EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Parallel processing of distributed beamforming and multichannel linear prediction for speech denoising and deverberation in wireless acoustic sensor networks

Zhe Han[1,2], Yuxuan Ke[1], Xiaodong Li[1,2] and Chengshi Zheng[1,2*]

Abstract

More and more smart home devices with microphones come into our life in these years; it is highly desirable to connect these microphones as wireless acoustic sensor networks (WASNs) so that these devices can be better controlled in an enclosure. For indoor applications, both environmental noise and room reverberation may severely degrade speech quality, and thus both of them need to be removed to improve users' experience. For this goal, this paper proposes a parallel processing framework of distributed beamforming and multichannel linear prediction (DB-BFMCLP), which consists of generalized sidelobe canceler and multichannel linear prediction for simultaneous speech dereverberation and noise reduction in WASNs. By sharing a common desired response vector, the proposed DB-BFMCLP can provide a significant reduction in communication bandwidth without sacrificing performance. The convergence guarantee of the DB-BFMCLP to its centralized implementation is derived mathematically. Simulation results verify the superiority of the proposed method to the existing related methods in noisy and reverberant scenarios.

Keywords Wireless acoustic sensor networks, Speech enhancement, Microphone arrays

1 Introduction

Recent progress in micro-electro-mechanical systems (MEMS) and wireless communications enable the development and popularization of low-cost and low-power wireless sensor networks (WSNs) [1]. A WSN usually consists of multiple nodes connected by wireless links, which has been applied to various fields including speech extraction, acoustic source localization, and acoustic event detection [2, 3]. In general, wireless acoustic sensor networks (WASNs) that can be applied for smart home devices equipped with at lease one microphone in each device. Each node also has an individual signal processing unit and a communication module to achieve monitoring, processing, and broadcasting, respectively.

Compared with conventional compact microphone arrays, a WASN comprises several nodes that are placed dispersedly and/or randomly, so that it can cover a much larger area. Besides, WASNs can enhance the robustness and the extensibility of the system by the decentralized operation [4]. The often-studied problems of WASNs are synchronization acquisition and transmission. The main factor of the synchronization problem is the offset of clocks oscillators, and many efforts have been made to solve the clock synchronization problem [5]. The other factor is the asynchronous or synchronous updating of

*Correspondence:
Chengshi Zheng
cszheng@mail.ioa.ac.cn
[1] Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190 Beijing, China
[2] University of Chinese Academy of Sciences, 100049 Beijing, China

signals and parameters in each node; it has been proved that appropriate updating would more likely lead to optimal estimators [6, 7].

For WASNs, centralized methods need to gather all observations in a fusion center. In theory, centralized methods can achieve the best performance for complete information, but these methods usually require a large communication bandwidth and computational power. Due to limited transmission bandwidth and energy resources in practice, the optimal centralized methods are often difficult if not impossible for practical applications. An alternative solution is to use distributed methods, which can achieve nearly the same performance as centralized methods, while requiring much fewer broadcast channels [6, 8–12].

Recently, many speech enhancement methods have been proposed for WASNs to reduce environmental noise. In [10], a distributed multichannel Wiener filter (DB-MWF) method was proposed for binaural hearing aids. This method considered the case of a single speech source under stationary noise scenarios assumption and only one-channel signal was transmitted from each node to the other. In [6], the coexistence of multiple speakers was considered, a distributed adaptive node-specific signal estimation (DANSE) method was proposed, which aims to obtain different outputs in each node. In [11], a linearly constrained distributed adaptive node-specific signal estimation (LC-DANSE) method was proposed, which uses a node-specific linearly constrained minimum variance (LCMV) beamformer. A distributed generalized sidelobe canceler (DB-GSC) with multiple constraints was presented for speech enhancement in [12], where the convergence of the DB-GSC to the centralized generalized sidelobe canceler (GSC) was proved. Note that the DB-GSC method was based on a specific transformation that allows reformulating the centralized beamformer as a sum of all local GSC.

Apart from noise, room reverberation may also degrade speech quality severely in an enclosure [13, 14]. For indoor speech communication applications, such as hands-free telephony, speaking to smart home devices, and conference call, microphones are often placed at a certain distance from the desired speaker. In these circumstances, microphones can receive not only the direct sound but also the reflections because of the surrounding objects and walls, where the late reflections are referred to as reverberation. It has been shown that these undesired reverberant components degrade both the performance of automatic speech recognition (ASR) systems and speech perceptual quality. To solve this problem, many dereverberation methods have been proposed [15–18]. Among these methods, the multi-channel linear prediction (MCLP) proposed in [19] is widely used for its promising performance. The weighted recursive least squares (RLS) method was introduced to accelerate the convergence rate of the filtering parameters in [19]. In addition, [20] demonstrated that the MCLP can suppress reverberation without assuming specific acoustic conditions, although it was originally proposed for single-source dereverberation under noise-free scenarios. For WASNs, dereverberation is also very important for speech enhancement. In [21, 22], two multi-channel dereverberation approaches in ad hoc microphone arrays were introduced, in which the reverberation was reduced by selecting a subset of microphones with a relatively lower level of reverberation. Unlike noise reduction methods, dereverberation methods for WASNs often ignore the constraints, e.g., the limited transmission bandwidth and energy resource.

In reverberant and noisy environments, dereverberation and noise reduction should be integrated in a parallel processing framework or in a serial processing framework [23, 24]. In [23], a system was proposed that employs multiple-output MCLP followed by the minimum variance distortionless response (MVDR) beamformer. However, the cascade architecture of the system has high computational complexity and is difficult to extend to the WASNs. In [24], the sidelobe-cancelation (SC) filter was combined with the linear prediction (LP) filter to a unified framework named integrated sidelobe cancelation and linear prediction (ISCLP), where the two filters are estimated jointly by a Kalman filter. However, the GSC performance is highly dependent on the quality of the estimated relative early transfer functions (RETFs). To prevent the self-cancelation phenomenon caused by inaccurate RETFs, the filter coefficients of the GSC update only when the speakers are all inactive. Therefore, the filter of the GSC and that of the MCLP cannot update their coefficients simultaneously, especially when considering that the MCLP needs to update its filter coefficients when the speakers are active [12, 19, 25]. A joint optimization of the two filters is still unsolved for WASNs.

To solve the above difficulty, we unify the GSC and the MCLP together into a beamforming and multichannel linear prediction (BFMCLP) framework, which can achieve the independent update for both filters, to deal with reverberant speech in noisy scenarios. Besides, by sharing the common response vector and deriving the distributed RLS method, we extend the BFMCLP to a distributed implementation (DB-BFMCLP) which is potential for the WASNs. The DB-BFMCLP method needs much fewer signals to be broadcasted in each node than centralized methods.

The remainder of this paper is organized as follows. In Section 2, the problem formulation is presented. In Section 3, the centralized BFMCLP method is described. The

DB-BFMCLP is presented in Section 4, and its convergence to the BFMCLP is also included in this section. In Section 5, we evaluate the performance by simulations. Finally, some conclusions are given in Section 6.

2 Problem formulation

In this section, we consider that a fully connected WASN with M microphones contains J nodes ($M \geq J$), and the number of speech sources observed by this WASN is N. M_j denotes the number of microphones in the jth node, and we have $\sum_{j=1}^{J} M_j = M$. This paper focuses on the situation that each node is equipped with more than one microphone. It should be noted that no communication-bandwidth reduction can be obtained in the node with only one microphone, since at least one-channel signal needs to be transmitted if all nodes are used for better performance instead of using only partial nodes. In some previous studies, more attention is paid to the problems about sensor subset selection, source location, or network topology in the area of the WASNs consisted of several signal-microphone nodes [8, 26, 27]. These problems are out of the scope of this paper.

In the short-time Fourier transform (STFT) domain, the reverberant observation of the speech signal from the nth speaker captured by the mth microphone can be modeled as

$$x_{nm}(k,t) = \sum_{l=0}^{L_h-1} a_{nm}(k,l) s_n(k,t-l), \tag{1}$$

where t and k denote the time-frame and frequency-bin indices, respectively. $a_{nm}(k,l)$ denotes the time-invariant acoustic transfer function (ATF) between the nth source and the mth microphone, and L_h depends on the reverberation time and the length of the STFT window. In this paper, we treat all frequency sub-bands independently, the frequency-bin index k is hereafter omitted for brevity. In WASNs, we use the vector notation:

$$\mathbf{x}_n(t) = [x_{n1}(t), x_{n2}(t), ..., x_{nM}(t)]^T, \tag{2}$$

with $(\cdot)^T$ denoting the transpose. By dividing ATFs coefficients, the reverberant speech components from the nth speaker $\mathbf{x}_n(t)$ may be decomposed into the direct and early reflected components $\mathbf{x}_{n|e}$ and late reverberant components $\mathbf{x}_{n|l}$, given by:

$$\mathbf{x}_n(t) = \mathbf{x}_{n|e}(t) + \mathbf{x}_{n|l}(t). \tag{3}$$

In practice, the ATFs are difficult to estimate without the knowledge of the acoustic sources. Instead, the RETFs are often chosen to characterize the relative relationship of the desired source signals received by microphones:

$$\mathbf{x}_{n|e}(t) = \mathbf{h}_n \left[\mathbf{x}_{n|e}(t)\right]_1, \tag{4}$$

where $[\cdot]_1$ denotes the first item of the vector, $\mathbf{h}_n$ denotes an $M \times 1$ RETF between the nth speaker and M microphones in the WASN. It is obvious that $[\mathbf{h}_n]_1 = 1$. Consider all the N speakers and the $M \times 1$ vector $\mathbf{v}(t)$ represents the environmental noise, the stacked $M \times 1$ vector of received signals by all microphones is given by:

$$\begin{aligned} \mathbf{y}(t) =& \mathbf{x}_e(t) + \mathbf{x}_l(t) + \mathbf{v}(t) \\ =& \sum_{n=1}^{N} \mathbf{x}_{n|e}(t) + \sum_{n=1}^{N} \mathbf{x}_{n|l}(t) + \mathbf{v}(t), \end{aligned} \tag{5}$$

where $\mathbf{x}_e(t) = \mathbf{H}\left[\left[\mathbf{x}_{1|e}(t)\right]_1, \left[\mathbf{x}_{2|e}(t)\right]_1, ..., \left[\mathbf{x}_{N|e}(t)\right]_1\right]^T$, and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N]$ is the $M \times N$ RETFs matrix for all the N speakers.

In the Jth node WASN, the vectors $\mathbf{y}(t)$ and $\mathbf{h}_n$, and the matrix $\mathbf{H}$ can be stacked by all nodes:

$$\mathbf{y}(t) = \left[\bar{\mathbf{y}}_1^T(t), \bar{\mathbf{y}}_2^T(t), ..., \bar{\mathbf{y}}_J^T(t)\right]^T, \tag{6}$$

$$\bar{\mathbf{y}}_j(t) = \left[y_{j1}(t), y_{j2}(t), ..., y_{jM_j}(t)\right]^T, \tag{7}$$

$$\mathbf{h}_n = \left[\bar{\mathbf{h}}_{n1}^T, \bar{\mathbf{h}}_{n2}^T, ..., \bar{\mathbf{h}}_{nJ}^T\right]^T, \tag{8}$$

$$\mathbf{H} = \begin{bmatrix} \bar{\mathbf{H}}_1 \\ \bar{\mathbf{H}}_2 \\ \vdots \\ \bar{\mathbf{H}}_J \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{h}}_{11} & \bar{\mathbf{h}}_{21} & \cdots & \bar{\mathbf{h}}_{N1} \\ \bar{\mathbf{h}}_{12} & \bar{\mathbf{h}}_{22} & \cdots & \bar{\mathbf{h}}_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{h}}_{1J} & \bar{\mathbf{h}}_{2J} & \cdots & \bar{\mathbf{h}}_{NJ} \end{bmatrix}, \tag{9}$$

where $\bar{(\cdot)}$ denotes the local data belonging to one node, $y_{ji}(t)$ denotes the ith microphone signal of the jth node. The vectors $\bar{\mathbf{y}}_j(t) \in \mathbb{C}^{M_j \times 1}$ and $\bar{\mathbf{h}}_{nj} \in \mathbb{C}^{M_j \times 1}$ denote the signal captured by the jth node and the RETF from the nth speaker to the jth node, respectively.

3 BFMCLP

In this section, we develop the parallel processing of the BFMCLP for simultaneous speech dereverberation and noise reduction. We introduce the BFMCLP at the beginning, and then investigate its stability.

3.1 Framework

The parallel processing framework of the BFMCLP is shown in Fig. 1. It consists of GSC and MCLP, and the microphone signal vector $\mathbf{y}(t)$ is used as input to both parallel branches. As shown in the block diagram, the GSC consists of three components: a fixed beamformer (FB) $\mathbf{f}$ steers a beam to a desired speaker and reduces

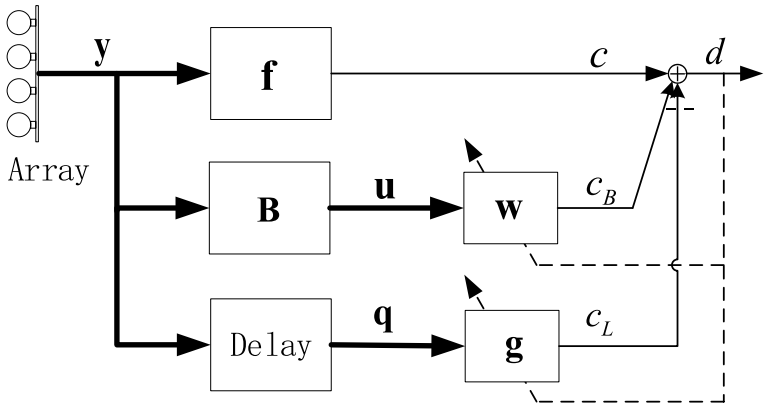

Fig. 1 Block-diagram of the BFMCLP

the other competing speakers, a blocking matrix (BM) **B** which is orthogonal to the target signal cancels the desired speaker, and a data-dependent adaptive filter **w** filters the output of **B**. The difference between the signals from the FB path and the adaptive filter **w** filter path is the original GSC output [28].

The accuracy of the estimated RETFs matrix has a significant impact on the performance of the GSC. If the desired speech can be completely canceled in the BM, the GSC performs well in suppressing noise, interferences, and late reverberant components without distorting the desired speech. An estimation of the RETF for one speaker can be obtained by performing eigenvector decomposition on the corresponding covariance matrix and the eigenvector associated with the maximum eigenvalue is then extracted [25, 29]. Beforehand, the desired covariance matrix needs to be computed by subtracting the noise covariance matrix from the noisy covariance matrix. We assume that the activity patterns of the speakers are non-overlapping, and an ideal voice activity detector (VAD) is employed in this paper. In this way, the desired covariance matrices can be obtained at the initialization stage of the WASN.

However, the accuracy of estimated RETFs will decrease significantly with the increase of the reverberation time. To prevent the speech cancelation problem caused by inaccurate RETFs, the adaptive filter $\mathbf{w}(t)$ only updates when the desired speaker is inactive, whereas such an update strategy may lead to performance degradation for dereverberation. To overcome this problem, the MCLP is introduced to suppress reverberation by deconvolution in the second branch, which consists of a delay module and an estimated room regression vector **g**. Note that Eq. (1) indicates that the reverberation effect can be modeled as the output of a multi-channel autoregressive (MCAR) system. It is the theoretical basis of the adaptive dereverberation method, where the microphone array signals can be expressed as the model of MCLP [30, 31]. In this section, we propose the BFMCLP method, in which the GSC and the MCLP are performed in parallel. In this way, we can achieve much better performance when the speech degrades by both reverberation and noise. The details of the BFMCLP method are presented below.

The FB **f** can be defined with the following constraints set:

$$\mathbf{H}^H\mathbf{f} = \mathbf{p}, \tag{10}$$

where $(\cdot)^H$ in the following denote conjugate transpose, and **p** is an $N \times 1$ desired response vector consisting of ones and zeros. The desired output $d(t)$ of the BFMCLP is the sum of the direct and early reflected components of the desired speakers which correspond to 1 in the vector **p**:

$$d(t) = \left[x_{1|e}(t), x_{2|e}(t), ..., x_{N|e}(t)\right]\mathbf{p}. \tag{11}$$

A closed-form solution of Eq. (10) is $\mathbf{f} = \mathbf{H}\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{p}$, and the output of the FB is:

$$c(t) = \mathbf{f}^H\mathbf{y}(t) = d(t) + \mathbf{f}^H(\mathbf{x}_l(t) + \mathbf{v}(t)). \tag{12}$$

Let the BM $\mathbf{B} \in \mathbb{C}^{M\times(M-N)}$ be defined as a basis for the orthogonal complement of the space spanned by the columns of matrix **H**, it is designed to cancel the desired speakers, given by

$$\mathbf{B}^H\mathbf{H} = \mathbf{0}_{(M-N)\times N}, \tag{13}$$

and a closed-form solution of Eq. (13) can be written as $\mathbf{B} = \left[\mathbf{I} - \mathbf{H}\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\right]_{:,1:M-N}$. The output of the BM can be given by:

$$\mathbf{u}(t) = \mathbf{B}^H\mathbf{y}(t) = \mathbf{B}^H(\mathbf{x}(t) + \mathbf{v}(t)). \tag{14}$$

In the MCLP branch, $\mathbf{q}(t)$ denotes the delayed signal of $\mathbf{y}(t)$:

$$\mathbf{q}(t) = \left[\mathbf{q}_1^T(t), ..., \mathbf{q}_M^T(t)\right]^T, \tag{15}$$

$$\mathbf{q}_m(t) = \left[y_m(t-\tau), ..., y_m\left(t-\tau-\left(L_g-1\right)\right)\right]^T, \tag{16}$$

where L_g depends mainly on L_h, and τ denotes the prediction delay in the MCLP model which can prevent the over whitening problem [31]. As shown in Fig. 1, the output of BFMCLP $\hat{d}(t)$ can be given by:

$$\begin{aligned}\hat{d}(t) =& c(t) - c_B(t) - c_L(t)\\ =& c(t) - \mathbf{w}^H(t-1)\mathbf{u}(t) - \mathbf{g}^H(t-1)\mathbf{q}(t),\end{aligned} \tag{17}$$

where $\hat{d}(t)$ denotes the estimation of the desired speaker, $\mathbf{w}(t)$ and $\mathbf{g}(t)$ update independently over time. In the BFMCLP method, the two branches are designed for joint dereverberation and noise reduction.

The filter coefficients $\mathbf{w}(t)$ and $\mathbf{g}(t)$ update iteratively with the normalized least mean squares (NLMS) [32] and RLS [33], respectively, and the details are summarized in Table 1.

As shown in Table 1, $\mathbf{k}(t)$ is the gain vector, $\mathbf{P}(t)$ is the inverse correlation matrix of the input signal $\mathbf{q}(t)$, $0 < \alpha < 1$ and $0 < \rho < 1$ denote the forgetting factors, $\lambda(t)$ denotes the variance of the desired signal, and μ is the step size. It is to be emphasized that $\mathbf{w}(t)$ only updates when all the speakers are inactive, while $\mathbf{g}(t)$ updates continuously. In this way, instead of estimating both filters simultaneously, the BFMCLP can prevent the self-cancelation problem effectively.

3.2 Stability of the BFMCLP

Especially note that there is no distortion for the desired signals in the output of the BFMCLP. Because of the existence of the BM $\mathbf{B}$ and the prediction delay τ, we have $\mathrm{E}\{\mathbf{u}(t)d^*(t)\} = \mathbf{0}$ and $\mathrm{E}\{\mathbf{q}(t)d^*(t)\} = \mathbf{0}$ in which $\mathrm{E}\{\cdot\}$ denotes the expectation, indicating that $c_b(t)$ and $c_L(t)$ are all uncorrelated with the desired signal $d(t)$.

In this subsection, we will further prove that the independent update of the two paths will not cause divergence of the system. We assume that the microphone signals are composed of speakers and one interference radiating from a specific direction:

$$\mathbf{y}(t) = \mathbf{x}_e(t) + \mathbf{x}_l(t) + \mathbf{v}_e(t) + \mathbf{v}_l(t), \tag{18}$$

where $\mathbf{v}_l(t)$ and $\mathbf{v}_e(t)$ are the early-reflected components and late-reverberant components of the noise, respectively. We assume the RETFs are known. In the following, we analyze the system in two situations: the desired speaker is active or inactive.

Table 1 The details of the BFMCLP method

1 Initialization $\rightarrow \mathbf{g}(0) = [\mathbf{0}]_{ML_g\times 1}$, $\mathbf{P}(0) = [\mathbf{0}]_{ML_g\times ML_g}$, $\mathbf{w}(0) = [\mathbf{0}]_{(M-N)\times 1}$, $P(0) = \|\mathbf{u}(1)\|$.

2 The estimation of the desired signal and the update of $\mathbf{w}(\mathbf{t})$ and $\mathbf{g}(\mathbf{t})$ for $t = 1, 2, \cdots$ in all sub-bands.

(a) Calculate the desired signal as

$$c_B(t) = (\mathbf{B}\mathbf{w}(t-1))^H \mathbf{y}(t), \tag{57}$$

$$c_L(t) = \mathbf{g}^H(t-1)\mathbf{q}(t), \tag{58}$$

$$c(t) = \mathbf{f}^H \mathbf{y}(t), \tag{59}$$

$$\hat{d}(t) = c(t) - c_B(t) - c_L(t). \tag{60}$$

(b) Estimate $\lambda(t)$

$$\lambda(t) = \left\|[\mathbf{y}(t)]_1\right\|^2. \tag{61}$$

(c) Update the adaptive filter $\mathbf{w}(t)$ and room regression vector $\mathbf{g}(t)$

$$P(t) = \rho P(t-1) + (1-\rho)\|\mathbf{u}(t)\|^2, \tag{62}$$

$$\mathbf{w}(t) = \mathbf{w}(t-1) + \mu\frac{\mathbf{u}(t)\hat{d}^*(t)}{P(t)}, \tag{63}$$

$$\mathbf{k}(t) = \frac{\mathbf{P}(t-1)\mathbf{q}(t)}{\alpha\lambda(t) + \mathbf{q}^H(t)\mathbf{P}(t-1)\mathbf{q}(t)}, \tag{64}$$

$$\mathbf{g}(t) = \mathbf{g}(t-1) + \mathbf{k}(t)\hat{d}^*(t), \tag{65}$$

$$\mathbf{P}(t) = \frac{\mathbf{P}(t-1) - \mathbf{k}(t)\mathbf{q}^H(t)\mathbf{P}(t-1)}{\alpha}. \tag{66}$$

(d) $t = t+1$, update $\mathbf{q}(t)$.

3 End.

3.2.1 Speaker active

When the speaker is active, the filter coefficients of the GSC branch are all fixed. Because the RETFs are estimated in advance and $\mathbf{w}(t)$ updates when only the noise exists, the GSC can suppress the early-reflected components of the noise without distorting the desired speaker. Thus, the output of the GSC branch is

$$\begin{aligned} c_{\mathrm{GSC}}(t) =& c(t) - c_B(t) \\ =& (\mathbf{f} - \mathbf{B}\mathbf{w})^H \mathbf{y}(t) \\ =& d(t) + (\mathbf{f} - \mathbf{B}\mathbf{w})^H (\mathbf{x}_l(t) + \mathbf{v}_l(t)), \end{aligned} \tag{19}$$

The input of the MCLP branch $\mathbf{q}(t)$ is always correlated to $(\mathbf{x}_l(t) + \mathbf{v}_l(t))$, and the MCLP aims to suppress the late-reverberation components by making the output $\hat{d}(t) = c_{\mathrm{GSC}}(t) - c_L(t)$ temporally uncorrelated [20].

3.2.2 Speaker inactive

When the speaker is inactive, the vector $\mathbf{y}(t) = \mathbf{v}_e(t) + \mathbf{v}_l(t)$ needs to be canceled completely. In other words, the system should minimize the $\mathrm{E}\left\{\left|\hat{d}(t)\right|^2\right\}$, where the $\hat{d}(t)$ denotes the residual in this subsection. And the filters $\mathbf{w}(t)$ and $\mathbf{g}(t)$ update independently and simultaneously over time.

The filter $\mathbf{w}(t)$, which updates by the NLMS method, should minimize the cost function [28]:

$$J_{\mathbf{w}}(t) = \|\mathbf{w}(t) - \mathbf{w}(t-1)\|^2 + \mathrm{Re}\{\lambda_{\mathbf{w}}^* \hat{d}(t)\}, \tag{20}$$

where $\lambda_{\mathbf{w}}$ is the Lagrange multiplier, and $\mathrm{Re}\{\cdot\}$ extracts the real part of a complex variable. Differentiate $J_{\mathbf{w}}(t)$ with the respect to $\mathbf{w}(t)$:

$$\frac{\partial J_{\mathbf{w}}(t)}{\partial \mathbf{w}^H(t)} = 2(\mathbf{w}(t) - \mathbf{w}(t-1)) - \lambda_{\mathbf{w}}^* \mathbf{u}(t). \tag{21}$$

By setting Eq. (21) to zero, we can obtain the optimal filter coefficients:

$$\mathbf{w}(t) = \mathbf{w}(t-1) + \frac{1}{2}\lambda_{\mathbf{w}}^* \mathbf{u}(t). \tag{22}$$

We set the constraint

$$c(t) = \mathbf{w}^H(t)\mathbf{u}(t) + \mathbf{g}^H(t-1)\mathbf{q}(t), \tag{23}$$

and solve for the $\lambda_{\mathbf{w}}$ by substituting Eq. (22) into Eq. (23), given by

$$c(t) = \mathbf{w}^H(t-1)\mathbf{u}(t) + \frac{1}{2}\lambda_{\mathbf{w}}\|\mathbf{u}(t)\|^2 + \mathbf{g}^H(t-1)\mathbf{q}(t), \tag{24}$$

then we obtain:

$$\lambda_{\mathbf{w}} = \frac{2\hat{d}(t)}{\|\mathbf{u}(t)\|^2}, \tag{25}$$

where $\hat{d}(t)$ is defined in Eq. (60) in Table 1. Thus, Eq. (63) in Table 1 can be obtained by substituting Eq. (25) into Eq. (22) and introducing a scaling factor denoted by μ, where $P(t)$ is a recursive average of $\|\mathbf{u}(t)\|^2$.

Next, we consider the filter $\mathbf{g}(t)$. In the method of least squares, the optimized $\mathbf{g}(t)$ in BFMCLP should satisfy the principle of orthogonality:

$$\mathrm{E}\{\mathbf{q}(t)\hat{d}^*(t)\} = \mathbf{0}. \tag{26}$$

Then we can get [28, 31, 33]:

$$\Phi\mathbf{g}=\mathbf{z}, \tag{27}$$

where $\Phi = \mathrm{E}\{\mathbf{q}(t)\mathbf{q}^H(t)\}$ denotes the correlation matrix of the input $\mathbf{q}(t)$, and $\mathbf{z} = \mathrm{E}\left[\mathbf{q}(t)c_{\mathrm{GSC}}^*(t)\right]$ denotes the cross-correlation vector of $\mathbf{q}(t)$ and $c_{\mathrm{GSC}}(t) = c(t) - c_B(t)$. In the RLS method, the recursive computations of Φ and $\mathbf{z}$ are given by:

$$\Phi(t) = \lambda_{\mathbf{g}}\Phi(t-1) + \mathbf{q}(t)\mathbf{q}^H(t), \tag{28}$$

$$\mathbf{z}(t) = \lambda_{\mathbf{g}}\mathbf{z}(t-1) + \mathbf{q}(t)c_{\mathrm{GSC}}^*(t), \tag{29}$$

where $\lambda_{\mathbf{g}}$ is the *forgetting factor*. Then, the matrix inversion lemma can be used to obtain the recursive computation of $\mathbf{g}(t)$, which is

$$\begin{aligned} \mathbf{g}(t) =& \Phi^{-1}(t)\mathbf{z}(t) \\ =& \mathbf{P}(t)\mathbf{z}(t) \\ =& \lambda_{\mathbf{g}}\mathbf{P}(t)\mathbf{z}(t-1) + \mathbf{P}(t)\mathbf{q}(t)c_{\mathrm{GSC}}^*(t), \end{aligned} \tag{30}$$

using $\mathbf{P}(t) = \lambda_{\mathbf{g}}^{-1}\mathbf{P}(t-1) - \lambda_{\mathbf{g}}^{-1}\mathbf{k}(t)\mathbf{q}^H(t)\mathbf{P}(t-1)$ and $\mathbf{k}(t) = \mathbf{P}(t)\mathbf{u}(t)$ [28], $\mathbf{g}(t)$ in Eq. (65) in Table 1 can be obtained:

$$\begin{aligned} \mathbf{g}(t) =& \mathbf{P}(t-1)\mathbf{z}(t-1) + \mathbf{k}(t)\mathbf{u}^H(t)\mathbf{P}(t-1)\mathbf{z}(t-1) + \mathbf{P}(t)\mathbf{q}(t)c_{\mathrm{GSC}}^*(t) \\ =& \mathbf{g}(t-1) + \mathbf{k}(t)\left[c_{\mathrm{GSC}}^*(t) - \mathbf{q}^H(t)\mathbf{g}(t-1)\right] \\ =& \mathbf{g}(t-1) + \mathbf{k}(t)\hat{d}^*(t). \end{aligned} \tag{31}$$

In summary, using the output of the BFMCLP $\hat{d}(t)$ as the residual for updating the two branches, the whole system can converge towards the optimal solution.

4 DB-BFMCLP

In this section, we extend the BFMCLP for use in the WASNs. A simple estimation is obtained by utilizing only local signals, and the sub-optimal solution can be obtained by doing so, reducing both bandwidth and power consumption.

4.1 Framework

As illustrated in Fig. 2, the input $\mathbf{y}_j(t) \in \mathbb{C}^{(M_j+N-N_j)\times 1}$ of the jth node is the stacked vector of local signals $\bar{\mathbf{y}}_j(t) \in \mathbb{C}^{M_j\times 1}$ and the transmitted signals $\dot{\mathbf{r}}_j(t) \in \mathbb{C}^{(N-N_j)\times 1}$ from other nodes:

$$\mathbf{y}_j(t) = \left[\, \bar{\mathbf{y}}_j^T(t) \;\; \dot{\mathbf{r}}_j^T(t) \,\right]^T. \tag{32}$$

At the same time, the jth node transmits the shared signals $\mathbf{r}_j(t) \in \mathbb{C}^{N_j\times 1}$ to other nodes. $\mathbf{r}_j(t)$ is defined in such a way as follows. In a typical application scenario of WASNs, the M microphones and the N speakers are all placed randomly and dispersedly; therefore, the signal-to-noise ratios (SNRs) of microphones for each source are different. When the positions of all the speakers are fixed and the activity patterns of the speakers are non-overlapping, we can estimate the distances between each speaker and the nodes in WASN at system initialization stage by using the ideal VAD and the magnitude of the signal received by the first microphone in each node. We choose the microphone with the highest energy for the nth speaker as the reference of the nth speaker. We assume that the j_1th, j_2th ... j_{N_j}th microphones (N_j in total) in the jth node have speakers, then $\mathbf{r}_j(t)$ is written as:

$$\mathbf{r}_j(t) = \mathbf{T}_j \bar{\mathbf{y}}_j(t), \tag{33}$$

$$\mathbf{T}_j = \begin{bmatrix} \mathbf{t}_{jj_1} \\ \mathbf{t}_{jj_2} \\ \vdots \\ \mathbf{t}_{jj_{N_j}} \end{bmatrix}, \tag{34}$$

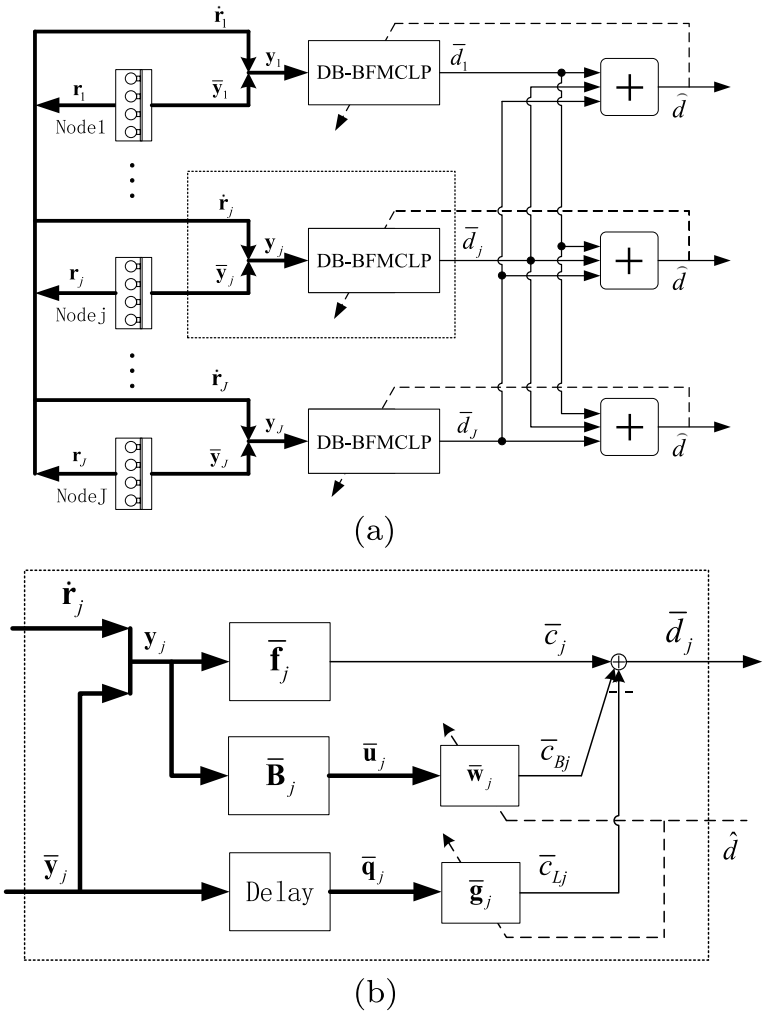

Fig. 2 Block-diagram of the DB-BFMCLP. **a** Overall block-diagram. **b** Details at the jth node

$$\mathbf{t}_{jj_i} = \left[\underbrace{0 \ \ldots \ 0}_{j_i-1} \quad 1 \quad \underbrace{0 \ \ldots \ 0}_{M_j-j_i} \right], \tag{35}$$

where $\mathbf{T}_j \in \mathbb{N}^{N_j \times M_j}$ and $\mathbf{t}_{jj_i} \in \mathbb{N}^{1 \times M_j}$. The $N \times 1$ vector $\mathbf{r}(t)$ denotes the stacked vector of all $\mathbf{r}_j$ and the $(N - N_j) \times 1$ vector $\dot{\mathbf{r}}_j$ denotes the received signal of the jth node, which can be written as:

$$\mathbf{r}(t) = \left[\mathbf{r}_1^T(t), \mathbf{r}_2^T(t), ..., \mathbf{r}_J^T(t) \right]^T, \tag{36}$$

$$\dot{\mathbf{r}}_j(t) = \left[\mathbf{r}_1^T(t), ..., \mathbf{r}_{j-1}^T(t), \mathbf{r}_{j+1}^T(t), ..., \mathbf{r}_J^T(t) \right]^T. \tag{37}$$

Note that $\sum_{j=1}^{J} N_j = N$, $0 \le N_j \le M_j$, and a microphone being selected as the reference of one speaker cannot be a reference for another. Similar to $\mathbf{y}_j(t)$, the RETFs belonging to the jth node are:

$$\mathbf{H}_j = \left[\bar{\mathbf{h}}_{1j}, \bar{\mathbf{h}}_{2j}, ..., \bar{\mathbf{h}}_{Nj} \right], \\ \dot{\mathbf{h}}_{1j}, \dot{\mathbf{h}}_{2j}, ..., \dot{\mathbf{h}}_{Nj} \tag{38}$$

$$\mathbf{h}_{nj} = \mathbf{T}_j \bar{\mathbf{h}}_{nj}, \tag{39}$$

$$\dot{\mathbf{h}}_{nj} = \left[\mathbf{h}_{n1}^T, ..., \mathbf{h}_{n(j-1)}^T, \mathbf{h}_{n(j+1)}^T, ... \mathbf{h}_{nJ}^T \right]^T. \tag{40}$$

As illustrated in Fig. 2(b), when the constraints set $\mathbf{p}$ is consistent across the WASN, the parameters of the jth node are given by:

$$\bar{\mathbf{f}}_j = \frac{1}{J} \mathbf{H}_j \left(\mathbf{H}_j^H \mathbf{H}_j \right)^{-1} \mathbf{p}, \tag{41}$$

$$\bar{\mathbf{B}}_j = \left[\mathbf{I} - \mathbf{H}_j \left(\mathbf{H}_j^H \mathbf{H}_j \right)^{-1} \mathbf{H}_j \right]_{:,1:M_j-N_j}, \tag{42}$$

$$\bar{\mathbf{u}}_j(t) = \bar{\mathbf{B}}_j^H \mathbf{y}_j(t). \tag{43}$$

Note that the input of the MCLP branch in the jth node is still $\bar{\mathbf{y}}_j$ rather than $\mathbf{y}_j$:

$$\bar{\mathbf{q}}_j(t) = \left[\mathbf{q}_{j1}^T(t), ..., \mathbf{q}_{jM_j}^T(t) \right]^T, \tag{44}$$

$$\mathbf{q}_{ji}(t) = \left[y_{ji}(t-\tau), ..., y_{ji}\left(t - \tau - \left(L_g - 1\right)\right) \right]^T. \tag{45}$$

In addition, we provide more details of the implementation of the DB-BFMCLP in one node as an example in Table 2. Note that all the signals in vector $\mathbf{r}(t)$ can be obtained in each node of the WASN. Without loss of generality, we choose the first item of $\mathbf{r}(t)$ in Eq. (72) in Table 2.

4.2 Convergence proof

In this part, we will show the convergence property of the proposed DB-BFMCLP to the BFMCLP. As mentioned in Section 3, the filters $\mathbf{w}(t)$ and $\mathbf{g}(t)$ update their coefficients independently in the BFMCLP method. Because the full convergence proof of the DB-GSC to the centralized GSC has been provided in [12], only the convergence of the MCLP branch is presented in this paper. We assume $\hat{d}(t) = c(t) - c_L(t)$ without considering the BM of the GSC and the filter $\mathbf{w}(t)$. Some parameters are introduced for clarification, for example, $\hat{d}_{cen}(t)$ represents the output of the centralized method and $\hat{d}_{dis}(t)$ denotes that of the distributed one.

In the RLS method, there are two different estimation errors, where one is the a priori estimation error and the other is the a posteriori estimation error [28]. The a priori estimation error in the BFMCLP is introduced when estimating the desired speech signal:

$$\begin{aligned} \hat{d}_{\text{cen}}(t) &= c(t) - \mathbf{g}^H(t-1)\mathbf{q}(t) \\ &= \mathbf{f}^H \mathbf{y}(t) - \mathbf{g}^H(t-1)\mathbf{q}(t). \end{aligned} \tag{46}$$

And the a posteriori estimation error is given by [28]:

$$\begin{aligned} \hat{z}_{\text{cen}}(t) &= c(t) - \mathbf{g}^H(t)\mathbf{q}(t) \\ &= c(t) - \left[\mathbf{g}(t-1) + \mathbf{k}(t)\hat{d}_{\text{cen}}^*(t) \right]^H \mathbf{q}(t) \\ &= \left(1 - \mathbf{k}^H(t)\mathbf{q}(t)\right)\hat{d}_{\text{cen}}(t), \end{aligned} \tag{47}$$

further, the ratio of the a posteriori estimation error $\hat{z}_{\text{cen}}(t)$ to the a priori estimation error $\hat{d}_{\text{cen}}(t)$ is the conversion factor $\gamma_{\text{cen}}(t)$, given by:

$$\begin{aligned} \gamma_{\text{cen}}(t) &= \frac{\hat{z}_{\text{cen}}(t)}{\hat{d}_{\text{cen}}(t)} \\ &= 1 - \mathbf{k}^H(t)\mathbf{q}(t) \\ &= 1 - \frac{\mathbf{q}^H(t)\mathbf{P}(t-1)\mathbf{q}(t)}{\alpha\lambda(t) + \mathbf{q}^H(t)\mathbf{P}(t-1)\mathbf{q}(t)}, \end{aligned} \tag{48}$$

which is determined by the input signal $\mathbf{q}(t)$ and the inverse correlation matrix $\mathbf{P}$. Note that the cost function in RLS is minimized based on the a posteriori estimation error $\hat{z}_{\text{cen}}(t)$, and it does not depend on the a priori estimation error $\hat{d}_{\text{cen}}(t)$ [28]. Obviously, $\alpha\lambda(t) > 0$ and $\mathbf{q}^H(t)\mathbf{P}(t-1)\mathbf{q}(t) > 0$ always hold, which is because $\mathbf{P}$ is a positive definite matrix. Therefore, $\gamma_{\text{cen}}(t)$ is less than 1 on average, leading to the convergence property of the RLS.

Because the common desired response vector $\mathbf{p}$, as shown in Eq. (41), is shared in the WASN, it is obvious that:

Table 2 The details of the DB-BFMCLP method at the *j*th node

1 Initialization $\bar{\mathbf{g}}_j(0) = [\mathbf{0}]_{M_j L_g \times 1}$, $\bar{\mathbf{P}}_j(0) = [\mathbf{0}]_{M_j L_g \times M_j L_g}$, $\bar{\mathbf{w}}_j(0) = [\mathbf{0}]_{(M_j - N_j)\times 1}$ and $P_j(0) = \|\bar{\mathbf{u}}_j(1)\|$ for $j = 1, \ldots, J$.

2 The estimation of the desired signal and the update of $\mathbf{w}(\mathbf{t})$ and $\mathbf{g}(\mathbf{t})$ for $t = 1, 2, \cdots$ in all sub-bands.

(a) Each node broadcasts $\mathbf{r}_j(t)$ to all other nodes and receives $\dot{\mathbf{r}}_j(t)$

(b) Calculate the desired signal as

$$c_{Bj}(t) = \left(\bar{\mathbf{B}}_j \bar{\mathbf{w}}_j(t-1)\right)^H \mathbf{y}_j(t), \tag{67}$$

$$c_{Lj}(t) = \bar{\mathbf{g}}_j^H(t-1)\,\bar{\mathbf{q}}_j(t), \tag{68}$$

$$c_j(t) = \bar{\mathbf{f}}_j^H \mathbf{y}_j(t), \tag{69}$$

$$\bar{d}_j(t) = c_j(t) - c_{Bj}(t) - c_{Lj}(t). \tag{70}$$

(c) Sum up the local outputs of all the nodes

$$\hat{d}(t) = \sum_{j=1}^{J} \bar{d}_j(t). \tag{71}$$

(d) Estimate $\lambda(t)$

$$\lambda(t) = \left\|[\mathbf{r}(t)]_1\right\|^2. \tag{72}$$

(e) Update the room regression vector $\bar{\mathbf{g}}_j(t)$ and adaptive filter $\bar{\mathbf{w}}_j(t)$ for $j = 1, \ldots, J$

$$P_j(t) = \rho P_j(t-1) + (1-\rho)\|\bar{\mathbf{u}}_j(t)\|^2, \tag{73}$$

$$\bar{\mathbf{w}}_j(t) = \bar{\mathbf{w}}_j(t-1) + \mu \frac{\bar{\mathbf{u}}_j(t)\,\hat{d}^*(t)}{J P_j(t)}, \tag{74}$$

$$\bar{\mathbf{k}}_j(t) = \frac{\bar{\mathbf{P}}_j(t-1)\,\bar{\mathbf{q}}_j(t)}{\alpha\lambda(t) + \bar{\mathbf{q}}_j^H(t)\,\bar{\mathbf{P}}_j(t-1)\,\bar{\mathbf{q}}_j(t)}, \tag{75}$$

$$\bar{\mathbf{g}}_j(t) = \bar{\mathbf{g}}_j(t-1) + \bar{\mathbf{k}}_j(t) \times \frac{\hat{d}^*(t)}{J}, \tag{76}$$

$$\bar{\mathbf{P}}_j(t) = \frac{\bar{\mathbf{P}}_j(t-1) - \bar{\mathbf{k}}_j(t)\,\bar{\mathbf{q}}_j^H(t)\,\bar{\mathbf{P}}_j(t-1)}{\alpha}. \tag{77}$$

(f) $t = t+1$, update $\bar{\mathbf{q}}_j(t)$.

3 End.

$$\sum_{j=1}^{J} \mathbf{H}_j^H \bar{\mathbf{f}}_j = \frac{1}{J}\sum_{j=1}^{J} \mathbf{p} = \mathbf{p}, \tag{49}$$

and then

$$\sum_{j=1}^{J} \bar{\mathbf{f}}_j^H \mathbf{y}_j(t) = \mathbf{f}^H \mathbf{y}(t) = c(t). \tag{50}$$

As shown in Table 2, the local output of each node in the DB-BFMCLP method can be written as:

$$\begin{aligned} \hat{d}_{\text{dis}}(t) =& \sum_{j=1}^{J} \left(\bar{c}_j(t) - \bar{\mathbf{g}}_j^H(t-1)\bar{\mathbf{q}}_j(t)\right) \\ =& \sum_{j=1}^{J} \bar{\mathbf{f}}_j^H \mathbf{y}_j(t) - \mathbf{g}^H(t-1)\mathbf{q}(t) \\ =& c(t) - \mathbf{g}^H(t-1)\mathbf{q}(t). \end{aligned} \tag{51}$$

The desired recursive equation for updating the room regression vector $\bar{\mathbf{g}}_j^H(n)$ with $j = \{1, \cdots, J\}$ is

$$\bar{\mathbf{g}}_j(t) = \bar{\mathbf{g}}_j(t-1) + \bar{\mathbf{k}}_j(t)\frac{\hat{d}_{dis}^*(t)}{J}, \tag{52}$$

where $\bar{\mathbf{k}}_j(t)$ is the gain vector of the *jth* node denoted by Eq. (75) in Table 2. For the sake of analysis, we assume that only the room regression vector of the first node updates. Then, the a posteriori output of distributed method can be denoted as

$$\begin{aligned} \hat{z}_{dis}(t) =& \left(\bar{c}_1(t) - \bar{\mathbf{g}}_1^H(t)\bar{\mathbf{q}}_1(t)\right) + \sum_{j=2}^{J}\left(\bar{c}_j(t) - \bar{\mathbf{g}}_j^H(t-1)\bar{\mathbf{q}}_j(t)\right) \\ =& c(t) - \left[\bar{\mathbf{g}}_1^H(t), \bar{\mathbf{g}}_2^H(t-1), \ldots, \bar{\mathbf{g}}_J^H(t-1)\right]\mathbf{q}(t). \end{aligned} \tag{53}$$

By substituting Eq. (51) and Eq. (52) into Eq. (53), can be further written as

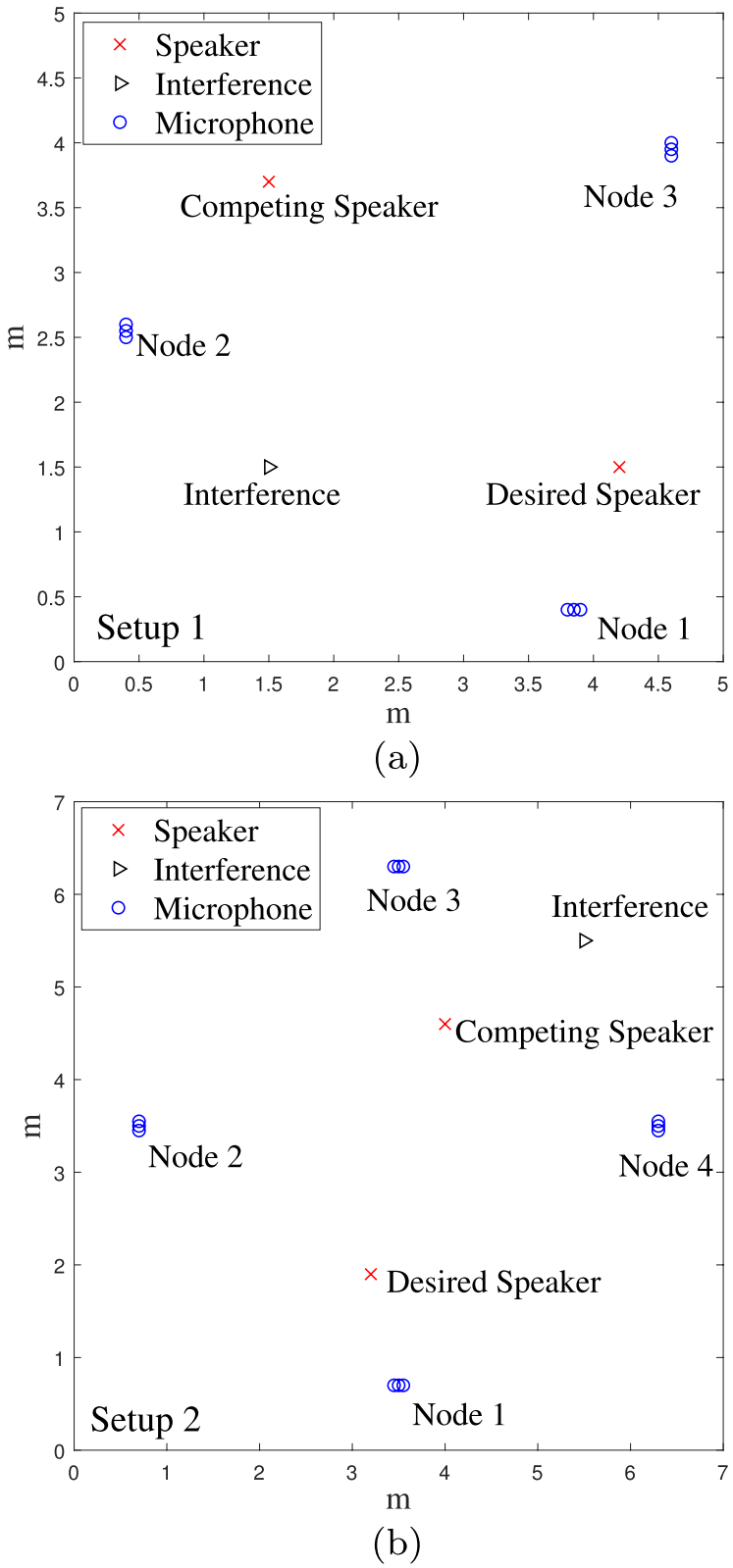

Fig. 3 Parameters of two simulated rooms. **a** T_{60} = 450ms, **b** T_{60} = 610ms, 720ms, 830ms, 940*ms*

$$\hat{z}_{\mathrm{d}is}(t) = \hat{d}_{\mathrm{d}is}(t)\left(1 - \frac{\bar{\mathbf{k}}_1^H(t)\bar{\mathbf{q}}_1(t)}{J}\right). \tag{54}$$

It is obvious that the conversion factor can be written as

$$\gamma_{\mathrm{d}is}(t) = \frac{\hat{z}_{\mathrm{d}is}(t)}{\hat{d}_{\mathrm{d}is}(t)} = 1 - \frac{\bar{\mathbf{k}}_1^H(t)\bar{\mathbf{q}}_1(t)}{J}. \tag{55}$$

Considering Eq. (55), by using the final output $\hat{d}_{\mathrm{d}is}(t)$ for updating the local prediction filter $\bar{\mathbf{g}}_j(t)$ of all nodes, the relationship between the a posteriori output and the a priori output of the distributed method is similar to the centralized method, and the conversion factor is determined by the delayed signal $\bar{\mathbf{q}}_1(t)$ and the gain vector $\bar{\mathbf{k}}_1(t)$. In contrast, if the local room regression vector updates using local output $\hat{d}_j(n)$, it is difficult to analyze the relationship. In addition, when all nodes update simultaneously, the conversion factor of distributed structure can be represented as:

$$\begin{aligned}\gamma_{\mathrm{d}is}(t) =& \frac{\hat{z}_{\mathrm{d}is}(t)}{\hat{d}_{\mathrm{d}is}(t)} \\ =& 1 - \sum_{j=1}^{J} \frac{\bar{\mathbf{k}}_j^H(t)\bar{\mathbf{q}}_j(t)}{J} \\ =& 1 - \frac{1}{J}\sum_{j=1}^{J} \frac{\bar{\mathbf{q}}_j^H(t)\bar{\mathbf{P}}_j(t-1)\bar{\mathbf{q}}_j(t)}{\alpha\lambda(t) + \bar{\mathbf{q}}_j^H(t)\bar{\mathbf{P}}_j(t-1)\bar{\mathbf{q}}_j(t)}.\end{aligned} \tag{56}$$

It is obvious that $\gamma_{\mathrm{d}is}(t)$ is less than 1 on average. Thus, the convergence of the proposed DB-BFMCLP can be guaranteed. It will be demonstrated in the following section that $\left[\bar{\mathbf{g}}_1^T, \bar{\mathbf{g}}_2^T, ..., \bar{\mathbf{g}}_J^T\right]^T$ would converge to the optimal solution of the centralized method after enough iterations.

5 Simulations

In this section, to validate the proposed BFMCLP method and the convergence of the proposed DB-BFMCLP method, the two methods are evaluated in the noisy environments with varying degrees of reverberation.

5.1 Simulation setup

The sizes of two simulated rooms are 5 m×5 m×3 m and 7 m×7 m×3 m, respectively. The reverberation time of the small room is set to $T_{60} = 450$ ms. For the big room, $T_{60} = 610$ ms, 720 ms, 830 ms, and 940 ms are considered.

Besides, each node in WASNs has 3 microphones with the distance of two adjacent microphones 5 cm. The positions of nodes, speakers, and interferences relative to the room are illustrated in Fig. 3. We select 40 speakers

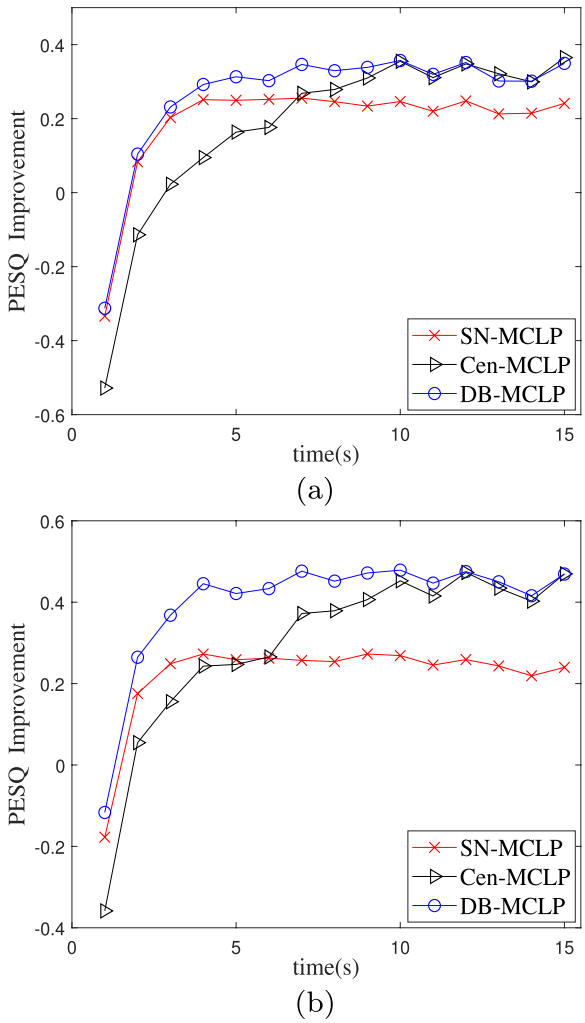

Fig. 4 Convergence of the evaluated methods along time in the term of PESQ improvement. (a) $T_{60} = 450$ms, (b) $T_{60} = 830$ms

(20 males and 20 females) from the TIMIT database as the clean speech signals. The performance shown as follow is all averaged over several experiments. Each signal of one speaker is set to 30 s, and the simulated signals are obtained by convolving simulated room impulse responses (RIRs). The RIRs are simulated with an efficient implementation of the image source model [34]. A

Table 3 The number of channels transmitted of each method per TF-bin at the *j*th node

Methods	MCLP/GSC/ LCMV/BFMCLP	LC-DANSE	DB-GSC	DB-BFMCLP
Number of channels broadcast	M_j	N	N_j	N_j
Number of channels received	$M - M_j$	$(J-1)N$	$N - N_j$	$N - N_j$

Table 4 Computational complexity of the three methods per TF-bin at the *jth* node

	Computational complexity(FLOPs)
GSC	$N^3+3MN^2+NM^2+5MN-\frac{1}{2}M^2-\frac{3}{2}N^2+\frac{13}{2}M-\frac{11}{2}N-1$
BFMCLP	$N^3+3MN^2+NM^2+5MN-\frac{1}{2}M^2-\frac{3}{2}N^2+\frac{13}{2}M-\frac{11}{2}N+2M^3{L_g}^3+\frac{11}{2}M^2{L_g}^2+\frac{11}{2}ML_g$
DB-BFMCLP	$N^3+3Q_jN^2+NQ_j^2+5Q_jN-\frac{1}{2}Q_j^2-\frac{3}{2}N^2+\frac{13}{2}Q_j-\frac{11}{2}N+2M_j{}^3{L_g}^3+\frac{11}{2}M_j{}^2{L_g}^2+\frac{11}{2}M_jL_g+J+1$

stationary noise is also located in each simulated room. To focus on measuring the performance of the proposed methods, we assume that the clocks of the sensors are synchronized. We further test whether the distributed methods can converge to the optimal solution or not by comparing the results with the centralized methods. Accordingly, we uniformly update the signals and parameters simultaneously.

The sampling rate is 16 kHz. The STFT uses a square-root Hanning window, and the frame length is set to

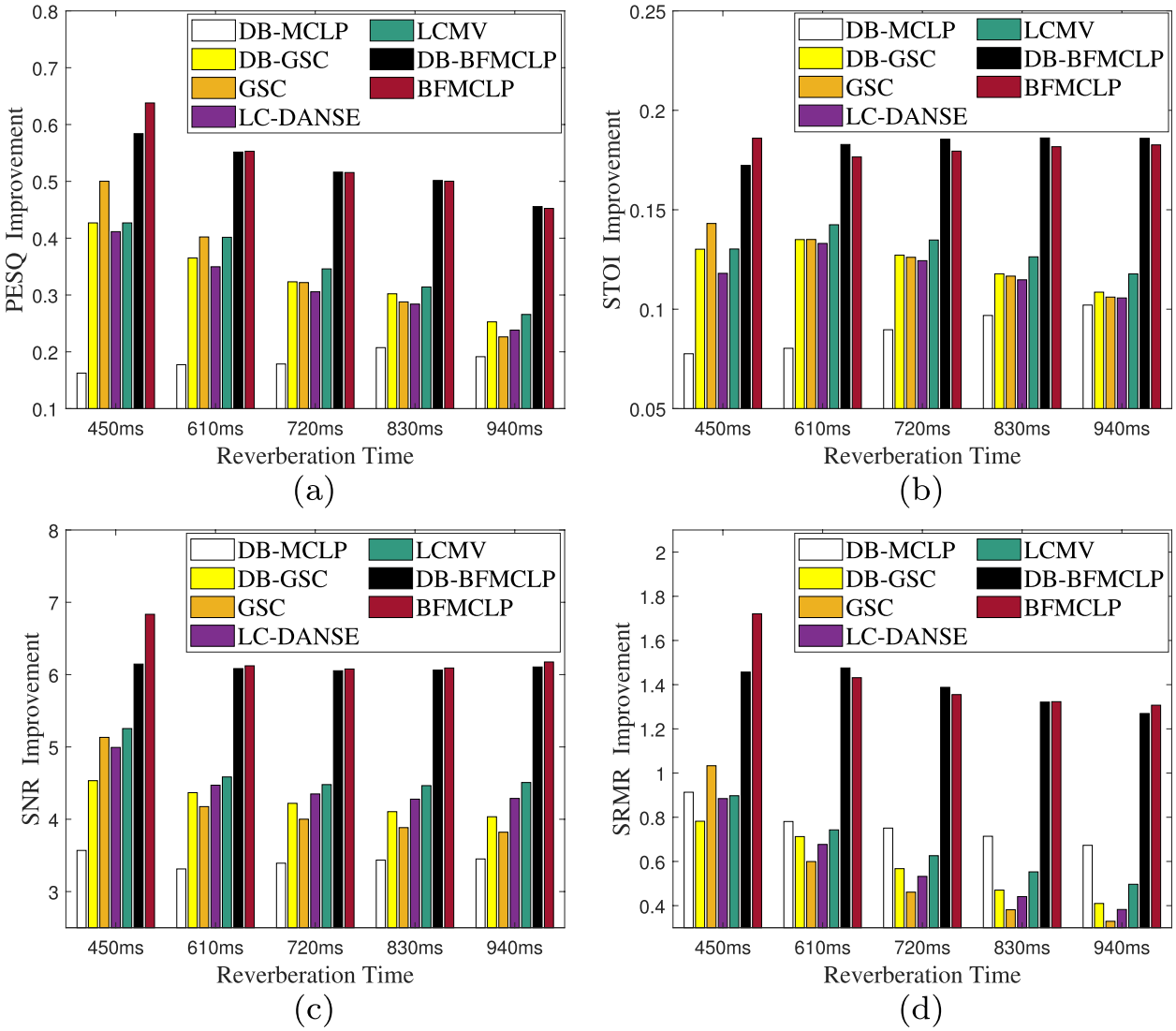

Fig. 5 Performance comparison of the evaluated methods with varying degrees of reverberation (SNR = 13 dB). **a** PESQ improvement, **b** STOI improvement, **c** SNR improvement, **d** SRMR improvement

1024 with the frame shift 512 to balance the performance and the real time of the methods in reverberant and noisy scenarios. The performance is evaluated by four often-used objective measurements including the perceptual evaluation of speech quality (PESQ) [35], the short-time objective intelligibility (STOI) [36], the SNR, and the speech-to-reverberation modulation energy ratio (SRMR) [37].

5.2 Evaluation of the distributed MCLP

We first test the convergence of the distributed MCLP in the DB-BFMCLP in reverberant scenarios, where the first setup (a) with $T_{60} = 450$ ms and the second setup (b) with $T_{60} = 830$ ms are considered. Without the GSC branch, the DB-BFMCLP and BFMCLP become the distributed MCLP (DB-MCLP) and MCLP, respectively. In the circumstances, we choose the first microphone as the reference of the single speaker, and $\mathbf{h} = [1, 0, ..., 0]^T$. The speech signals are located in the position of the desired speaker, and $L_g = 8$ and $\tau = 1$ are set in this evaluation. The PESQ improvements of the outputs of the single node MCLP (SN-MCLP), the centralized MCLP (Cen-MCLP), and the DB-MCLP versus time are depicted in Fig. 4. One can see that the performance of the Cen-MCLP and that of the DB-MCLP is closed when they are both in a convergent state and both outperform the SN-MCLP, and the convergence speed of the distributed approach is faster [38]. This is because the room regression vector $\mathbf{g}$ is separated into lower-dimension ones in the DB-MCLP.

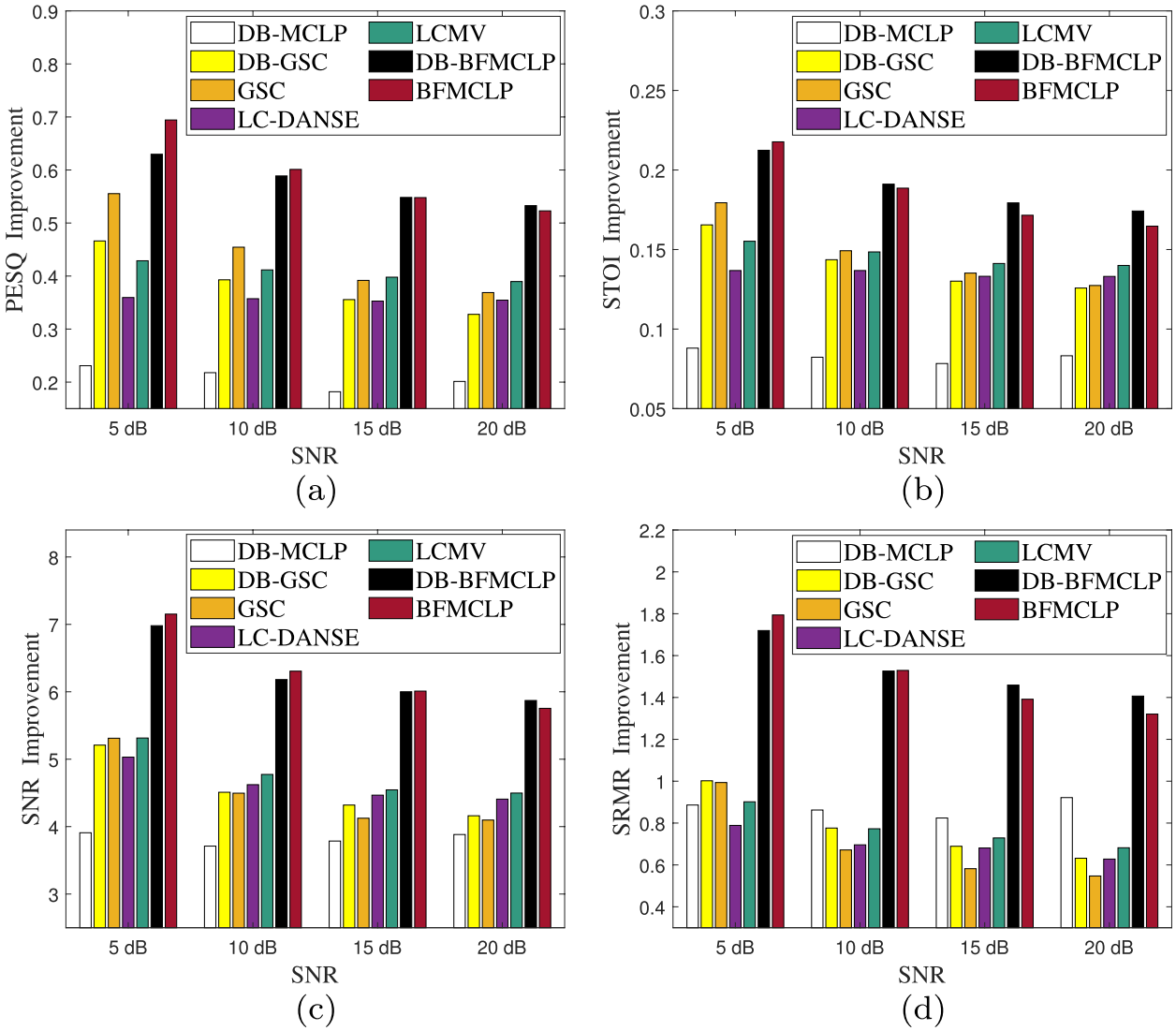

Fig. 6 Performance comparison of the evaluated methods with varying degrees of noise ($T_{60} = 610$ ms). **a** PESQ improvement, **b** STOI improvement, **c** SNR improvement, **d** SRMR improvement

5.3 Evaluation of the BFMCLP and the DB-BFMCLP

We investigate the performance of the BFMCLP and the DB-BFMCLP in noisy and reverberant scenarios by twenty runs. We compare the proposed two methods with five existing related ones. In sum, we use the following seven methods in total for complete comparison: the MCLP, the GSC, the DB-GSC (the distributed structure of the GSC), the LCMV method, the LC-DANSE method (the distributed structure of the LCMV), the BFMCLP method, and the DB-BFMCLP method. In addition, $L_g = 4$ and $\tau = 1$ are chosen in this evaluation.

The signal-to-interference ratio (SIR), which measures the power ratio between the received desired speaker and the competing speaker, is set to 0 dB. The SNR, which defines the power ratio between the speakers and the noise, is set to 13 dB in the cases when studying the influence of the reverberation time. The SNR is set to 5 dB, 10 dB, 15 dB, and 20 dB to evaluate the influence of the noise.

The channel numbers of each method per TF-bin are presented in Table 3. One can see that all of the three distributed methods need fewer channels than their centralized structures. The DB-GSC and DB-BFMCLP require that the number of speakers should not be more than the total number of microphones in the WASN; the two methods are more robust to the number of speakers because $N < M_j$ needs to be satisfied in the LC-DANSE [11].

We also show the computational complexity of the BFMCLP and the DB-BFMCLP in Table 4, where both a scalar complex addition and a scalar complex multiplication are counted as one floating point operation (FLOP) [39]. For simplicity of expression, we set $Q_j = (M_j + N - N_j)$. As a comparison, we also present the computational complexity of the existing GSC method. It can be observed from Table 4 that, because of the smaller number of filter dimensions, the complexity of the DB-BFMCLP is reduced significantly.

The improvements of the above mentioned methods with the four objective measures are presented in Figs. 5 and 6. It is clear that the performance of the DB-BFMCLP and the BFMCLP are closed in most cases, which further verifies the convergence of the DB-BFMCLP to the BFMCLP. An observation in Fig. 5 is that the impact of reverberation on speech quality gradually exceeds that of noise when the reverberation time increases, which causes the performance degradation to the existing related beamformers. Instead, the MCLP can maintain a stable performance. It demonstrates that reverberation can limit the performance of the related beamformers. However, the BFMCLP and the DB-BFMCLP have obvious advantages in all measurements under reverberant and noisy environments,

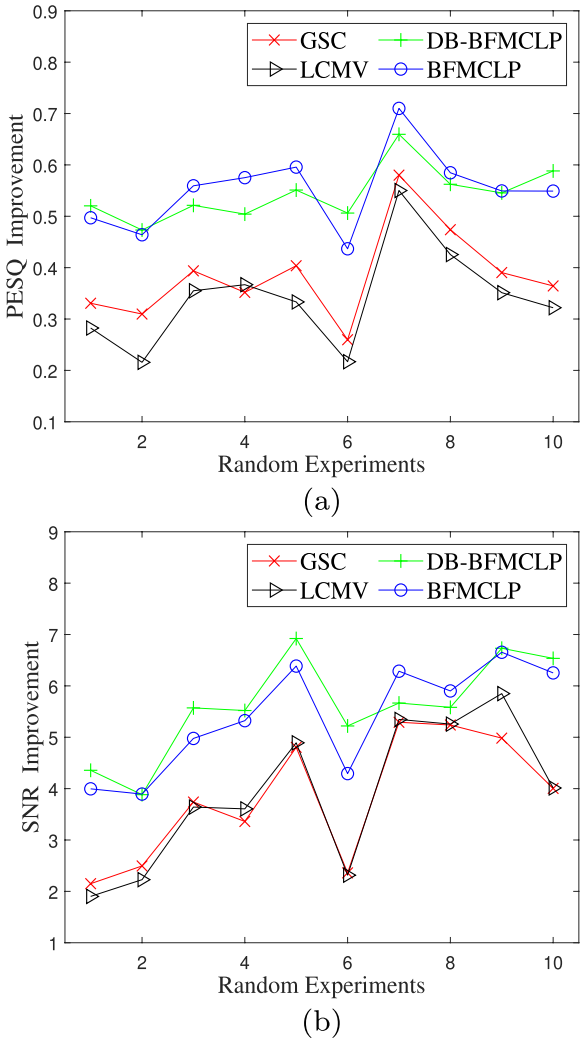

Fig. 7 Performance comparison in more general experiments. **a** PESQ improvement, **b** SNR improvement

demonstrating the superiority of the parallel structure proposed in this paper.

Furthermore, we perform ten random experiments to verify the stability of the system, where in each experiment the room size $S \in [25, 72]\,\mathrm{m}^2$, SIR $\in [-2, 2]$ dB, SNR $\in [10, 20]$ dB, and reverberation time $T_{60} \in [400, 900]$ ms are chosen randomly. Two speakers, one interference and a four-node WASN, are randomly and dispersedly arranged in the room, and the microphone constellation in each node remains fixed as in Section 5.1. The improvements depicted in Fig. 7 indicate the robustness of the DB-BFMCLP and the BFMCLP.

5.4 Evaluation of the influence of VAD errors

An ideal VAD has been used in the previous studies, and the filters and parameters are updated when speakers inactive in speech enhancement methods. In this part, we

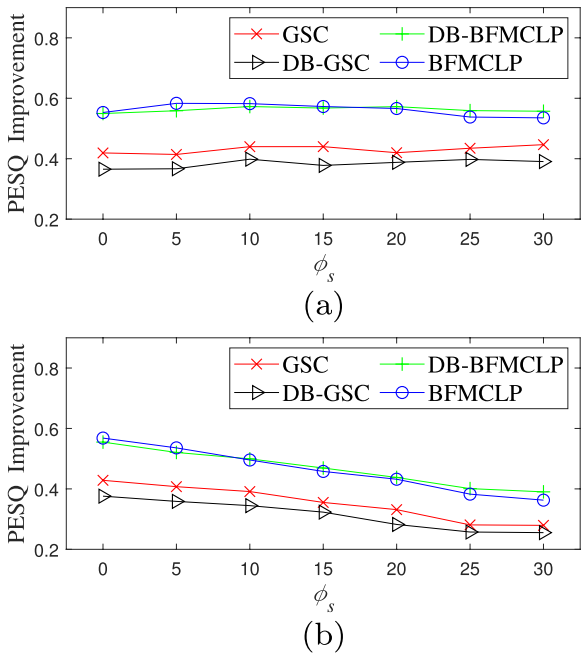

Fig. 8 PESQ improvement of the evaluated methods with varying degrees of VAD errors. **a** Accurate RETFs, **b** inaccurate RETFs

further study the influence of VAD errors on the performance of the GSC, BFMCLP, and their distributed structures for completeness. Here, ϕ_s indicates the percentage of the speech-and-noise frames that are error detected as noise-only frames.

The influence of ϕ_s on the performance of the four methods is studied in two scenarios using the simulated room depicted in Fig. 3b, with $T_{60} = 610$ ms and SNR = 13 dB. In the first scenario, we assume that the accurate $\mathbf{H}$ still has been known to all nodes; the inaccurate noise frames are only used to update the filter $\mathbf{w}$; the PESQ improvements in this scenario are depicted in Fig. 8a. In the second scenario, the inaccurate noise frames are simultaneously used to estimate the RETF $\mathbf{H}$ and the filter $\mathbf{w}$, and the results are shown in Fig. 8b. The four methods are obviously more sensitive to the estimation error of the RETFs, and the superiority of the two parallel structures to the two GSC-methods can be concluded from the Fig. 8 in either of the two scenarios.

6 Conclusion

In this paper, for speech enhancement in reverberant and noisy environments, the parallel implementation of BFMCLP method has been proposed and extended for WASNs. The proposed methods suppress reverberation and noise by exploiting the property that the delayed signal in the MCLP and the blocked signal in GSC are all uncorrelated with the desired signal. The parallel architecture has two advantages: one is that the two filters can be updated independently to prevent the self-cancelation problem effectively due to the estimation error of the RETFs, which can improve the stability of the system, and the other is that the parallel architecture can be easily extended to distributed systems. We provide the details of the two parallel methods and prove the convergence of the DB-BFMCLP method. Finally, we test the BFMCLP and the DB-BFMCLP in reverberant and noisy scenarios; simulation results indicate that the two proposed methods outperform the existing methods, and the DB-BFMCLP provides a performance comparable to the centralized BFMCLP, while it significantly reduces both the computational and the transmission cost.

Abbreviations

WSN	Wireless sensor network
WASN	Wireless acoustic sensor network
ASR	Automatic speech recognition
MEMS	Micro-electro-mechanical system
DB-MWF	Distributed multichannel Wiener filter
DANSE	Distributed adaptive node-specific signal estimation
LCMV	Linearly constrained minimum variance
LC-DANSE	Linearly constrained distributed adaptive node-specific signal estimation
GSC	Generalized sidelobe canceler
DB-GSC	Distributed generalized sidelobe canceler
MCLP	Multi-channel linear prediction
MVDR	Minimum variance distortionless response
SC	Sidelobe-cancelation
LP	Linear prediction
ISCLP	Integrated sidelobe cancelation and linear prediction
BFMCLP	Beamforming and multichannel linear prediction
DB-BFMCLP	Distributed beamforming and multichannel linear prediction
STFT	Short-time Fourier transform
RETF	Relative early transfer function
ATF	Acoustic transfer function
FB	Fixed beamformer
BM	Blocking matrix
VAD	Voice activity detector
MCAR	Multi-channel autoregressive
RLS	Recursive least squares
NLMS	Normalized least mean squares
RIR	Room impulse response
PESQ	Perceptual evaluation of speech quality
STOI	Short-time objective intelligibility
SRMR	Speech-to-reverberation modulation energy ratio
SNR	Signal-to-noise ratio
SIR	Signal-to-interference ratio
FLOP	Floating point operation

Acknowledgements
Not applicable.

Authors' contributions
Zhe Han: software and writing original draft. Yuxuan Ke: platform and writing——review and editing. Chengshi Zheng and Xiaodong Li: supervision and writing-review and editing. All authors read and approved the final manuscript.

Funding
This work was supported in part by the National Natural Science Foundation of China under Grant 62101550.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to that all of them can be generated by readers themselves according to the simulation setup in Section 5 but are available from the corresponding author on reasonable request if they have difficulties.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 25 February 2022 Accepted: 21 April 2023
Published online: 22 May 2023

References

1. D. Estrin, L. Girod, G. Pottie, M. Srivastava, in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Instrumenting the world with wireless sensor networks, vol. 4 (2001), pp. 2033–2036. https://doi.org/10.1109/ICASSP.2001.940390
2. O.M. Bouzid, G.Y. Tian, J. Neasham, B. Sharif, Investigation of sampling frequency requirements for acoustic source localisation using wireless sensor networks. Appl. Acoust. **74**(2), 269–274 (2013). https://doi.org/10.1016/j.apacoust.2010.12.013
3. R. Ali, T. van Waterschoot, M. Moonen, An integrated mvdr beamformer for speech enhancement using a local microphone array and external microphones. EURASIP J. Audio Speech Music Process. 10 (2021). https://doi.org/10.1186/s13636-020-00192-2
4. X. Guo, M. Yuan, Y. Ke, C. Zheng, X. Li, Distributed node-specific block-diagonal LCMV beamforming in wireless acoustic sensor networks. Signal Process. **185**, 108085(2021). https://doi.org/10.1016/j.sigpro.2021.108085. www.sciencedirect.com/science/article/pii/S0165168421001237
5. A. Bertrand, in *2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*. Applications and trends in wireless acoustic sensor networks: a signal processing perspective (2011). pp. 1–6. https://doi.org/10.1109/SCVT.2011.6101302
6. A. Bertrand, M. Moonen, Distributed adaptive node-specific signal estimation in fully connected sensor networks–part i: Sequential node updating. IEEE Trans. Sig. Process. **58**(10), 5277–5291 (2010). https://doi.org/10.1109/TSP.2010.2052612
7. A. Bertrand, M. Moonen, Distributed adaptive node-specific signal estimation in fully connected sensor networks–part ii: Simultaneous and asynchronous node updating. IEEE Trans. Signal Process. **58**(10), 5292–5306 (2010). https://doi.org/10.1109/TSP.2010.2052613
8. J. Zhang, R. Heusdens, R.C. Hendriks, Rate-distributed spatial filtering based noise reduction in wireless acoustic sensor networks. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(11), 2015–2026 (2018). https://doi.org/10.1109/TASLP.2018.2851157
9. S. Markovich-Golan, A. Bertrand, M. Moonen, S. Gannot, Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks. Signal Process. **107**, 4–20 (2015). https://doi.org/10.1016/j.sigpro.2014.07.014
10. S. Doclo, M. Moonen, T. Van den Bogaert, J. Wouters, Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids. IEEE Trans. Audio Speech Lang. Process. **17**(1), 38–51 (2009). https://doi.org/10.1109/TASL.2008.2004291
11. A. Bertrand, M. Moonen, Distributed node-specific LCMV beamforming in wireless sensor networks. IEEE Trans. Signal Process. **60**(1), 233–246 (2012). https://doi.org/10.1109/TSP.2011.2169409
12. S. Markovich-Golan, S. Gannot, I. Cohen, Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks. IEEE Trans. Audio Speech Lang. Process. **21**(2), 343–356 (2013). https://doi.org/10.1109/TASL.2012.2224454
13. P.A. Naylor, N.D. Gaubitch, *Speech* dereverberation. Springer London. (2010). https://doi.org/10.1007/978-1-84996-056-4
14. Z. Honghu, Y. Jia, P. Jianxin, Chinese speech intelligibility of elderly people in environments combining reverberation and noise. Appl. Acoust. **150**, 1–4 (2019). https://doi.org/10.1016/j.apacoust.2019.02.002
15. K. Lebart, J.M. Boucher, P. Denbigh, A new method based on spectral subtraction for speech dereverberation. Acta Acustica U. Acustica. **87**, 359–366 (2001)
16. A. Schwarz, K. Reindl, W. Kellermann, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, A two-channel reverberation suppression scheme based on blind signal separation and wiener filtering (2012), pp. 113–116. https://doi.org/10.1109/ICASSP.2012.6287830
17. E.A.P. Habets, J. Benesty, A two-stage beamforming approach for noise reduction and dereverberation. IEEE Trans. Audio Speech Lang. Process. **21**(5), 945–958 (2013). https://doi.org/10.1109/TASL.2013.2239292
18. M. Miyoshi, Y. Kaneda, Inverse filtering of room acoustics. IEEE Trans. Acoust. Speech Signal Process. **36**(2), 145–152 (1988). https://doi.org/10.1109/29.1509
19. T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, B. Juang, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation (2008), pp. 85–88. https://doi.org/10.1109/ICASSP.2008.4517552
20. T. Yoshioka, T. Nakatani, Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening. IEEE Trans. Audio Speech Lang. Process. **20**(10), 2707–2720 (2012). https://doi.org/10.1109/TASL.2012.2210879
21. S. Gergen, A. Nagathil, R. Martin, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Audio signal classification in reverberant environments based on fuzzy-clustered ad-hoc microphone arrays (2013), pp. 3692–3696. https://doi.org/10.1109/ICASSP.2013.6638347
22. S. Pasha, C. Ritz, in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. Clustered multi-channel dereverberation for ad-hoc microphone arrays (2015). pp. 274–278. https://doi.org/10.1109/APSIPA.2015.7415519
23. M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, T. Nakatani, Strategies for distant speech recognitionin reverberant environments (2015). https://doi.org/10.1186/s13634-015-0245-7
24. T. Dietzen, S. Doclo, M. Moonen, T. van Waterschoot, Integrated sidelobe cancellation and linear prediction kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 740–754 (2020). https://doi.org/10.1109/TASLP.2020.2966869
25. T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, T. van Waterschoot, Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(3), 544–558 (2019). https://doi.org/10.1109/TASLP.2018.2886743
26. Y. Chan, K. Ho, A simple and efficient estimator for hyperbolic location. IEEE Trans. Signal Process. **42**(8), 1905–1915 (1994). https://doi.org/10.1109/78.301830
27. Y. Zeng, R.C. Hendriks, Distributed delay and sum beamformer for speech enhancement via randomized gossip. IEEE/ACM Trans. Audio Speech, and Language Processing **22**(1), 260–273 (2014). https://doi.org/10.1109/TASLP.2013.2290861
28. S. Haykin, *Adaptive Filter Theory* (Prentice Hall, 2002)
29. I. Kodrasi, S. Doclo, in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods (2017). pp. 116–120. https://doi.org/10.1109/HSCMA.2017.7895573
30. K. Abed-Meraim, E. Moulines, P. Loubaton, Prediction error method for second-order blind identification. IEEE Trans. Signal Process. **45**(3), 694–705 (1997). https://doi.org/10.1109/78.558487
31. T. Yoshioka, T. Nakatani, K. Kinoshita, M. Miyoshi, *Speech Dereverberation and Denoising Based on Time Varying Speech Model and Autoregressive Reverberation Model* (Springer, Berlin Heidelberg, 2010), pp.151–182
32. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001). https://doi.org/10.1109/78.934132
33. T. Yoshioka, H. Tachibana, T. Nakatani, M. Miyoshi, in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Adaptive dereverberation of speech signals with speaker-position change detection (2009), pp. 3733–3736. https://doi.org/10.1109/ICASSP.2009.4960438

34. J. Allen, D. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**, 943–950 (1979). https://doi.org/10.1121/1.382599
35. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, vol. 2 (2001), pp. 749–752. https://doi.org/10.1109/ICASSP.2001.941023
36. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011). https://doi.org/10.1109/TASL.2011.2114881
37. J.F. Santos, M. Senoussaoui, T.H. Falk, in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. An improved non-intrusive intelligibility metric for noisy and reverberant speech (2014). pp. 55–59. https://doi.org/10.1109/IWAENC.2014.6953337
38. C. Zheng, A. Deleforge, X. Li, W. Kellermann, Statistical analysis of the multichannel wiener filter using a bivariate normal distribution for sample covariance matrices. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(5), 951–966 (2018). https://doi.org/10.1109/TASLP.2018.2800283
39. H. Raphael, *Floating point operations in matrix-vector calculus* (Technische Universität München, Tech. rep, 2007)

Publisher's Note