MODELING INTER-NODE ACOUSTIC DEPENDENCIES WITH RESTRICTED BOLTZMANN MACHINE FOR DISTRIBUTED MICROPHONE ARRAY BASED BSS

Keisuke Kinoshita, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan

ABSTRACT

An accurate estimation of a source activity information is essential for many speech enhancement algorithms including blind source separation (BSS). In this paper, we propose a novel BSS method that accurately models and estimates the source activity in distributed microphone array (DMA) scenarios. In DMA scenarios, microphones (or in more general term, microphone-nodes) are often spatially distributed to a great degree. If there are multiple source signals in such an environment, the level of each source signal at each microphone-node varies significantly, thus the source activities observable at one microphone-node should be significantly different from those of other nodes. Therefore, it is essential to assume *node-specific* source activities in DMA scenarios. In the proposed method, the estimation of the node-specific source activities are done by integrating node-wise clustering-based BSS processings based on inter-node acoustic dependencies, i.e., a *co-occurrence* of the source activities among nodes. To model the co-occurrence relationship, we employ Restricted Boltzmann Machine (RBM) in a similar manner as it is used for collaborative filtering. This paper introduces a probabilistic formulation of the proposed method, and experimentally demonstrates how essential it is to estimate the node-specific source activities for distributed microphone array based BSS.

Index Terms— Distributed microphone array, blind source separation, node-specific source activity, restricted Boltzmann machine.

1. INTRODUCTION

In recent years, portable devices equipped with microphones, such as PDA, smartphones and laptop computers, have rapidly spread in our daily life. By making them work collaboratively through sensor network technologies, we may be able to form a virtual microphone array that can solve challenging speech processing tasks [1–8]. In this paper, such microphone array will be referred to as a distributed microphone array (DMA), and each independent recording device will be referred to as a microphone-node within the DMA. Although there are many challenges to be overcome [9], DMA has recently started attracting a lot of attention as a promising alternative to conventional microphone array. In this paper, we propose a novel blind source separation (BSS) approach that can be considered as an extension of conventional clustering-based BSS algorithms [10–13] to appropriately deal with DMA scenarios.

For last decades, considerable research has been undertaken to achieve better speech enhancement using a co-located (i.e., concentrated at one place) microphone array. For example, they include studies on beamforming techniques such as the delay-and-sum beamformer [14], generalized side-lobe canceler (GSC) [15] and minimum variance distortionless response (MVDR) beamformer [16], multichannel Wiener filter [16–18], and BSS techniques such as Independent Component Analysis [19] and clustering-based BSS algorithms [10–13] like DUET. These studies have shown that multi-channel speech enhancement algorithms provide superior performances in adverse environments compared with single-channel approaches. For example, the clustering-based BSS algorithms [11–13], which we focus on in this paper, are shown to perform effectively in various BSS scenarios including underdetermined cases [20]. It first calculates a direction-of-arrival (DOA) type of feature for each time-frequency (TF) bin, and then, with the sparseness assumptions [10], it clusters these features into clusters associated with different speakers. Finally, the clustering results which indicate activity patterns of target speakers (hereafter, source activities) are directly used to form a TF separation mask for each target speaker. Although the sparseness assumption may hold just approximately, meaning that often more than 2 sources may indeed exist at one TF bin, it achieves very good BSS performance just by considering that a TF bin belongs to the source with the highest energy. Since the clustering-based BSS is very effective in various environments and its probabilistic formulation provides high flexibility for extension, one may think it is quite interesting and attractive to extend it to DMA scenarios.

However, there is a large gap between co-located scenarios and DMA scenarios in terms of clustering-based BSS, which should be carefully taken into account. In conventional clustering-based BSS approaches that use co-located microphones, the source activities are assumed to be common to all microphones. This assumption is natural and adequate in co-located microphone scenarios, since all microphones are located close to each other. However, this assumption can be violated in DMA environments. As the term DMA shows, microphones, or in more general term microphone-nodes are quite often spatially distributed to a great degree (cf. Fig. 2). If so, the level of each target signal at each microphone-node varies significantly, hence the source activities that are observable at a microphone-node may be significantly different from those of other microphone-nodes. Although the estimation of the *node-specific* source activities has not been well studied in past literatures [1–5] yet, some studies in fact assumes that they are known a priori and proposed the distributed implementation of multi-channel linear minimum mean square error (MMSE) filtering [2].

Apparently, one easiest way to estimate the node-specific source activities is to apply the clustering-based BSS to each node separately, i.e., node-wise processing. However, an obvious limitation in such case is that the node-wise BSS cannot benefit from the other nodes, thus cannot work effectively by making good use of DMA.

In this paper, we propose to extend the clustering-based BSS algorithm to deal with the DMA scenarios. Specifically, we focus on the inter-node dependency that, when a source signal is significantly active at a TF bin of one microphone-node, it tends to be active at the same TF bin of neighboring nodes. By modeling this *co-occurrence* relationship among nodes using an appropriate probabilistic model, it estimates the node-specific source activities more accurately than

the simple node-wise processing. To model the co-occurrence relationship, we employ Restricted Boltzmann Machine (RBM) in a similar manner as it is used for collaborative filtering [21].

2. NOTATIONS

The followings are the variables we use in this paper. Frequency indices are omitted from all variables, since all processing will be performed independently for each frequency bin.

I:	Number of nodes
J:	Number of clusters in a node
K:	Number of sources (In this paper, $K = J$.)
x_i :	Observed feature at i-th node
$\mathbf{x}$:	Observed feature of all nodes, i.e., $\mathbf{x} = [x_1, \cdots, x_I]^T$
$n_{i,j}$:	Source activity of j-th cluster at i-th node ($n_{i,j} = 1$ if the cluster is active, and 0 otherwise)
$\mathbf{n}_i$:	Source activity vector at i-th node i.e., $\mathbf{n}_i = [n_{i,1}, \cdots, n_{i,J}]^T$
$\mathbf{n}$:	Source activity vector of all nodes i.e., $\mathbf{n} = [\mathbf{n}_1^T, \cdots, \mathbf{n}_I^T]^T$
a_k :	Variable that indicates common latent source activity
$\mathbf{a}$:	Vector of the common latent source activity i.e., $\mathbf{a} = [a_1, \cdots, a_K]^T$

3. NODE-WISE CLUSTERING-BASED BSS

Let us first review the conventional node-wise clustering-based BSS which serves as a basis of the proposed method. The following explanation will be given by taking the i-th node case as an example.

As it was partly mentioned in the previous section, thanks to the property of speech sparseness, the TF components of observed sound mixtures can be assigned to clusters belonging to different speakers. In statistical signal processing, this operation can be done by first defining a hidden variable $n_{i,j}$ as an indicator of the dominant signal in the mixtures, and then determining its posterior probability $p(n_{i,j}|x_i)$ given an observed feature vector x_i. Here j indicates the cluster/speaker indices.

These posterior probabilities $p(n_{i,j}|x_i)$ can be obtained by estimating the parameters of the generative model of the observed mixture, which is often denoted in a form of a mixture model as:

$$p(x_i; \theta^{(\mathbf{n}_i)}) = \sum_{j=1}^{J} p(n_{i,j}) p(x_i|n_{i,j}; \theta^{(n_{i,j})}), \tag{1}$$

where $\theta^{(n_{i,j})}$ is a set of parameters to determine a shape of the distribution, e.g., mean and variance. The feature vector x_i can be an DOA feature [10], complex normalized observation vector [11, 12], and $p(x_i|n_{i,j}; \theta^{(n_{i,j})})$ can take also various forms such as (complex) Gaussian-like distribution [10, 11] and complex Watson distribution [12]. In many cases, the parameters $\theta^{(\mathbf{n}_i)}$ are estimated in the maximum likelihood sense based on the observed signal x_i. After the parameter estimation, the posterior probability $p(n_{i,j}|x_i)$ indicates a source activity of the j-th speaker at the i-th node, thus can be directly used as a separation mask.

4. THE PROPOSED METHOD

Although the node-wise clustering-based BSS solely can estimate the node-specific source activities $p(n_{i,j}|x_i)$, it has an obvious limitation that it cannot benefit from the other nodes to estimate the source activities. In the proposed method, we employ a probabilistic model that can well capture the relationship among the node-specific source activities, i.e., *co-occurrence* relationship, and integrate the model with the node-wise clustering-based BSS to improve its performance.

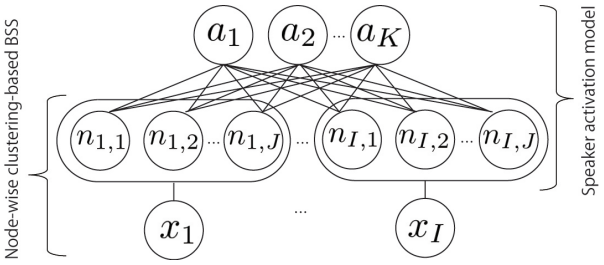

Fig. 1. Generative model for DMA scenario

4.1. Overview of the proposed model

Figure 1 shows a graphical model of the proposed method. The bottom 2 layers indicate a node-specific observation model which basically consists of I node-wise clustering-based BSS, while the top 2 layers correspond to the probabilistic model that models the co-occurrence relationship among node-specific source activities $\mathbf{n}_1, \ldots, \mathbf{n}_I$. To capture the co-occurrence relationship, we introduce a hidden variable $\mathbf{a}$, that can be physically interpreted as a variable that inherently represents the common latent source activities behind a target acoustic scene. By doing so, we can directly apply RBM and model joint probability $p(\mathbf{n}, \mathbf{a}; \theta^{(w)})$. Note that RBM is an appropriate model to capture the co-occurrence relationship as it can be used for, for example, collaborative filtering [21]. As you can see from Fig. 1, the observation at each node and node-specific source activities are now connected to the other nodes via the hidden variable $\mathbf{a}$, enabling information exchange among nodes.

Note that we employ sparseness assumption at each node i, which means that only one component within $\mathbf{n}_i$ equals to 1 and the others are 0. To model this situation, we use Bernoulli-Bernoulli RBM with softmax visible units, utilized, for example, in [21, 22].

4.2. Likelihood function

Overall likelihood function of the proposed method can be formulated as:

$$L(\theta) = \sum_{\mathbf{a}} \sum_{\mathbf{n}} p(\mathbf{x}, \mathbf{n}, \mathbf{a}; \theta), \tag{2}$$

$$= \sum_{\mathbf{a}} \sum_{\mathbf{n}} p(\mathbf{n}, \mathbf{a}; \theta^{(w)}) p(\mathbf{x}|\mathbf{n}; \theta^{(n)}),$$

$$= \sum_{\mathbf{a}} \sum_{\mathbf{n}} p(\mathbf{n}, \mathbf{a}; \theta^{(a)}) \prod_i p(x_i|n_{i,j}; \theta^{(n_i)}), \tag{3}$$

where $\theta = \{\theta^{(w)}, \theta^{(n)}\}$. $\theta^{(w)}$ indicates the parameters of the model that represents the co-occurrence of source activities among nodes, while $\theta^{(n)}$ is the parameters of the node-wise clustering-based BSS. Hereafter, $p(\mathbf{n}, \mathbf{a}; \theta^{(w)})$ will be referred to as source activity model.

By comparing eq. (3) with eq. (1), we can intuitively understand how the proposed method can be seen as an extension of the conventional clustering-based BSS. The term corresponding to the prior distribution of the speaker activity in eq. (1) (i.e., $p(n_{i,j})$) is now replaced with $p(\mathbf{n}, \mathbf{a}; \theta^{(w)})$ in eq. (3) to take the co-occurrence relationship among the node-specific source activities $\mathbf{n}_1, \ldots, \mathbf{n}_I$ into account.

4.3. Source activity model

The source activity model $p(\mathbf{n}, \mathbf{a}; \theta^{(w)})$ is formulated by using RBM denoted as follows.

$$p(\mathbf{n}, \mathbf{a}; \theta^{(w)}) = \frac{1}{Z} \exp(-E(\mathbf{n}, \mathbf{a})),$$

where

$$\begin{aligned} E(\mathbf{n}, \mathbf{a}) &= -\sum_{i=1}^{I} \left(\sum_{j=1}^{J} b_{i,j} n_{i,j} + \sum_{j=1}^{J} \sum_{k=1}^{K} n_{i,j} w_{i,j,k} a_k \right) \\ &\quad - \sum_{k=1}^{K} c_k a_k, \\ &= -\sum_{i=1}^{I} \left(\mathbf{b}_i^T \mathbf{n}_i + \mathbf{n}_i^T W_i \mathbf{a} \right) - \mathbf{c}^T \mathbf{a}, \end{aligned}$$

where $\theta^{(w)} = \{W_i, \mathbf{b}_i, \mathbf{c}\}$ and Z is a normalization term.

Conditional probabilities between the visible unit $\mathbf{n}$ and the hidden unit $\mathbf{a}$ can be represented as follows.

$$p(a_k = 1|\mathbf{n}) = \sigma(c_k + \sum_{i=1}^{I} \sum_{j=1}^{J} n_{i,j} w_{i,j,k}), \tag{4}$$

$$p(n_{i,j} = 1|\mathbf{a}) = \frac{\exp(b_{i,j} + \sum_k w_{i,j,k} a_k)}{\sum_{j'} \exp(b_{i,j'} + \sum_k w_{i,j',k} a_k)}, \tag{5}$$

where $\sigma(\cdot)$ corresponds to the sigmoid funciton $\sigma(x) = 1/(1 + \exp(-x))$. Equation (5) is different from the standard RBM, because of the sparseness assumption we made [21, 22].

One important modification that we have to make to the standard RBM is the following additional conditional probability. While it is possible for the standard RBM to directly observe features for the visible unit, the proposed method observes the features via the node-wise clustering-based BSS (cf. Fig. 1). It means that the visible unit has to be treated a latent variable that cannot be directly observable. Thus, the input feature to the visible unit has to be determined/sampled by considering contributions from both the bottom layer (i.e., node-wise clustering-based BSS) and the top layer (i.e., $p(n_{i,j} = 1|\mathbf{a})$). Finally, we have an additional conditional probability denoted as:

$$p(n_{i,j} = 1|\mathbf{a}, \mathbf{x}) = \frac{p(x_i|n_{i,j} = 1) p(n_{i,j} = 1|\mathbf{a})}{\sum_{j'=1}^{J} p(x_i|n_{i,j'} = 1) p(n_{i,j'} = 1|\mathbf{a})}. \tag{6}$$

Now let us describe how the parameters of the source activity model are estimated. As in the standard RBM, we estimate the parameters $\theta^{(w)}$ by gradient decent using contrastive divergence [23, 24]. The gradient of each parameter can be written as follows.

$$\begin{aligned} \frac{\partial L_R(\theta)}{\partial w_{i,j,k}} &= \frac{1}{T} \sum_t \hat{n}_{t,i,j} \sigma(c_k + \sum_{i'} \sum_{j'} \hat{n}_{t,i',j'} w_{i',j',k}) \\ &\quad - \frac{1}{T} \sum_t \tilde{n}_{t,i,j} \sigma(c_k + \sum_{i'} \sum_{j'} \tilde{n}_{t,i',j'} w_{i',j',k}), \\ \frac{\partial L_R(\theta)}{\partial b_{i,j}} &= \frac{1}{T} \sum_t \hat{n}_{t,i,j} - \frac{1}{T} \sum_t \tilde{n}_{t,i,j}, \\ \frac{\partial L_R(\theta)}{\partial c_k} &= \frac{1}{T} \sum_t \sigma(c_k + \sum_{i',j'} \hat{n}_{t,i',j'} w_{i',j',k}) \\ &\quad - \frac{1}{T} \sum_t \sigma(c_k + \sum_{i',j'} \tilde{n}_{t,i',j'} w_{i',j',k}), \end{aligned}$$

where $L_R(\theta)$ simply indicates log of the likelihood function introduced in eq. (2), and t is the time frame index. In the standard RBM training, $\hat{n}_{t,i,j}$ corresponds directly to the input visible features, and $\tilde{n}_{t,i,j}$ to the reconstructed input feature through RBM.

In the proposed method, while the way of calculating $\tilde{n}_{t,i,j}$ remains exactly the same as in the standard RBM, $\hat{n}_{t,i,j}$ has to be sampled by Gibbs sampling, since it is a latent variable that is not directly observable. The following is the procedure to calculate these 2 variables.

- Estimation of $\hat{n}_{t,i,j}$
  - 0. Estimate $p(n_{i,j} = 1|x_j)$ by using a conventional clustering-based BSS.
  - 1. Sample an initial value of $\hat{n}_{t,i,j}$ based on $p(n_{i,j} = 1|x_j)$.
  - 2. Iterate the following operation m times ($m = 1$ in this paper)
    - 2-(a). Sample $\hat{a}_{t,k}$ using $p(a_k = 1|\mathbf{n}; \theta^{(w)})$ and current $\hat{n}_{t,i,j}$.
    - 2-(b). Sample new $\hat{n}_{t,i,j}$ using $p(n_{i,j} = 1|\mathbf{a}, \mathbf{x}; \theta^{(w)})$, $\hat{a}_{t,k}$ and $\mathbf{x}_t$.
  - 3. Use $\hat{n}_{t,i,j}$ obtained at 2-(b) for the gradient calculation.
- Estimation of $\tilde{n}_{t,i,j}$
  - 1. Sample an initial value of $\tilde{n}_{t,i,j}$ using $p(n_{i,j} = 1|x_{t,j})$
  - 2. Iterate the following operation m times ($m = 1$ in this paper)
    - 2-(a). Sample $\tilde{a}_{t,k}$ using $p(a_k = 1|\mathbf{n}; \theta^{(w)})$ and current $\tilde{n}_{t,i,j}$
    - 2-(b). Sample new $\tilde{n}_{t,i,j}$ using $p(n_{i,j} = 1|\mathbf{a}; \theta^{(w)})$ and $\tilde{a}_{t,k}$ obtained at the previous step
  - 3. Use $\tilde{n}_{t,i,j}$ obtained at 2-(b) for the gradient calculation

4.4. Overall parameter estimation procedure

The following summarizes the parameter estimation procedure of the proposed method.

0. (Initializaton): Estimate $p(n_{i,j} = 1|x_j)$ using a conventional clustering-based BSS.
1. Iterate the following steps until convergence
   - 1-(a). Calculate $\partial L_R(\theta)/\partial w_{i,j,k}$, $\partial L_R(\theta)/\partial b_{i,j}$, $\partial L_R(\theta)/\partial c_k$ as in the section 4.3.
   - 1-(b). Update $w_{i,j,k}$, $b_{i,j}$, c_k as follows.

$$\begin{aligned} w_{i,j,k} &\leftarrow w_{i,j,k} + \mu \frac{\partial L_R(\theta)}{\partial w_{i,j,k}} \\ b_{i,j} &\leftarrow b_{i,j} + \mu \frac{\partial L_R(\theta)}{\partial b_{i,j}} \\ c_k &\leftarrow c_k + \mu \frac{\partial L_R(\theta)}{\partial c_k} \end{aligned}$$

2. Use $p(n_{i,j} = 1|\hat{\mathbf{a}}_t, \mathbf{x}_t)$ as soft masks to obtain enhanced signals

Optionally, we can also iteratively update the parameters $\theta^{(n)}$ by solving $\partial L_R(\theta)/\partial \theta^{(n_i)} = 0$ after the step 1-(b), but this part is omitted from this paper because of space limitations.

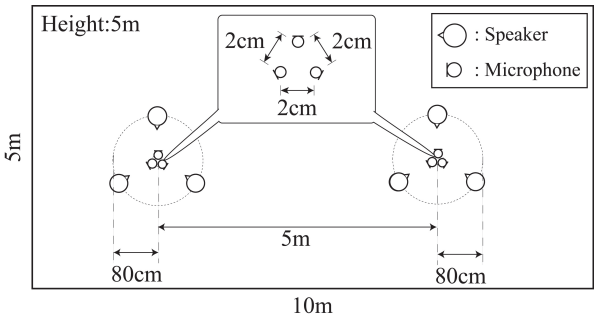

Fig. 2. Experimental condition

5. EXPERIMENT

In this section, we evaluate the effectiveness of the proposed method in comparison with conventional methods.

5.1. Acoustic conditions

To evaluate the proposed method, we simulated a DMA environment depicted in Fig. 2 by using the image method [25]. This scenario simulates a situation where 2 groups, each of which involves 3 people, are having conversations at the coffee tables (i.e., $J = 6$ and $K = 6$). On top of each coffee table, there is a co-located 3-element microphone array that should be regarded as a microphone node (i.e., $I = 2$) within this DMA scenario. The size of the simulated room was 10 m (W)×5 m (D)×5 m (H), and 2 microphone nodes are separated by 5 m.

We simulated 4 reverberant conditions with reverberation time (T_{60}) of 0.2, 0.4, 0.6 and 0.8 seconds, respectively. White noise are added to each microphone with SNR of 10 dB.

5.2. Tasks and other conditions

Our objective is to separate 6 simultaneous speakers. For comparison with the proposed method, we employed a state-of-the-art clustering-based BSS algorithm [11], and performed BSS using all 6 microphones to obtain the source activities common to all nodes. It will be referred to as "global clustering". We also applied the same method separately to each node to obtain the node-specific source activities. This method will be referred to as "node-wise clustering". The node-wise clustering was performed with 2 different initialization schemes. The first scheme is to simply initialize all the parameters randomly. The second scheme is to first perform the global clustering and then use obtained posterior $p(n_j|\mathbf{x})$ as an initial value of the corresponding posterior of the node-wise clustering, and run expectation-maximization iteration of the node-wise clustering several times (3 times in this paper). These 2 types of node-wise clustering results were respectively used for initialization of the proposed method. In total, we have 5 different BSS schemes to be compared. For all the 5 methods described above, the separated signals are generated by applying soft masks to the signals observed at the closest microphone-node to each speaker. In this experiment, we assumed that the co-occurrence patterns of the source activities are frequency independent, thus one common latent source activity model were estimated for all the frequency bins. The step size parameter μ was set to 0.01. The sampling frequency was 8 kHz.

The results are evaluated based on the Signal-to-Interference Ratio (SIR) [26]. Twenty random combinations of speech utterances of different speakers from the TIMIT database [27] are used. The results shown below are obtained by averaging over all combinations.

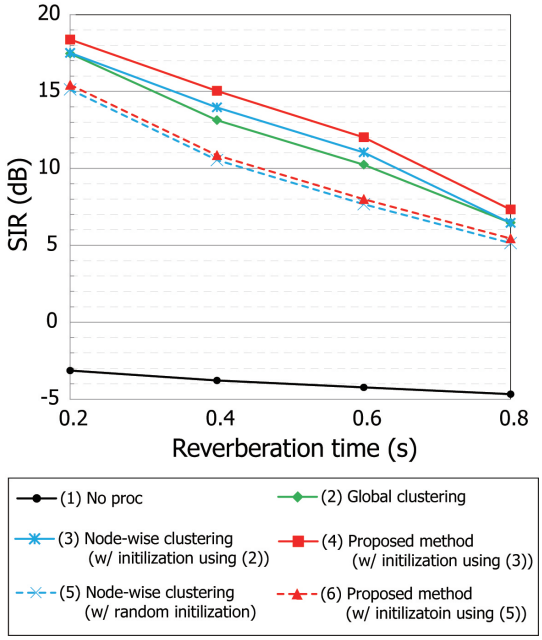

Fig. 3. Experimental result

5.3. Results

Figure 3 shows SIRs of the processed signals and unprocessed signal . The first thing we can notice is that all methods achieved higher SIRs than the unprocessed signal. Now, let us compare 3 solid lines. By comparing the performance of "(3) node-wise clustering (w/ initialization using global clustering)" with that of the "(2) global clustering", we can see better or at least comparable performances. This result partly shows the advantage of assuming node-specific source activities. If we take a look at the performance of the "(4) proposed method (w/ initialization using (3))", we can confirm the benefit of incorporating co-occurrence relationship among nodes. This benefit can also be confirmed by comparing two dashed lines obtained with random initialization. In both cases, the proposed method successfully outperformed the node-wise clustering, and showed the potential importance of modeling the co-occurrence relationship under DMA environments.

6. CONCLUSION

This paper proposed an extension of the clustering-based BSS algorithm to deal with DMA scenarios. In DMA scenarios with multiple source signals, the level of each source signal at each microphone-node tends to vary significantly, thus accordingly source activity observable at each microphone-node differs from node to node. To model this situation, we proposed a method to estimate node-specific source activities by integrating node-wise clustering-based BSS in a probabilistic manner based on a co-occurrence of the activities among nodes. The co-occurrence relationship was modeled by RBM as it was used for collaborative filtering. Experimental results showed the advantage of the proposed method over the conventional node-wise clustering and global-clustering in adverse environments, and demonstrated the potential importance of modeling the co-occurrence relationship to improve BSS performance under DMA scenarios.

7. REFERENCES

[1] S. Doclo, T. Bogaert, M. Moonen, and J. Wouters, "Reduced bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, pp. 38–51, 2009.

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

[3] I. Himawan, I. Mccowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19(4), pp. 661–676, May 2011.

[4] F. Nesta and M. Omologo, "Cooperative Wiener-ICA for source localization and separation by distributed microphone arrays," in *Proc. IEEE Int'l Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 181–184.

[5] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Distributed microphone array processing for speech source separation with classifier fusion," in *Proc. of IEEE Int'l Workshop on Machine Learning for Signal Processing*, September 2012.

[6] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. of IEEE Workshop on Applications of Signal Process. to Audio and Acoust.*, 2009, pp. 161–164.

[7] S. Wehr R. Lienhart, I. Kozintsev and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *Proc. IEEE Int'l Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. 14–17.

[8] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," in *Proc. Int'l Workshop on Acoust. Echo and Noise Control (IWAENC).*, 2003, pp. 14–17.

[9] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *Proc. IEEE Symposium on Communications and Vehicular Technology (SCVT)*, 2011, pp. 1–6.

[10] O. Yilmaz and S. Rickard, "Blind separation of speech mixture via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[11] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 516–527, March 2011.

[12] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE Int'l Conf. Acoust., Speech, Signal Process. (ICASSP)*, September 2010, pp. 241–244.

[13] T. Yoshioka M. Delcroix T. Nakatani, S. Araki and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21(12), pp. 2516–2531, Dec. 2013.

[14] J. L. Flanagan, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78(11), pp. 1508–1518, 1985.

[15] L. J. Grifiths and C. W. Jim, "An alternative approach to linear constrained adaptive beamforming," *IEEE Trans. Antennas propagat.*, vol. AP-30(1), pp. 27–34, 1982.

[16] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

[17] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, pp. 636–656, 2007.

[18] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 1368–1381, 2011.

[19] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Springer-Verlag, New York, NY, 2005.

[20] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[21] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proc. International conference on Machine learning (ICML)*, 2007, pp. 791–798.

[22] H. Larochelle G. E. Dahl, R. P. Adams, "Training restricted boltzmann machines on word observations," in *Proc. International conference on Machine learning (ICML)*, 2012.

[23] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 1711?1800, pp. 1928–1936, 2002.

[24] G. E. Hinton, "A practical guide to training restricted boltzmann machines," *Univ. of Toronto, Toronto, ON, Canada, Tech. Rep.*, 2010.

[25] J. B. Allen and D. A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65(4), pp. 943–950, 1979.

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.,*, vol. 14(5), pp. 1462–1469, 2006.

[27] W.M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.