Model-Based Distributed Node Clustering and Multi-Speaker Speech Presence Probability Estimation in Wireless Acoustic Sensor Networks

Zhao, Yingke; Nielsen, Jesper Kjær; Chen, Jingdong; Christensen, Mads Græsbøll

Published in:
The Journal of the Acoustical Society of America

DOI (link to publication from Publisher):
10.1121/10.0001449

Publication date:
2020

Document Version
Early version, also known as pre-print

Citation for published version (APA):
Zhao, Y., Nielsen, J. K., Chen, J., & Christensen, M. G. (2020). Model-Based Distributed Node Clustering and Multi-Speaker Speech Presence Probability Estimation in Wireless Acoustic Sensor Networks. *The Journal of the Acoustical Society of America*, *147*(6), 4189-4201. https://doi.org/10.1121/10.0001449

Model-Based Distributed Node Clustering and Multi-Speaker Speech Presence Probability Estimation in Wireless Acoustic Sensor Networks

Yingke Zhao

Center of Intelligent Acoustics and Immersive Communications
and School of Marine Science and Technology,
Northwestern Polytechnical University,
127 Youyi West Road, Xi'an 710072, China

Jesper Kjær Nielsen

Audio Analysis Lab, CREATE, Department of Architecture, Design,
and Media Technology, Aalborg University, Aalborg 9000, Denmark.

Jingdong Chen*

Center of Intelligent Acoustics and Immersive Communications,
Northwestern Polytechnical University,
127 Youyi West Road, Xi'an 710072, China

Mads Græsbøll Christensen

Audio Analysis Lab, CREATE, Department of Architecture, Design,
and Media Technology, Aalborg University, Aalborg 9000, Denmark

(Dated: April 15, 2020)

Abstract

A great challenge in the wireless acoustic sensor network (WASN) based signal processing is to develop robust speech presence probability (SPP) estimation methods, which can work at each time frame and each frequency band. The knowledge of SPP plays an essential role in speech enhancement and noise estimation. Single channel SPP estimation and centralized multi-channel SPP estimation have been well studied. However, few efforts can be found for the distributed SPP estimation for WASN applications with multiple speakers. Accordingly, this paper presents a distributed model-based SPP estimation method for multi-speaker detection, which does not need any fusion center. A distributed k-means clustering method is first used to cluster the nodes into subnetworks, which target at detecting different speakers. For each node in the subnetwork, the speech and noise power spectral densities (PSD) are estimated locally by using a model-based method, then a distributed SPP estimator is developed in each subnetwork. A distributed consensus method is used to obtain the distributed clustering and the distributed SPP estimation. The results show that the proposed distributed clustering method can assign nodes into subnetworks based on their noisy observations. Moreover, the proposed distributed SPP estimator achieves robust speech detection performance under different noise conditions.

* Electronic mail: jingdongchen@ieee.org.

I. INTRODUCTION

A wireless acoustic sensor network (WASN) can be formed by microphones, which are randomly placed in the environment. Each node in the WASN can be a single microphone or any conventional microphone array. Compared to conventional microphone arrays, such as linear arrays, circular arrays, spherical arrays, etc., WASNs are more flexible and scalable. Another disadvantage of conventional microphone arrays is that they only sample the sound field locally. When the array is far away from the source signal, the low signal-to-noise ratio (SNR) makes satisfactory signal processing performance hard to achieve. In contrast, WASNs are able to capture more spatial information, since they can physically cover a larger space. However, WASN encounters some difficult challenges. Different nodes have different clocks and dealing with clock skew is a challenging problem. Meanwhile, the amplitude response of the acoustic transfer function between sources and different nodes may be different. Additionally, the received signal quality, such as the input signal-to-noise ratio (iSNR), is different from node to node, which may dramatically degrade the performance of traditional methods. Another challenge in WASN based signal processing method is to develop in-network processing, which is scalable regarding communication bandwidth requirements and computational complexity (Bertrand, 2011). The development of distributed optimization methods (Boyd *et al.*, 2006; Zhang and Kwok, 2014; Zhang and Heusdens, 2017) makes WASN more attractive in audio applications. Distributed speech enhancement methods, such as distributed signal estimation (Bertrand and Moonen, 2012; Szurley *et al.*, 2016), distributed Wiener filtering(de la Hucha Arce *et al.*, 2017), distributed maximum SINR filtering(Tavakoli *et al.*, 2017) and distributed minimum-variance beamforming (Markovich-Golan *et al.*, 2015), need to estimate the noise covariance matrix across nodes in order to form the optimal filter. Usually, the estimation of the noise covariance matrix is obtained in a recursive manner, and the updating is performed when the speech is absent. Therefore, speech enhancement algorithms rely on an accurate speech detection method to make the decision on whether the speech signal is present or absent. As the speech signal is always contaminated by noise, robust detection of speech from noisy observations is non-trivial, especially with non-stationary noise. The appearance of multiple speakers in the environment, which is not uncommon in real scenarios, makes the detection even more difficult. In terms of multichannel speech enhancement for different speakers, a source specific SPP needs to be obtained at each time frame and each frequency band. Although single channel SPP estimators and centralized multi-channel SPP estimators have been extensively studied (Gerkmann *et al.*, 2008; Momeni *et al.*, 2014; Souden *et al.*, 2010;Souden *et al.*, 2011; Taseska and Habets, 2014), few references can be found in the distributed case with a WASN (Hamaidi *et al.*, 2017; Hamaidi *et al.*, 2017; Bahari *et al.*, 2017). Besides, most of the existing speech detection methods only work at time segments level (Sohn *et al.*, 1999; Ramirez *et al.*, 2004; Hamaidi *et al.*, 2017; Hamaidi *et al.*, 2017; Bahari *et al.*, 2017), and most are for batch mode case.

By using a WASN, the signal processing methods can be developed either in a centralized or a distributed manner. Unlike centralized solutions, the distributed solutions do not depend on a fusion center. The long distance communication and large communication bandwidth requirements are reduced with distributed solutions in WASN, since each node only need to communicate and exchange information with its neighbours (Bertrand, 2011). With the distributed solution, the computational burden is distributed over the WASN, which avoids large amount of data processing in a fusion center (Bertrand, 2011). In (Souden *et al.*, 2011), a multichannel noise tracking method was developed, in which the multichannel speech presence probability (MC-SPP) was estimated. The experiments showed that the speech detection performance becomes better with an increasing number of microphones. Even though the results are promising, the noise tracking method needs careful initialization, and it is difficult to determine the optimal parameters which are the forgetting factors in the updating of the signal statistics and the smoothing parameter of MC-SPP. Moreover, the algorithm only functions in a centralized manner. In (Taseska and Habets, 2014), the MC-SPP estimation is applied in sound extraction by using distributed microphone arrays. However, the proposed algorithm is still a centralized solution. With the objective to develop distributed speech enhancement techniques, a robust distributed SPP estimation at each time frame and each frequency band is needed. In (Hamaidi *et al.*, 2017; Bahari *et al.*, 2017), the multi-speaker VAD problem with WASN is formed as a node clustering problem first, and then the VADs for different speakers are obtained at the clustered nodes. However, the proposed method needs a distributed eigenvalue decomposition (EVD) to enumerate the source number as well as to obtain the node clustering result, which is computationally expensive, and the distributed EVD only works in the network with a tree topology. In (Gergen *et al.*, 2015), the authors proposed a node clustering method based on fuzzy c-Means algorithm with the MFCCs and their modulation spectra of the noisy signal segments as features. The node clustering method was then applied to source separation problem in ad hoc arrays (Gergen *et al.*, 2018). However, a in-network processing derivation is missing. In (Szurley *et al.*, 2016), a topology-independent distributed adaptive node-specific signal estimation (TI-DANSE) algorithm is introduced. Compared to the distributed adaptive node-specific signal estimation (DANSE) (Bertrand and Moonen, 2010; Bertrand and Moonen, 2011; Szurley *et al.*, 2015), the TI-DANSE overcomes the problems of changing topologies and scalability of DANSE method.

In (Zhao *et al.*, 2018), we have proposed a distributed solution for a single speaker voice activity detection (VAD). A model-based noise PSD estimation method is first performed at each node locally. Based on the estimated noise PSDs, we apply the generalized likelihood ratio test (GLRT) to obtain a global decision. In this case, we find that the GLRT can be solved by applying distributed consensus methods (Zhao *et al.*, 2018). In this paper, we introduce a distributed model-based node clustering method and a distributed model-based SPP estimation method. The proposed distributed detection method, which is an extension of the distributed VAD method in (Zhao *et al.*, 2018), can get a SPP estimate per time frame and frequency bin for multiple speakers. Furthermore, the model-based SPP estimation method maintains robust detection performance even under non-stationary noise conditions. The network is first divided into subnetworks. Each subnetwork is interested in detecting a certain speaker. For distributed node clustering, we utilize a consensus based distributed k-means type method (Qin *et al.*, 2017) with distributed cluster number enumeration. In the distributed SPP estimation step, the SPP is formulated as a function of generalized likelihood ratio (GLR). In order to obtain the GLR, the noise PSD is estimated at each node locally. We can use any noise PSD estimation method in this step. Conventional PSD estimators such as the minimum statistics (MS) based method (Martin, 2001) and the minimum mean-square error (MMSE) based method (Hendriks *et al.*, 2010; Gerkmann and Hendriks, 2012) are developed to track stationary noise. However, they have limited performance under non-stationary noise conditions. In (Nielsen *et al.*, 2018), a model-based noise PSD estimator was proposed. By using a statistical model to the speech signal and noise signal, the introduced noise estimation method is able to take into account the prior spectral information of speech and different types of noise (Kavalekalam *et al.*, 2018). Due to its robust noise estimation performance with non-stationary noise, we generalize the PSD estimation method introduced in (Nielsen *et al.*, 2018) to WASN in this paper. Based on the estimated signal PSDs, the SPP estimate can be obtained by using the GLR within each subnetwork. Under this circumstance, we find that the calculation of the GLR involves a distributed averaging problem (Zhao *et al.*, 2018), which can be solved by utilizing the distributed consensus methods, such as the random gossip method (Boyd *et al.*, 2006), the alternating direction method of multipliers (ADMM) (Zhang and Kwok, 2014), or the primal-dual method of multipliers (PDMM) (Zhang and Heusdens, 2017). In the distributed SPP estimation step, besides taking the inter-band information into account, we further consider the inter-frame information to improve the detection performance.

The rest of this paper is organized as follows. Section II depicts the signal model and the problem formulation. Section III reviews the centralized detection in WASN. Section IV introduces the distributed node clustering and the distributed SPP estimation. Section V reviews the model-based signal statistics estimation method. Experimental results are then presented in Section VI. Section VII concludes the paper.

II. SIGNAL MODEL AND PROBLEM FORMULATION

The problem encountered in this paper is to detect the speech signals by using a WASN with M microphones randomly placed in a room environment, i.e., each node in the WASN is a single microphone and is interested in a specific speaker. We have Q different speakers. At time t, the signal received at the mth microphone is expressed as

$$y_m(t) = x_m(t) + v_m(t), \tag{1}$$

where $x_m(t)$ is the clean speech, $v_m(t)$ is the noise signal, where we consider the interference signal as part of the noise.

A frame of an observed signal at the mth microphone in a vector form is written as

$$\begin{aligned} \mathbf{y}_m(t) &= [y_m(t) \;\; \cdots \;\; y_m(t-T+1)]^T \\ &= \mathbf{x}_m(t) + \mathbf{v}_m(t), \end{aligned} \tag{2}$$

where $\mathbf{x}_m(t)$ and $\mathbf{v}_m(t)$ are speech signal vector and noise signal vector, respectively, which are defined similarly to $\mathbf{y}_m(t)$. As in (Nielsen *et al.*, 2018), we introduce U_x autoregressive (AR) processes to describe the speech signal $\mathbf{x}_m(t)$ and U_v AR processes to describe the noise signal $\mathbf{v}_m(t)$. The excitation variances are assumed to be unknown and the AR spectral envelopes are pre-trained and stored in the speech and noise codebooks. The speech and noise codebooks are trained by using a variation of the LPC-VQ method (Paliwal and Atal,1998; Gersho and Gray, 2012). By selecting one AR process from the speech codebook and one AR process from the noise codebook as a statistical model $\mathcal{M}_u, u = 1 \, \ldots \, U$, we have $U = U_x U_v$ statistical models in total. With the statistical model $\mathcal{M}_u, u = 1 \, \ldots \, U$, the speech signal and the noise signal can be expressed as multivariate Gaussian distributions, i.e.,

$$p(\mathbf{x}_m(t)|\sigma_{x,u}^2, \mathcal{M}_u) = \mathcal{N}(\mathbf{0}, \sigma_{x,u}^2 \mathbf{Q}_x(\mathbf{a}_u)), \tag{3}$$

$$p(\mathbf{v}_m(t)|\sigma_{v,u}^2, \mathcal{M}_u) = \mathcal{N}(\mathbf{0}, \sigma_{v,u}^2 \mathbf{Q}_v(\mathbf{b}_u)), \tag{4}$$

where $\sigma_{x,u}^2$ and $\sigma_{v,u}^2$ represent the excitation variances, and $\mathbf{Q}_x(\mathbf{a}_u)$ and $\mathbf{Q}_v(\mathbf{b}_u)$ are the gain normalized covariance matrixes, $\mathbf{a}_u = [1 \; a_u(1) \; \ldots \; a_u(P)]^T$ and $\mathbf{b}_u = [1 \; b_u(1) \; \ldots \; b_u(P)]^T$ are AR parameters of the speech signal and noise signal, respectively, and P is the AR order. The matrix $\mathbf{Q}_x(\mathbf{a}_u)$ which is the covariance matrix of an AR-process asymptotically behaves as a circulant matrix as frame length goes to infinite (Gray, 2006). Since the frame length T is much larger than the AR order P, it is reasonable to treat $\mathbf{Q}_x(\mathbf{a}_u)$ as a circulant matrix (Srinivasan *et al.*, 2007). A circulant matrix can then be diagonalized by the DFT matrix (Gray, 2006), i.e.,

$$\mathbf{Q}_x(\mathbf{a}_u) = \mathbf{F}\mathbf{D}_x(\mathbf{a}_u)\mathbf{F}^H, \tag{5}$$

where $\mathbf{F}$ is the DFT matrix with its (k,t)th element being

$$\mathbf{F}_{k,t} = \frac{1}{\sqrt{T}} \exp(\jmath 2\pi kt/T), \;\; t, k = 0 \ldots T-1, \tag{6}$$

and $[\cdot]^H$ denotes the conjugate transpose operator. $\mathbf{D}_x(\mathbf{a}_u)$ is a diagonal matrix which is given by

$$\mathbf{D}_x(\mathbf{a}_u) = (\mathbf{\Lambda}_x^H(\mathbf{a}_u)\mathbf{\Lambda}_x(\mathbf{a}_u))^{-1}, \tag{7}$$

where

$$\mathbf{\Lambda}_x(\mathbf{a}_u) = \mathrm{diag}\left(\sqrt{T}\mathbf{F}^H \begin{bmatrix} \mathbf{a}_u \\ \mathbf{0} \end{bmatrix}\right). \tag{8}$$

The matrix $\mathbf{Q}_v(\mathbf{b}_u)$ can be diagonalized in a similar way (Nielsen *et al.*, 2018). In the following sections, the detection problem is formed in the frequency domain. The fast Fourier transform (FFT) length is equal to the frame length.

A. The speech presence probability

The detection includes two parts. First, we intend to get the node clustered near one specific speaker, and then the distributed speech detection is introduced within the clustered nodes for a certain speaker.

The problem considered in this section is to develop an SPP estimate per time frame and frequency band within the clustered nodes which are near a certain speaker. We assume that the network is divided into Q subnetworks, each subnetwork is represented as a node cluster $C_q, q = 1 \dots Q$, and the nodes in cluster C_q observe source q as their dominant speech signal. The collaboration between the nodes within the cluster intends to get the SPP for a specific speech signal.

Mathematically, a speech detector is a two-state model selection problem. At frequency bin k and time frame n, we have one hypothesis $H_{C_q,0}(k,n)$ denoting that speech from the qth speaker is absent at the clustered nodes C_q, and one hypothesis $H_{C_q,1}(k,n)$ denoting that speech is present at the clustered nodes, i.e.,

$$\begin{aligned} H_{C_q,0}(k,n):\ & \bar{\mathbf{y}}_{C_q}(k,n) = \bar{\mathbf{v}}_{C_q}(k,n), \\ H_{C_q,1}(k,n):\ & \bar{\mathbf{y}}_{C_q}(k,n) = \bar{\mathbf{x}}_{C_q}(k,n) + \bar{\mathbf{v}}_{C_q}(k,n), \end{aligned} \tag{9}$$

where

$$\bar{\mathbf{y}}_{C_q}(k,n) = \left[\bar{\mathbf{y}}^T_{C_q,1}(k,n)\ \bar{\mathbf{y}}^T_{C_q,2}(k,n)\ \ldots\ \bar{\mathbf{y}}^T_{C_q,M_q}(k,n)\right]^T \tag{10}$$

contains the noisy observations in the node cluster C_q, and we have $M = \Sigma_{q=1}^{Q} M_q$. Moreover, $\bar{\mathbf{x}}_{C_q}(k,n)$ and $\bar{\mathbf{v}}_{C_q}(k,n)$ are the clean speech vector and the additive noise vector, respectively. The noisy signal vector at the m_qth node contains the N past time segments as

$$\bar{\mathbf{y}}_{C_q,m_q}(k,n) = [\mathbf{y}^T_{C_q,m_q}(k,n)\ \ldots\ \mathbf{y}^T_{C_q,m_q}(k,n-N+1)]^T, \tag{11}$$

where $\mathbf{y}_{C_q,m_q}(k,n)$ is a vector of length $2K'+1$ containing the frequency bands centered at frequency index k as

$$\mathbf{y}_{C_q,m_q}(k,n) = [Y_{C_q,m_q}(k-K',n)\ \ldots\ Y_{C_q,m_q}(k+K',n)]^T, \tag{12}$$

where $Y_{C_q,m_q}(k,n)$ is the STFT coefficient of the observation signal. Parameter K' controls the number of frequency bands which are used in the detection. Thus, $\bar{\mathbf{y}}_{C_q,m_q}(k,n)$ contains both the inter-frame and inter-band information. For the special case, $K' = 0$ and $N = 1$, $\bar{\mathbf{y}}_{C_q,m_q}(k,n)$ only has the current band and the current frame information. $\bar{\mathbf{x}}_{C_q}(k,n)$ and $\bar{\mathbf{v}}_{C_q}(k,n)$ are formed in a same way as $\bar{\mathbf{y}}_{C_q}(k,n)$. The SPP of the qth speaker is defined as

$$p_{C_q}(k,n) \triangleq p(H_{C_q,1}(k,n)|\bar{\mathbf{y}}_{C_q}(k,n)). \tag{13}$$

In order to compute (13), we use a complex Gaussian statistical model for each noisy signal STFT coefficient which can be obtained from (3) and (4). This model has been extensively used in the noise PSD estimation methods (Gerkmann and Hendriks, 2012; Cohen and Berdugo, 2002; Hendriks *et al.*, 2010). The model is given by

$$p(Y_{C_q,m_q}(k,n)|H_{C_q,0}(k,n)) = \frac{1}{\pi\phi_{V_{C_q,m_q}}(k,n)} \exp\left\{-\frac{|Y_{C_q,m_q}(k,n)|^2}{\phi_{V_{C_q,m_q}}(k,n)}\right\}, \tag{14}$$

and

$$p(Y_{C_q,m_q}(k,n)|H_{C_q,1}(k,n)) = \frac{1}{\pi(\phi_{X_{C_q,m_q}}(k,n)+\phi_{V_{C_q,m_q}}(k,n))}\exp\left\{-\frac{|Y_{C_q,m_q}(k,n)|^2}{\phi_{X_{C_q,m_q}}(k,n)+\phi_{V_{C_q,m_q}}(k,n)}\right\}, \quad (15)$$

where $\phi_{X_{C_q,m_q}}(k,n)$ and $\phi_{V_{C_q,m_q}}(k,n)$ are the speech PSD and noise PSD, respectively. In Section V, the signal PSDs will be estimated by using the model-based method (Nielsen *et al.*, 2018). We further make the assumption that $Y_{C_q,m_q}(k+\kappa, n-\eta), m_q = 1, ..., M_q, \kappa = -K', ..., K', \eta = 0, ..., N-1$ are independent given $H_{C_q,0}(k,n)$ or $H_{C_q,1}(k,n)$. Then we have

$$p(\bar{\mathbf{y}}_{C_q}(k,n)|H_{C_q,0}(k,n)) = \prod_{m_q=1}^{M_q}\prod_{\kappa=-K'}^{K'}\prod_{\eta=0}^{N-1} p(Y_{C_q,m_q}(k+\kappa,n-\eta)|H_{C_q,0}(k,n)), \quad (16)$$

$$p(\bar{\mathbf{y}}_{C_q}(k,n)|H_{C_q,1}(k,n)) = \prod_{m_q=1}^{M_q}\prod_{\kappa=-K'}^{K'}\prod_{\eta=0}^{N-1} p(Y_{C_q,m_q}(k+\kappa,n-\eta)|H_{C_q,1}(k,n)). \quad (17)$$

The GLR is defined as

$$L_G(\bar{\mathbf{y}}_{C_q}(k,n)) = \frac{p(H_{C_q,1}(k,n))}{1-p(H_{C_q,1}(k,n))}\frac{p(\bar{\mathbf{y}}_{C_q}(k,n)|H_{C_q,1}(k,n))}{p(\bar{\mathbf{y}}_{C_q}(k,n)|H_{C_q,0}(k,n))}, \quad (18)$$

where $p(H_{C_q,1}(k,n))$ is a prior SPP. By using Bayes rule, the SPP in (13) can be rewritten as

$$p_{C_q}(k,n) = \frac{L_{\mathrm{G}}(\bar{\mathbf{y}}_{C_q}(k,n))}{1+L_{\mathrm{G}}(\bar{\mathbf{y}}_{C_q}(k,n))}. \quad (19)$$

In the case of WASN, we can apply a distributed method to solve the two-model selection problem in (9). In the next section, we first introduce the centralized node clustering and centralized SPP estimation before discussing their distributed solutions.

III. CENTRALIZED DETECTION IN WASN

The appearance of multiple speakers is not uncommon in real acoustic scenarios. The WASN based signal processing method gives us an alternative way to solve the multi-speaker detection problem. The detection contains two steps: the first step is to cluster the nodes into subnetworks with each of the subnetworks interested in processing the speech signal from a certain speaker. The second step is to apply the SPP estimation within the clustered nodes to collaboratively achieve the detection objective for different speakers.

A. Centralized node clustering with source enumeration

We apply a k-means clustering method (Hartigan and Wong, 1979) to get the nodes near a certain sound source clustered as a subnetwork. We have the number of $U' = U_x + U_v$ AR spectral envelopes stored in each columns of the matrix $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_{U'}]$, $\mathbf{D}$ is also called dictionary or codebook. The AR spectral envelope $\mathbf{d}_{u'} = [d_{u'}(0)\ d_{u'}(1)\ \dots\ d_{u'}(T-1)]^T, u' = 1\ \dots\ U'$ is obtained as:

$$d_{u'}(k) = \frac{1}{\left|1 + \sum_{p=1}^{P} a_{u'}(p) \exp(\frac{-\jmath 2\pi pk}{T})\right|^2}, \tag{20}$$

where $a_{u'}(p)$ is the AR parameter. The feature used in clustering is based on the Itakura-Saito (IS) divergence between the noisy signal PSD and the PSD of each AR model in the codebook. It is shown in (Kavalekalam *et al.*, 2019) that the maximum likelihood estimates of the excitation variances for a given set of speech and noise AR coefficients is equal to maximising the IS divergence between the modelled spectrum and the noisy signal spectrum. The feature for the mth node is

$$\check{\mathbf{b}}_m(n) = \left[D_{\text{IS}}\left(\boldsymbol{\phi}_{y_m}(n), \mathbf{d}_1\right)\ \dots\ D_{\text{IS}}(\boldsymbol{\phi}_{y_m}(n), \mathbf{d}_{U'})\right]^T, \tag{21}$$

where $D_{\text{IS}}(\boldsymbol{\phi}_{y_m}(n), \mathbf{d}_{u'}),\ u' = 1, \dots, U'$ is the IS divergence, with

$$\boldsymbol{\phi}_{y_m}(n) = \frac{1}{T}\left[|Y_m(0,n)|^2\ \cdots\ |Y_m(T-1,n)|^2\right] \tag{22}$$

being the periodogram spectral estimate of the noisy signal (without loss of generality, we assume that the FFT length is equal to the signal frame length). The objective of k-means clustering is to divide the M features $\{\check{\mathbf{b}}_m(n)\}_{m=1}^{M}$ into Q clusters in which each observation is assigned to the cluster with the nearest mean. This is achieved by initializing the algorithm with Q cluster centers first. The clustering result is then obtained by iterating between the following two steps: 1) feature $\check{\mathbf{b}}_m(n)$ is assigned to its nearest cluster center $\mathbf{c}_q$; 2) the cluster center $\mathbf{c}_q$ is then recomputed as the mean of the data which is assigned to the qth cluster. Iterating between step 1) and step 2) until convergence gives the final clustering result. One of the main issues with k-means clustering is to find the proper cluster number which is usually not available in practice. In the problem encountered in this paper, the optimal number of cluster reveals the number of sources in the acoustic environment. The Calinski-Harabasz criterion (Caliński and Harabasz, 1974), which is also called the variance ratio criterion (VRC), can be utilized as a cluster validity measure to find the optimal number of clusters. We run the k-means clustering for different cluster numbers Q, and the optimal Q is then obtained by choosing the one which gives the largest VRC (Caliński and Harabasz, 1974), i.e.,

$$\mathrm{VRC}(Q) = \frac{\mathrm{BGSS}(M-Q)}{\mathrm{WGSS}(Q-1)}, \tag{23}$$

where BGSS is the between-group (cluster) sum of squares, and WGSS is the within-group (cluster) sum of squares. These are given by

$$\mathrm{BGSS} = \sum_{q=1}^{Q} M_q \|\mathbf{c}_q - \mathbf{c}(n)\|^2 \tag{24}$$

and

$$\mathrm{WGSS} = \sum_{q=1}^{Q}\sum_{m=1}^{M} \mu_{m,q} \|\check{\mathbf{b}}_m(n) - \mathbf{c}_q\|^2, \tag{25}$$

where $\mathbf{c}(n) = (1/M)\sum_{m=1}^{M} \check{\mathbf{b}}_m(n)$ indicates the mean of all the features in the WASN, and

$$\mu_{m,q} = \begin{cases} 1, \text{ if } \check{\mathbf{b}}_m(n) \ \in C_q, \ q = 1 \ \ldots \ Q \\ 0, \text{ otherwise} \end{cases} . \tag{26}$$

From the definitions of WGSS and BGSS, we can notice that compact and separated clusters have small WGSS as well as large BGSS which leads to large value of VRC.

After the node clustering, the nodes which have their received signal dominated by a certain speaker are clustered as a subnetwork. The collaboration between nodes within the subnetwork achieves the SPP estimate for a certain speaker.

B. Centralized SPP estimation

As nodes in the network have been clustered into subnetworks by using the method introduced in Section III A, SPP estimation is then applied within each subnetwork to detect a certain speaker. This section formulates the centralized SPP estimation problem in the subnetwork.

By taking the logarithm in (18) and with (16), (17), we have

$$\ln L_{\mathrm{G}}(\bar{\mathbf{y}}_{C_q}(k,n)) = \\ \sum_{m_q=1}^{M_q} \sum_{\kappa=-K'}^{K'} \sum_{\eta=0}^{N-1} \ln \left[\frac{p(Y_{C_q,m_q}(k+\kappa, n-\eta)|H_{C_q,1}(k,n))}{p(Y_{C_q,m_q}(k+\kappa, n-\eta)|H_{C_q,0}(k,n))} \right] + \ln \left[\frac{p(H_{C_q,1}(k,n))}{1-p(H_{C_q,1}(k,n))} \right]. \tag{27}$$

(27) shows that the log GLR function is the summation of local information at each node in the subnetwork. By using the centralized method, every node in the network send their local information to a fusion center, the calculation of $\ln L_{\mathrm{G}}(\bar{\mathbf{y}}_{C_q}(k,n))$ and SPP in (19) is then performed in the fusion center.

IV. DISTRIBUTED DETECTION IN WASN

In Section III, we have introduced the main procedure of detecting a certain speaker in the WASN, but the derivation is carried out in a centralized way. In this section, we will discuss the distributed node clustering and distributed SPP estimation by rewriting them into averaging problems which can be solved by using distributed optimization.

A. Distributed node clustering

As discussed in Section III A, a k-means algorithm can be used to cluster the nodes by using the feature of the noisy observation signal at each node. For applications with a WASN, such as distributed noise reduction and distributed beamforming, a distributed clustering algorithm is needed. The main issue with the k-means algorithm is to update the new centers at each iteration. To update the new center, we need to calculate the mean of the features which are assigned to a certain cluster. This can be obtained by solving the distributed averaging problem. In order to get the means of the clusters, we need the sum of the features in each cluster as well as the number of nodes which are assigned to that cluster. To do that, we introduce a matrix $\mathbf{R}_m$ and a vector $\mathbf{r}_m$ which are held by each node m. If the feature hold by a certain node is assigned to cluster q, the matrix of size $U' \times Q$ has the following form

$$\mathbf{R}_m = [\mathbf{0} \ \ldots \ \check{\mathbf{b}}_m \ \ldots \ \mathbf{0}], \tag{28}$$

with its qth column being the feature at node m, and the other entries being zeros, where $\check{\mathbf{b}}_m$ is defined in (21). Moreover,

$$\mathbf{r}_m = [0 \ \ldots \ 1 \ \ldots \ 0]^T \tag{29}$$

is a vector with Q elements with its qth element being 1 and zeros elsewhere. In each iteration of the k-means clustering, the average of matrix $\mathbf{R}_m$ in the whole network will give us the scaled sum of the data in each cluster, i.e.,

$$\begin{aligned}\mathbf{R} &= \frac{1}{M}\sum_{m=1}^{M}\mathbf{R}_m \\ &= \frac{1}{M}\left[\sum_{m\in C_1}\check{\mathbf{b}}_m \ \ldots \ \sum_{m\in C_Q}\check{\mathbf{b}}_m\right],\end{aligned} \tag{30}$$

and the average of vector $\mathbf{r}_m$ will have the scaled number of the nodes at each cluster, i.e.,

$$\begin{aligned}\mathbf{r} &= \frac{1}{M}\sum_{m=1}^{M}\mathbf{r}_m \\ &= \frac{1}{M}[M_1 \ \dots \ M_Q]^T.\end{aligned} \tag{31}$$

By dividing the qth column of matrix $\mathbf{R}$ by the qth element of $\mathbf{r}$ gives us the updated center of the qth cluster.

Since each update in the k-means clustering iteration can be obtained by calculating averages in the network, we then briefly summarize the solution of averaging problem with distributed optimization in the following part. The network can be described as a graph $G = (\mathcal{V}, \mathcal{E})$ which has sets of nodes (vertices) $\mathcal{V}$ connected by edges $\mathcal{E}$. Equations (30) and (31) can be obtained by solving an averaging problem in the graph, i.e.,

$$e_{\text{ave}} = \frac{1}{M}\sum_{i\in\mathcal{V}} e_i, \tag{32}$$

where e_{ave} is the average of the local values $e_i,\ i = 1, \dots, M$. In (30), the local value e_i is matrix $\mathbf{R}_m$. Similarly, in (31), the local value e_i is vector $\mathbf{r}_m$. Standard consensus propagation algorithms, such as random gossip (Boyd *et al.*, 2006), ADMM (Zhang and Kwok, 2014) and PDMM (Zhang and Heusdens, 2017), can be used to obtain an estimate of e_{ave} distributedly. Since PDMM converges faster than random gossip and ADMM (Zhang and Heusdens, 2017), we apply the asynchronous PDMM method in this paper. With the asynchronous updating scheme, only the variables associated with one node in the graph update their estimates while all other variables keep their estimates fixed (Zhang and Heusdens, 2017). The averaging problem in (32) is equivalent to solving a quadratic optimization problem as follows:

$$\min_{\chi_i}\sum_{i\in\mathcal{V}}\frac{1}{2}(\chi_i - e_i)^2 \quad \text{s.t.} \quad \chi_i = \chi_j \quad \forall(i,j)\in\mathcal{E}. \tag{33}$$

The optimal solution to (33) is $\chi_1^\star = \chi_2^\star = \dots = \chi_M^\star = e_{\text{ave}}$. With e_i being $\mathbf{R_m}$ in (33), the solution is $\chi_1^\star = \dots = \chi_M^\star = \mathbf{R}$. Similarly, with e_i being $\mathbf{r_m}$, the solution to (33) is $\chi_1^\star = \dots = \chi_M^\star = \mathbf{r}$. The PDMM method first constructs an augmented primal-dual

Lagrangian function for the original optimization problem in the graph, and then iteratively approaches one saddle point of the constructed function (Zhang and Heusdens, 2017). At iteration $g+1$, the updating of the asynchronous PDMM to solve the problem in (33) can be derived as

$$\hat{\chi}_i^{g+1} = \frac{p_i + \sum_{j\in\mathcal{N}_i}\left(\gamma_1\hat{\chi}_j^g + A_{ij}\hat{\lambda}_{j|i}^g\right)}{1+|\mathcal{N}_i|\gamma_1} \quad i \in \mathcal{V}, \tag{34}$$

$$\hat{\lambda}_{i|j}^{g+1} = \hat{\lambda}_{j|i}^g - \frac{1}{\gamma_2}\left(A_{ji}\hat{\chi}_j^k + A_{ij}w_i^{g+1}\right) \quad \forall j \in \mathcal{N}_i, \tag{35}$$

where

$$w_i^{g+1} = \frac{\sum_{j\in\mathcal{N}_i}\left(\hat{\chi}_j^g + \gamma_2 A_{ij}\hat{\lambda}_{j|i}^g\right) + \gamma_2 e_i}{|\mathcal{N}_i| + \gamma_2}, \tag{36}$$

where $\mathcal{N}_i$ denotes the set of all the neighbouring nodes of node i. In the following of this paper, the neighbouring nodes of a node are selected as its on-hop neighbours with a certain maximum communication distance. The auxiliary node variables $\hat{\lambda}_{i|j}$ and $\hat{\lambda}_{j|i}$ are node related, $\hat{\lambda}_{i|j}$ is owned by node i and it is related to node j. The parameters γ_1 and γ_2 are primal scalar and dual scalar, respectively. With the averaging problem in (33), the edge-function is $\chi_i = \chi_j$, the variables A_{ij} and A_{ji} are related to the edge-function which are $(A_{ij}, A_{ji}) = (1, -1) \ \ \forall(i,j) \in \mathcal{E}, i < j$. More details can be found in (Zhang and Heusdens, 2017). The asynchronous PDMM method is briefly reviewed as follows: 1) the estimate of e_{ave}, i.e., $\hat{\chi}_i$, is initialized as e_i at the ith node; 2) in each time slot, node i is randomly selected to be active; 3) node i updates its estimate of e_{ave} and the node variables by using (34) and (35); 4) node i then send $(\hat{\chi}_i, \hat{\lambda}_{i|j})$ to its corresponding one-hop neighbours $j \in \mathcal{N}_i$. After the convergence of the PDMM, each node will obtain an accurate estimate of the average. The distributed node clustering based on PDMM is summarized in Algorithm 1. After applying the distributed node clustering for different cluster number Q, the optimal value of Q is chosen as the one which gives the largest VRC. It can be noticed from (30) that $\mathbf{c}(n)$ is actually the sum of each row of matrix $\mathbf{R}$. Since $\mathbf{R}$ is available at each node after distributed node clustering, then BGSS can be obtained locally after the k-means clustering has converged. In (25), the calculation of WGSS is an averaging problem in the WASN

Algorithm 1 Node clustering with distributed k-means

Description:

1: Randomly choose data from $e_i, i \in \mathcal{V}$ to initialized the cluster centers at each node.
2: **for** $h = 1 \ldots H$
3: Each node assigns its feature $\check{\mathbf{b}}_m$ to the nearest cluster center, and generates $\mathbf{R}_m$ and $\mathbf{r}_m$ based on the local assignment result.
   Apply PDMM to calculate (30) and (31):
4: **for** $g = 1, 2, 3, ..., G^{'}$
5: Randomly select a node i to active and communicate with its neighbours.
6: Node i updates its estimate $\hat{\chi}_i$ and variable $\hat{\lambda}_{i|j}$ following (34) and (35).
7: Node i sends $(\hat{\chi}_i, \hat{\lambda}_{i|j})$ to its neighbour $j \in \mathcal{N}_i$.
8: **end for**
9: Get $\mathbf{R}$ and $\mathbf{r}$ at each node.
10: Each node updates the cluster centers by using the information in step 9.
11: **end for**

which can be solved by using the PDMM method.

As shown in Algorithm 1, we need to run a distributed averaging at each iteration of the k-means clustering to make the clustering work in a distributed manner. Besides, we also need to select a proper cluster number to obtain the optimal clustering results. This may seem to be time- and communication- consuming at the first glance, but we should notice that as the network is set up, the structure of it will be settled, and in most of the applications the positions of the sound sources will not change very fast. The node clustering does not need to be done very frequently, so the delay caused by the distributed averaging in the clustering step is typically acceptable for a distributed detection system. In the rest of the paper, we assume the acoustic scene does not change much. So the distributed node clustering only need to be performed once before we apply the SPP estimation.

B. Distributed SPP estimation in the subnetwork

As mentioned in Section III B, the log GLR is a summation of local values. Similar to the distributed node clustering in Section IV A, we can obtain the log GLR by solving the distributed averaging problem (Zhao *et al.*, 2018). To obtain the GLR in (19), we need to

Algorithm 2 Distributed SPP estimation within the subnetwork C_q

Description:

Estimate PSDs at each node in cluster C_q:

for $m_q = 1 \,...\, M_q$
    Estimate $\phi_{X_{C_q,m_q}}(k+\kappa, n-\eta)$, $\phi_{V_{C_q,m_q}}(k+\kappa, n-\eta)$, $\kappa = -K', ..., K',\ \eta = 0, ..., N-1$ using the model-based noise PSD estimator (see Section V).
    Get the local information in (27), i.e.,
    $e_{m_q} = \sum_{\kappa=-K'}^{K'} \sum_{\eta=0}^{N-1} \ln\left[\frac{p(Y_{C_q,m_q}(k+\kappa,n-\eta)|H_{C_q,1}(k,n))}{p(Y_{C_q,m_q}(k+\kappa,n-\eta)|H_{C_q,0}(k,n))}\right]$.
end for

Apply PDMM to calculate $\ln L_{\mathrm{G}}(\bar{\mathbf{y}}_{C_q}(k,n))$:

for $g = 1, 2, 3, ..., G^{'}$
    Randomly select a node i in cluster C_q to active and communicate with its neighbours.
    Node i updates its estimate $\hat{\chi}_i$ and variable $\hat{\lambda}_{i|j}$ by following (34) and (35),
    Node i sends $(\hat{\chi}_i, \hat{\lambda}_{i|j})$ to its neighbour j.
end for

Get a global solution of the log GLR at each node in cluster C_q.

Calculate SPP of the qth speaker in (19) at each node in cluster C_q.

first compute

$$e_{m_q} = \sum_{\kappa=-K'}^{K'} \sum_{\eta=0}^{N-1} \ln\left[\frac{p(Y_{C_q,m_q}(k+\kappa, n-\eta)|H_{C_q,1}(k,n))}{p(Y_{C_q,m_q}(k+\kappa, n-\eta)|H_{C_q,0}(k,n))}\right] \tag{37}$$

locally at node m_q. The averaging of e_{m_q} within the subnetwork with $\ln\left[\frac{p(H_{C_q,1}(k,n))}{1-p(H_{C_q,1}(k,n))}\right]$ gives us the log GLR. The PDMM method is applied to obtain (27) distributedly. We summarize the distributed SPP estimation in Algorithm 2.

V. MODEL-BASED SIGNAL STATISTICS ESTIMATION

In Section II A, the SPP are computed given the PSDs. In practice, however, we need to estimate the signal statistics. We use the noise PSD estimator introduced in (Nielsen *et al.*, 2018) which is able to track non-stationary noise. A brief description of the model-based noise estimation method is summarized in this section.

As the signal statistics are estimated at each node independently, the cluster index is omitted for clarity from now on. Since the autoregressive (AR) processes are sufficient to model the generation of speech and noise (Nielsen *et al.*, 2018), we use the AR processes to model the speech and noise signals as described in Section II. In practice, the AR-parameters are pre-trained and stored in speech and noise codebooks. The training of the AR-parameters is explained in Section VI. Mathematically, the noise PSD mentioned in (14) and (15) at each node can be defined as (Stoica and Moses, 2005)

$$\phi_{V_m}(k,n) = \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left[|V_m(k,n)|^2|\mathbf{y}_m(t)\right]. \tag{38}$$

The conditional expectation in (38) is the second moment of the density $p(|V_m(k,n)|^2|\mathbf{y}_m(t))$. We can get another form of (38) as

$$\phi_{V_m}(k,n) = \lim_{T\to\infty} \frac{1}{T}\left[\int_{\mathbb{R}^{T\times 1}} |V_m(k,n)|^2 p(\mathbf{v}_m(t)|\mathbf{y}_m(t))d\mathbf{v}_m(t)\right]. \tag{39}$$

To compute the posterior $p(\mathbf{v}_m(t)|\mathbf{y}_m(t))$, we use the statistical models $\{\mathcal{M}_u\}_{u=1}^U$, which were introduced in Section II to explain the data. These models can be incorporated into (39). Then the model-based PSD can be expressed as

$$\begin{aligned}\phi_{V_m}(k) \approx& \frac{1}{T}\sum_{u=1}^{U} q(\mathcal{M}_u|\mathbf{y}_m)\left[\int_{\mathbb{R}^{T\times 1}} |V_m(k)|^2 p(\mathbf{v}_m|\mathbf{y}_m,\mathcal{M}_u)d\mathbf{v}_m\right] \\ =& \sum_{u=1}^{U} q(\mathcal{M}_u|\mathbf{y}_m)\phi_{V_m}(k|\mathcal{M}_u),\end{aligned} \tag{40}$$

and the time index is omitted for clarity.

The excitation noise variances are treated as unknown random variables with the prior

$$p(\sigma_{x,u}^2|\mathcal{M}_u) = \mathrm{Inv}\mathcal{G}(\alpha_{x,u},\beta_{x,u}) \tag{41}$$

and

$$p(\sigma_{v,u}^2|\mathcal{M}_u) = \mathrm{Inv}\mathcal{G}(\alpha_{v,u},\beta_{v,u}), \tag{42}$$

where $\mathrm{Inv}\mathcal{G}[\cdot,\cdot]$ denotes inverse Gamma density.

The posteriors which are needed to estimate the noise PSD have no closed-form. The variational Bayesian (BS) framework (Bishop, 2006; Jordan *et al.*, 1999) can be used to produce analytical approximation. In (Nielsen *et al.*, 2018), the full joint posterior can be factorised as

$$p(\mathbf{v}_m, \sigma^2_{x,u}, \sigma^2_{v,u}|\mathbf{y}_m, \mathcal{M}_u)p(\mathcal{M}_u|\mathbf{y}_m) \approx q(\mathbf{v}_m|\mathbf{y}_m, \mathcal{M}_u)q(\sigma^2_{x,u}, \sigma^2_{v,u}|\mathbf{y}_m, \mathcal{M}_u)q(\mathcal{M}_u|\mathbf{y}_m). \quad (43)$$

According to (Nielsen *et al.*, 2018) and its supplementary document, the posterior factor $q(\mathbf{v}_m|\mathbf{y}_m, \mathcal{M}_u)$ is given by

$$q(\mathbf{v}_m|\mathbf{y}_m, \mathcal{M}_u) = \mathcal{N}(\hat{\mathbf{v}}_{m,u}, \hat{\mathbf{\Sigma}}_u), \quad (44)$$

where

$$\hat{\mathbf{\Sigma}}_u = \left[\frac{\breve{a}_{x,u}}{\breve{b}_{x,u}}\mathbf{Q}_x^{-1}(\mathbf{a}_u) + \frac{\breve{a}_{v,u}}{\breve{b}_{v,u}}\mathbf{Q}_v^{-1}(\mathbf{b}_u)\right]^{-1}, \quad (45)$$

$$\hat{\mathbf{v}}_{m,u} = \frac{\breve{a}_{x,u}}{\breve{b}_{x,u}}\hat{\mathbf{\Sigma}}_u\mathbf{Q}_x^{-1}(\mathbf{a}_u)\mathbf{y}_m. \quad (46)$$

The scalars $\breve{a}_{x,u}$, $\breve{b}_{x,u}$, $\breve{a}_{v,u}$, and $\breve{b}_{v,u}$ are obtained from

$$q(\sigma^2_{x,u}, \sigma^2_{v,u}|\mathbf{y}_m, \mathcal{M}_u) = \mathrm{Inv}\mathcal{G}(\breve{a}_{x,u}, \breve{b}_{x,u})\mathrm{Inv}\mathcal{G}(\breve{a}_{v,u}, \breve{b}_{v,u}), \quad (47)$$

where

$$\breve{a}_{x,u} = \alpha_{x,u} + T/2, \quad (48)$$

$$\breve{b}_{x,u} = \beta_{x,u} + \left[\hat{\mathbf{x}}^T_{m,u}\mathbf{Q}_x^{-1}(\mathbf{a}_u)\hat{\mathbf{x}}_{m,u} + \mathrm{tr}\left(\mathbf{Q}_x^{-1}(\mathbf{a}_u)\hat{\mathbf{\Sigma}}_u\right)\right]/2, \quad (49)$$

$$\breve{a}_{v,u} = \alpha_{v,u} + T/2, \quad (50)$$

$$\breve{b}_{v,u} = \beta_{v,u} + \left[\hat{\mathbf{v}}^T_{m,u}\mathbf{Q}_v^{-1}(\mathbf{a}_u)\hat{\mathbf{v}}_{m,u} + \mathrm{tr}\left(\mathbf{Q}_v^{-1}(\mathbf{a}_u)\hat{\mathbf{\Sigma}}_u\right)\right]/2, \quad (51)$$

$$\hat{\mathbf{x}}_{m,u} = \mathbf{y}_m - \hat{\mathbf{v}}_{m,u}. \tag{52}$$

The parameters of the posterior factors are computed iteratively, and the VB framework guarantees that the algorithm converges. Convergence of the VB algorithm can be controlled by the variational lower bound $\mathfrak{L}_u$. The posterior model probabilities has the following relation with the variational lower bound $\mathfrak{L}_u$:

$$q(\mathcal{M}_u|\mathbf{y}_m) \propto \exp(\mathfrak{L}_u)p(\mathcal{M}_u), \tag{53}$$

where $\propto$ denotes proportional to. The variational lower bound consists of many terms. For more details, we refer the interested reader to reference (Nielsen *et al.*, 2018) and the supplementary document. With the model probabilities $\{q(\mathcal{M}_u|\mathbf{y}_m)\}_{u=1}^{U}$, the models explaining the data well are given more weight than the other models. Since the posterior factor $q(\mathbf{v}_m|\mathbf{y}_m, \mathcal{M}_u)$ is a normal distribution, its second moment is

$$\mathrm{E}\left[\mathbf{v}_m\mathbf{v}_m^T|\mathbf{y}_m, \mathcal{M}_u\right] = \hat{\mathbf{v}}_{m,u}\hat{\mathbf{v}}_{m,u}^T + \hat{\mathbf{\Sigma}}_u, \tag{54}$$

then we have

$$\int_{\mathbb{R}^{T\times 1}} |V_m(k)|^2 p(\mathbf{v}_m|\mathbf{y}_m, \mathcal{M}_u)d\mathbf{v}_m = |\mathbf{f}_k^H\hat{\mathbf{v}}_{m,u}|^2 + \mathbf{f}_k^H\hat{\mathbf{\Sigma}}_u\mathbf{f}_k, \tag{55}$$

where $\mathbf{f}_k$ is the kth column of DFT matrix $\mathbf{F}$. Inserting (55) in (40), we get a model-averaged version of the MMSE estimator (Gerkmann and Hendriks, 2012; Hendriks *et al.*, 2010) as

$$\hat{\phi}_{V_m}(k,n) = \frac{1}{T}\sum_{u=1}^{U} q(\mathcal{M}_u|\mathbf{y}_m)\left[|\mathbf{f}_k^H\hat{\mathbf{v}}_{m,u}|^2 + \mathbf{f}_k^H\hat{\mathbf{\Sigma}}_u\mathbf{f}_k\right]. \tag{56}$$

A more detailed derivation of the model-based noise PSD estimation is available in (Nielsen *et al.*, 2018). The estimated speech PSD can be obtained in a similar way. Inserting (56) and the speech PSD estimate in (14) and (15), with the distributed estimation of $\ln L_{\mathrm{G}}(\bar{\mathbf{y}}(k,n))$, the SPP is obtained by using (19).

In practice, the speech and noise codebooks are trained by using a variation of the LPC-

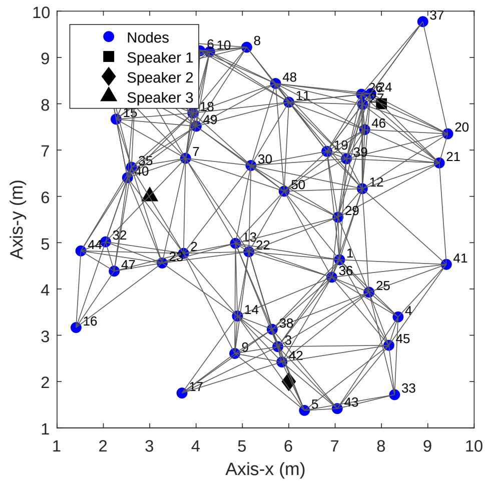

FIG. 1. (Color online) Room setup. The room is of size 10 m $\times$ 10 m $\times$ 3 m. We have 50 nodes randomly placed in the room. The maximum communication distance is set to 2.5 m.

VQ method (Paliwal and Atal,1998; Gersho and Gray, 2012). More specifically, by passing the training signal as input to the vector quantizer, the linear prediction coefficients, which are converted into line spectral frequency coefficients are extracted from the windowed frames of the signal. Once we get the trained AR processes, the spectral envelopes are computed according to (20).

VI. SIMULATIONS

In this section, simulations are performed to demonstrate the performance of the distributed detection in simulated room acoustics. We simulate a room of size 10 m$\times$10 m$\times$3 m with the room impulse response (RIR) generated by using the image source model method (Allen and Berkley, 1979). The reverberation time is $T_{60} \approx 200$ ms. As shown in Fig. 1, we have 50 nodes (microphones) randomly placed in the room. The solid lines indicate edges, and the two nodes connected by the edge can communicate with each other. The maximum communication distance is set to 2.5 m. Three speakers are located at (8 m, 8 m, 1.5 m), (6 m, 2 m, 1.5 m) and (3 m, 6 m, 1.5 m). The speech signals are scaled to have the same power before convolving the RIRs. In all the experiments, the speech and noise codebooks consist of AR vectors of order 14. The AR model order for both the speech and noise signal was empirically chosen (Nielsen *et al.*, 2018; Kavalekalam *et al.*, 2019). We train a speech codebook with 64 entries (32 entries for male speakers and 32 for female speakers). The

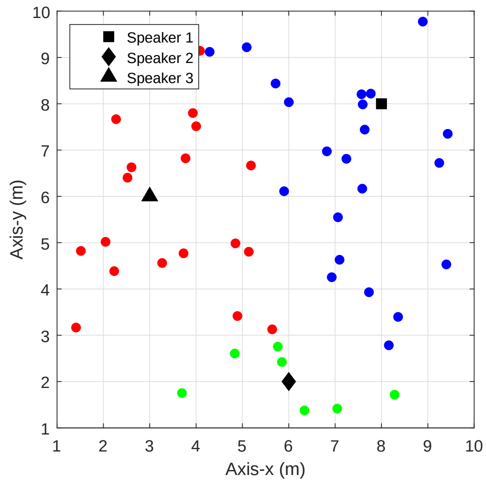

FIG. 2. (Color online) The result of the node clustering when there are three speakers and babble noise as background noise (iSNR = 10 dB). The different colored nodes indicate the divided subnetworks. We set 100 iteration for PDMM.

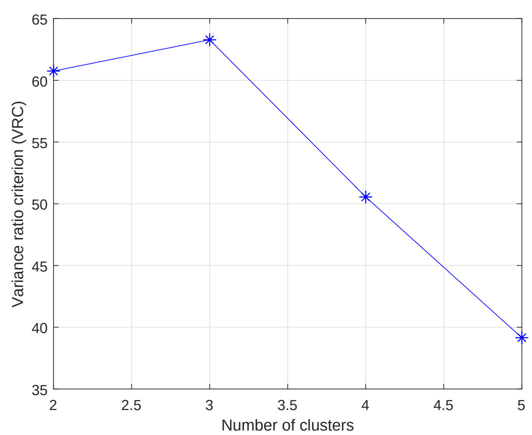

FIG. 3. (Color online) The evaluation result of the distributed k-means clustering for different cluster numbers. We set 100 iteration for PDMM.

noise codebook contains 16 entries (4 entries for babble, restaurant, exhibition, and 2 entries for street and station noise, respectively). The speech training data is from the TIMIT database (Lyons, 1990) and the noise training data is from the AURORA database (Hirsch and Pearce, 2000). The testing speech is taken from the CHiME corpus (Christensen *et al.*, 2010), and the testing noise is from part of the NOISEX-92 database ,i.e., babble.wav and factory1.wav, which is not contained in the training process. All the signals are downsampled to 8 kHz. The noisy signal is transformed into the frequency domain using the STFT, with a Hanning window of length 256 and a 50% overlap. A 256-point FFT is used to

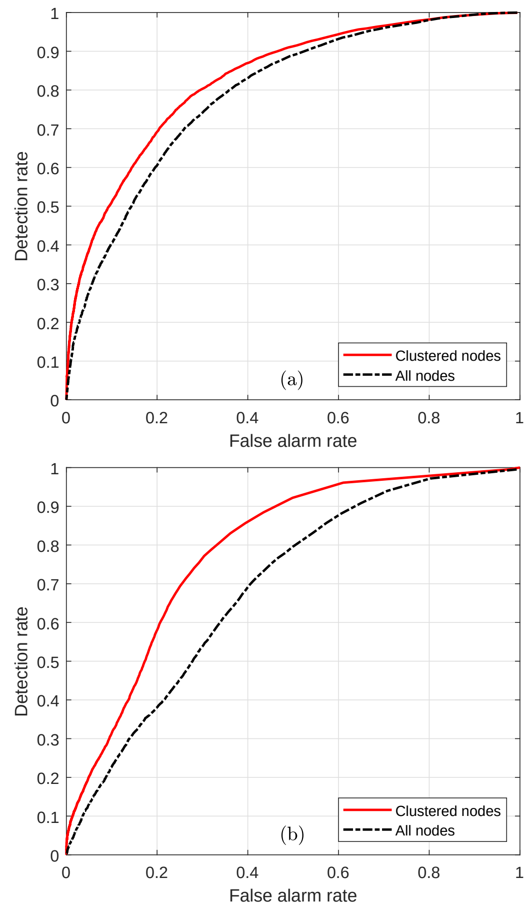

FIG. 4. (Color online) The detection performance for different speakers by using the clustered nodes and all nodes. We set $K' = 1$ and $N = 2$. The iteration number for PDMM is set to 100. (a) The ROC curve for speaker 1. (b) The ROC curve for speaker 2.

transform each frame into the STFT domain.

The first experiment intends to show the performance of distributed node clustering method which is introduced in Section IV A. We consider babble noise with 10 dB iSNR here. The distributed clustering is designed to work in an online way, but only the result for one frame (256 points with 8 kHz sampling frequency) is shown in Fig. 2. For a certain frame of data, we set 100 iterations for the PDMM. We see that the nodes near a certain sound source are clustered together. With the clustered nodes forming a subnetwork which is interested in a certain speaker, detection is then applied by using the observed signal in the clustered nodes. We also evaluate the clustering performance by using the variance ratio criterion, and the result is illustrated in Fig. 3. For the experimental setup in this case, the optimal clustering number is chosen as 3 which gives the highest VRC. The optimal

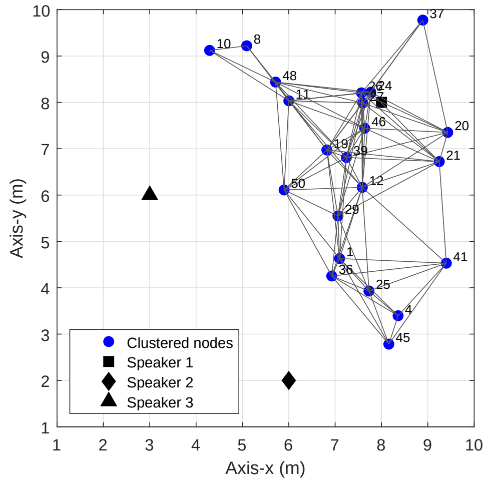

FIG. 5. (Color online) The clustered nodes near speaker 1 and their connection conditions.

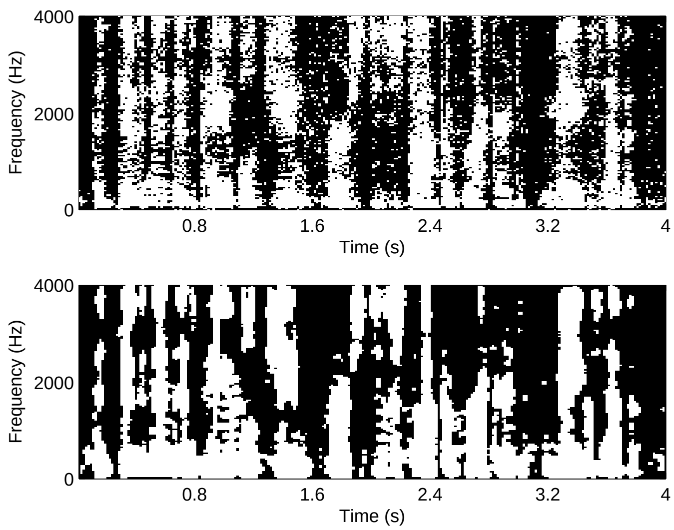

FIG. 6. (Color online) The detection result for speaker 1 with the false alarm rate being 0.2. We set $K' = 1$ and $N = 2$. The iteration number for PDMM is set to 50. The white area indicates speech is present and the dark area indicates speech is absent. The upper figure is the ground truth decision matrix, the lower figure is the detection result we get by using the model-based SPP estimation method.

clustering number also reveals the number of sound sources in the environment.

Next, we will explain the detection performance. In detection problems, it is common to utilize the receiver operating characteristic (ROC) to evaluate the performance of a detector. The second experiment is to study the necessity of applying the nodes clustering before detection. The background noise is set as babble noise with iSNR being 10 dB. Since the

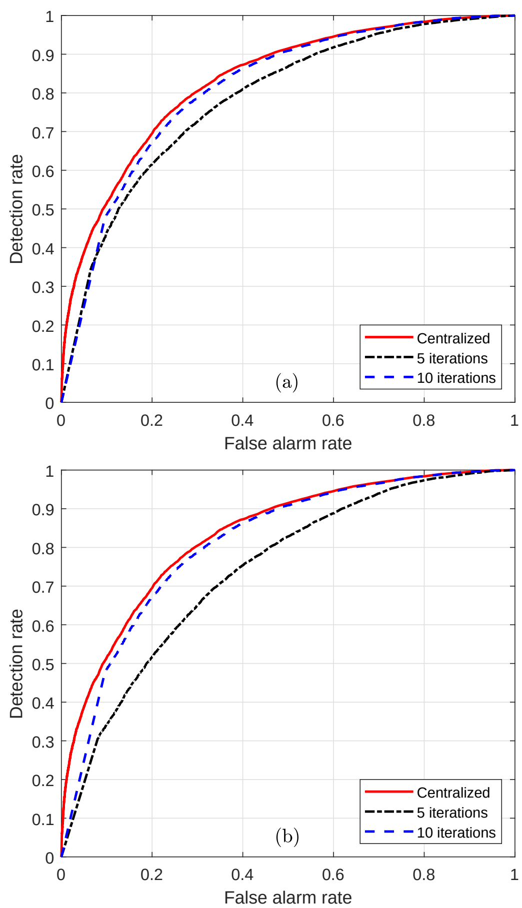

FIG. 7. (Color online) The distributed VAD convergence performance at different nodes near speaker 1. We set $K' = 1$ and $N = 2$. (a) The ROC curve for different PDMM iterations at node 12. (b) The ROC curve for different PDMM iterations at node 25.

noise covariance matrix can be updated when a speech signal is absent or the observation signal is dominated by noise, we set an iSNR threshold to the subband noisy signal to get a ground truth decision matrix. The desired signal at each subnetwork is the clean speech received by one of the nodes in each subnetwork. More specifically, the frequency bands with higher iSNR than the iSNR threshold are marked as speech presence, and the others are marked as speech absence. For speaker 1, we choose node 39 as the reference node and node 3 is set as the reference node for speaker 2. The iSNR threshold is set to be -5 dB. The prior SPP is set to be $p(H_1(k, n)) = 0.5$. Figure 4 shows the results of the detection performance for speaker 1 and speaker 2. We set $K' = 1$ and $N = 2$. We set 100 iterations for PDMM to make sure that the distributed detection method converges. By means of comparing the detection performance with subnetwork between using all nodes in the network, the result

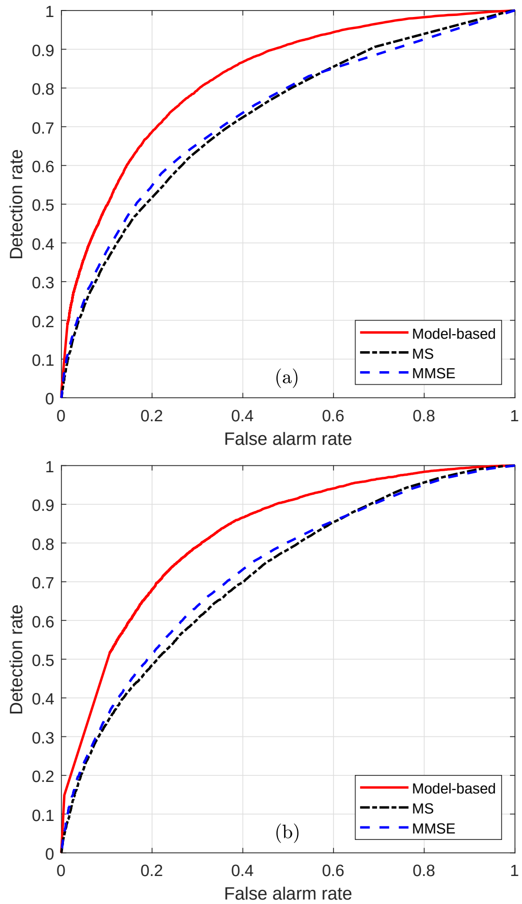

FIG. 8. (Color online) The distributed VAD performance under babble noise condition (iSNR = 10 dB) with different noise PSD estimators at node 25. We set 50 iterations for PDMM. (a) The ROC curve under babble noise. $K' = 0$ and $N = 1$. (b) The ROC curve under babble noise. $K' = 1$ and $N = 2$.

shows that the detection can benefit from the node clustering. It is seen that both Fig. 4 (a) and Fig. 4 (b) that better detection performance can be achieved by using the clustered nodes. This is simply because the sound propagation attenuation makes the received signal at the nodes faraway from the interested source contain less useful information of the desired signal.

The next experiment is to study the convergence performance of the distributed detection. As nodes have been clustered into subnetworks, the distributed detection is applied within the nodes near a certain speaker. We assume that the acoustic scene does not change too frequently, the locations of the nodes and sound sources are settled during the whole procedure of detection, so the same node clustering result is applied for online detection.

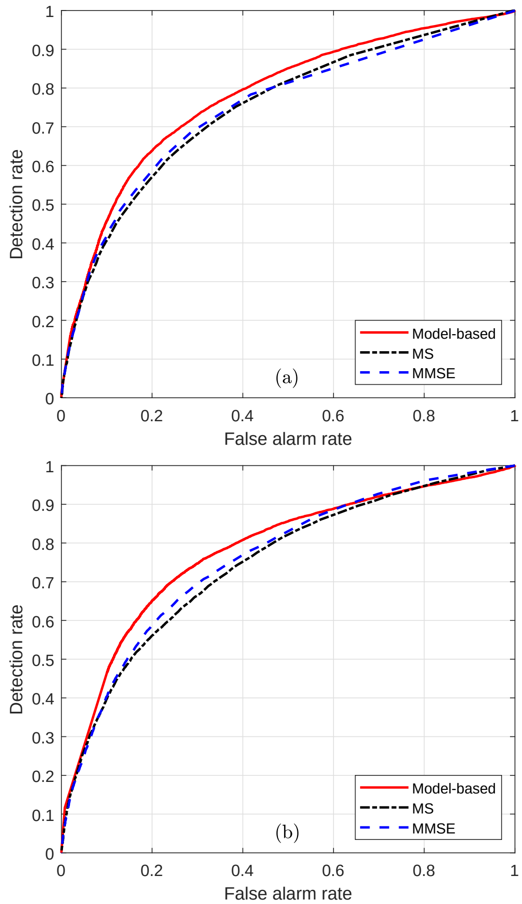

FIG. 9. (Color online) The distributed VAD performance under factory noise condition (iSNR = 10 dB) with different noise PSD estimators at node 25. We set 50 iterations for PDMM. (a) The ROC curve under factory noise. $K' = 0$ and $N = 1$. (b) The ROC curve under factory noise. $K' = 1$ and $N = 2$.

We show the clustered nodes and their connection conditions in Fig. 5 for speaker 1. The detection performance under babble noise condition is shown in Fig. 6. We choose the proper threshold of GLRT to get 0.2 false alarm rate. We then evaluate the convergence performance of the distributed detection at different nodes. Babble noise is considered here, inter-band information and inter-frame information are used in the detection ($K' = 1$, $N = 2$). In the distributed consensus step, we apply the PDMM method to get the distributed averaging result. And the corresponding detection results for speaker 1 is illustrated in Fig. 7. Figure 7 (a) plots the ROC curve of the 12th node with different number of iterations of the PDMM, and Fig. 7 (b) illustrates the performance of the 25th node. We notice from Fig. 7 that the convergence speed of the distributed detection is different at different nodes. The node with higher iSNR converges faster than the one with lower iSNR. The reason is that the higher iSNR at the nodes near the desired signal will lead to better speech PSD estimate, which will contribute to better detection performance.

In the last experiment, the detection performance with different noise estimators is studied for speaker 1. We consider babble noise and factory noise here. The ROC curve at the 25th node are plotted in Fig. 8 and Fig. 9. The number of iterations of the PDMM method is set to be 50. The proposed distributed detection is able to maintain robust performance under different noise conditions. Moreover, the model-based detection outperforms the MS and MMSE based methods. Furthermore, under the condition that the factory noise information is not included in the codebook, the model-based method still outperforms the MS and MMSE based methods in detection performance. We also test the detection performance by taking into account different number of time frames and frequency bins. Comparing Fig. 9 (a) and Fig. 9 (b), one can see that the detection performance is improved by using neighbouring frames and frequency bins for different methods.

VII. CONCLUSIONS

In this paper, we proposed a distributed multi-speaker speech presence probability estimation method by using WASN. A node clustering was first applied to assign the nodes into subnetworks. We formulated the node clustering as a model-based clustering problem, and a distributed k-means method was used to make the clustering work in a distributed manner. It was noticed from the experimental results that the detector obtained better performance with clustered nodes compared to using the observations from all nodes. We also proposed a distributed detector with WASN. By taking advantage of the model-based noise PSD estimation method, the proposed distributed detection method was able to obtain robust performance under non-stationary noise condition. We formed the distributed detector by using the GLRT theory. The global decision was made by considering the likelihood functions at all channels in the subnetwork. Finally, the distributed detection can be obtained by solving the distributed averaging problem. We utilized the PDMM as consensus method to obtain the distributed optimization. The proposed detection method does not need any fusion center. We studied the performance of the distributed detection method under different noise conditions. The experimental results showed that the distributed detection method converged efficiently to the centralized solution, and the performance was quite robust under different types of non-stationary noise with the appearance of competing speakers.

ACKNOWLEDGMENT

This work was supported in part by the Key Program of National Science of Foundation of China (NSFC) under Grant No. 61831019, the NSFC and Israel Science Foundation (ISF) joint research program under Grant No. 61761146001, and the NSFC "Distinguished Young Scientists Fund" under Grant No. 61425005.

Allen, J. B. and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Amer. **65**,943–950.

Bahari, M. H., Hamaidi, L. K., Muma, M., Plata-Chaves, J., Moonen, M., Zoubir, A. M., and Bertrand, A. (**2017**). "Distributed multi-speaker voice activity detection for wireless acoustic sensor networks," arXiv preprint arXiv:1703.05782.

Bertrand, A. (**2011**). "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," Proc. IEEE Symp. Commun. Veh. Technol. (SCVT), 1–6.

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

Bertrand, A. and Moonen, M. (**2011**). "Distributed adaptive estimation of node-specific signals in wireless sensor networks with a tree topology," IEEE Trans. Signal Process. **59**,2196–2210.

Bertrand, A. and Moonen, M. (**2012**). "Distributed signal estimation in sensor networks where nodes have different interests," Signal Process. **92**,1679–1690.

Bishop, C. M (**2006**). "Pattern Recognition and Machine Learning," New York, NY, USA: Springer

Boyd, S., Ghosh, A, Prabhakar, B., and Shah, D (**2006**). "Randomized gossip algorithms," IEEE Trans. Info. Theory **52**, 2508–2530.

Caliński, T. and Harabasz, J. (**1974**). “A dendrite method for cluster analysis,” Commun. Stat. **3**,1–27.

Christensen, H., Barker, J., Ma, N., and Green, P. D. (**2010**). “The CHiME corpus: a resource and a challenge for computational hearing in multisource environments,” Proc. Interspeech, 1918–1921.

Cohen, I. and Berdugo, B. (**2002**). “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” IEEE Signal Process. Lett. **9**,12–15.

de la Hucha Arce, F., Moonen, M., Verhelst, M., and Bertrand, A. (**2017**). “Adaptive quantization for multichannel Wiener filter-based speech enhancement in wireless acoustic sensor networks,” Wireless Commun. Mobile Comput. **2017**.

Gergen, S., Nagathil, A., and Martin, R. (**2015**). “Classification of reverberant audio signals using clustered ad hoc distributed microphones Signal Processing,” Signal Process. **107**,21–32.

Gergen, S., Martin, R., and Madhu, N. (**2018**). “Source Separation by Feature-Based Clustering of Microphones in Ad Hoc Arrays,” Proc. IWAENC 530–534.

Gerkmann, T., Breithaupt, C., and Martin, R. (**2008**). “Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors,” IEEE Trans. Audio, Speech, Lang. Process. **16**,910–919.

Gerkmann, T. and Hendriks, R. C. (**2012**). “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” IEEE Trans. Audio, Speech, Lang. Process. **20**,1383–1393.

Gersho, A. and Gray, R. M. (**2012**). *Vector Quantization and Signal Compression* (Springer Science Business Media).

Gray, R. M. (**2006**). “Toeplitz and circulant matrices: A review,” Found. Trends Commun. Inf. Theory **2**,155–239.

Hamaidi, L. K., Muma, M., and Zoubir, A. M. (**2017**). “Robust distributed multi-speaker voice activity detection using stability selection for sparse non-negative feature extraction,” Proc. IEEE EUSIPCO, 161–165.

Hamaidi, L. K., Muma, M., and Zoubir, A. M. (**2017**). “Multi-speaker voice activity detection by an improved multiplicative non-negative independent component analysis with sparseness constraints,” Proc. IEEE ICASSP, 4611–4615.

Hartigan, J. A. and Wong, M. A. (**1979**). "Algorithm AS136: A k-means clustering algorithm," Applied Statistics **28**,100–108.

Hendriks, R. C., Heusdens, R., and Jensen, J. (**2010**). "MMSE based noise PSD tracking with low complexity," Proc. IEEE ICASSP, 4266–4269.

Hirsch, H. G. and Pearce, D. (**2000**). "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proc. ISCA ITRW ASR, 181–188.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K (**1999**). "An introduction to variational methods for graphical models," Mach. Learn., **37**,183–233.

Kavalekalam, M. S., Nielsen, J. K., Christensen, M. G., and Boldt, J. B. (**2018**). "A Study of noise PSD estiamtors for single channel speech enhancement," Proc. IEEE ICASSP, 5464–5468.

Kavalekalam, M. S., Nielsen, J. K., Boldt, J. B., and Christensen, M. G. (**2019**). "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," IEEE/ACM Trans. Audio, Speech, Lang. Process **27**,99–113.

Lyons, J. W. (**1990**). "DARPA TIMIT acoustic-phonetic continuous speech corpus," Technical Report NISTIR 4930, National Institute of Standards and Technology.

Markovich-Golan, S., Bertrand, A., Moonen, M., and Gannot, S. (**2015**). "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," Signal Process. **107**,4–20.

Martin, R. (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process. **9**,504–512.

Momeni, H., Habets, E. A., and Abutalebi, H. R. (**2014**). "Single-channel speech presence probability estimation using inter-frame and inter-band correlations," Proc. IEEE ICASSP, 2903–2907.

Nielsen, J. K., Kavalekalam, M. S., Christensen, M. G., and Boldt, J. B. (**2018**). "Model-based noise PSD estimation from speech in non-stationary noise," Proc. IEEE ICASSP, 5424–5428.

Paliwal, K. K. and Atal, B. S. (**1998**). "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. Speech Audio Process., **1**,3–14.

Qin, J., Fu, W., Gao, H., and Zheng, W. X. (**2017**). "Distributed k-means algorithm and fuzzy c-means algorithm for sensor networks based on multiagent consensus theory," IEEE

Trans. Cybern. **47**,772–783.

Ramirez, J., Segura, J. C., Benitez, C., Torre, A., and Rubio, A. (**2004**). "Efficient voice activity detection algorithms using long-term speech information," Speech Commun. **42**,271–287.

Sohn, J., Kim, N. S., and Sung, W. (**1999**). "A statistical model-based voice activity detection," IEEE Signal Process. Lett. **6**,1–3.

Souden, M., Chen, J., Benesty, J., and Affes, S. (**2010**). "Gaussian model-based multichannel speech presence probability," IEEE Trans. Audio, Speech, Lang. Process. **18**,1072–1077.

Souden, M., Chen, J., Benesty, J., and Affes, S. (**2011**). "An integrated solution for online multichannel noise tracking and reduction," IEEE Trans. Audio, Speech, Lang. Process. **19**,2159–2169.

Srinivasan, S., Samuelsson, J., and Kleijn, W. B. (**2007**). "Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments," IEEE Trans. Audio, Speech, Lang. Process. **15**,441–452.

Stoica, P. and Moses, R. L. (**2005**). *Spectral Analysis of Signals* (Upper Saddle River, NJ: Prentice-Hall).

Szurley, J., Bertrand, A., and Moonen, M. (**2015**). "Distributed adaptive node-specific signal estimation in heterogeneous and mixed-topology wireless sensor networks," Signal Process. **117**,44–60.

Szurley, J., Bertrand, A., and Moonen, M. (**2016**). "Topology-independent distributed adaptive node-specific signal estimation in wireless sensor networks," IEEE Trans. Signal Inf. Process. Netw. **3**,130–144.

Taseska, M. and Habets, E. A. (**2014**). "Informed spatial filtering for sound extraction using distributed microphone arrays," IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**,1195–1207.

Tavakoli, V. M., Jensen, J. R., Heusdens, R., Benesty, J., and Christensen, M. G. (**2017**). "Distributed max-SINR speech enhancement with ad hoc microphone arrays," Proc. IEEE ICASSP, 151–155.

Zhang, G. and Heusdens, R. (**2017**). "Distributed optimization using the primal-dual method of multipliers," IEEE Trans. Signal Inf. Process. Netw. **4**,173–187.

Zhang, R. and Kwok, J. (**2014**). “Asynchronous distributed ADMM for consensus optimization,” Proc. Int. Conf. Mach. Learn., 1701–1709.

Zhao, Y., Nielsen, J. K., Christensen, M. G., and Chen, J. (**2018**). “Model-based voice activity detection in wireless acoustic sensor networks,” Proc. IEEE EUSIPCO, 425–429.