Model-Based Voice Activity Detection in Wireless Acoustic Sensor Networks

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

Published in:
2018 26th European Signal Processing Conference (EUSIPCO)

DOI (link to publication from Publisher):
10.23919/EUSIPCO.2018.8553457

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Model-Based Voice Activity Detection in Wireless Acoustic Sensor Networks

Yingke Zhao[1,2], Jesper Kjær Nielsen[2], Mads Græsbøll Christensen[2] and Jingdong Chen[1]
[1]CIAIC and School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
[2]Audio Analysis Lab, CREATE, Aalborg University, Aalborg, Denmark

***Abstract*—One of the major challenges in wireless acoustic sensor networks (WASN) based speech enhancement is robust and accurate voice activity detection (VAD). VAD is widely used in speech enhancement, speech coding, speech recognition, etc. In speech enhancement applications, VAD plays an important role, since noise statistics can be updated during non-speech frames to ensure efficient noise reduction and tolerable speech distortion. Although significant efforts have been made in single channel VAD, few solutions can be found in the multichannel case, especially in WASN. In this paper, we introduce a distributed VAD by using model-based noise power spectral density (PSD) estimation. For each node in the network, the speech PSD and noise PSD are first estimated, then a distributed detection is made by applying the generalized likelihood ratio test (GLRT). The proposed global GLRT based VAD has a quite general form. Indeed, we can judge whether the speech is present or absent by using the current time frame and frequency band observation or by taking into account the neighbouring frames and bands. Finally, the distributed GLRT result is obtained by using a distributed consensus method, such as random gossip, i.e., the whole detection system does not need any fusion center. With the model-based noise estimation method, the proposed distributed VAD performs robustly under non-stationary noise conditions, such as babble noise. As shown in experiments, the proposed method outperforms traditional multichannel VAD methods in terms of detection accuracy.**

***Index Terms*—Wireless acoustic sensor networks, noise PSD estimation, distributed voice activity detection**

I. Introduction

With the development of distributed optimization methods, WASN is becoming more and more popular in audio signal processing applications. Compared to the traditional microphone arrays, WASNs are more flexible and scalable, and are able to physically cover a larger space and capture more spatial information. The distributed speech enhancement methods, such as the distributed Wiener filter [1], the distributed maximum SNR filter [2], the distributed beamforming [3], need an estimate of the second-order statistics of the noise before forming the linear filter. Usually, the noise covariance matrix is estimated in a recursive way, and the estimated covariance matrix is updated only when the speech is absent. Therefore, an accurate global voice activity detection (VAD) is needed to detect the presence/absence of human speech. In terms of frequency domain multichannel speech enhancement, the global VAD needs to be obtained at each time frame and each frequency band. Single channel VAD has been extensively studied [4], [5], but not in the multichannel and the WASN cases. In [6], the VAD problem with WASN is formed as a node clustering problem first, and then the VAD is obtained locally at the clustered nodes. However, the development of distributed VAD method with global decision making strategy is not well studied.

For a WASN, the VAD can be developed in a centralized way or in a distributed manner. The distributed method is more flexible than the centralized one, since each node can leave or join the network and does not depend on a fusion center. Besides, in terms of reducing the data traffic and the communication bandwidth in the network, the distributed way is usually preferred [7]. In [8], the authors developed a multichannel noise estimation method based on multichannel speech presence probability (MC-SPP) estimation, which can also be used in making the multichannel VAD decision. The results show that the detection performance increases by using multiple microphones. Even though the results are promising, this method needs to be initialized carefully, and the optimal parameters are difficult to find. Moreover, the whole system only works in a centralized way. In order to implement distributed speech enhancement techniques, it is essential to develop a robust VAD method that works in a distributed way.

This paper introduces a distributed model-based VAD method. The proposed method is able to obtain a global decision distributedly per frame and band. Additionally, the distributed VAD maintains robust detection performance even with non-stationary noise. For the distributed VAD, the noise PSD estimation is performed at each node independently. Any noise PSD estimation method can be applied here. The traditional noise estimation methods, such as the minimum mean-square error (MMSE) noise PSD estimator [9], [10] and the minimum statistics (MS) noise PSD estimator [11] are widely used, but these methods have limited performance when dealing with non-stationary noise. In [12], the authors introduced a model-based noise PSD estimator by applying a statistical model to the speech and noise signals. The proposed noise PSD estimator is able to include prior spectral information about speech and different types of non-stationary noise [13]. Based on the estimated noise PSDs, we apply the GLRT to obtain the global decision. In this case, we find that the GLRT involves a distributed averaging problem, which can easily be solved by applying distributed consensus methods, such as the random gossip method [14], the alternating di-

This work was supported in part by the Villum Foundation and the NSFC Distinguished Young Scientists Fund under grant No. 61425005. The work of Y. Zhao was supported in part by the China Scholarship Council.

ISBN 978-90-827970-1-5

rection method of multipliers (ADMM) [15], the primal-dual method of multipliers (PDMM) [16]. In the distributed VAD, besides considering the inter-band information, we further take the inter-frame information into account to improve the VAD performance.

II. Signal Model and Problem Formulation

The problem considered in this paper is to develop a distributed VAD per time frame and frequency band in a WASN. Mathematically, this is a two-state model selection problem. At frequency bin k and time frame n, we have one hypothesis $H_0(k,n)$ denoting that speech is absent at all microphones, and one hypothesis $H_1(k,n)$ denoting that speech is present at all microphones, i.e.,

$$\begin{aligned} H_0(k,n):&\ \bar{\mathbf{y}}(k,n) = \bar{\mathbf{v}}(k,n), \\ H_1(k,n):&\ \bar{\mathbf{y}}(k,n) = \bar{\mathbf{s}}(k,n) + \bar{\mathbf{v}}(k,n), \end{aligned} \tag{1}$$

where

$$\bar{\mathbf{y}}(k,n) = \left[\bar{\mathbf{y}}_1^T(k,n),\ \bar{\mathbf{y}}_2^T(k,n),\ \ldots,\ \bar{\mathbf{y}}_M^T(k,n)\right]^T \tag{2}$$

is the network wide noisy observation. $\bar{\mathbf{s}}(k,n)$ and $\bar{\mathbf{v}}(k,n)$ are the clean speech vector and the additive noise vector respectively, and $[\cdot]^T$ denotes the transpose operator. The observation vector at the mth microphone contains the N past time segments as

$$\begin{aligned} &\bar{\mathbf{y}}_m(k,n) = \\ &[\mathbf{y}_m^T(k,n),\ \mathbf{y}_m^T(k,n-1),\ \ldots,\ \mathbf{y}_m^T(k,n-N+1)]^T, \end{aligned} \tag{3}$$

where $\mathbf{y}_m(k,n)$ contains $2K+1$ frequency bands centered at frequency index k, i.e.,

$$\mathbf{y}_m(k,n) = [Y_m(k-K,n), ..., Y_m(k+K,n)]^T, \tag{4}$$

where $Y_m(k,n)$ is the short-time-Fourier-transform (STFT) coefficient of the time domain noisy signal. Thus, $\bar{\mathbf{y}}_m(k,n)$ contains both the inter-band and inter-frame information. For the special case, $K=0$ and $N=1$, $\bar{\mathbf{y}}_m(k,n)$ only contains the current frame and current band. $\bar{\mathbf{s}}(k,n)$ and $\bar{\mathbf{v}}(k,n)$ have the same form as $\bar{\mathbf{y}}(k,n)$.

In order to solve the VAD problem, we assume a complex Gaussian statistical model to each STFT coefficient, this model has been widely used in the noise PSD tracking methods [9], [10], [18], and is given by

$$\begin{aligned} &p(Y_m(k,n)|H_0(k,n)) = \\ &\qquad \frac{1}{\pi\phi_{V_m}(k,n)} \exp\left\{-\frac{|Y_m(k,n)|^2}{\phi_{V_m}(k,n)}\right\}, \end{aligned} \tag{5}$$

$$\begin{aligned} p(Y_m(k,n)|H_1(k,n)) = &\frac{1}{\pi(\phi_{S_m}(k,n)+\phi_{V_m}(k,n))} \\ &\times \exp\left\{-\frac{|Y_m(k,n)|^2}{\phi_{S_m}(k,n)+\phi_{V_m}(k,n)}\right\}, \end{aligned} \tag{6}$$

where $\phi_{S_m}(k,n)$ and $\phi_{V_m}(k,n)$ are speech PSD and noise PSD respectively. We further assume that $Y_m(k+\kappa, n-\eta), m = 1,...,M, \kappa = -K,...,K, \eta = 0,...,N-1$ are independent given $H_0(k,n)$ or $H_1(k,n)$.

In a WASN, the two-model selection problem in (1) can be solved in a distributed way. Before going into the distributed solution, the centralized VAD is first introduced in the next section.

III. Centralized VAD

This section formulates the centralized detection problem. We apply the GLRT method to solve the VAD problem in (1). Based on the detection theory, the GLRT makes the decision with the following function:

$$L_G(\bar{\mathbf{y}}(k,n)) = \frac{p(\bar{\mathbf{y}}(k,n)|H_1(k,n))}{p(\bar{\mathbf{y}}(k,n)|H_0(k,n))} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \tag{7}$$

where $L_G(\bar{\mathbf{y}}(k,n))$ is called the generalized likelihood ratio, $p(\bar{\mathbf{y}}(k,n)|H_1(k,n))$ and $p(\bar{\mathbf{y}}(k,n)|H_0(k,n))$ are the likelihood functions, and $\gamma > 0$ is a threshold which is found by $P_{FA} = \int_{\{\bar{\mathbf{y}}(k,n):L_G(\bar{\mathbf{y}}(k,n))>\gamma\}} p(\bar{\mathbf{y}}(k,n)|H_0(k,n))d\bar{\mathbf{y}}(k,n)$ [19], where P_{FA} is the false alarm rate. With the independency assumption in Section II, the likelihood functions in (7) can be written as

$$\begin{aligned} &p(\bar{\mathbf{y}}(k,n)|H_0(k,n)) = \\ &\qquad \prod_{m=1}^{M}\prod_{\kappa=-K}^{K}\prod_{\eta=0}^{N-1} p(Y_m(k+\kappa,n-\eta)|H_0(k,n)), \end{aligned} \tag{8}$$

$$\begin{aligned} &p(\bar{\mathbf{y}}(k,n)|H_1(k,n)) = \\ &\qquad \prod_{m=1}^{M}\prod_{\kappa=-K}^{K}\prod_{\eta=0}^{N-1} p(Y_m(k+\kappa,n-\eta)|H_1(k,n)), \end{aligned} \tag{9}$$

By taking the logarithm in (7) and with (8), (9), we have

$$\begin{aligned} &\ln L_G(\bar{\mathbf{y}}(k,n)) = \\ &\sum_{m=1}^{M}\sum_{\kappa=-K}^{K}\sum_{\eta=0}^{N-1} \ln\left[\frac{p(Y_m(k+\kappa,n-\eta)|H_1(k,n))}{p(Y_m(k+\kappa,n-\eta)|H_0(k,n))}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \ln\gamma. \end{aligned} \tag{10}$$

As shown in (10), the GLRT function is the summation of local information which is held by each microphone.

IV. Distributed VAD

As mentioned in section III, the GLRT function in (10) is nothing but a summation of local information. Therefore, the GLRT function can be obtained by solving the distributed averaging problem [14], i.e., $a = (1/M)\sum_{m=1}^{M} p_m$, where p_m indicates the local value at the mth node. The problem in (10) involves the computation of the following quantity:

$$p_m = \sum_{\kappa=-K}^{K}\sum_{\eta=0}^{N-1} \ln\left[\frac{p(Y_m(k+\kappa,n-\eta)|H_1(k,n))}{p(Y_m(k+\kappa,n-\eta)|H_0(k,n))}\right]. \tag{11}$$

Standard consensus propagation algorithms, such as random gossip [14], ADMM [15] and PDMM [16], can be used to

ISBN 978-90-827970-1-5

compute (10) in a distributed way. In a gossip algorithm, each node communicates with one neighbor in each time slot. More specifically, in a certain time slot, node i will contact some neighboring node j with probability $P_{i,j}$. In this paper, we uniformly at random choose a neighbouring node of the ith node. Then nodes i and j set their estimates of a equal to the average of their current values [14]. After random gossip has converged, each node will get an accurate estimate of a. We apply the random gossip method to obtain (10) distributedly. The distributed detection procedure is shown in Algorithm 1.

Algorithm 1 Distributed VAD

Description:

1: **for** $m = 1, 2, 3, ..., M$
2: Input $\bar{\mathbf{y}}_m(k, n)$.
3: Estimate $\phi_{S_m}(k+\kappa, n-\eta)$, $\phi_{V_m}(k+\kappa, n-\eta)$, $\kappa = -K, ..., K$, $\eta = 0, ..., N-1$ using the model-based noise PSD estimator (see Section V).
4: Get the local information in (10), i.e., $p_m = \sum_{\kappa=-K}^{K} \sum_{\eta=0}^{N-1} \ln \left[\frac{p(Y_m(k+\kappa,n-\eta)|H_1(k,n))}{p(Y_m(k+\kappa,n-\eta)|H_0(k,n))} \right]$.
5: **end for**
Apply random gossip to caculate $\ln L_G(\bar{\mathbf{y}}(k, n))$:
6: **for** $g = 1, 2, 3, ..., G$
7: At the gth iteration, randomly select a node i and activate one of its neighbours, i.e., node j.
8: Node i and node j update their estimations by averaging their current values.
9: **end for**
10: Repeat step 6-step 9 until convergence. We can make a global solution of the GLRT function in each node.
11: Make the decision about whether it is $H_0(k, n)$ or $H_1(k, n)$ based on (10) in each node.

V. Model-based noise PSD estimation

So far, we have assumed that the speech PSD and the noise PSD are known. In practice, these PSDs need to be estimated. For noise PSD tracking, the well-known MMSE [9], [10] method and the MS [11] method work well with stationary noise. For non-stationary noise, however, they have limited performance [13]. In this section, we briefly summarize a model-based noise PSD estimation method which are able to track non-stationary noise, such as babble noise. A detailed description can be found in [12].

At each microphone, the T time domain samples of the noisy signal are observed as $\mathbf{y}'_m = \mathbf{s}'_m + \mathbf{v}'_m$. The noise PSD mentioned in (5) and (6) is defined as [17]

$$\phi_{V_m}(k, n) = \lim_{T\to\infty} \frac{1}{T} E[|V_m(k, n)|^2 | \mathbf{y}'_m]. \quad (12)$$

The conditional expectation in (12) is the second moment of the density $p(|V_m(k, n)|^2|\mathbf{y}'_m)$ which leads to another form of (12), i.e.,

$$\phi_{V_m}(k, n) = \lim_{T\to\infty} \frac{1}{T} \left[\int_{\mathbb{R}^{T\times 1}} |V_m(k, n)|^2 p(\mathbf{v}'_m|\mathbf{y}'_m) d\mathbf{v}'_m \right]. \quad (13)$$

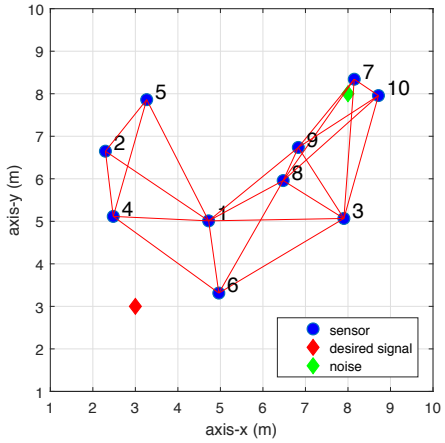


Fig. 1: Room setup. The blue circles indicate microphones, red diamond is the desired signal, green diamond denotes the noise signal, and the red lines are edges.

To compute the posterior $p(\mathbf{v}'_m|\mathbf{y}'_m)$, we introduce statistical models which are denoted as $\{\mathcal{M}_l\}_{l=1}^{L}$ to explain the data. These models can be incorporated into (13). Then the model based PSD can be expressed as

$$\begin{aligned} &\phi_{V_m}(k, n) \\ \approx &\frac{1}{T} \sum_{l=1}^{L} p(\mathcal{M}_l|\mathbf{y}'_m) \left[\int_{\mathbb{R}^{T\times 1}} |V_m(k, n)|^2 p(\mathbf{v}'_m|\mathbf{y}'_m, \mathcal{M}_l) d\mathbf{v}'_m \right] \\ = &\sum_{l=1}^{L} p(\mathcal{M}_l|\mathbf{y}'_m) \phi_{V_m}(k, n|\mathcal{M}_l) \end{aligned} \quad (14)$$

With the model probabilities $\{p(\mathcal{M}_l|\mathbf{y}'_m)\}_{l=1}^{L}$, the models explaining the data well are given more weight than the other models. We use the autoregressive models to model the speech and noise signals. In practice, the AR-parameters are pre-trained and stored in speech and noise codebooks. Finally, we get a model-averaged version of the MMSE estimator [9], [10] as

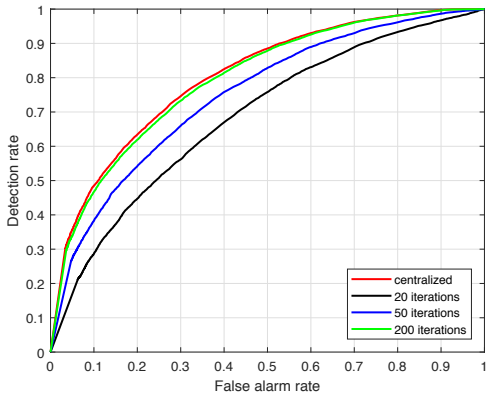
$$\hat{\phi}_{V_m}(k, n) = \frac{1}{T} \sum_{l=1}^{L} p(\mathcal{M}_l|\mathbf{y}'_m)[|\mathbf{f}^H \hat{\mathbf{v}}'_{m,l}|^2 + \mathbf{f}^H \hat{\Sigma}_l \mathbf{f}] \quad (15)$$

The definitions of $\mathbf{f}$, $\hat{\mathbf{v}}'_{m,l}$ and $\hat{\Sigma}_l$ can be found in [12]. A more detailed derivation of the model-based noise PSD estimation is available in [12]. The estimated speech PSD can be obtained in a similar way. Inserting (15) and the speech PSD estimate in (5) and (6), and with the distributed estimation of $\ln L_G(\bar{\mathbf{y}}(k, n))$, the decision is made by using (10).

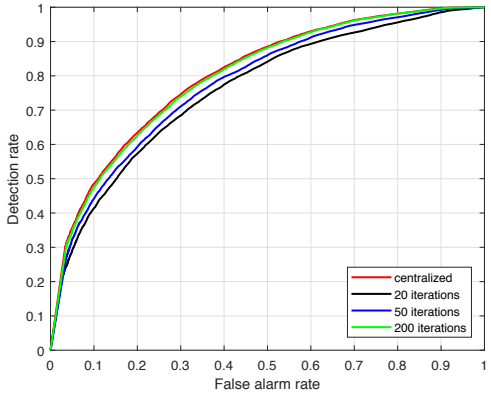
TABLE I: Location of the microphone and the corresponding input SNR. The axis z is 1.5 m for all microphones.

mic index	1	2	3	4	5
(x,y)	(4.7, 5.0)	(2.3, 6.6)	(7.9, 5.1)	(2.5, 5.1)	(3.3, 7.9)
iSNR(dB)	4.3	2.8	-4.1	7.5	-0.3
mic index	6	7	8	9	10
(x,y)	(5.0, 3.3)	(8.1, 8.3)	(6.5, 6.0)	(6.8, 6.7)	(8.7, 8.0)
iSNR(dB)	7.5	-22.7	-3.6	-6.9	-17.4

ISBN 978-90-827970-1-5

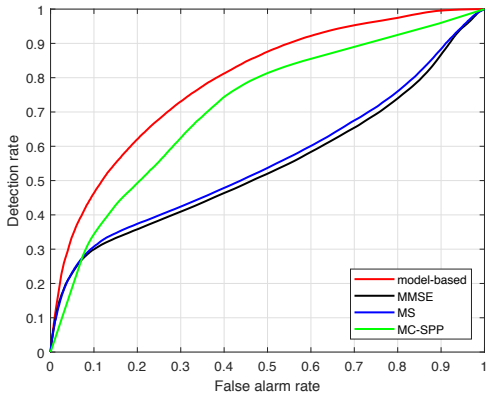


(a) The ROC curve at microphone 7 with different number of iterations, $K = 1$ and $N = 1$.

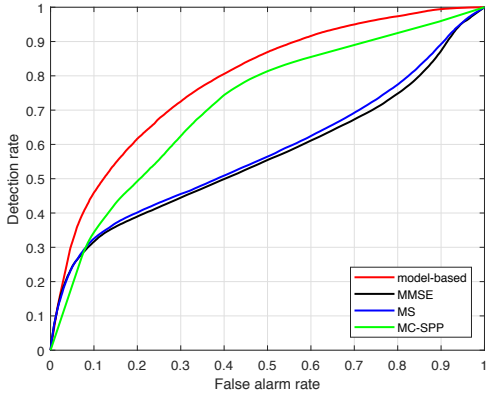


(b) The ROC curve at microphone 8 with different number of iterations, $K = 1$ and $N = 1$.

Fig. 2: The convergence performance of the distributed VAD.



(a) The ROC curve by only using neighbouring frequency information ($K = 1$ and $N = 1$).



(b) The ROC curve by using neighbouring frequency information ($K = 1$) and neighbouring frame information ($N = 2$).

Fig. 3: The VAD performance by using different noise PSD estimation methods, and the MC-SPP based method.

VI. Experimental results

In this section, numerical experiments are used to demonstrate the performance of the distributed VAD in a simulated room acoustic. We simulate a room of size $10\ \mathrm{m} \times 10\ \mathrm{m} \times 3\ \mathrm{m}$, with the room impulse response generated by using the image source model method [20]. The reverberation time is $T_{60} \approx 150\ \mathrm{ms}$. As shown in Fig. 1, we have 10 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 communication distance is set to be $3.5\ \mathrm{m}$. The desired signal is located at (3,3,1.5). The noise is simulated as a point source located at (8,8,1.5). The noise signal is scaled to have the same power as the desired signal. The position of the nodes and corresponding input SNR information are shown in Table I. The model-based noise PSD estimator which is introduced in Section V needs the codebooks being trained in advance. In the experiments, the codebooks are trained by using the LPC-VQ method [21]. We train a speech codebook with 64 entries (32 entries for male speaker and 32 for female speaker). The noise codebook contains 16 entries (4 entries for babble, restaurant and exhibition noise and 2 entries for street and station noise). The speech training data is from the TIMIT database [22] and the noise training data is from the NOIZEUS database [23]. The testing speech is taken from the CHiME corpus [24], and the testing noise is from part of the NOIZEUS database which are not used in the training step. We set the clean signal received by the first microphone as the desired signal, i.e., $X_1(k, n)$. In order to get the ROC curve, we set a power threshold to the normalized subband energy of the desired signal to get a ground truth decision matrix. More specifically, the frequency bands with higher energy than the threshold are marked as speech presence, and the others are marked as speech absence. For the distributed consensus step, we apply the random gossip method to get the distributed averaging result. For comparison, we also evaluate the performance of MMSE and MS based VAD methods, i.e., apply the MMSE noise estimator or the MS noise estimator instead of the model-based method in step 3 of the Algorithm 1. When applying the MMSE estimator and the MS estimator, the estimation of the speech PSD is obtained by subtracting the estimated noise PSD from the noisy signal PSD.

In the first experiment, we evaluate the convergence perfor-

ISBN 978-90-827970-1-5

mance of the distributed VAD at different microphones. We consider babble noise here, and only inter-band information is used in the VAD. Fig. 2 (a) shows the ROC curve of the 7th microphone with different number of iterations of the random gossip, and Fig. 2 (b) illustrates the performance of the 8th microphone. We notice from Fig. 2 that the convergence speed of the distributed VAD is different at different microphones. The microphone with higher input SNR converges faster than the one with lower input SNR. The reason is that the higher input SNR at the microphones near the desired signal will lead to better speech PSD estimate which will contribute to better VAD performance.

In the next experiment, the VAD performance when using different methods in the noise PSD estimate step is studied. Each experiment is repeated 5 times by adding different types of noise (babble, restaurant, exhibition, street and station). The mean ROC curves at the 7th microphone are illustrated in Fig. 3. The number of iterations of the random gossip method is set to be 200. As comparison, we also test the centralized MC-SPP based multichannel VAD. We set the parameters the same as in [8]. As shown in Fig. 3, the model-based distributed VAD performs better than the other three methods. This is because the model-based noise PSD estimation outperforms the MMSE and MS methods in tracking non-stationary noise, which also contributes to a better VAD performance. By comparing Fig. 2 (a) with Fig. 2 (b), we can notice that the detection performance gets slightly better by taking into account the inter-frame information, especially with the MMSE and MS based VAD methods.

VII. Conclusions

In this paper, we proposed a distributed multichannel VAD by using the WASN. By taking advantage of the model-based noise PSD estimation method, the proposed method are able to obtain robust performance under non-stationary noise condition. We formed the distributed VAD by using the GLRT theory. And the global decision can be made by considering the likelihood functions at all channels. Finally, the distributed VAD can be obtained by solving the distributed averaging problem. We utilized the random gossip 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 VAD 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. It was also worthwhile noticing that the proposed method outperformed the MMSE, MS based VAD as well as the MC-SPP based method.

References

[1] F. de la Hucha Arce, M. Moonen, M. Verhelst, and A. Bertrand, "Adaptive quantization for multi-channel Wiener filter-based speech enhancement in wireless acoustic sensor networks," *Wireless Commun. Mobile Comput.*, vol. 2017, 2017, Art. no. 3173196.

[2] V. M. Tavakoli, J. R. Jensen, R. Heusdens, J. Benesty, and M. G. Christensen, "Distributed max-SINR speech enhancement with ad hoc microphone arrays," *IEEE Int. Conf. Acoust. Acoustics, Speech, Signal Process. (ICASSP)*, pp. 151–155 2017.

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

[4] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing lett.*, vol. 6, pp. 1–3, 1999.

[5] J. Ramirez, J. C. Segura, C. Benitez, A. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech commun.*, vol. 42, no. 3-4, pp. 271–287, 2004.

[6] L. K. Hamaidi, M. Muma, C. Benitez, and A. M. Zoubir, "Robust distributed multi-speaker voice activity detection using stability selection for sparse non-negative feature extraction," in *Proc. 25th IEEE Eur. Signal. Process. Cof. (EUSIPCO)*, pp. 161–165, 2017.

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

[8] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE Int. Conf. Acoust. Acoustics, Speech, Signal Process. (ICASSP)*, pp. 4266–4269, 2010.

[10] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[12] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *IEEE Int. Conf. Acoust. Acoustics, Speech, Signal Process. (ICASSP)*, 2018.

[13] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, and J. B. Boldt, "A study of noise PSD estiamtors for single channel speech enhancement" in *IEEE Int. Conf. Acoust. Acoustics, Speech, Signal Process. (ICASSP)*, 2018.

[14] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE trans. inf. theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[15] R. Zhang and J. Kwok, "Asynchronous distributed ADMM for consensus optimization," in *Int. Conf. Mach. Learn.*, pp. 1701–1709, 2014.

[16] G. Zhang and R. Heusdens, "Distributed optimization using the primal-dual method of multipliers," *IEEE Trans. Signal Inf. Process. Netw.*, 2017.

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

[18] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing II: Detection Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[21] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Springer Science & Business Media, vol. 159, 2012.

[22] J. W. Lyons, "DARPA TIMIT acoustic-phonetic continuous speech corpus," *Technical Report NISTIR 4930, National Institute of Standards and Technology*, 1993.

[23] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech commun.*, vol. 49, no. 7-8, pp. 588–601, 2007.

[24] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Interspeech*, pp. 1918-1921, 2010.

ISBN 978-90-827970-1-5 © EURASIP 2018