AN INTEGRATION OF SOURCE LOCATION CUES FOR SPEECH CLUSTERING IN DISTRIBUTED MICROPHONE ARRAYS

Mehrez Souden, Keisuke Kinoshita, and Tomohiro Nakatani

NTT Communication Science Laboratories, Kyoto, Japan

ABSTRACT

We propose a new approach for clustering competing speech sources using distributed microphone arrays. In this approach, we first define two feature vectors where the first captures the intra-node location information while the second captures the level difference of speech energy recorded at different nodes. Then, we introduce Watson and Dirichlet mixture models to model the first and second features, respectively. We integrate both types of information in an expectation maximization algorithm to cluster the simultaneous speech sources. The performance of the proposed approach is superior to best node selection and comparable to centralized processing in terms of conventional blind source separation metrics.

***Index Terms*—** Distributed microphone array, source clustering, blind source separation, expectation maximization.

1. INTRODUCTION

Distributed microphone array (DMA) processing is emerging as an alternative to the conventional treatment of microphone arrays with co-located elements, and has the potential to solve challenging tasks thanks to the extended spatial coverage and scalability of DMAs. A good illustration of this fact is represented by meeting applications where participants have personal communication devices (PCDs). A DMA is formed when these PCDs, considered to be nodes of the DMA, record audio signals, and collaborate to process them. A summary of other potential applications can be found in [1].

Central to the design of DMA-based algorithms are proper definitions of the roles of nodes, the information to be processed locally, and the data shared in the network to achieve a global processing goal. In the particular context of blind source separation (BSS) of speech using DMAs, earlier contributions include a distributed independent component analysis (ICA)-based algorithm proposed in [2]. Therein it is shown that successful BSS can be achieved when different nodes apply different ICA adaptations and exchange speech sparseness-based [3, 4] regularization factors. In [5], an approach where every node detects its local neighboring sources and separates them using ICA is proposed. To apply ICA in both methods, every node must have at least as many microphones as detected sources. This constraint may not be practical due to hardware constraints, for instance. In contrast, clustering-based BSS does not require such a strong assumption [4, 6]. In [7], we proposed a DMA-based approach for speech clustering using the local normalized recordings of the nodes. The problem was cast into a distributed expectation maximization (EM) procedure. Since different nodes collect different spatial characteristics of the sources, it was concluded that it is more natural to fuse the estimated posterior probabilities than the model parameters, as in other distributed clustering techniques [8,9].

In this paper, we extend our approach for speech source clustering and separation in DMAs that we first proposed in [7]. Indeed, using the local normalized recordings as feature vectors for clustering does not allow the algorithm to capture the internode interactions. Particular nodes may be located near some sources, and this information can be valuable for clustering. For example, in the scenarios with multiple PCDs mentioned above, it is natural to assume that every device is located nearer to its user than to other participants. Our focus in this paper is then to exploit this information jointly with normalized local recordings. Specifically, we introduce a new feature vector, which captures the relative internode speech attenuation. To model this information, we propose to use the Dirichlet mixture model (DMM) [10, 11] that we combine with the Watson mixture model (WMM) [6, 12–14] for the local normalized recordings in an EM algorithm to obtain improved clustering performance.

2. DATA MODEL

We are interested in scenarios where $L > 1$ competing speech signals are recorded by a DMA of N nodes. At time frame t and frequency $k = 1, ..., K$, where K is the number of frequency bins, the nth node recordings are expressed in the short time Fourier transform domain as

$$\mathbf{y}_n(k,t) \approx \sum_{l=1}^{L} \mathbf{x}_{n,l}(k,t) + \mathbf{v}_n(k,t). \tag{1}$$

$\mathbf{y}_n(k,t) = [Y_{n,1}(k,t) \ \cdots \ Y_{n,M_n}(k,t)]^T$ contains the M_n multichannel recordings, $\mathbf{x}_{n,l}(k,t) = \mathbf{h}_{n,l}(k)S_l(k,t)$ contains the reverberant microphone observations of the lth speech signal, $S_l(k,t)$, and M_n denotes the number of microphones at the nth node. $\mathbf{h}_{n,l}(k) = [H_{n,1l}(k) \ \cdots \ H_{n,M_n l}(k)]^T$ contains the channel transfer functions between the lth source and the microphone elements of the nth node, and $\mathbf{v}_n(k,t) = [V_{n,1}(k,t) \ \cdots \ V_{n,M_n}(k,t)]^T$ represents the additive acoustic noise. We further define the global observation vector $\mathbf{y}(k,t) = \left[\mathbf{y}_1^T(k,t) \ \cdots \mathbf{y}_N^T(k,t)\right]^T$. We do not mention the explicit dependence on frequency, k, in our notations next since all our processing is performed frequency-bin-wise.

3. CONVENTIONAL CENTRALIZED PROCESSING

This section revisits the conventional centralized clustering approach, which uses the following feature vector [6]

$$\boldsymbol{\psi}(t) \triangleq \frac{\mathbf{y}(t)}{\|\mathbf{y}(t)\|}. \tag{2}$$

The normalization of the recording vector reduces the effect of speech energy fluctuations and maps the recordings on the complex unit hypersphere. It turns out that this feature vector can be accurately modeled using a multimodal distribution thanks to the property of sparseness of speech [3, 4]. Each of the modes of such a distribution is concentrated around the normalized propagation vector

(centroid) of one of the L competing sources [6, 14]. Subsequently, separating the speech sources amounts to defining a latent variable, $\mathcal{H}$, that identifies the most likely mode at a given time-frequency slot. Since we have L sparse signals, $\mathcal{H}$ can take L discrete values denoted as $\mathcal{H}_1, ..., \mathcal{H}_L$. When $\mathcal{H} = \mathcal{H}_l, l = 1, ..., L$, the lth speaker dominates the mixture. In other words, the clustering and separation of recorded sounds becomes possible by determining the L posterior probabilities $p(\mathcal{H}_l|\boldsymbol{\psi}(t)), l = 1, ..., L$.

In conventional location-based clustering [6, 14], $\boldsymbol{\psi}(t)$ is modeled as

$$p(\boldsymbol{\psi}(t); \boldsymbol{\theta}) = \sum_{l=1}^{L} w_l p(\boldsymbol{\psi}(t)|\mathcal{H}_l). \tag{3}$$

$\boldsymbol{\theta}$ contains all model parameters, $\sum_l w_l = 1$, $0 \leq w_l \leq 1$, and $w_l = P(\mathcal{H}_l)$. In [6], a Gaussian-like distribution was used for $p(\boldsymbol{\psi}(t)|\mathcal{H}_l)$ following the idea of line orientation [15, 16]. This is rather an approximation of the Watson distribution since $\|\boldsymbol{\psi}(t)\| = 1$, and we have [13, 14, 17]

$$p(\boldsymbol{\psi}(t)|\mathcal{H}_l; \mathbf{a}_l, \kappa_l) = \frac{\Gamma(M)}{2\pi^M \mathcal{M}(1, M, \kappa_l)} \exp\left(\kappa_l \left|\mathbf{a}_l^H \boldsymbol{\psi}(t)\right|^2\right) \tag{4}$$

that we equivalently denote as $\boldsymbol{\psi}(t)|\mathcal{H}_l \sim \mathcal{W}_M(\mathbf{a}_l, \kappa_l)$. κ_l and $\mathbf{a}_l$ represent the concentration parameter and centroid of the distribution [17], $M = \sum_{n=1}^{N} M_n$, $\Gamma(\cdot)$ is the Gamma function, and $\mathcal{M}(\cdot, \cdot, \cdot)$ is Kummer's confluent hypergeometric function [18]. The EM algorithm can then be used to determine $\boldsymbol{\theta} \triangleq \left[w_1 \ \cdots \ w_L \ \mathbf{a}_1^T \ \cdots \ \mathbf{a}_L^T \ \kappa_1 \ \cdots \ \kappa_L\right]^T$, and the posterior probabilities of the L clusters are obtained as a byproduct. Detailed expressions of the model parameters can be found in [12–14].

It should be noted that in the definition of $\boldsymbol{\psi}(t)$, all signal observations are jointly processed by a central unit. In a DMA, however, nodes constituting the network can have a certain autonomy and process their local recordings. Furthermore, they can have some transmit/receive capability to communicate and collaborate with each other following some protocol. It is then important to appropriately define the feature vectors that can be used by these nodes, their associated statistical models, and the data exchanged in the network to achieve collaboration. For simplicity, we assume that all nodes communicate with each other, although it is possible to use the same principle with some other topologies, e.g., [1].

4. DISTRIBUTED LOCATION-BASED SPEECH CLUSTERING

This section discusses two types of feature vectors that we design to capture the *intra-* and *inter*-node location information. We define these feature vectors and provide their statistical models. We then integrate them into an EM algorithm to find speech clusters.

4.1. Intra-Node Feature Vector

We assume that each of the nodes has at least two microphones, thereby allowing it to capture the location information about the participating speakers by using the local normalized recordings vector. In other words, the nth node computes its local location feature $\boldsymbol{\psi}_n(t) \triangleq \frac{\mathbf{y}_n(t)}{\|\mathbf{y}_n(t)\|}$, which is distributed on the complex unit hypersphere similar to $\boldsymbol{\psi}(t)$, and it is reasonable to assume that it has a multimodal distribution due to the various contributions of the speech sources, with $\boldsymbol{\psi}_n(t)|\mathcal{H}_l \sim \mathcal{W}_{M_n}(\mathbf{a}_{n,l}, \kappa_{n,l})$. Hence, the set of unknown parameters of our model is $\tilde{\boldsymbol{\theta}} \triangleq \left[w_1 \ \cdots \ w_L \ \mathbf{a}_{1,1}^T \ \cdots \ \mathbf{a}_{N,L}^T \ \kappa_{1,1} \ \cdots \ \kappa_{N,L}\right]^T$. Every node can cluster its local recordings regardless of other nodes. Nevertheless, it is beneficial to consider exchanging some information between nodes to reach a global consensus on the activities of the speech sources (equivalently, the posterior probabilities of the L clusters) from the recordings. The exchanged information can be raw recordings, or some processing results, e.g., a node decision on the local observed data, which we consider here. In our case, we are interested in sharing the posterior probabilities between the nodes since they can be transmitted at a lower rate than raw data [7]. To achieve distributed processing, we assume the following

$$p\left(\tilde{\boldsymbol{\psi}}(t)|\mathcal{H}_l\right) = \prod_{n=1}^{N} p(\boldsymbol{\psi}_n(t)|\mathcal{H}_l), \tag{5}$$

where $\tilde{\boldsymbol{\psi}}(t) \triangleq \left[{\boldsymbol{\psi}_1}^T(t) \ \cdots \ {\boldsymbol{\psi}_N}^T(t)\right]^T$. The above conditional independence assumption has been commonly used to obtain a "workable approximation of the reality which may be more complex" [19] in distributed sensor networks. For instance, a similar assumption was made in the context of unsupervised learning of Gaussian mixture models [8, 9]. It is worthwhile noting that besides the fact that we are interested in speech clustering and separation using the location information in DMAs, which has not yet been investigated, the proposed model has independent parameters per node (centroids and concentration parameters), and only *a posteriori* and *a priori* probabilities of speech presence, which describe the speakers' activities, are common to all nodes. Conversely, in most existing literature on the use of distributed EM to estimate mixture model parameters, the latter are common to all nodes [8, 9].

By virtue of assumption (5) and Bayes' rule, we can express the lth global posterior probability as [7, 19]

$$\begin{aligned} \tilde{\wp}\left(t, l, \tilde{\boldsymbol{\theta}}\right) &\triangleq p\left(\mathcal{H}_l|\tilde{\boldsymbol{\psi}}(t); \tilde{\boldsymbol{\theta}}\right) \\ &= \zeta\left(t, l, \tilde{\boldsymbol{\theta}}\right) \cdot \chi\left(t, \tilde{\boldsymbol{\theta}}\right). \end{aligned} \tag{6}$$

where

$$\zeta\left(t, l, \tilde{\boldsymbol{\theta}}\right) \triangleq w_l^{1-N} \prod_{n=1}^{N} p\left(\mathcal{H}_l|\boldsymbol{\psi}_n(t); \tilde{\boldsymbol{\theta}}\right), \tag{7}$$

which we define as a *consensus decision* [7], and $\chi\left(t, \tilde{\boldsymbol{\theta}}\right)$ acts as a normalization term that can be ignored [7, 19]. Hence, only posterior probabilities, $p\left(\mathcal{H}_l|\boldsymbol{\psi}_n(t); \tilde{\boldsymbol{\theta}}\right)$, $l = 1, ..., L$ and $n = 1, ...N$, need to be shared between nodes (the global posterior probabilities are obtained by fusing the local estimates and normalizing the *consensus decision*). In [7], we argued that, in practice, it may be more beneficial to approximate the product-rule based consensus decision using the following sum rule [7, 19], which is used in this paper

$$\zeta\left(t, l, \tilde{\boldsymbol{\theta}}\right) \approx (1 - N) w_l + \sum_{n=1}^{N} p\left(\mathcal{H}_l|\boldsymbol{\psi}_n(t); \tilde{\boldsymbol{\theta}}\right). \tag{8}$$

4.2. Inter-Node Feature Vector

It is known that the energy of the recorded speech varies significantly with respect to the source-sensor distance (proportional to inverse square distance in a free field). This property can be exploited in DMAs where some nodes may be spatially closer to a set of sources than others as in Figure 1, for instance. Unfortunately, the definition of $\boldsymbol{\psi}_n(t)$ within every node does not allow us to capture the energy attenuation effect due to the node-source distance. It is therefore necessary to find another feature that captures this effect, and

include it in our source clustering algorithm. To this end, we define the feature vector

$$\boldsymbol{\rho}(t) \triangleq [\rho_1(t) \ \cdots \ \rho_N(t)]^T, \tag{9}$$

where for $n = 1, ..., N$ $\rho_n(t) \triangleq \frac{\|\mathbf{y}_n(t)\|^2}{\|\mathbf{y}(t)\|^2}$, which clearly captures the amplitude attenuation between nodes. Note that to form this feature vector, node n has to share only its synthesized signal, $\|\mathbf{y}_n\|^2$, with other nodes, then the denominator is formed by summing all terms. This is in a way similar to [20, 21] and some references therein, where the idea is to share only some synthesized signals (output of the noise reduction filter) with other nodes in an iterative way to obtain a near-optimal centralized solution.

To find the probabilistic model, which better fits the distribution of $\boldsymbol{\rho}(t)$, we first note that by assuming that all nodes detect all sources $\sum_{n=1}^{N-1} \rho_n(t) < 1$, $0 < \rho_n(t) < 1$, and $\rho_N(t) = 1 - \sum_{n=1}^{N-1} \rho_n(t)$. Furthermore, $\boldsymbol{\rho}(t)$ results from the contributions of all sources. Hence, it is natural to use the multimodal DMM [11]

$$p(\boldsymbol{\rho}(t); \boldsymbol{\alpha}) = \sum_{l=1}^{L} w_l p(\boldsymbol{\rho}(t)|\mathcal{H}_l; \boldsymbol{\alpha}_l) \tag{10}$$

where [10, 11]

$$p(\boldsymbol{\rho}(t)|\mathcal{H}_l; \boldsymbol{\alpha}_l) = \frac{\Gamma\left(\sum_{n=1}^{N} \alpha_{n,l}\right)}{\prod_{n=1}^{N} \Gamma(\alpha_{n,l})} \prod_{n=1}^{N} \rho_n^{\alpha_{n,l}-1}(t), \tag{11}$$

$\boldsymbol{\alpha} \triangleq \left[\boldsymbol{\alpha}_1^T \cdots \boldsymbol{\alpha}_L^T\right]^T$, and $\boldsymbol{\alpha}_l \triangleq [\alpha_{1,l} \cdots \alpha_{N,L}]^T$.

4.3. Integration and Parameter Estimation Using EM

Note that $\boldsymbol{\rho}(t)$ and $\boldsymbol{\psi}_n(t), n = 1, ..., N$, capture two complementary types of information in the DMA, and it is reasonable to assume that they are independent. Consequently, we can combine both cues to compute $\wp\left(t, l, \tilde{\boldsymbol{\theta}}\right) \triangleq p(\mathcal{H}_l|\boldsymbol{\rho}(t), \tilde{\boldsymbol{\psi}}(t); \tilde{\boldsymbol{\theta}})$ as

$$\wp\left(t, l, \check{\boldsymbol{\theta}}\right) = \frac{\tilde{\wp}\left(t, l, \tilde{\boldsymbol{\theta}}\right) p(\boldsymbol{\rho}(t)|\mathcal{H}_l; \boldsymbol{\alpha}_l)}{\sum_{l=1}^{L} \tilde{\wp}\left(t, l, \tilde{\boldsymbol{\theta}}\right) p(\boldsymbol{\rho}(t)|\mathcal{H}_l; \boldsymbol{\alpha}_l)}. \tag{12}$$

Using the law of total probability, we have

$$p\left(\tilde{\boldsymbol{\psi}}(t), \boldsymbol{\rho}(t); \check{\boldsymbol{\theta}}\right) = \sum_{l=1}^{L} w_l p\left(\tilde{\boldsymbol{\psi}}(t), \boldsymbol{\rho}(t)|\mathcal{H}_l; \check{\boldsymbol{\theta}}\right), \tag{13}$$

where $\tilde{\boldsymbol{\theta}}$ contains all model parameters, which can be determined by maximizing (13), or equivalently its associated auxiliary function

$$\begin{aligned} Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}') &= \sum_{t=1}^{T} \sum_{l=1}^{L} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right) \ln\left(p\left(\tilde{\boldsymbol{\psi}}(t), \boldsymbol{\rho}(t), \mathcal{H}_l; \check{\boldsymbol{\theta}}\right)\right) \\ &= Q_1(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}') + Q_2(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}') + Q_3(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}') \end{aligned} \tag{14}$$

where[1] $\check{\boldsymbol{\theta}}'$ is some prior estimate of $\check{\boldsymbol{\theta}}$

$$Q_1(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}') = \sum_{t,l} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right) \ln(w_l), \tag{15}$$

$$Q_2(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}') = \sum_{t,l,n} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right) \ln(p(\boldsymbol{\psi}_n(t)|\mathcal{H}_l; \mathbf{a}_{n,l}, \kappa_{n,l})), \tag{16}$$

[1] $\sum_{t=1}^{T} \sum_{l=1}^{L} \sum_{n=1}^{N}$ was denoted as $\sum_{t,l,n}$ due to space constraints.

and

$$Q_3(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}') = \sum_{t,l} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right) \ln(p(\boldsymbol{\rho}(t)|\mathcal{H}_l; \boldsymbol{\alpha}_l)). \tag{17}$$

Now, we can determine all the model parameters. Indeed, by taking account of the constraint that $\sum_{l=1}^{L} w_l = 1$ and maximizing $Q_1(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')$ with respect to (w.r.t.) w_l, we find that

$$w_l = \sum_{t=1}^{T} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right)/T. \tag{18}$$

Similarly, by setting the derivative of $Q_2(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')$ w.r.t. $\kappa_{n,l}$ and $\mathbf{a}_{n,l}$ to 0, we can demonstrate that $\mathbf{a}_{n,l}$ is given by the eigenvector corresponding to the maximum eigenvalue, $r_{n,l}$, of the matrix

$$\mathbf{R}_{n,l} = \frac{\sum_{t=1}^{T} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right) \boldsymbol{\psi}_n(t) \boldsymbol{\psi}_n^H(t)}{\sum_{t=1}^{T} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right)} \tag{19}$$

and $\kappa_{n,l}$ satisfies

$$\frac{\frac{\partial \mathcal{M}\left(1, M_n, \kappa_{n,l}\right)}{\partial \kappa_{n,l}}}{\mathcal{M}(1, M_n, \kappa_{n,l})} = r_{n,l}, \tag{20}$$

which has no closed-form solution for $\kappa_{n,l}$. This particular problem was intensively studied in [12, 13], and approximations and bounds were provided. Here, we consider the simple approximation [13]

$$\kappa_{n,l} \approx \frac{M_n r_{n,l} - 1}{r_{n,l}(1 - r_{n,l})}(1 + r_{n,l}). \tag{21}$$

Finally, we note that determining the DMM model parameters by maximizing $Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')$ does not lead to a closed form solution [10,11]. Nonetheless, a simple and fast Newton-Raphson algorithm combined with the method of moments can lead to very accurate estimates of these parameters [10]. To guarantee the positiveness of the estimated $\alpha_{n,l}$, we follow the approach described in [11], by re-parameterizing $\alpha_{n,l} = \exp(\beta_{n,l})$ for some real $\beta_{n,l}$. Let $\boldsymbol{\beta}_l = [\beta_{1,l} \cdots \beta_{N,l}]^T$, then by performing the following iteration a few times (a maximum of five times in our experiments)

$$\boldsymbol{\beta}_l^{(j)} = \boldsymbol{\beta}_l^{(j-1)} - \nabla^{-1}\left(\boldsymbol{\beta}_l^{(j-1)}\right) \Delta\left(\boldsymbol{\beta}_l^{(j-1)}\right) \tag{22}$$

for iteration j, we can accurately estimate $\boldsymbol{\alpha}_l = \exp(\boldsymbol{\beta}_l)$. In (22), $\Delta(\boldsymbol{\beta}_l)$ is the gradient of $Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')$ w.r.t. $\boldsymbol{\beta}$, and is analytically expressed in Appendix. Furthermore, $\nabla(\boldsymbol{\beta}_l)$ is the Hessian of $Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')$, which can be easily inverted as explained in Appendix.

To sum up, the E-step of our algorithm implements (6), (8), (11), and (12), while its M-step implements (18), (19), (21), and (22). Clustering is then achieved using the final posterior probabilities.

5. EXPERIMENTAL RESULTS

Our experiments were conducted using the setup shown in Figure 1, where we have three speakers in a reverberant room, and assume that there are three nodes to record and process them. Every node has a pair of microphones with a 0.2 m spacing. The image method [22] was used to simulate the propagation environment with a reverberation time of 240 ms. Twenty random combinations of speech utterances of different speakers from the TIMIT database [23] were used in our experiments, and the BSS results shown below were obtained by averaging over all combinations. The speech signals were convoluted with the channel impulse responses and computer generated white Gaussian noise was added to the signals such that the

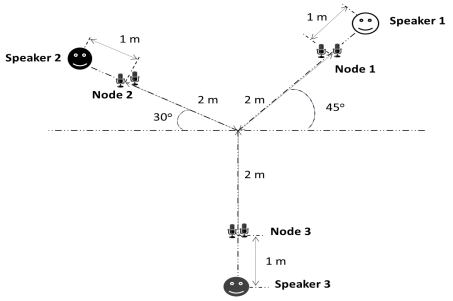

Fig. 1. Investigated propagation scenario.

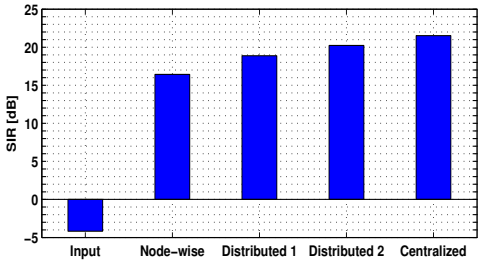

Fig. 2. Performance in terms of SIR.

input SNR was 30 dB. We compare the proposed source clustering using both feature vectors (denoted as *Distributed 2*) with our initial approach using only intra-node information (denoted as *Distributed 1*) [7], local clustering with Oracle best node selection (denoted as *Node-wise*), and the centralized source clustering (denoted as *Centralized*) described in Section 3. We evaluated the performance of all the investigated methods using the method proposed in [24] with the BSS metrics being the output signal-to-interference ratio (SIR), the signal-to-artifacts ratio (SAR), and the signal-to-distortion ratio (SDR).

In Figure 2, we confirm that our initial posterior probability fusion reported in [7] allows us to achieve larger output SIR than the best-node selection approach thanks to information sharing between nodes. The inclusion of the new feature vector leads to a slight increase in the SIR and brings the performance closer to that of the centralized processing. In Figure 3, we see that *Distributed 1* leads to increased signal distortion, which seems to come at the price of the increased SIR observed in Figure 2. However, it is also clear that *Distributed 2* allows us to better preserve the desired signals thanks to the integration of the inter-node location information. The overall separation results are summarized in Figure 4 in terms of SDR, where we see that the integration of both feature vectors outperforms both best node selection and *Distributed 1*, and approaches the centralized solution performance.

6. CONCLUSION

In this paper, we proposed a source clustering approach for DMAs, which integrated two distinct location features. The first is the normalized local recording vector, which captures the acoustic channel diversity within each node, while the second models the source-node distance effect on speech energy, and, hence, captures the inter-node effect. We modeled the two feature vectors using Watson and Dirichlet mixture models respectively. Furthermore, we proposed an approach for integrating the contributions of all nodes in the estimation of the posterior probabilities of the speech clusters. Our

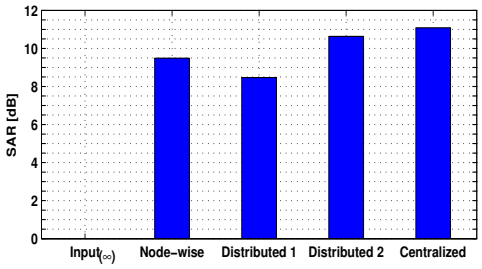

Fig. 3. Performance in terms of SAR.

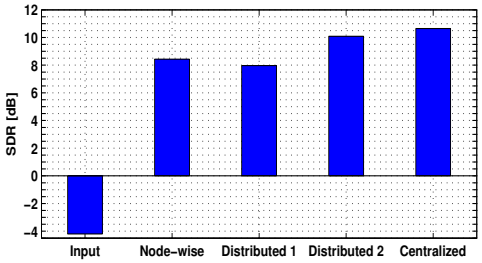

Fig. 4. Performance in terms of SDR.

evaluations demonstrated that including the recorded energy level difference between the nodes improves the DMA-based BSS performance.

Appendix

For clarity, we define $\gamma_{n,l} = \sum_{t=1}^{T} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right) \ln\left(\rho_n(t)\right)$ and $\tau_l = \sum_{t=1}^{T} \wp\left(t, l, \check{\boldsymbol{\theta}}'\right)$. Then, we find that the nth entry of $\Delta\left(\boldsymbol{\beta}_l\right)$ is

$$\frac{\partial Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')}{\partial \beta_{n,l}} = \alpha_{n,l}\gamma_{n,l} + \alpha_{n,l}\left(\Psi\left(\sum_{n=1}^{N}\alpha_{n,l}\right) - \Psi\left(\alpha_{n,l}\right)\right)\tau_l. \tag{23}$$

$\Psi(\cdot)$ is the digamma function [18]. The diagonal elements of $\nabla(\boldsymbol{\beta}_l)$ are expressed as

$$\frac{\partial^2 Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')}{\partial \beta_{n,l}^2} = \frac{\partial Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')}{\partial \beta_{n,l}} + \alpha_{n,l}^2\left(\Psi'\left(\sum_{n=1}^{N}\alpha_{n,l}\right) - \Psi'\left(\alpha_{n,l}\right)\right)\tau_l, \tag{24}$$

while its off-diagonal terms are given by

$$\frac{\partial^2 Q(\check{\boldsymbol{\theta}}, \check{\boldsymbol{\theta}}')}{\partial \beta_{n,i}\partial \beta_{n,l}} = \alpha_{n,l}\alpha_{n,i}\Psi'\left(\sum_{n=1}^{N}\alpha_{n,l}\right)\tau_l. \tag{25}$$

$\Psi'(\cdot)$ is the trigamma function [18]. Then, it is easy to see that $\nabla(\boldsymbol{\beta}_l)$ can be compactly expressed as

$$\begin{aligned}\nabla(\boldsymbol{\beta}_l) \quad &= \quad \operatorname{diag}\left[\Delta\left(\boldsymbol{\beta}_l\right) - \tau_l\boldsymbol{\alpha}_l \odot \boldsymbol{\alpha}_l \odot \Psi'\left(\boldsymbol{\alpha}_l\right)\right] \\ &\quad + \tau_l\Psi'\left(\sum_{n=1}^{N}\alpha_{n,l}\right)\boldsymbol{\alpha}_l\boldsymbol{\alpha}_l^T \end{aligned} \tag{26}$$

where $\odot$ is the element-wise multiplication and $\operatorname{diag}[\cdot]$ is the diagonal matrix with input vector on its diagonal. The inverse of this matrix can be very easily computed using the Sherman-Morrisson formula [25].

7. REFERENCES

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *Proc. IEEE SCVT*, 2011, pp. 1–6.

[2] F. Nesta and M. Omologo, "Cooperative Wiener-ICA for source localization and separation by distributed microphone arrays," in *Proc. IEEE ICASSP,* 2010, pp. 181–184.

[3] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda,"Sound source segregation based on estimating the incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. Technol.*, vol. 22, pp. 149–157, February 2001.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, pp. 1830–1847, July 2004.

[5] Y. Hioka and W.B. Klein, "Distributed blind source separation with an application to audio signals," in *Proc. IEEE ICASSP*, 2011, pp. 233–236.

[6] 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.

[7] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Distributed microphone array processing for speech source separation with classifier fusion," in *Proc. IEEE MLSP,* 2012, pp. 1–6.

[8] P. Forero, A. Cano, and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE J. Selec. Top. Signal Process.*, vol. 5, pp. 1–18, Aug. 2011.

[9] D. Gu, "Distributed EM algorithm for Gaussian mixtures in sensor networks," *IEEE Trans. Neural Networks,* vol. 19, 1154–1166, July 2008.

[10] T.P. Minka, "Estimating a Dirichlet distribution," *Technical report*, Microsoft Research, Cambridge, 2003.

[11] N. Bouguila, D. Ziou, and J. Vaillancourt, "Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application," *IEEE Trans. Image Process.,* vol. 13, pp. 1533–1543, Nov. 2004.

[12] A.S. Bijral, M. Breitenbach, and G. Grudic, "Mixture of Watson distributions: a generative model for hyperspherical embedding," in *J. Machine Learning Research*, pp. 35–42, 2007.

[13] S. Sra and D. Karp, "The multivariate Watson distribution: maximum-likelihood estimation and other aspects," preprint: arXiv:1104.4422v2, May 2012.

[14] D.H. Tran and R. Haeb-Umbach, "Blind separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE ICASSP,* 2010, pp. 241–244.

[15] P.D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. ICA*, Sept. 2004, pp. 430–436.

[16] P.D. O'Grady and B. A. Pearlmutter, "The LOST algorithm: Finding lines and separating speech mixtures," *EURASIP J. Adv. Signal Process.,* pp. 1–17, 2008.

[17] K. V. Mardia and I. L. Dryden, "The complex Watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B*, vol. 61, pp. 913–926, 1999.

[18] I.S. Gradshteyn and I.M. Ryzhik, *Table of integrals, series, and products,* seventh edition, Academic Press, MA, USA, 2007.

[19] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intel.,* vol. 20, pp. 226–239, March 1998.

[20] S. Doclo, M. Moonen, T.V.D. Bogaert, 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, Jan. 2009.

[21] 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.

[22] J.B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943-950, 1979.

[23] 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.

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

[25] J. Sherman and J. W. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *Annals of Mathematical Statistics,* vol. 21, pp. 124–127, 1950.