F} \tag{15}$$

$$\mathbf{F} = (\mathbf{B}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{H}. \tag{16}$$

where the filter $\mathbf{F}$ operates over the loudspeaker signals and effectively serves as an AEC filter which cancels the echo components in the so-called fixed beamformer outputs corresponding to $\mathbf{H}$. In practice, $\mathbf{F}$ can also be implemented adaptively via a normalized least mean square (NLMS) algorithm as shown in [12, 2].

The prior knowledge speech correlation matrix $\hat{\mathbf{R}}_{\mathbf{ss}}$ i.e. the solution to (10), is [1, 8]

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \mathbf{H}\hat{\mathbf{Q}}\mathrm{diag}\{\sigma_{\hat{y}_1} - \sigma_{\hat{n}_1}, 0, \dots, 0\}\hat{\mathbf{Q}}^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}, \tag{17}$$

where $\sigma_{\hat{y}_1}$ and $\sigma_{\hat{n}_1}$ are the 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. (12)), the filter $\check{\mathbf{w}}_k$ can finally be expressed as [1]

$$\check{\mathbf{w}}_k = \mathbf{C}\check{\mathbf{W}}_{GEVD}\mathbf{H}^{\mathrm{H}}\mathbf{e}_{d_k} \tag{18}$$

$$\check{\mathbf{W}}_{GEVD} = \hat{\mathbf{Q}}^{-H}\mathrm{diag}\left\{1 - \frac{\sigma_{\hat{n}_1}}{\sigma_{\hat{y}_1}}, 0, \dots, 0\right\}\hat{\mathbf{Q}}^{\mathrm{H}}. \tag{19}$$

5. DISTRIBUTED COMBINED AEC AND NR WITH PRIOR KNOWLEDGE

In the distributed processing approach, the prior knowledge GEVD-based DANSE (PK-GEVD-DANSE) algorithm [1] is used and each node instead of broadcasting M_k microphone and loudspeaker signals, broadcasts 2 fused signals, a desired signal reference and a noise reference signal. In the context of AEC, the second fused signal is a fused loudspeaker signal. Each node performs the estimations based on its local microphone and loudspeaker signals and the received fused signals from the other nodes. The fused signals for node k are

$$z_k = \mathbf{p}_k^H\mathbf{y}_k \tag{20}$$

$$\underline{z}_k = \boldsymbol{\lambda}_k^H\mathbf{B}_k^H\mathbf{y}_k = \boldsymbol{\lambda}_k^H\mathbf{u}_k \tag{21}$$

where $\mathbf{p}_k$ is an M_k- dimensional fusion vector and $\boldsymbol{\lambda}_k$ is an l_k-dimensional fusion vector. Then each node has access to the following signal vector $\tilde{\mathbf{y}}_k = \left[\mathbf{y}_k^H \ \mathbf{z}_{-k}^H \ \underline{\mathbf{z}}_{-k}^H\right]^H$, where the subscript $_{-k}$ refers to the concatenation of the fused signals of nodes other than k, which can be expressed as $\mathbf{z}_{-k} = [z_1^H \dots z_{k-1}^H z_{k+1}^H \dots z_K^H]^H$ 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

$$\tilde{\mathbf{H}}_k = \begin{bmatrix} \mathbf{H}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{K-1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \ \tilde{\mathbf{B}}_k = \begin{bmatrix} \mathbf{B}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{K-1} \end{bmatrix} \tag{22}$$

where $\tilde{\mathbf{H}}_k$ is an $(M_k + 2K - 2) \times (m_k + K - 1)$ matrix and $\tilde{\mathbf{B}}_k$ is an $(M_k + 2K - 2) \times (l_k + K - 1)$ matrix. Then equations (15) and (16) become respectively

$$\mathbf{C}_k = \tilde{\mathbf{H}}_k - \tilde{\mathbf{B}}_k\mathbf{F}_k \tag{23}$$

$$\mathbf{F}_k = (\tilde{\mathbf{B}}_k^H\hat{\mathbf{R}}_{\tilde{\mathbf{n}}_k\tilde{\mathbf{n}}_k}\tilde{\mathbf{B}}_k)^{-1}\tilde{\mathbf{B}}_k^H\hat{\mathbf{R}}_{\tilde{\mathbf{n}}_k\tilde{\mathbf{n}}_k}\tilde{\mathbf{H}}_k \tag{24}$$

where $\hat{\mathbf{y}}_k = \mathbf{C}_k^H\tilde{\mathbf{y}}_k$, $\hat{\mathbf{R}}_{\hat{\mathbf{n}}_k\hat{\mathbf{n}}_k} = \mathbf{C}_k^H\hat{\mathbf{R}}_{\tilde{\mathbf{n}}_k\tilde{\mathbf{n}}_k}\mathbf{C}_k$ and $\hat{\mathbf{R}}_{\hat{\mathbf{y}}_k\hat{\mathbf{y}}_k} = \mathbf{C}_k^H\hat{\mathbf{R}}_{\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k}\mathbf{C}_k$.The fusion vectors are defined as [1]

$$\boldsymbol{\lambda}_k = [\mathbf{I}_{l_k} \ \mathbf{0}](\tilde{\mathbf{B}}_k^H\hat{\mathbf{R}}_{\tilde{\mathbf{n}}_k\tilde{\mathbf{n}}_k}\tilde{\mathbf{B}}_k)^{-1}\tilde{\mathbf{B}}_k^H\hat{\mathbf{R}}_{\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k}\tilde{\mathbf{w}}_k. \tag{25}$$

$$\mathbf{p}_k = [\mathbf{I}_{M_k} \ \mathbf{0}]\tilde{\mathbf{w}}_k \tag{26}$$

where

$$\tilde{\mathbf{w}}_k = \mathbf{C}_k\check{\mathbf{W}}_{GEVD,k}\tilde{\mathbf{H}}^H[1 \ \mathbf{0}]^H. \tag{27}$$

with $\check{\mathbf{W}}_{GEVD,k}$ defined similar to (19) 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 (20) and (21) using their current fusion vectors. One node then updates its fusion vectors by means of (22)-(27). When the nodes update sequentially in a round-robin fashion (e.g. one node updates per time frame) the local signal estimates $\tilde{\mathbf{w}}_k^H\tilde{\mathbf{y}}_k$ have been shown to converge in each node to the centralized signal estimates obtained with (18) [1].

6. SIMULATIONS

This section outlines the simulations carried out using the MWF without PK, PK-MWF and PK-GEVD-DANSE algorithms described in sections 3, 4 and 5, respectively. The scenario depicted in Figure 1 was used. The performance of the algorithm is 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 [13] 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 distances of the arrays was set to 20cm for all the nodes. The microphone signals have been created such that the signal-to-interference ratio at microphone 1 in node 2 was equal to -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. For the batch-mode simulations, the full-length signals were used to estimate the correlation matrices in (8), (12) and $\{\hat{\mathbf{R}}_{\hat{\mathbf{y}}_k\hat{\mathbf{y}}_k}, \hat{\mathbf{R}}_{\hat{\mathbf{n}}_k\hat{\mathbf{n}}_k}\}$ in section 5 by selecting the time frames where the desired speech signal was active and not active, respectively, based on a perfect VAD. All nodes in Figure 1 had a loudspeaker reproducing a speech signal which were simultaneously active when the desired speech signal was not. The second loudspeaker in node 3 was reproducing a voiceless 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 reproducing babble noise was simulated as well.

PK-GEVD-DANSE updates every node simultaneously with a relaxation factor $\alpha_{rS} = 0.9$, to guarantee convergence as suggested in [10, 14]. The ERLE was computed with non-overlapping windows of 1024 samples and the average over the time frames is shown in Figure 2. The SNR is shown in Figure 3. Both metrics were computed for the first microphone in each node. The GEVD acronym was not included for brevity in the legends. The SNR was computed by filtering the noise component at each microphone signal with the filter obtained at the end of the iterations for each implementation. Figure 4 shows the mean squared error for the three discussed algorithms at the first microphone of each node, and it can be seen that in all nodes including the PK reduces the error in the estimation of the desired speech signal. It can also be seen that the PK-GEVD-DANSE algorithm converges to the PK-MWF algorithm when simultaneous updates in the nodes are performed.

7. CONCLUSIONS

It has been shown that PK-GEVD-DANSE can achieve a good performance for distributed combined AEC and NR in WASN. When GEVD-DANSE is applied and no prior knowledge is included in the formulation, this fundamentally consists of applying NR using the loudspeaker signals as extra sensor signals without considering that they do not have a desired speech component. The GEVD estimation is performed with l_k fewer signals in PK-GEVD-DANSE than in GEVD-DANSE [14], but PK-GEVD-DANSE needs to broadcast one signal more than GEVD-DANSE. It has also been shown that PK-GEVD-DANSE algorithm performing simultaneous updates in the nodes converges to the PK-MWF algorithm.

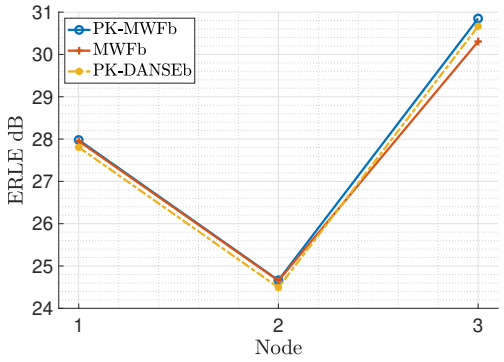

Fig. 2. Average ERLE computed at the first microphone of each node.

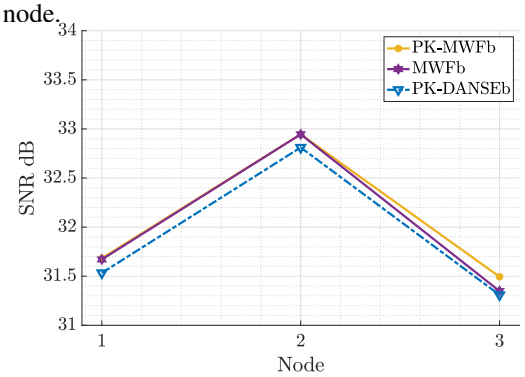

Fig. 3. SNR computed at the first microphone of each node.

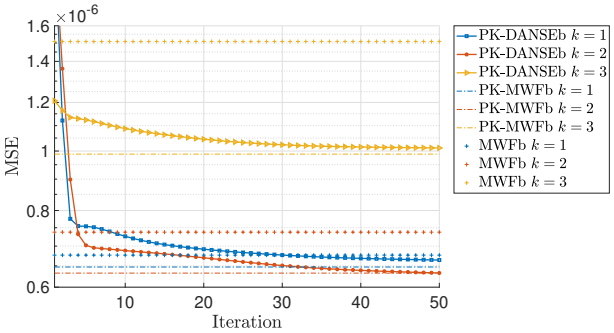

Fig. 4. Mean squared error for MWFb, PK-MWFb and PK-DANSEb at each node.

8. REFERENCES

[1] Robbe Van Rompaey and Marc Moonen, "Distributed adaptive node-specific signal estimation in a wireless sensor network with partial prior knowledge of the desired source steering vector," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019.

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

[3] Jacob Benesty, Jesper Rindom Jensen, Mads Graesboll Christensen, and Jingdong Chen, *Speech enhancement: A signal subspace perspective*, Elsevier, 2014.

[4] Eric Böhmler, Jürgen Freudenberger, and Sebastian Stenzel, "Combined echo and noise reduction for distributed microphones," in *2011 Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*. IEEE, 2011, pp. 98–103.

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

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

[7] Alexander Bertrand, *Signal processing algorithms for wireless acoustic sensor networks*, Ph.D. thesis, Katholieke Universiteit Leuven, 2011.

[8] Romain Serizel, Marc Moonen, Bas Van Dijk, and Jan 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.

[9] Jacob Benesty, Jingdong Chen, Yiteng Arden Huang, and Simon Doclo, "Study of the wiener filter for noise reduction," in *Speech Enhancement*, pp. 9–41. Springer, 2005.

[10] Alexander Bertrand and Marc Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 12, 2009.

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

[12] Simon Doclo, Sharon Gannot, Marc Moonen, and Ann Spriet, *Handbook on array processing and sensor networks*, chapter Acoustic beamforming for hearing aid applications, pp. 269–302, Wiley Online Library, 2008.

[13] Enzo De Sena, Niccolò Antonello, Marc Moonen, and Toon Van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 774–786, 2015.

[14] Alexander Bertrand and Marc Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks—part ii: Simultaneous and asynchronous node updating," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5292–5306, 2010.