Blind source separation using spatially distributed microphones based on microphone-location dependent source activities

Keisuke Kinoshita, Mehrez Souden, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Japan

{kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp, msouden6@mail.gatech.edu

10.21437/Interspeech.2013-235

Abstract

Distributed microphone array (DMA) processing has recently been gathering increasing research interest due to its various applications and diverse challenges. In many conventional multi-channel speech enhancement algorithms that use co-located microphones, such as the multi-channel Wiener filtering and mask-based blind source separation (BSS) approaches, statistics of the target and interference signals are required if we are to design an optimal enhancement filter. To obtain such statistics, we estimate activity information regarding source and interference signals (hereafter, source activity information), that is generally assumed to be *common to all the microphones*. However, in DMA scenarios, the source activities observable at any given microphone may be significantly different from those of others when the microphones are spatially distributed to a great degree, and the level of each signal at each microphone varies significantly. Thus, to capture such source activity information appropriately and thereby achieve optimal speech enhancement in DMA environments, in this paper we propose an approach for estimating *microphone-dependent* source activity, and for performing blind source separation based on such information. The proposed method estimates the activity of each source signal at each microphone, which can be explained by the microphone-independent speech log power spectra and microphone-location dependent source gains. We introduce a probabilistic formulation of the proposed method, and an efficient algorithm for model parameter estimation. We show the efficacy of the proposed method experimentally in comparison with conventional methods in various DMA scenarios.

Index Terms: Distributed microphone array, blind source separation, source activities, microphone-dependent, node-specific,

1. Introduction

Distributed microphone array (DMA) processing is emerging as a promising approach with the potential to solve challenging speech processing tasks efficiently, such as signal enhancement [1–5] and localization [6]. To accomplish such tasks in a DMA framework, diverse challenges have to be addressed [7]. In this paper, we propose a BSS approach that estimates microphone-dependent source activities, which we think is essential for DMA-based speech enhancement as we explain below.

For last decades, considerable research has been undertaken to achieve better speech enhancement. This research has shown that multi-channel speech enhancement algorithms [8–18] provide superior performance in adverse environments compared with single-channel approaches by taking advantage of the spatial information. For example, noise reduction can be achieved using a delay-and-sum beamformer, minimum variance distortionless response (MVDR) beamformer, or generalized sidelobe canceler (GSC) when the source propagation vector, referred to as the steering vector, is known or estimated [8–10]. However, since the estimation of the steering vector is challenging in practice, mainly because of the reverberation [9], strong alternative approaches such as the multichannel Wiener filter (MWF), or more specifically the parameterized, speech-distortion-weighted, MWF (PMWF) [1], were proposed [10–13]. One of the biggest advantages of these approaches compared with steering-vector-based beamforming is that they can construct an optimal enhancement filter if only the statistics of interference and observed signals are available. To obtain these statistics reliably, it is essential to appropriately estimate the activities of target and interference signals. For example, if they can be obtained with reasonable accuracy, PMWF is able to jointly perform source separation and noise reduction efficiently [14]. In contrast to the above beamforming approaches, some researchers have proposed multi-channel speech enhancement approaches that can work effectively even in an underdetermined scenario [15–18]. Interestingly, an essential task for these approaches is also the estimation of the activities of target and interference signals in each time-frequency bin. Based on these estimations, we can estimate a time-frequency mask that extracts target signals assuming speech sparseness [15].

It is important to note that, in the above multi-channel speech enhancement approaches that use co-located microphones, the activities of target and interference signals are assumed to be *global*, i.e., common to all the microphones. This assumption is quite natural and adequate in co-located microphone scenarios, since all the microphones are located close to each other. However, this assumption can be violated in DMA environments. That is, the activities of multiple sources that are observable at a microphone may be significantly different from those of other microphones, when the microphones are spatially distributed to a great degree, and the level of each signal at each microphone varies significantly. Therefore, it is essential to estimate the microphone-dependent source activities to achieve optimal signal enhancement in DMA environments, although this issue has not been well addressed in the literature [1–5]. In fact, in some studies, assuming that these microphone-dependent source activities, or equivalently the source statistics, are known, the distributed implementation of multi-channel linear minimum mean square error (MMSE) filtering was proposed, which can achieve microphone-dependent signal extraction [2].

In this paper, we propose an approach that estimates microphone-dependent source activities and utilizes them to achieve BSS in DMA environments. In contrast to many DMA approaches that assume that a node within the DMA has to consist of multiple microphones [1–5], the proposed method can cope with a node composed of only one microphone. In this paper, we introduce accurate probabilistic models of the multiple speech source activities at each node, i.e., at each microphone. These models are devised based on underlying microphone-independent speech log power spectra and microphone-dependent acoustic channel gains. An efficient model parameter estimation algorithm is also derived based on the expectation-maximization (EM) algorithm. This paper is organized as follows. We first introduce an observation data model in section 2, and then explain the basic idea behind the

[1] The MVDR (or GSC) is a variant of the PMWF.

proposed method in section 3. After introducing an efficient parameter estimation algorithm in section 4, we evaluate the proposed method in various DMA scenarios in comparison with conventional BSS methods and show the effectiveness of the proposed method in section 5.

2. Observation data model

In this section, we introduce data models for the observed signal and their probabilistic formulation to enable us to incorporate them later into an efficient model parameter estimation algorithm.

2.1. A model for a single source at a microphone

Let t $(= 1, \ldots, \mathrm{N}_t)$ and f $(= 1, \ldots, \mathrm{N}_f)$ be time and frequency indices of a time-frequency bin, and $\mathrm{N}_s \geq 2$ be the number of point source signals. Then the l-th source recorded at the m $(= 1, \ldots, \mathrm{N}_m)$-th microphone at a time frame t and a frequency f, $x_{t,f}^{(l,m)}$, can be written in the log power spectrum domain as

$$\begin{aligned} x_{t,f}^{(l,m)} &= \log |S_{t,f}^{(l)} H_f^{(l,m)}|^2 + e_{t,f}^{(l,m)}, \\ &= \log |S_{t,f}^{(l)}|^2 + \log |H_f^{(l,m)}|^2 + e_{t,f}^{(l,m)}, \\ &= s_{t,f}^{(l)} + \beta_f^{(l,m)} + e_{t,f}^{(l,m)}, \end{aligned} \tag{1}$$

where $S_{t,f}^{(l)}$ and $s_{t,f}^{(l)}$, respectively, correspond to the clean speech signal of the l-th source in the short-term Fourier transform (STFT) domain and the log power spectrum domain, and $H_f^{(l,m)}$ and $\beta_f^{(l,m)}$ are the transfer function in the STFT domain and the log power spectrum domain, respectively. Hereafter, $\beta_f^{(l,m)}$ is referred to as the "microphone-location dependent source gain", and $x_{t,f}^{(l,m)}$ as the l-th source image at the m-th microphone. The error term $e_{t,f}^{(l,m)}$ is introduced to compensate for the mismatch between $x_{t,f}^{(l,m)}$ and $\log |S_{t,f}^{(l)} H_f^{(l,m)}|^2$ that can be caused, for instance, by the fluctuation of the transfer functions. Here we assume that $e_{t,f}^{(l,m)}$ is a white Gaussian noise with a zero mean and a variance $\sigma_f^{(l,m)}$ equal to the average square of $e_{t,f}^{(l,m)}$. Following these definitions, the relationship between $s_{t,f}^{(l)}$ and $x_{t,f}^{(l,m)}$ can be modeled as a Gaussian probability density function (pdf) as follows.

$$p(x_{t,f}^{(l,m)}; \theta^{(l)}) = \mathcal{N}(s_{t,f}^{(l)} + \beta_f^{(l,m)}, \sigma_f^{(l,m)}). \tag{2}$$

where $\theta^{(l)}$, a set of model parameters, is composed of $s_{t,f}^{(l)}$, $\beta_f^{(l,m)}$ and $\sigma_f^{(l,m)}$ for all t, f and m.

2.2. Observation model in multiple source scenario

To model the observation of multiple sources in the log power spectrum domain at the m-th microphone, we adopt LogMax approximation [19]. With this approximation, we assume that the observed log spectrum, $o_{t,f}^{(m)}$, can be represented as the maximum among all the source images at the microphone, thus equalling that of the *dominant* source as:

$$o_{t,f}^{(m)} = \max\{x_{t,f}^{(1,m)}, \ldots, x_{t,f}^{(\mathrm{N}_s,m)}\}. \tag{3}$$

In this model, non-dominant sources are allowed to take any values less than the observed log power spectrum. The above LogMax model can also be defined in a probabilistic form as [17]:

$$p(o_{t,f}^{(m)} | L_{t,f}^{(m)}, x_{t,f}^{(1,m)}, x_{t,f}^{(2,m)}) = \delta(o_{t,f}^{(m)} - x_{t,f}^{(L_{t,f}^{(m)},m)}), \tag{4}$$

$$p(L_{t,f}^{(m)} | x_{t,f}^{(1,m)}, \ldots, x_{t,f}^{(\mathrm{N}_s,m)}) = \begin{cases} 1 \text{ if } L_{t,f}^{(m)} = \arg\max_l x_{t,f}^{(l,m)}, \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

where $L_{t,f}^{(m)}$ indicates the index of the *dominant* source at a time-frequency bin at the m-th microphone. Hereafter, the index $L_{t,f}^{(m)}$ is referred to as the dominant source index (DSI) and its superscript and subscript are occasionally omitted for the simplicity sake. The first equation above ensures that $o_{t,f}^{(m)}$ is equal to the microphone image of the dominant source, $x_{t,f}^{(L,m)}$, as indexed by the DSI, and the second equation ensures that the DSI is equal to the index of the dominant source. Note that here we assume a different source activity pattern, i.e., DSI, at each microphone, which leads directly to the estimation of the microphone-dependent source activity as described in the next section.

Now, using the above probabilistic models, the joint pdf of $o_{t,f}^{(m)}$ and L can be derived as:

$$\begin{aligned} p(o_{t,f}^{(m)}, L; \theta) = & p(x_{t,f}^{(L,m)} = o_{t,f}^{(m)}; \theta^{(L)}) \\ & \times \prod_{l \neq L} \int_{-\infty}^{o_{t,f}^{(m)}} p(x_{t,f}^{(l,m)}; \theta^{(l)}) dx_{t,f}^{(l,m)}. \end{aligned} \tag{6}$$

where θ is a set of $\theta^{(l)}$ for all l. The derivation of this equation is detailed in [20].

3. Basic idea behind the proposed method

In this section, we explain the basic idea behind the proposed method by intuitively explaining the behavior of the important parameters in the above observation model, and show how the proposed algorithm can estimate the microphone-dependent source activity information while taking into account the contributions from the other microphones.

One of the key parameters of the proposed algorithm is the DSI, $L_{t,f}^{(m)}$, which can take different values for *each* microphone. Since the DSI indicates the activity of each source, and is microphone dependent, it implies that determining the DSI at each microphone allows us to estimate the microphone-dependent source activities.

The other key parameters of the proposed method are those that inherently support the DSI, namely the microphone-location dependent source gain $\beta_f^{(l,m)}$, which is time-invariant, and the *microphone independent time-varying* source log power spectrum $s_{t,f}^{(l)}$ defined in eq. (1). For instance, if a source signal is captured at the m-th microphone with a high SNR, the microphone-location dependent source gain tends to take a relatively large value and consequently the source signal would be observed as the dominant source at that microphone through the LogMax observation model. When a source signal turns out to be dominant at certain time-frequency bins, the parameter corresponding to the log power spectra of the source can be estimated. In contrast, if a source signal is captured at the m-th microphone with a lower level, the microphone-location dependent source gain tends to take a smaller value and consequently the source signal would be observed as a *non*-dominant source at the microphone. Needless to say, since the log power spectrum of a non-dominant source cannot be directly observed through the LogMax observation model, the update of the source log power spectrum will be omitted and left for cases where it can be updated reliably, i.e., high SNR cases at other microphones. In other words, the proposed algorithm estimates the log power spectra of a source, $s_{t,f}^{(l)}$, mainly using microphones that are closer to the source location, because a source

can be more frequently dominant at closer microphones. This mechanism works effectively for the estimation of source activity especially in DMA environments.

Now, it is important to note that although up to this point we have described a rather heuristic parameter update rule to promote a better understanding of our algorithm, the actual update will be carried out in a probabilistically optimal way, which means that the algorithm will automatically decide in a soft manner depending on each situation whether or not such parameters as the source log power spectrum can be updated. Indeed, as will be discussed in the following subsections, the model parameter estimation will be carried out using the EM algorithm where the DSIs are handled as hidden variables. In the optimization procedure, the posterior of the DSI is updated in the E-step to express more accurately which source dominates each time-frequency bin at each microphone. In the M-step, $\beta_f^{(l,m)}$ and $s_{t,f}^{(l)}$ of a source are updated based mainly on the time-frequency bins dominated by that source, which are indicated by the DSI posterior.

4. Parameter estimation

In this section, we introduce an efficient parameter estimation algorithm based on the aforementioned probabilistic models using a maximum a posteriori (MAP) criterion.

4.1. MAP parameter estimation using EM algorithm

4.1.1. Auxiliary function

Using the above models, we can estimate the model parameters, θ, assuming the DSIs, L, to be hidden variables. For a efficient MAP parameter estimation, we employ the EM algorithm and maximize the following auxiliary function.

$$\begin{aligned} Q(\theta|\hat{\theta}) &= E\{\log p(\{o_{t,f}^{(m)}\},\{L_{t,f}^{(m)}\};\theta)|\hat{\theta}\} \\ &= \sum_t \sum_m \sum_f \sum_l \hat{M}_{t,f}^{(l,m)} \log p(o_{t,f}^{(m)}, L_{t,f}^{(m)} = l;\theta), \\ &= \sum_m \sum_l Q^{(l,m)}(\theta^{(l)}|\hat{\theta}), \end{aligned}$$

$$\begin{aligned} Q^{(l,m)}(\theta^{(l)}|\hat{\theta}) = \sum_t \sum_f \Bigg[& \hat{M}_{t,f}^{(l,m)} p(x_{t,f}^{(l,m)} = o_{t,f}^{(m)};\theta^{(l)}) \\ & + (1-\hat{M}_{t,f}^{(l,m)}) \int_{-\infty}^{o_{t,f}^{(m)}} p(x_{t,f}^{(l,m)};\theta^{(l)}) dx_{t,f}^{(l,m)} \Bigg] \end{aligned} \quad (7)$$

where $\hat{M}_{t,f}^{(l,m)}$ indicates the posterior probability $p(L_{t,f}^{(m)} = l|o_{t,f}^{(m)},\hat{\theta})$ defined as:

$$\hat{M}_{t,f}^{(l,m)} = \frac{p(o_{t,f}^{(m)}, L_{t,f}^{(m)} = l;\hat{\theta})}{\sum_{l'} p(o_{t,f}^{(m)}, L_{t,f}^{(m)} = l';\hat{\theta})}, \quad (8)$$

The derivation of eq. (7) can be partially found in [20].

Note that eq. (7) cannot be analytically maximized in the M-step owing to the complexity of the second term. Instead, we can employ the Newton-Raphson method for its efficient maximization as discussed in [18].

4.1.2. Optimization procedure

Based on the EM algorithm and the Newton-Raphson method, the optimization procedure of the proposed method can be derived as follows.

1. Initialize $\hat{\theta}$
2. Iterate the following until convergence for each f

   (E-step) Update $\hat{M}_{t,f}^{(l,m)}$ for all l, m and t as in eq. (8)

   (M-step) Update $\hat{s}_{t,f}^{(l)}$, $\hat{\beta}_f^{(l,m)}$ and $\hat{\sigma}_f^{(l,m)}$ as follows:

$$\hat{s}_{t,f}^{(l)} \leftarrow \hat{s}_{t,f}^{(l)} - \left(\frac{\partial^2 Q^{(l,m)}(\theta^{(l)}|\hat{\theta})}{\partial {s_{t,f}^{(l)}}^2} \right)^{-1} \left(\frac{\partial Q^{(l,m)}(\theta^{(l)}|\hat{\theta})}{\partial s_{t,f}^{(l)}} \right),$$

$$\hat{\beta}_f^{(l,m)} \leftarrow \hat{\beta}_f^{(l,m)} - \left(\frac{\partial^2 Q^{(l,m)}(\theta^{(l)}|\hat{\theta})}{\partial {\beta_f^{(l,m)}}^2} \right)^{-1} \left(\frac{\partial Q^{(l,m)}(\theta^{(l)}|\hat{\theta})}{\partial \beta_f^{(l,m)}} \right),$$

$$\hat{\sigma}_f^{(l,m)} \leftarrow \frac{\sum_t \hat{\kappa}_{t,f}^{(l,m)} (o_{t,f}^{(l,m)} - (\hat{s}_{t,f}^{(l)} + \hat{\beta}_f^{(l,m)}))^2}{\sum_t \hat{M}_{t,f}^{(l,m)}}.$$

3. Estimate the l-th source image at the m-th microphone as in the next section

Here, $\hat{\kappa}_{t,f}^{(l,m)} = \hat{M}_{t,f}^{(l,m)}$ if $o_{t,f}^{(l,m)} > (\hat{s}_{t,f}^{(l)} + \hat{\beta}_f^{(l,m)})$, and $\hat{\kappa}_{t,f}^{(l,m)} = 1$ otherwise. Note that the update equations for $s_{t,f}^{(l)}$ and $\beta_f^{(l,m)}$ look very similar. Their main difference lies in the averaging operation. That is, $s_{t,f}^{(l)}$ can be obtained by averaging the statistics over the microphone index, while $\beta_f^{(l,m)}$ is obtained by averaging the statistics over the time index. The update equation for $\hat{\sigma}_f^{(l,m)}$ follows the equation reported in [17].

4.1.3. Incorporation of a regularization term

Because some sources can never be dominant at certain microphones, the optimal value for $\beta_f^{(l,m)}$ can be minus infinity, thus making the iterative estimation unstable. To prevent such a problem, we define a regularization term (i.e., prior pdf) for the speech log power spectrum $s_{t,f}^{(l)}$ as follows, and add it to the auxiliary function described in previous subsection with a certain weight ρ.

$$\log p(\{s_{t,f}^{(l)}\}) = \sum_f \log \mathcal{N}(s_{t,f}^{(l)}; \bar{\mu}_f^{(l)}, \bar{\sigma}_f^{(l)}). \quad (9)$$

4.2. MMSE estimation of a source image

After the parameter estimation with the EM algorithm, we can estimate the l-th source image at the m-th microphone, $\hat{x}_{t,f}^{(l,m)}$, based on MMSE estimation, similarly to the procedure described in [17]. The estimated log power spectrum can be given as:

$$\begin{aligned} \hat{x}_{t,f}^{(l,m)} = & M_{t,f}^{(l,m)} o_{t,f}^{(m)} \\ & + (1 - M_{t,f}^{(l,m)}) \frac{\int_{-\infty}^{o_{t,f}^{(m)}} \tilde{x}_{t,f}^{(l,m)} p(x_{t,f}^{(l,m)};\hat{\theta}^{(l)}) dx_{t,f}^{(l,m)}}{\int_{-\infty}^{o_{t,f}^{(m)}} p(x_{t,f}^{(l,m)};\hat{\theta}^{(l)}) dx_{t,f}^{(l,m)}}. \end{aligned} \quad (10)$$

where $\tilde{x}_{t,f}^{(l,m)} = s_{t,f}^{(l)} + \beta_f^{(l,m)}$.

5. Experiment

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

5.1. Acoustic conditions: 4 DMA scenarios

To evaluate the proposed method, we simulated the four different DMA environments depicted in Fig. 1 by using the image method [21]. The size of the simulated room was 10 m

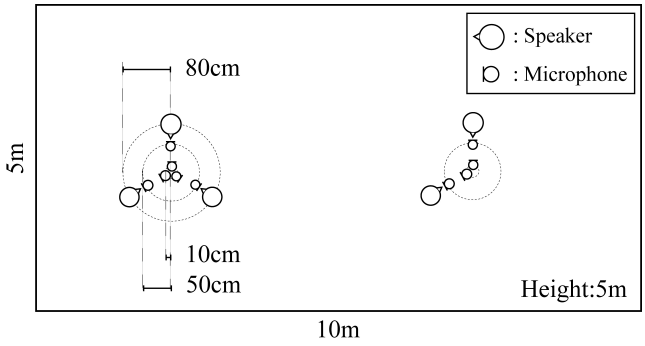

Figure 1: Experimental setup: positions of speakers and microphones for 4 DMA scenarios

(W)×5 m (D)×5 m (H), and the reverberation time was around 100 ms. The first and second scenarios simulate the simultaneous conversation of 3 speakers sitting in a circle with a radius of 80 cm (Please see only the left half of the depicted room in Fig. 1 and assume there is nothing in the right half). While for the first scenario, the separation is accomplished using 3 microphones placed concentrically in a circle with a radius of 10 cm, for the second scenario, these 3 microphones are placed in a circle with a radius of 50 cm. These scenarios simulate a situation where 3 people are having a conversation at a coffee table with their own single-channel recording devices placed on the table. The third and fourth scenarios simulate the conversation of two individual groups involving 3 and 2 people respectively, as depicted in Fig. 1. In the third scenario, the separation is accomplished using 5 microphones placed in a concentric fashion in a circle with a radius of 10 cm, and in the fourth scenario, these 5 microphones are placed in a circle with a radius of 50 cm. Note that, with these setups, the acoustic diversity embedded within the microphone observations tends to increase from scenario 1 to 4, and thereby we can expect to see an increasingly clear difference between the conventional and proposed methods.

5.2. Tasks and other conditions

Our objective is to separate 3 simultaneous speakers (first and second scenarios) and 5 speakers (third and fourth scenarios). For comparison with the proposed method, we tested the state-of-the-art conventional method [16], which estimates a *global* source activity pattern common for all the microphones and uses them as a soft mask for separation. For the conventional method, the separated signals are generated by applying the soft masks to the signals observed at the closest microphone to each speaker. To initialize the EM iteration for the proposed method, we employed the results of this conventional method. To calculate the regularization term defined in eq. (9), we also used the output of the conventional method. This regularization term was calculated for each test utterance and added to the auxiliary function with a weight (ρ) of 0.00001.

The results are evaluated based on the cepstral distance that measures the difference between the target signal (e.g., the processed speech) and the image of each source at the closest microphone to each speaker. Twenty random combinations of speech utterances of different speakers from the TIMIT database [22] are used for each scenario, and the results shown below are obtained by averaging over all combinations.

5.3. Results

Figure 2 shows the cepstral distances that we obtained with the observed signal denoted as "Obs.", the conventional method denoted as "Conv." and the proposed method denoted as "Prop." Note that here the observed signal indicates the signal observed

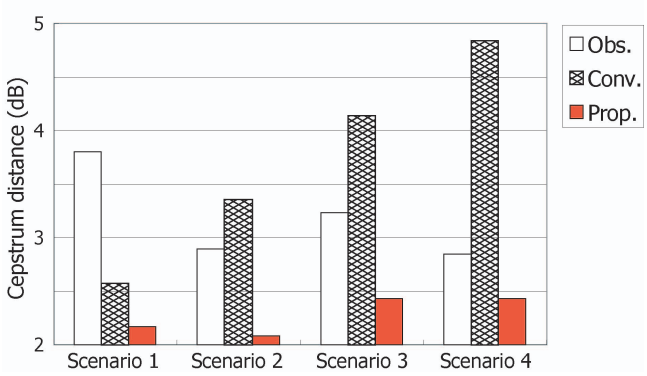

Figure 2: Results: cepstral distances (dB) of observed signal (oracle best microphone selection), conventional method and proposed method

at the closest microphone to each speaker, thus it can be considered as a result obtained with *oracle* best microphone selection.

In scenario 1, we can first confirm the efficacy of the conventional method because we can observe a substantial reduction in the cepstral distance from the best microphone selection, i.e., "Obs." We can also observe some further improvement with the proposed method even in this co-located microphone scenario. This improvement can be explained by the fact that the proposed method optimizes all the parameters in the log power spectrum domain, which is essentially the same domain as that of the employed evaluation metric.

In contrast to the 1st scenario, we can no longer see the improvement with the conventional method in terms of this evaluation metric, although we can actually confirm the separation effect in terms of audible quality. It is likely that the conventional method tends to perform separation rather aggressively in these cases, and causes substantial distortions. We can also observe that as expected the performance of the conventional method degrades as the distance between microphones increases from scenario 1 to 4. In contrast to the conventional method, the proposed method successfully reduces the cepstral distances in all the DMA scenarios by comparison with the best microphone selection, i.e., "Obs."

6. Summary

In this paper, we proposed an approach that optimally estimates microphone-dependent source activities and utilizes them to achieve BSS in DMA environments. Using the EM algorithm, the proposed method efficiently estimates the activity of each source signal at each microphone, which can be fully characterized by the microphone-independent speech log power spectra and microphone-location dependent source gains. We experimentally demonstrated that the proposed method can achieve better performance in various DMA scenarios than the conventional method, which assumes *global* source activity, and the best microphone selection. Future work will include an evaluation of the proposed method in the presence of an inevitable mismatch in sampling frequencies between microphones/recording devices, i.e., drift error. We can expect the proposed method to work robustly in such environments, since it does not employ the phase information between microphones, which is known to be sensitive to drift error.

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] 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.

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

[9] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 49, pp. 1614–1626, August 2001.

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

[11] 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.

[12] 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.

[13] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 18, pp. 260–276, 2010.

[14] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for joint blind source separation and noise reduction," in *Proc. IEEE Int'l Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 109–112.

[15] 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.

[16] 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.

[17] T. Nakatani, S. Araki, T. Yoshioka, and M. Fujimoto, "Joint unsupervised learning of hidden markov source models and source location models for multichannel source separation," in *Proc. of ICASSP*, 2011, pp. 237 – 240.

[18] T. Nakatani, T. Yoshioka, S. Araki, M. Delcroix, and M. Fujimoto, "LogMax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise," in *Proc. of ICASSP*, 2012, pp. 4029–4032.

[19] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, 2003, pp. 1009–1012.

[20] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech and Language*, vol. 27(3), pp. 851–873, 2013.

[21] 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.

[22] W. 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.