Simultaneous Tracking and Separation of Multiple Sources Using Factor Graph Model

Koby Weisberg, *Student Member, IEEE*, Bracha Laufer-Goldshtein, *Student Member, IEEE* and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—In this paper, we present an algorithm for direction of arrival (DOA) tracking and separation of multiple speakers with a microphone array using the factor graph statistical model. In our model, the speakers can be located in one of a predefined set of candidate DOAs, and each time-frequency (TF) bin can be associated with a single speaker. Accordingly, by attributing a statistical model to both the DOAs and the associations, as well as to the microphone array observations given these variables, we show that the conditional probability of these variables given the microphone array observations can be modeled as a factor graph. Using the loopy belief propagation (LBP) algorithm, we derive a novel inference scheme which simultaneously estimates both the DOAs and the associations. These estimates are used in turn for separating the sources, by directing a beamformer towards the estimated DOAs, and then applying a TF masking according to the estimated associations. A comprehensive experimental study demonstrates the benefits of the proposed algorithm in both simulated data and real-life measurements recorded in our laboratory.**

***Index Terms*—Speaker tracking, speaker separation, factor graphs, loopy belief propagation (LBP)**

I. Introduction

Multiple-speaker separation is a well-known problem in the speech processing community, aiming to separate the measured microphone signal to its different sources. Another problem of substantial interest is tracking of a moving speaker, which can be used for separation tasks, and is also required in other applications, including navigation, target acquisition and beamforming. Both problems become challenging when multiple moving speakers are concurrently active, as well as when additive interference signals are also captured by the microphone array.

Among the most common DOA estimation methods are the steered response power (SRP)-phase transform (PHAT) algorithm [1] and the multiple signals classification (MUSIC) algorithm [2]. However, these techniques are not optimal in the multiple-speaker case, and do not address dynamic scenarios where the sources are moving during the recording. For the separation task, existing algorithms can be roughly divided into four groups: independent component analysis (ICA) algorithms that assume independence of the original source signals [3]; beamforming methods based on the spatial diversity of the speakers [4]; algorithms based on nonnegative matrix factorization (NMF) of the speech power spectral density (PSD); and methods that rely on the sparsity of speech signals in the TF domain [5], [6]. In the latter, the main assumption is that each TF bin is dominated by a single active speaker. These algorithms usually estimate a separation mask that assigns each TF bin to the active speaker, and use it for separation by applying a mask to the PSD of the measured signal. Comprehensive surveys of separation methods can be found in [7]–[9].

Several algorithms address the joint problem of localization and separation. In [10], the expectation-maximization (EM) algorithm is implemented for estimating both the DOAs and the separation masks of multiple static speakers with a single microphone pair. The algorithm is based on a Mixture of Gaussians (MoG) model defining a grid of possible DOA candidates. Assuming a single dominant speaker in each TF bin, the interaural phase differences (IPDs) from all TF bins are clustered into groups associated with a particular speaker from a candidate DOA. The E-step in the proposed EM iterations provides a soft assignment of each observation to both speaker and DOA. By marginalizing over the DOAs, a separation mask is obtained. The weights of the Gaussians, obtained by the M-step, define a probability distribution on the candidate DOAs, and the DOAs of the active speakers are estimated by selecting the candidates with the highest probabilities. In [11], the algorithm was extended using a Markov random field (MRF) model to promote smoothness of the separation mask in both time and frequency, which was shown to improve the separation results. In [12], a dynamic scenario was addressed by two recursive EM (REM) variants, applied to a multichannel extension of the model in [10]: one based on Titterington recursive EM (TREM) [13] and the second based on Cappé and Moulines recursive EM (CREM) [14]. The separation task was not addressed in this paper.

Tracking and separation of moving speakers was addressed in [15]. In this paper, the basic model assumes static sources, and the tracking is applied as a post processing step following the static localization procedure. Here also, the IPDs are used as feature vectors, and are modeled using wrapped distributions. The DOA of each source is computed using circular linear regression, which in the multiple-speaker case, is solved by the EM algorithm. Similar to [10], the E-step is used for estimating the separation mask, and the slopes of the IPDs are transformed to DOAs using the prior knowledge on the inter-channel delay. A dynamic scenario is addressed by first finding the DOAs for each time-step, and then using

Koby Weisberg, Bracha Laufer-Goldshtein and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: kobyavi@gmail.com; bracha.laufer@biu.ac.il; Sharon.Gannot@biu.ac.il). This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreement No. 871245.

the estimated DOAs as observations for a factorial wrapped Kalman filter.

The above mentioned papers use the IPD features for the localization task, however with these features, the presence of additive measurement noise is not directly addressed. In [16]–[18], the phase-related feature vectors were substituted by the raw short-time Fourier transform (STFT) observations. In addition, the noise (or reverberation) was explicitly modeled, resulting in improved performance in noisy (or reverberant) scenarios. The observations at the microphone array were modelled as a mixture of multivariate complex-Gaussians with zero-mean, and a spatial covariance matrix consisting of both the speech and the noise PSDs. Furthermore, it was shown in [17] that the PSDs of the candidate speakers can be estimated in advance (prior to the application of the EM algorithm) from the outputs of a set of minimum variance distortionless response (MVDR)-beamformers (BFs).

The algorithm in [17] was extended in [19] to address dynamic scenarios. First, it was shown that the raw observation features can be substituted by new features, which are the likelihood ratio test (LRT) at each candidate DOA indicating whether the MVDR-BF output at this DOA dominated by either speech or noise. The utilization of these new features, results in a lower computational burden that is beneficial in online and real-time tracking applications. Second, a tracking procedure was proposed by applying the CREM algorithm.

The above algorithms do not provide an explicit DOA estimate, but rather a probability map over the candidate DOAs. While for the static localization task the actual DOA can be found relatively easily by finding the peaks in the probability map, in a dynamic case the peaks should be calculated for each time-step rendering the explicit trajectory inference difficult.

Another approach to address the tracking task is to substitute the MoG model with an Hidden Markov Model (HMM). In this approach, the DOAs of the speakers are also discretized to a finite set of candidates. The model assumes that the dynamics of the sources is governed by a Markov process, with higher probability for switching from one candidate to an adjacent candidate at each time-step, thus allowing small changes in the DOA [6], [20], [21]. In [6] this model was extended to full 2D tracking rather than DOA-only tracking.

The tasks of tracking and separation depend on each other. The reason is that when the DOAs of the speakers are known, we can identify the dominating DOA in each TF bin and associate it with the corresponding speaker, and thus extract it by masking. In the opposite direction, given the association map that relates each TF bin to its dominating speaker, we can use the set of TF bins attached with each speaker to infer its corresponding DOA. Examples of using the outcomes of localization to perform separation can be found in [10], [12], [17], [22], and for the other direction in [23], [24]. In this paper, we handle the case where both the DOA and the associations are unknown, and solve both tasks simultaneously without pre-initialization.

A simultaneous tracking and separation algorithm was proposed in [6], [21], the latter apply a Bayesian approach. The definition of the hidden variables here is different from that defined in [10]. In [10] each TF observation is associated with both a DOA and a speaker, whereas in [21] each observation is associated only with a speaker, and the speaker is associated with a DOA. This approach uses fewer hidden variables, hence reducing the computational requirements, while modeling real scenarios more accurately. The continuous movement of the speakers is reflected by modelling the DOAs of the speakers as Markov processes. Since an exact inference of the hidden variables from the observations is intractable, a variational inference was applied.

In the current contribution, we present a novel algorithm for simultaneous tracking and separation of multiple speakers based on a factor graph model. Factor graph models [25] are used in many complex tasks in various signal processing fields, such as communication [26], sonar detection [27] and robotics [28], [29]. To the best of our knowledge, this model was not used for the task of speaker tracking and separation. In our paper, we define the hidden data as in [21] using two groups of latent variables. The first group consists of the DOA of the sources that are modeled as separated Markov chains for each source, where the transition probability is set to allow only small changes in the DOAs in subsequent time steps. The second group consists of the associations of the TF bins to the different sources, which can be modeled by an i.i.d. distribution or, following [11], using an MRF model to smooth the associations in time and frequency. Given both the DOAs and the associations, the observations are modeled as a complex-Gaussian distribution, which is a function of the unknown speech PSD that can be estimated using the maximum likelihood (ML) criterion. We then show that the posterior of the latent variables given the observations defines a factor graph, and derive a novel inference method for simultaneously estimating all latent variables, using the loopy belief propagation (LBP) inference algorithm [30].

The remainder of this paper is organized as follows. In Section II the problem of simultaneous tracking and separation is defined. In Section III the statistical model is defined and a factor graph model is formulated. The inference algorithm based on the LBP algorithm is derived in Section IV. An experimental study that demonstrates the capabilities of the proposed algorithm, can be found in Section V. Conclusions are outlined in Section VI.

II. Problem Formulation

In this section we formulate the problem addressed in this paper. Similar formulation can be found in [6], [21]. Consider an array of N microphones, receiving signals of J moving speakers. At each time step, each speaker is located at a specific DOA on a grid of M possible DOAs $[\vartheta_1, \ldots, \vartheta_M]$. Due to the dynamic nature of the problem, the DOAs may vary from one time step to the other. The proposed method is applied in the STFT domain with $t = 1, \ldots, T$ denoting the time index and $k = 1, \ldots, K$ denoting the frequency index. Let $d_t(j)$ be a categorical random variable denoting the DOA index of the jth speaker at time index t, i.e. $d_t(j) \in [1, \ldots, M]$. Assuming that only a few speakers are mixed and relying on the W-disjoint orthogonality (WDO)

property of speech signals in the STFT domain [5], each TF bin is dominated by a single active speaker. Let $a_{t,k}$ be a categorical random variable denoting the active speaker at the (t,k)th bin, i.e. $a_{t,k} \in [1,\ldots,J]$. Following these definitions, the nth microphone signal is given by:

$$z_{t,k}^{(n)} = g_k^{(n)}(d_t(a_{t,k}))s_{t,k}(d_t(a_{t,k})) + v_{t,k}^{(n)}, \tag{1}$$

where $d_t(a_{t,k}) \in [1,\ldots,M]$ is the DOA index of the active speaker at the (t,k)th bin, $g_k^{(n)}(m)$ is the relative transfer function (RTF) associated with the mth candidate DOA and defined between the nth microphone and the reference microphone, $s_{t,k}(m)$ is the speech signal from the mth candidate as measured by the reference microphone, and $v_{t,k}^{(n)}$ denotes a stationary ambient noise at microphone $n \in [1,\ldots,N]$.

In low-reverberation environments, the RTF approximately corresponds to the direct path between the source and the microphone:

$$g_k^{(n)}(m) = \exp\left(-\iota\frac{2\pi k}{K}\frac{\tau_{m,n}}{T_s}\right) \tag{2}$$

where T_s denotes the sampling period, and $\tau_{m,n}$ denotes the known time difference of arrival (TDOA) between the nth microphone and the reference microphone, associated with the mth candidate DOA. Note that in the algorithm derivation we assume a far-field and free-field scenario, such that the RTF of the different candidates can be solely determined by the TDOA. In the experimental study, we empirically demonstrate that the derived algorithm can also perform well in reverberant environments.

The measured signals (1) can be written in a vector form as:

$$\mathbf{z}_{t,k} = \mathbf{g}_k(d_t(a_{t,k}))s_{t,k}(d_t(a_{t,k})) + \mathbf{v}_{t,k} \tag{3}$$

where

$$\mathbf{z}_{t,k} = \left[z_{t,k}^{(1)}, z_{t,k}^{(2)}, \ldots z_{t,k}^{(N)}\right]^T$$
$$\mathbf{g}_k(m) = \left[1, g_k^{(2)}(m), \ldots, g_k^{(N)}(m)\right]^T$$
$$\mathbf{v}_{t,k} = \left[v_{t,k}^{(1)}, v_{t,k}^{(2)}, \ldots, v_{t,k}^{(N)}\right]^T.$$

assuming, without loss of generality, that the first microphone is chosen as the reference microphone. The generation of the observation by the defined model is illustrated in Figure 1.

Our goal is to estimate the DOAs of the speakers at each time-step and to separate the measured signals into the source signals of each of the speakers. To this end, we define a statistical model and present an inference scheme that simultaneously estimates the speaker associations $a_{t,k}$ and the DOAs $d_t(j)$.

III. Factor Graph Model

We consider the speaker associations $a_{t,k}$ and the DOAs $d_t(j)$ as latent variables that we would like to infer from the observations $\mathbf{z}_{t,k}$. Applying Bayes rule, the posterior of the latent variables is given by:

$$P(\mathbf{d},\mathbf{a}|\mathbf{z}) = \frac{P(\mathbf{z}|\mathbf{a},\mathbf{d})P(\mathbf{a})P(\mathbf{d})}{P(\mathbf{z})} \tag{4}$$

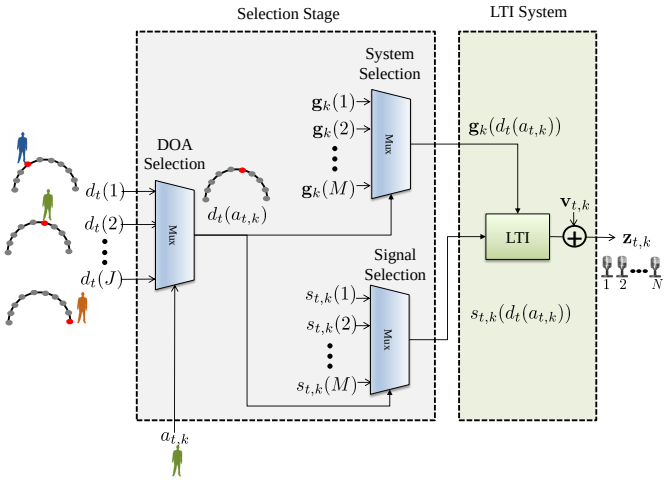


Fig. 1: An illustration of the generation of the observations by the presented model. The first part is the selection stage. The variable $a_{t,k}$ representing the active speaker, is used for selecting the DOA associated with the active speaker. The chosen DOA candidate is used for selecting both the RTF and the input speech signal that are associated with this candidate. The second part describes the actual generation of the observations by an LTI system model, in which the chosen speech signal is filtered by the chosen RTF and noise is added.

where $\mathbf{a} = \text{vec}_{t,k}\{a_{t,k}\}$, $\mathbf{d} = \text{vec}_{t,j}\{d_t(j)\}$, $\mathbf{z} = \text{vec}_{t,k}\{\mathbf{z}_{t,k}\}$, and we assume independence between the DOAs $\mathbf{d}$ and the associations $\mathbf{a}$.

The main task of this paper is to find the marginal posterior of the variables, namely $P(a_{t,k}|\mathbf{z})\ \forall t,k$, and $P(d_t(j)|\mathbf{z})\ \forall t,j$. However, an exact computation of these marginal distributions is intractable. In [21] this posterior was approximated by a product of probabilities from known families, and the variational inference was used for estimating the parameters of these probabilities. In the current paper, we present a statistical model in which the posterior is given in a form of a factor graph. We then propose to use the LBP inference algorithm in order to find the marginal posterior for each variable.

In this section, we define the prior probabilities of the hidden variables $P(\mathbf{a})$ and $P(\mathbf{d})$, as well as the probability of the observations given the hidden variables $P(\mathbf{z}|\mathbf{a},\mathbf{d})$, and use them to form the factor graph of the posterior probability (4). The inference algorithm that is applied to this factor graph model is described in Section IV. A brief general review on factor graph models and their inference methods is given in Appendix B.

A. The DOA model

Following [20], [21] the prior probabilities of the DOAs of each of the speakers are modeled as separated and independent Markov chains. The state of the Markov process associated with each speaker is the DOA index of the corresponding speaker at each time step. The transition probabilities are set in a way that allows the DOA of each speaker to vary smoothly

over time. Accordingly, the joint probability of $\mathbf{d}$ is given by:

$$P(\mathbf{d}) = \prod_{j=1}^{J} \left[\Omega_j(d_1(j)) \prod_{t=2}^{T} \Psi(d_{t-1}(j), d_t(j)) \right] \tag{5}$$

where we have defined the following *potential* functions:

$$\Psi(m_1, m_2) = P(d_t(j) = m_2 | d_{t-1}(j) = m_1) \tag{6a}$$

$$\Omega_j(m) = P(d_1(j) = m) \tag{6b}$$

where $P(d_t(j) = m_2 | d_{t-1}(j) = m_1)$ is the probability to switch from one DOA to another in subsequent time steps, and $P(d_1(j) = m)$ is the initial probability of the jth speaker at time $t = 1$. In order to achieve a continuous trajectory, the transition probability is set as:

$$P(d_t(j)|d_{t-1}(j)) \propto \begin{cases} 1 \text{ if } d_t(j) = d_{t-1}(j) \\ \exp(-\alpha) \text{ if } d_t(j) = d_{t-1}(j) \pm 1 \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

where $\alpha > 0$ is a hyper-parameter which controls the smoothness of the trajectory. The initial DOA probability is assumed to be known. However, we observed in our experiments (see Sec. V-C) for a case with three speakers) that it may also be randomly initialized, hence a prior knowledge on the initial DOA is in practice unnecessary.

B. The association model

For the prior probability of the association variables $\mathbf{a}$, we propose two alternative models. The simple model is an i.i.d. distribution where an independence between the associations in different TF bins is assumed, and each of them is uniformly distributed, namely:

$$P(\mathbf{a}) = \prod_{t,k} \frac{1}{J} = \frac{1}{J^{TK}} \tag{8}$$

which is a constant expression. In the following, we derive the inference algorithm for this model.

An alternative model is described in Section IV-F following [11]. This model takes into account the speech activity pattern across time and frequency, and represents the relation between adjacent TF bins using a Markov random field (MRF). The MRF model provides a more accurate description of the behavior of the association variables across time and frequency compared to the uniform model (8), at the cost of slightly increasing the complexity of the inference scheme. In the experimental part in Section V, we show that the MRF model has a slight advantage over the uniform model in terms of the actual performance. By describing both models, we would like to further demonstrate the flexibility of the proposed statistical framework that facilitates the use of various models for the associations with only small adjustments to the proposed inference algorithm.

C. The observation model

We will now define the statistical model of the observations given the hidden variables $P(\mathbf{z}|\mathbf{a}, \mathbf{d})$. The speech signal is modeled as a zero-mean complex-Gaussian random variable with a time-varying PSD:

$$P(s_{t,k}(d_t(a_{t,k}))) = \mathcal{N}\left(s_{t,k}(d_t(a_{t,k})); 0, \phi_{s,t,k}(d_t(a_{t,k}))\right) \tag{9}$$

where $\mathcal{N}(\cdot;\cdot,\cdot)$ denotes the complex-Gaussian probability and $\phi_{s,t,k}(d_t(a_{t,k}))$ is the unknown PSD of the speech signal received from the DOA of the active speaker at the (t,k)th bin. The noise is modeled as a zero-mean complex-Gaussian random vector with a time-invariant covariance matrix $\Phi_{\mathbf{v},k}$:

$$P(\mathbf{v}_{t,k}) = \mathcal{N}\left(\mathbf{v}_{t,k}; \mathbf{0}, \Phi_{\mathbf{v},k}\right). \tag{10}$$

It is assumed that the noise covariance matrix is known in advance, or can be estimated during speech-absent segments, due to the noise stationarity.

Following equations (3), (9) and (10), the conditional probability density function (p.d.f.) of the (t,k)th observation given the DOA of the active speaker at this bin can be expressed as

$$P(\mathbf{z}_{t,k}|d_t(a_{t,k})) = \mathcal{N}(\mathbf{z}_{t,k}, \mathbf{0}, \Phi_{\mathbf{z},t,k}(d_t(a_{t,k}))), \tag{11}$$

with:

$$\Phi_{\mathbf{z},t,k}(m) = \mathbf{g}_k(m)\mathbf{g}_k^H(m)\phi_{s,t,k}(m) + \Phi_{\mathbf{v},k}, \tag{12}$$

where the speech and noise signals are assumed to be statistically independent.

Since $\phi_{s,t,k}(m)$ does not directly depend on the identity of the active speaker but on its DOA, we can estimate it prior to the algorithm application using the maximum likelihood estimator (MLE) (as detailed in Appendix A1). Next, we factorize the conditional p.d.f. as follows:

$$P(\mathbf{z}_{t,k}|d_t(a_{t,k})) = T_{t,k}(d_t(a_{t,k})) \cdot G_{t,k} \tag{13}$$

where $T_{t,k}(m)$ consists of all terms that depend on the candidate DOA m, and $G_{t,k}$ consists of the remaining terms that are independent of m, and therefore is ignored in the following derivations that are based on non-normalized distributions. The function $T_{t,k}(m)$ represents the likelihood ratio test (LRT) that tests whether the signal from the mth candidate is associated with noise or with a speaker (see Appendix A1). The LRT is defined based on the output of an MVDR-BF directed towards the mth candidate:

$$\hat{s}_{\mathbf{w},t,k}(m) \equiv \mathbf{w}_k^H(m)\mathbf{z}_{t,k} \tag{14}$$

and the MVDR-BF is defined by:

$$\mathbf{w}_k(m) = \frac{\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)}{\mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)}. \tag{15}$$

The computation of $T_{t,k}(m)$ is described in Algorithm 1, and the full derivation can be found in Appendix A. Finally, assuming independence between the observations given the latent variables, the *likelihood* of the observations is given by:

$$P(\mathbf{z}|\mathbf{a}, \mathbf{d}) = \prod_{t,k} T_{t,k}(d_t(a_{t,k})) \cdot G_{t,k}. \tag{16}$$

Algorithm 1 Likelihood calculation

- Calculate the MVDR-BF $\mathbf{w}_k(m)\ \forall k,m$ using (15)
- Calculate the output of the MVDR-BF $\hat{s}_{\mathbf{w},t,k}(m)\ \forall t,k,m$ using (14)
- Calculate the PSD of the residual noise $\forall k,m$:

$$\phi_{v,k}(m) \equiv \frac{1}{\mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)}$$

- Calculate the SNR at the output of the MVDR-BF $\forall t,k,m$:

$$\eta_{t,k}(m) = \frac{|\hat{s}_{\mathbf{w},t,k}(m)|^2}{\phi_{v,k}(m)}$$

- Calculate the LRT $\forall t,k,m$:

$$T_{t,k}(m;\hat{\phi}_{s,t,k}(m)) = \frac{1}{\eta_{t,k}(m)}\exp\left(\eta_{t,k}(m)-1\right)$$

D. The observation factor

For the factor graph model, we need to explicitly define a factor for each observation as a function of all the associated latent variables. Thus, we rewrite (16) as:

$$P(\mathbf{z}|\mathbf{a},\mathbf{d}) = \frac{1}{C_z}\prod_{t,k}\Upsilon_{t,k}(a_{t,k},d_t(1)\dots d_t(J)) \tag{17}$$

where $\frac{1}{C_z} \equiv \prod_{t,k} G_{t,k}$ is a constant normalization and:

$$\Upsilon_{t,k}(a_{t,k},d_t(1)\dots d_t(J)) \equiv T_{t,k}(d_t(a_{t,k})). \tag{18}$$

We denote this function as the *observation* factor. Note that while $T_{t,k}(\cdot)$ is a function of a single variable $d_t(a_{t,k}) \in [1\dots M]$, the potential function $\Upsilon_{t,k}(\cdot,\dots,\cdot)$ is a function of $J+1$ variables, namely, $a_{t,k}$ and $d_t(1)\dots d_t(J)$. The definition of $\Upsilon_{t,k}(\cdot,\dots,\cdot)$ is necessary as the factor graph model requires that the factors are presented as direct functions of each of the individual hidden variables separately. Note also that in contrast to the DOA factor Ψ (7), which is fixed along time, the observation factor varies across time and frequency, since it is determined by the specific observation in each TF bin.

E. The Factor Graph

We can now express the posterior $P(\mathbf{a},\mathbf{d}|\mathbf{z})$ as a factor graph. Substituting (5), (8) and (17) into (4), we obtain:

$$P(\mathbf{d},\mathbf{a}|\mathbf{z}) = \frac{1}{C}\prod_{t,k}\Upsilon_{t,k}(a_{t,k},d_t(1)\dots d_t(J)) \times \prod_{j=1}^{J}\Omega_j(d_1(j))\prod_{t=2}^{T}\Psi(d_{t-1}(j),d_t(j)) \tag{19}$$

where the factors $\Psi(\cdot,\cdot)$, $\Omega_j(\cdot)$ and $\Upsilon_{t,k}(\cdot,\dots,\cdot)$ are defined in (6a),(6b) and (18) respectively, and $C \equiv C_z \cdot J^{TK} \cdot P(\mathbf{z})$ is a normalization constant. The factor graph model is illustrated in Fig. 2.

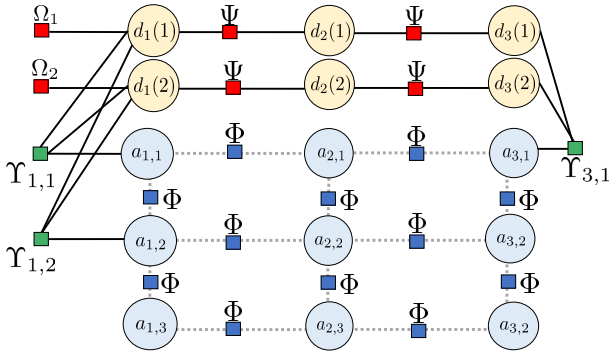


Fig. 2: The proposed factor graph. Here, $J=2$ speakers $K=3$ frequencies, and $T=3$ time-frames, for simplicity. Only three out of $T\times K$ observation factors are drawn. The gray dashed lines and the factors Φ correspond to the modified factor graph presented in Section IV-F, which is based on the MRF model for the associations. For the uniform distribution model of the associations (8) these connections and factors are ignored.

IV. Inference using the LBP

The obtained factor graph contains loops, as can be seen in the illustrative example in Fig. 2, and therefore the loopy belief propagation (LBP) [31] can be used for its inference. In this section, we derive the LBP algorithm to approximate the marginal posteriors of the latent variables given the observations. The final DOA trajectory and the separated signals are then obtained based on the computed marginals. In the LBP, messages are sent from the factors to the variables and vice versa (see Appendix B). In the proposed model there are three groups of factors: i) Ω (connected to $d_1(1),\dots,d_1(J)$); ii) Ψ (connected to $\mathbf{d}$); and iii) Υ (connected to all variables). The messages are functions of the corresponding variable (either source or destination), and are calculated using the general equations (54a) and (54b). However, these general equations can be simplified in our case to achieve more efficient formulas, as shown in the sequel.

A. Notation

In the following derivations we use a simplified set of notations. The messages from Ψ to $d_t(j)$ are denoted by $\overrightarrow{\psi}(d_t(j))$ and $\overleftarrow{\psi}(d_t(j))$ for the forward and backward messages, respectively. For the completeness of the notation we use this notation also for $t=1$ and $t=T$, where for $t=1$ the forward message of the factor Ψ is replaced with the corresponding Ω factor, and for $t=T$ the backward message is fixed to uniform, as there is no backward message to the last variable. For the observation factor, we use $\upsilon_{t,k}(\cdot)$ for the outgoing messages from the observation factor to each of the variables connected to it, where the destination variable is deduced from the term in the brackets, i.e. $\upsilon_{t,k}(d_t(j))$ refers to messages to the DOA variables and $\upsilon_{t,k}(a_{t,k})$ refers to messages to the association variables. The messages from $d_t(j)$ to the observations are denoted by $\delta_{t,k,j}(d_t(j))$. The different types of messages are illustrated in Fig. 3.

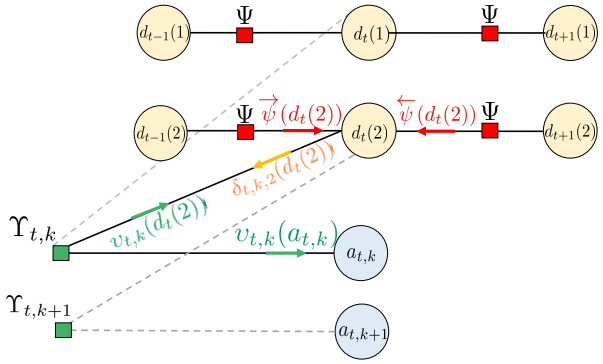


Fig. 3: The messages in the proposed LBP algorithm. The arrows are pointing from the sending variable/factor to the receiving variable/factor, and the notation of the associated message is written above/below the arrow.

B. Messages from the DOA factors

In general the factors send messages to their neighbor variables (the *outgoing messages*), where these messages depend on the incoming messages from variables to the factors (the *incoming messages*). However, for Ψ and Ω, the factor is a function of only a single or two variables and it has only a single incoming message. Therefore, we do not explicitly define the incoming messages for these factors. Instead, we substitute the incoming message with its definition (54a). As a result, each of the outgoing messages is expressed in terms of the outgoing messages of its neighbor factors to the corresponding variable.

The forward messages of Ψ for $t > 1$ are given by:

$$\overrightarrow{\psi}(d_t(j)) = \sum_{d_{t-1}(j)} \Psi(d_{t-1}(j), d_t(j))\overrightarrow{\psi}(d_{t-1}(j))\overline{\upsilon}_{t-1}(d_{t-1}(j)) \quad (20)$$

where

$$\overline{\upsilon}_t(d_t(j)) = \prod_k \upsilon_{t,k}(d_t(j)) \quad (21)$$

is the message of all K observations to $d_t(j)$. For $t = 1$ the message is given by:

$$\overrightarrow{\psi}(d_1(j)) = \Omega_j(d_1(j)). \quad (22)$$

The backward message $\overleftarrow{\psi}(d_t(j))$ is symmetric, where for $t = T$ it is set to uniform for completeness.

C. Message from and to the observation factors

The incoming messages from the DOA variables $d_t(j)$ to the observations Υ are given by the multiplication of the incoming messages of each DOA variable (54a), namely:

$$\delta_{t,k,j}(d_t(j)) = \overrightarrow{\psi}(d_t(j))\overleftarrow{\psi}(d_t(j)) \prod_{\tilde{k}\neq k} \upsilon_{t,k}(d_t(j)). \quad (23)$$

The full derivation of the outgoing messages from the observation factors to their neighbor variables can be found in Appendix C. In order to simplify the messages, we first define the correlation between $T_{t,k}(:)$ and the normalized incoming message $\delta_{t,k,j}(:)$ as:

$$\rho_{t,k}(j) = \sum_{m=1}^{M} T_{t,k}(m)\tilde{\delta}_{t,k,j}(m) \quad (24)$$

where $\tilde{\delta}_{t,k,j}(m) = \frac{\delta_{t,k,j}(m)}{\sum_m \delta_{t,k,j}(m)}$ is the normalized message. The correlation measures the similarity between $\delta_{t,k,j}(:)$, which is the current estimate of the jth speaker DOA, and $T_{t,k}(:)$, which is the (t,k)th bin DOA likelihood based on the observation. The obtained $\rho_{t,k}(j)$ is therefore a non-normalized association of the (t,k)th bin to a speaker based on the similarity between the observed DOA and the estimated DOA of each of the speakers, namely, a higher value is given to the speaker whose estimated DOA matches the observed DOA, and vice versa. This process is illustrated in Fig. 4.

Using the definition of $\rho_{t,k}(j)$, the message from the observation factor to the association variable is given by:

$$\upsilon_{t,k}(a_{t,k}) = \rho_{t,k}(a_{t,k}) \quad (25)$$

and the message from the observation factor to the DOA variables is given by:

$$\upsilon_{t,k}(d_t(j)) = T_{t,k}(d_t(j)) + \sum_{\ell\neq j} \rho_{t,k}(\ell). \quad (26)$$

The meaning of the message conveyed by Υ to the jth speaker DOA is as follows. The message consists of two terms: $T_{t,k}(d_t(j))$ that depends on the DOA value $d_t(j)$, and $\sum_{\ell\neq j}\rho_{t,k}(\ell)$, which is independent of $d_t(j)$. If one of the other speakers is active with high probability at this TF bin, then the value of the second term is high, and the message is close to uniform with respect to $d_t(j)$, i.e. does not indicate any preference to a certain DOA. Otherwise, the jth speaker is probably active at this TF bin, and the message is dominated by the first term $T_{t,k}(d_t(j))$, which is the DOA likelihood based on the (t,k)th bin observation.

In the next step, the messages from all frequencies are integrated together for each speaker in $\overline{\upsilon}_t(d_t(j))$ (21) to determine its new DOA. In this integration, uniform messages do not add any information. Therefore the integrated message for the jth speaker, contains only the information from the relevant frequencies where the jth speaker is active. The calculation of the messages $\overline{\upsilon}_t(d_t(j))$ is illustrated in Fig. 5.

Note that while the message to the variable $a_{t,k}$ depends on the incoming messages from all other variables $\rho_{t,k}(1),\ldots,\rho_{t,k}(J)$, the message to the DOA variable $d_t(j)$ of the jth speaker depends on the message from all other variables $\rho_{t,k}(1),\ldots,\rho_{t,k}(j-1),\rho_{t,k}(j+1),\ldots,\rho_{t,k}(J)$ except for the jth speaker message $\rho_{t,k}(j)$, since by the definition of the LBP algorithm, the message to a particular variable depends on all incoming messages except for the message from this variable itself.

Three additional notes on the differences between the general formulation of the message (54b) in Appendix B and the simplified message (26) are in place. 1) Instead of the raw incoming messages $\delta_{t,k,1}(:),\ldots,\delta_{t,k,J}(:)$, the outgoing messages use $\rho_{t,k}(1),\ldots,\rho_{t,k}(J)$ defined by the correlation between the incoming messages and the observations (24);

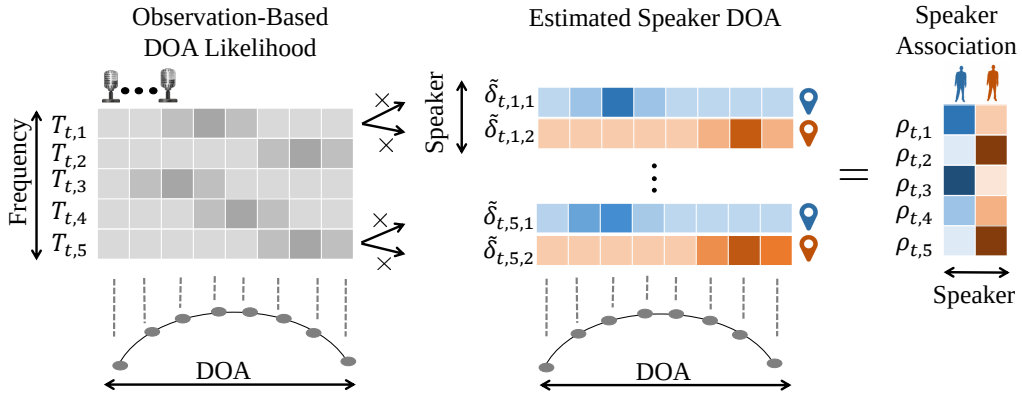


Fig. 4: Illustration of the calculation of $\rho_{t,k}(j)$. The vectors represent the probabilities over the candidate DOAs, where darker elements correspond to more probable candidates. The case of two speakers is illustrated by blue and orange vectors representing the current DOA estimate of the each of the speakers. The observation-based DOA likelihood vectors (in gray) are correlated with the estimated DOA of the speakers, resulting in $\rho_{t,k}(j)$, which represents the association of the (t,k)th bin to either of the speakers based on the observation.

2) The message from the association variable a does not appear here since this variable has no connected factor except the observation; and 3) The obtained messages involve only $T_{t,k}(:)$, and not the entire factor $\Upsilon_{t,k}$, since this is all the information that the factor contains (18).

D. The inference algorithm

The full inference algorithm is as follows. We first initialize all messages to be uniform, then we iterate over all the variables and update their incoming messages from their associated factors using equations (20, 22, 23, 25, 26). The iterations of the LBP algorithm are stopped when the following stopping criterion is satisfied: the maximum change in the log messages between subsequent iterations is smaller than ε or when the number of iterations reaches $N_{\max}$, which is defined as the maximum number of iterations.

The final stage is to compute the marginals, using the following equations:

$$P(d_t(j)|\mathbf{z}) \propto \overrightarrow{\psi}(d_t(j))\overleftarrow{\psi}(d_t(j))\overline{\upsilon}_t(d_t(j)) \tag{27a}$$

$$P(a_{t,k}|\mathbf{z}) \propto \upsilon_{t,k}(a_{t,k}) \tag{27b}$$

where $\overline{\upsilon}_t(d_t(j))$ is defined in (21) and the sign $\propto$ implies that an additional normalization step is required. The inference algorithm is summarized in Algorithm 2.

Algorithm 2 Loopy belief propagation (LBP) for simultaneous tracking and separation

```
Initialize all messages to uniform
while Stopping criterion not satisfied do
    for t=1:T do
        update Ψ messages ∀j using (20 or 22)
        compute δ̃_{t,k,j}(d_t(j)) ∀j,k using (23)
        compute υ_{t,k}(a_{t,k}) and υ_{t,k}(d_t(j)) ∀j,k using (25,26)
    end
end
compute the marginals using (27a,27b)
```

E. Tracking and separation

Applying the inference procedure, the marginals of all the hidden variables are computed. The trajectory of each speaker is obtained by selecting the most probable value for each $d_t(j)$:

$$\hat{d}_t(j) = \underset{m\in\{1,\ldots,M\}}{\operatorname{argmax}} P(d_t(j)=m|\mathbf{z}). \tag{28}$$

The association variables provide the separation mask, which can be used in order to separate the signal to its different sources. Following [17, Eq. (15)], the individual speech signal can be estimated by spatial multichannel filtering followed by single channel post-filtering (see e.g. [32]):

$$\widehat{S}_{t,k}(j) = P(a_{t,k}=j|\mathbf{z})\hat{s}_{\mathbf{w},t,k}(\hat{d}_t(j)) \tag{29}$$

where $P(a_{t,k}=j|\mathbf{z})$ is responsible for enhancing the jth speaker and attenuating the other speakers and $\hat{s}_{\mathbf{w},t,k}(\hat{d}_t(j))$ defined in (14) is the output of the MVDR-BF directed towards the estimated DOA of the jth speaker, and is responsible for reducing the ambient noise.

F. MRF model for the associations

In this section, we replace the uniform model of the association variables (8) by a more complex statistical model as suggested in [11], and describe the corresponding modifications to the factor graph and the inference algorithm. It was shown in [11] that in order to smooth the associations, and to reduce musical noise, it is more reasonable to model the dependency between the association variables in adjacent time and frequency indexes using the Markov random field (MRF) model. For this model, the joint probability of the association variables is given by:

$$P(\mathbf{a}) = \frac{1}{C_a}\prod_{t,k}\prod_{\tilde{t},\tilde{k}\in\mathcal{G}\{t,k\}}\Phi(a_{t,k},a_{\tilde{t},\tilde{k}}) \tag{30}$$

where $\mathcal{G}\{t,k\} = \{(t-1,k),(t+1,k),(t,k-1),(t,k+1)\}$ is the group of the indexes couples, C_a is a normalization constant, and $\Phi(j_1,j_2)$ is usually defined as:

$$\Phi(j_1,j_2) = \exp(\beta\delta_K(j_1,j_2)) \tag{31}$$

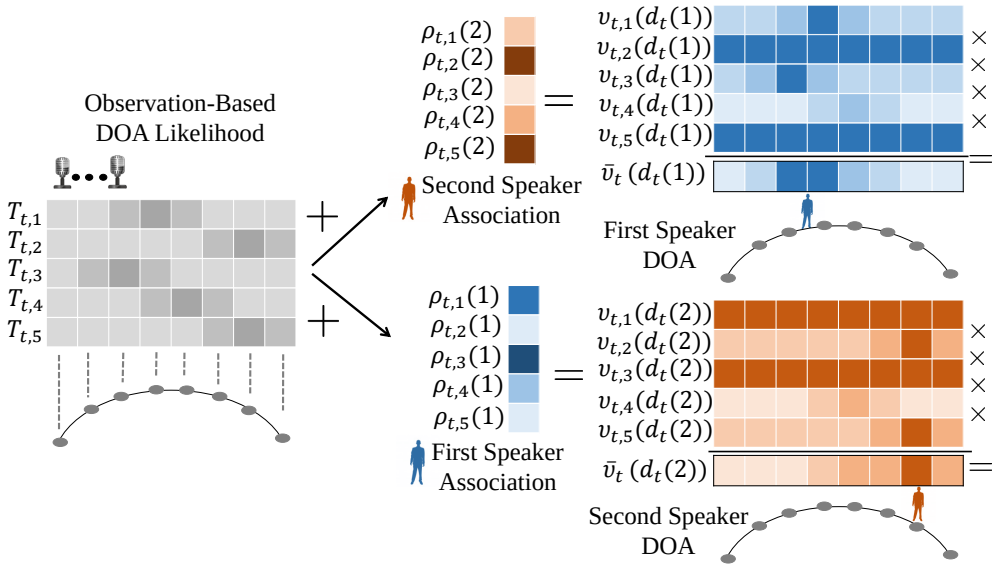


Fig. 5: Illustration of the calculation of $\bar{\upsilon}_t(d_t(j))$. Darker elements correspond to higher values. For each speaker the association to the other speakers (colored in orange or blue) is added to the DOA likelihood (colored in gray). The result is the per-frequency non-normalized probability for each speaker $\upsilon_{t,k}(d_t(j))$. Multiplication along the frequencies, results in the non-normalized DOA distribution.

where $\delta_K(\cdot,\cdot)$ is the discrete Kronecker delta function, and $\beta > 0$ is a hyper-parameter of the algorithm. This model encourages nearby TF bins to be associated to the same source, and makes the association map smoother. The parameter β controls this smoothness, where the map becomes smoother as β increases.

Incorporating this model, the factor graph is given by:

$$P(\mathbf{d},\mathbf{a}|\mathbf{z}) = \frac{1}{C}\prod_{t,k}\Upsilon_{t,k}(a_{t,k},d_t(1)\dots d_t(J)) \prod_{j=1}^{J}\Omega_j(d_1(j))\prod_{t=2}^{T}\Psi(d_{t-1}(j),d_t(j)) \prod_{t,k}\prod_{\tilde{t},\tilde{k}\in\mathcal{G}\{t,k\}}\Phi(a_{t,k},a_{\tilde{t},\tilde{k}}). \tag{32}$$

In the LBP we add $\overrightarrow{\phi}_t(a_{t,k})$ and $\overleftarrow{\phi}_t(a_{t,k})$, for the backward and forward messages of the MRF factors Φ in the time dimension and $\overrightarrow{\phi}_f(a_{t,k})$ and $\overleftarrow{\phi}_f(a_{t,k})$, for the messages in the frequency dimension. The outgoing messages of the factor Φ are given by:

$$\overrightarrow{\phi}_t(a_{t,k}) = \sum_{a_{t-1,k}}\Phi(a_{t-1,k},a_{t,k}) \overrightarrow{\phi}_t(a_{t-1,k})\overrightarrow{\phi}_f(a_{t-1,k})\overleftarrow{\phi}_f(a_{t-1,k})\upsilon_{t,k}(a_{t-1,k}). \tag{33}$$

The other three messages are defined similarly, and the edge messages are set to uniform. We also define the incoming message from the association variables to the observation:

$$q_{t,k}(a_{t,k}) = \overrightarrow{\phi}_t(a_{t,k})\overleftarrow{\phi}_t(a_{t,k})\overrightarrow{\phi}_f(a_{t,k})\overleftarrow{\phi}_f(a_{t,k}). \tag{34}$$

This modifies the incoming message (26) from the observation factor to the DOA variable $d_t(j)$ as follows:

$$\upsilon_{t,k}(d_t(j)) = T_{t,k}(d_t(j)) + \frac{\sum_{\ell\neq j}q_{t,k}(\ell)\rho_{t,k}(\ell)}{q_{t,k}(j)} \tag{35}$$

Compared to (26), the second constant additive term now measures the activity of the other speakers in the current TF bin based on both $\rho_{t,k}(j)$ that measures the association based on the current speaker DOA estimation, and $q_{t,k}(j)$ that measures the association based on the information from neighbor TF bins. The final inference of the DOA variables remains unchanged (27a), and the inference of the associations variable (27b) is modified to include also the MRF messages:

$$P(a_{t,k}|\mathbf{z}) \propto \overrightarrow{\phi}_t(a_{t,k})\overleftarrow{\phi}_t(a_{t,k})\overrightarrow{\phi}_f(a_{t,k})\overleftarrow{\phi}_f(a_{t,k})\upsilon_{t,k}(a_{t,t}). \tag{36}$$

G. Complexity and computation time

The complexity of the proposed algorithm depends on the number of microphones (N), number of DOA candidates (M), number of frequencies (K), number of time-frames (T), number of speakers (J) and number of the LBP iterations (denoted as N_{iter}). The algorithm is implemented in two stages. In the first, we calculate the likelihood ratio test (LRT) function $T_{t,k}(m)$ as described in Algorithm 1. Then, we run LBP inference procedure from Algorithm 2.

The calculation of $T_{t,k}(m)$ consist of:

1) Calculate the MVDR-BF: K times $N\times N$ matrix inversion and $K\cdot M$ multiplication of $N\times N$ matrix with $N\times 1$ vector, multiply the results with $N\times 1$ vector, and K scalar divisions - $\mathcal{O}(K\cdot N^3 + K\cdot M\cdot N^2 + K)$.
2) Apply the MVDR-BF on the signal: $T\cdot K\cdot M$ dot products of two $N\times 1$ vectors - $\mathcal{O}(T\cdot K\cdot M\cdot N)$.
3) Calculate the residual noise: Already calculated for the MVDR-BF.
4) Calculate the LRT: $\mathcal{O}(T\cdot K\cdot M)$ operations.

In total the order of magnitude of the required operations:

$$\mathcal{O}(K\cdot N^3 + K\cdot M\cdot N^2 + T\cdot K\cdot M\cdot N). \tag{37}$$

For each iteration in the LBP and for each time-step we have the following computations:

1) Compute the messages Ψ: $J\cdot(K+1)$ times element-wise multiplication of $M\times 1$ vectors. Multiply the results with $M \times M$ matrix - $\mathcal{O}(J \cdot K \cdot M + J \cdot M^2)$.
2) Compute $\tilde{\delta}_{t,k,j}(\cdot)$: $J \cdot (K+1)$ times element-wise multiplication of $M \times 1$ vectors - $\mathcal{O}(J \cdot K \cdot M)$.
3) Compute $\rho_{t,k}(\cdot)$: $K \cdot J$ dot product of two $M \times 1$ vectors - $\mathcal{O}(K \cdot J \cdot M)$
4) Compute $\upsilon_{t,k}(\cdot)$ for associations: Simple assignment. No computations required.
5) Compute $\upsilon_{t,k}(\cdot)$ for DOAs: $(J-1) \times K$ operations for the sum computation and then $K \cdot J$ additions of this sum to an $M \times 1$ vector - $\mathcal{O}(K \cdot J \cdot M)$.

In total the order of magnitude of the required operations:

$$\mathcal{O}(N_{\text{iter}} \cdot T \cdot (J \cdot K \cdot M + J \cdot M^2)). \tag{38}$$

The final inference algorithm consists of $M \cdot J \cdot T$ for (28) and $K \cdot J \cdot T$ for (29), which is included in the complexity of (38). The actual computation time for typical parameters, is reported in the experimental section V-D.

V. Experimental Study

The proposed algorithm was evaluated using both simulated time-varying scenes and real recordings carried out at the Bar-Ilan University (BIU) acoustic lab.

A. Parameters, evaluation methods and baseline algorithm

In our experiments we used a linear array, therefore the TDOA in (2) can be calculated in advance from the predefined grid of DOA candidates and the array constellation. Assuming that the sources are located far from the array (far-field condition), the TDOA in (2) is given by $\tau_{m,n} = \frac{1}{c_s} \cdot (r_n \cos(\vartheta_m))$, where ϑ_m is the mth candidate DOA, c_s is the sound velocity and r_n is the distance between the nth microphone and the first microphone. Note that we use the far-field assumption to analytically specify the RTF of the candidates, however, in the experiments we show that the proposed algorithm is not restricted to the reverberation-free far-field case, but can rather be applied in reverberant environments.

The parameters used in the implementation of our algorithm are as follows. The signals are sampled at 16 kHz. The STFT frame-length is set to 64 ms with 75% overlap. The grid of possible azimuth angles ranges between -90^0 and 90^0, with resolution of 2°. The noise PSD matrix was estimated in advance using a clean noise recording.

In our experiments, we observed that the optimal HMM parameter α highly depends on the signal to noise ratio (SNR) of the experiment. We therefore select the value of α in each experiment to be in the same order of magnitude of $T_{t,k}$, namely:

$$\alpha = \frac{\sum_{t,k} \left(\max_m \log T_{t,k}(m) - \min_m \log T_{t,k}(m)\right)}{T \cdot K}. \tag{39}$$

The parameter of the MRF model was set to $\beta = 0.5$, which was selected using a grid search. The LBP algorithm was stopped either after N_{max} iterations, or when the maximum change in the log messages between subsequent iterations was smaller than $\varepsilon = 10^{-3}$, where $N_{\text{max}} = 20$ or 50, for the simulation and lab experiments, respectively.

We have two options of how to define the initial DOA message $\Omega_j(m)$. The first option is to assume that the initial DOA is known, so in $\Omega_j(m)$ the known initial DOA is assigned with probability one and the other DOAs are assigned with zero probabilities. The second option is to assume that the initial DOA is unknown, to randomly generate the values of $\Omega_j(m)$, and to normalize them so they sum to one. In this option, we avoid using a uniform message since it may cause the estimates to collapse to one track.

In order to assess the performance of the algorithm, we evaluated both the tracking accuracy and the separation results. The tracking estimation error was first evaluated for each speaker using the root mean square error (RMSE) measure, namely $e_d(j) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{d}_t(j) - d_t(j))^2}$. The final score is obtained by averaging this value for all speakers. For the separation performance, we used the source to distortion ratio (SDR), source to interference ratio (SIR) and source to artifacts ratio (SAR) scores, evaluated by the BSS-Eval Toolbox [33].

As a baseline method we used the variational-based tracking algorithm proposed in [21]. In this algorithm the covariance matrix of the RTF is a priori defined, and we set it to $\mathbf{\Sigma}_a = 10\mathbf{I}$. The transition matrix was defined as in (7) with $\alpha = 0$. This algorithm requires the oracle initial DOA of the speakers for the RTF initialization. For fair comparison, we initialized both algorithms with the true DOA, and separately examined the performance of the proposed algorithm also with random initialization. For the same reason, we implemented the same separation procedure using (29) for both methods.

In addition, we report the separation results obtained using the oracle DOA in the construction of the MVDR-BF as well as the oracle separation mask, which was computed using the known separated speech signals. It is the best performance that may be achieved with the separation procedure defined in (29), and can therefore serve as an upper bound for the performance of the proposed algorithm.

B. Simulation experiment

For the simulated data, clean anechoic speech signals were drawn from the TIMIT database [34]. The speakers were randomly selected from a subset of 26 speakers. Speech utterances of the same speaker were concatenated to obtain a 5 s long speech signal. Note that the proposed method cannot perform well when long silence periods exist, since it stops tracking the speaker whenever he is inactive. However, the proposed method can tolerate small natural silence periods. Therefore, long silence segments were removed, so that all the speakers are almost simultaneously active during the entire signal.

To simulate moving sources, we used the signal generator.[1] The room dimensions were set to $6 \times 4 \times 3$ m with reverberation time $T_{60} \sim 200$ ms. The signals were captured by an eight-microphone linear array with inter-distances of

[1] www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator

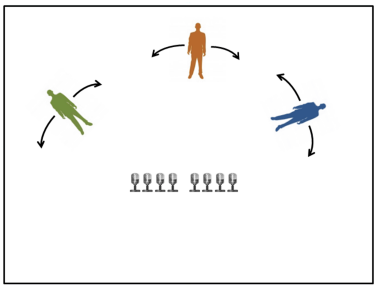


Fig. 6: An illustration of the simulation setup.

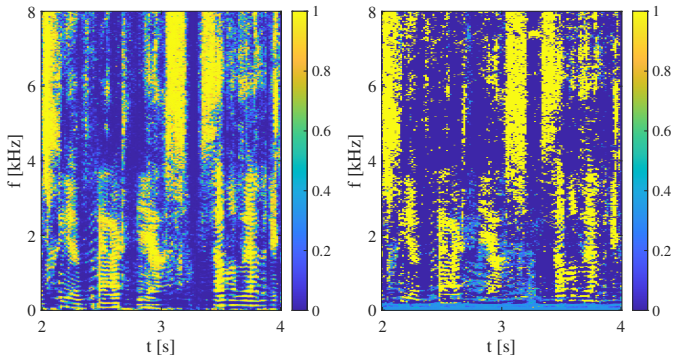


Fig. 7: Comparison of TF associations of the first speaker in the simulation experiment. The ground truth (left) and the estimated associations (right) are depicted.

$[3, 3, 3, 8, 3, 3, 3]$ cm. The array center was positioned in the center of the room, in coordinates $(3, 2, 1)$ m. The measured signals were contaminated by an additive babble diffuse noise with various SNR levels. The diffuse noise sound-field was generated using the noise generator software.[2]

Three moving speakers were simulated, with initial DOAs set to $36°$, $90°$ and $144°$, respectively. The speakers moved from their initial positions along an arc of a circle with a radius of 1 m from the array center. Their time-varying DOA has a sinusoidal form, with time period randomly selected between $1-2.5$s, and amplitude also randomly selected between $5°-8°$. The simulated setup is depicted in Fig. 6.

An example of the estimated TF associations as compared with the true associations of one of the speakers is given in Fig. 7. For the clarity of the demonstration we focus on a short segment of 2 s. We observe a good match between the true and the estimated associations, indicating that the proposed algorithm successfully recovers the TF activity of the speakers.

An example of the DOA estimation and the separation results obtained by the proposed algorithm is illustrated in Fig. 8. It can be seen that the proposed algorithm successfully recovers the trajectory of all speakers. True and estimated spectrograms of all the speakers are also depicted, demonstrating good separation performance.

The tracking and separation results were evaluated on 200 Monte-Carlo (MC) trials with different speakers and different trajectories for 3 SNR levels: 5 dB, 10 dB and 25 dB. The statistics of the obtained scores are reported in boxplots in the top row of Fig. 9 with outliers omitted for clarity. It can be seen that for the proposed algorithm the results of the uniform and the MRF models are comparable, and that they outperform the reference algorithm [21] on both tracking and separation tasks.

In addition, we examined the performance of the proposed method with respect to different room environments. Here, we fixed the SNR to 25 dB, and examined three reverberation times: 200 ms, 400 ms and 600 ms, and two source distances with respect to the center of the array: 1 m and 1.5 m. The tracking and separation results were averaged over 100 MC trials with different speakers and different trajectories. The results of this experiment are reported in Fig. 10. We observe a decrease in the separation scores and an increase in the DOA RMSE for higher reverberation levels or larger source-microphone distance. The difference in the performance between 1 m and 1.5 m distance becomes more significant for higher reverberation levels, apparently due to the fact that in high reverberation the direct-to-reverberant power ratio becomes much lower as the source-microphone distance increases.

C. Laboratory experiment

In addition to the simulated experiment, we evaluated the proposed algorithm using real recordings carried out at the BIU acoustic lab. We first defined two limited arcs on a circle with radius of $\sim$ 2 m: the first arc between $20°$-$75°$ and the other between $120°$-$165°$. Seven speakers participated in our experiment, five males and two females. Each speaker moved back and forth while speaking with a natural random trajectory on each of the defined arcs. The length of each recording was approximately 30s. The signals were captured by an eight-microphone linear array with inter-distances of $[3, 3, 3, 6, 3, 3, 3]$ cm. The array was located in the center of the designated circle, in a distance of approximately 1.5 meters from one of the walls. A photograph of the room configuration is given in Fig. 11. The reverberation time was set to $T_{60} \sim 450$ms by adjusting the controllable room panels. A diffuse babble noise was also separately recorded by the same array using 4 loudspeakers facing the room corners. Finally, after discarding few utterances due to technical problems in the recordings, we generated 29 combinations of different pairs of speakers with noise added with different SNR levels.

In order to evaluate the results we need both the clean speech for the separation evaluation, and the ground-truth trajectory for the tracking evaluation. For the separation evaluation we used the separately recorded speech signals in the first microphone as a reference. For the ground-truth DOA of the speakers we used Marvelmind indoor navigation system.[3] This system consists of a single mobile device and four stationary devices. The coordinates of the mobile device are reported w.r.t. the stationary devices with reported measurement error of ± 2 cm. In practice, we observed that occasionally this device introduces small glitches, apparently due to noise or

[2] www.audiolabs-erlangen.de/fau/professor/habets/software/noise-generators

[3] https://marvelmind.com/product/starter-set-hw-v4-9/

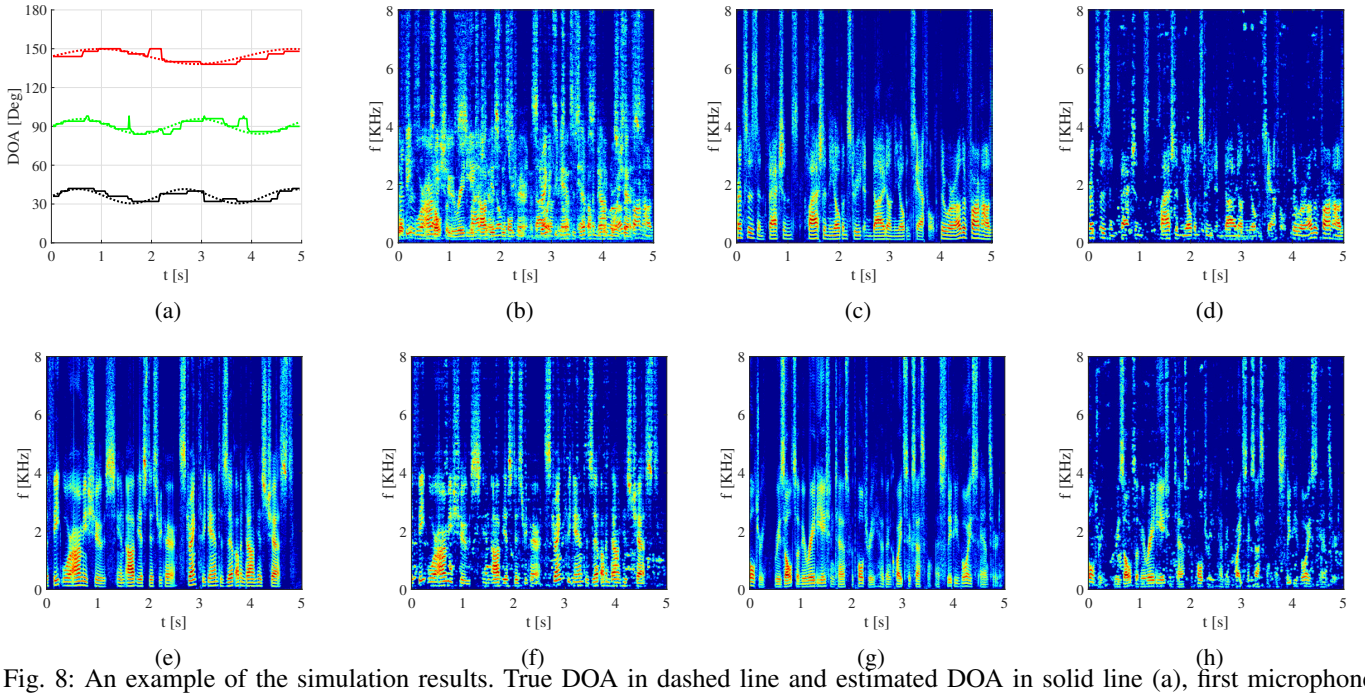


Fig. 8: An example of the simulation results. True DOA in dashed line and estimated DOA in solid line (a), first microphone mixed signal spectrogram (b), clean and estimated spectrogram of the first speaker (c+d), the second speaker (e+f) and the third speaker (g+h).

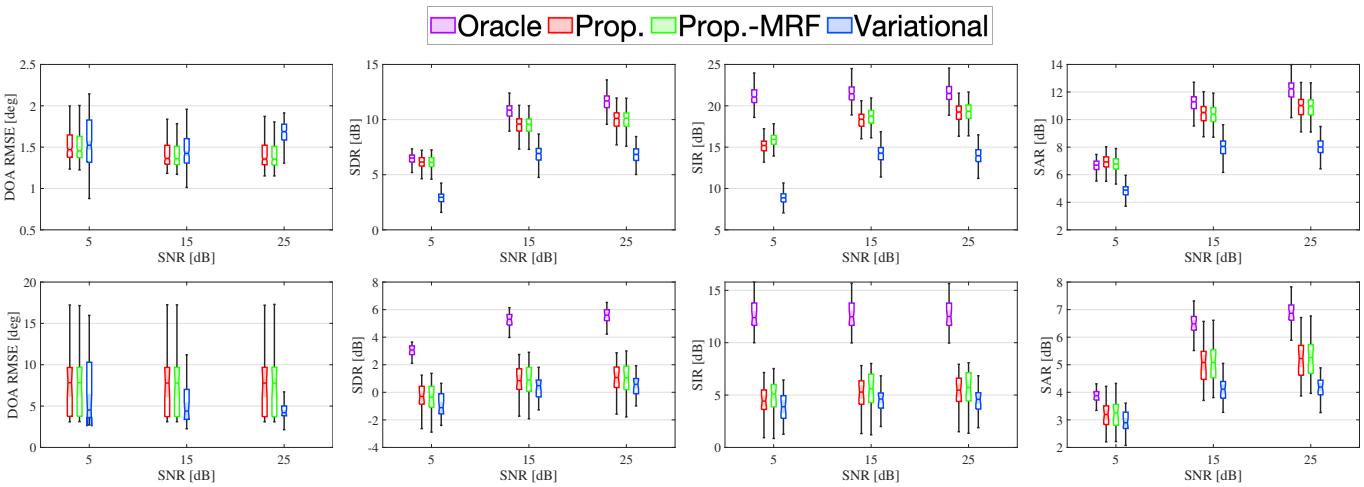


Fig. 9: Simulation and lab experiment: separation and tracking performance measures for various SNRs for simulation (top) and lab experiments (bottom). The results are reported for the reference variational method [21] and for the two versions of the proposed method, with the simple uniform prior of the associations (Prop.) and with the more complex MRF-model as described in Section IV-F (Prop.-MRF). In addition, we report the separation results obtained using the oracle DOA in the construction of the MVDR-BF as well as the oracle separation mask, which was computed using the known separated speech signals.

measurement instability. In the beginning of our experiment, we measured the microphone locations, and then each participant held the mobile device during his recording session. The ground-truth DOA is computed as the angle between the microphone array and the line connecting the center of the array and the speaker location.

An example of the DOA estimation obtained by the proposed method with random DOA initialization is shown in Fig. 12 (a). The estimated trajectory is close to the ground truth trajectory as measured by the indoor navigation system. Note that although the estimated DOAs of one of the speakers deviates from the true trajectory around $t = 25$s, the algorithm successfully traces back the true trajectory after few seconds. Figure 12 (b) shows an example of the DOA estimation obtained with random DOA initialization for a case with three speakers that two of them have close trajectories. It can be seen that the proposed algorithm successfully tracks the three speakers for almost the entire signal duration. The estimated trajectories deviate from the ground truth at the end of the signal when two speakers get closer to each other.

The statistics of the 29 different 2-speakers scenarios are reported in boxplots in the bottom row of Fig. 9. While

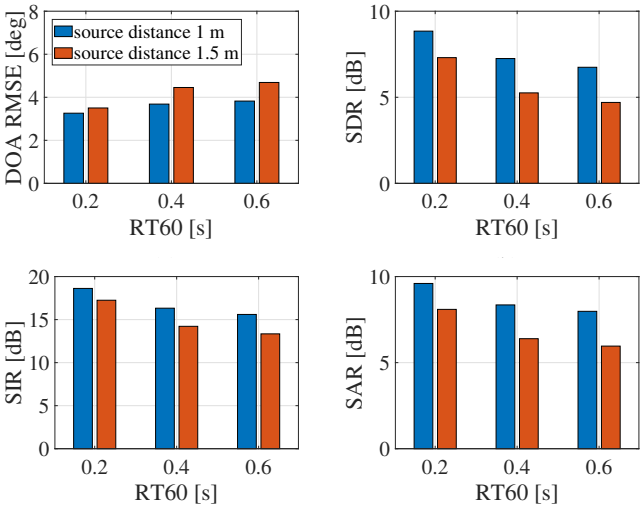


Fig. 10: Separation and tracking performance measures for three reverberation times: 200 ms, 400 ms and 600 ms, and two source distances with respect to the center of the array: 1 m and 1.5 m, averaged over 100 MC trials, with SNR= 25 dB.

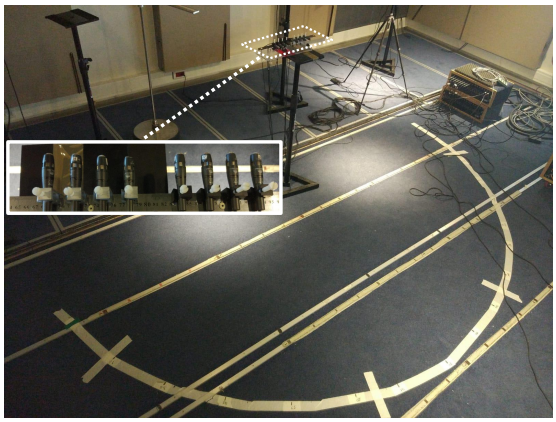


Fig. 11: A photo of the experimental setup at the BIU lab.

the proposed algorithm outperforms the reference algorithm in the separation task for all SNR values, in the tracking task it obtains higher errors. Comparing the uniform and the MRF models in the proposed algorithm, we observe a slight advantage to the latter in terms of the SIR measure as reflected from the median and the 75 percentile. This advantage is more pronounced in the 5 dB SNR case. Note that the DOA-RMSE might be biased due to measurement errors in the ground-truth DOA, as mentioned above. Note also that the ground-truth separated speech signals, taken as the measurements of the first microphone, cannot serve as a perfect reference as well, which may explain the relatively low separation scores. For subjective evaluation, the reader is referred to our website.[4]

We also examined the sensitivity of the proposed algorithm to the DOA initialization. A comparison of the DOA RMSE obtained by the proposed algorithm with either ground truth or random initialization is given in Fig. 13. It is observed that the error is increased by approximately 1 degree for most of the readings. This small increase in the error indicates that

[4]http://www.eng.biu.ac.il/gannot/speech-enhancement/

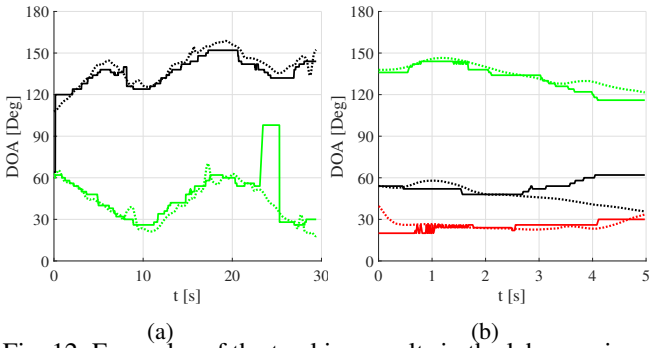


Fig. 12: Examples of the tracking results in the lab experiment: two distant speakers in a full 30 s recording (left) and three speakers, two of them close to each other, in a segment of 5 s recording (right). Dashed and solid lines correspond to ground truth (obtained by the indoor navigation system) and estimated trajectories, respectively. The initial DOAs were set randomly. In the three speakers case the estimated trajectories deviate from the ground truth at the end of the signal when two speakers get closer to each other.

the proposed algorithm can track the speakers without prior knowledge on their initial position.

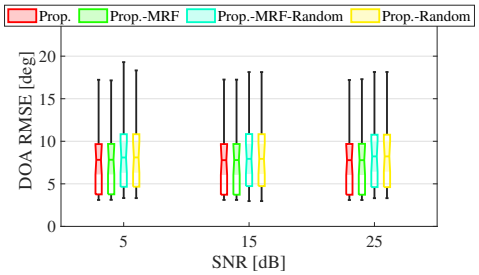


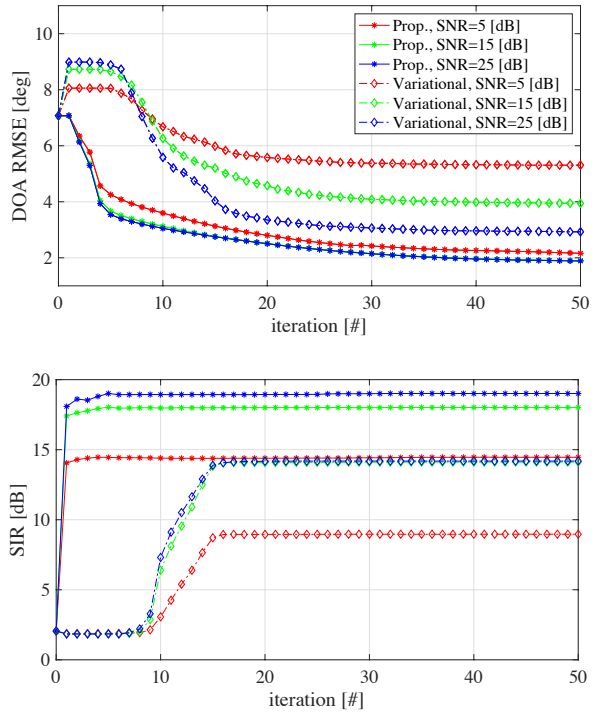
Fig. 13: Comparing DOA estimation performance with random and oracle initialization.

We also examined the dependency of separation quality measures on the gender of the speakers. We compared mixtures of same gender speakers, i.e. male and female, with mixtures of male and female speakers. Analyzing the results, did not show any significant differences. This conclusion might need further investigations, as the number of examples is small.

D. Computation time

In this section, we report the average computation time of each iteration and the performance of the proposed algorithm and the baseline algorithm as a function of the number of iterations for the simulation experiment. The computation time was calculated using 2.3 GHz Intel Core i9 single CPU, with 16 GB 2400 MHz DDR4 memory. The algorithm was implemented using Matlab©, without using the parallel computing utility. In our experiments, the recording length was 30 seconds. The parameters were: $N = 8$, $K = 513$, $J = 3$, $M = 91$, $T = 309$ and $N_{\max} = 50$. The average computation time was roughly 3.8 s per iteration per second of input signal, compared to an average of 6.6 s for the reference algorithm. Note also that the total computation time linearly

depends on the number of iterations. In Fig. 14, we report the tracking and separation performance measures as a function of the number of iterations. It is demonstrated that in terms of the separation performance, the proposed algorithm converges within 5 iterations, compared to 15 iterations required by the reference algorithm, and also obtains better SIR scores after convergence. For the DOA estimation, the proposed algorithm converges after 35 iterations to a lower RMSE compared to that achieved by the reference algorithm, which converges after 20 iterations. Note also that the DOA RMSE obtained by the proposed algorithm decreases to $3^\circ - 4^\circ$, already within 5 iterations. Therefore, when the available computation time is limited, we can run only 5 iterations of the proposed algorithm to obtain maximal separation performance and low DOA RMSE of less than 5°.



(b)

Fig. 14: Tracking and Separation performance of a particular scenario as function of the number of iterations.

VI. Conclusions

In this paper, we introduced an algorithm for simultaneous tracking and separation of multiple speakers using multiple microphone observations, utilizing the factor graph model. The difficulty of this problem rises from the cross dependency between the tasks, since the estimation of the DOA of each speaker relies on the set of TF bins associated with each speaker, and the associations of the TF bins to the speakers can be inferred when the DOA of each speaker is known. Many state-of-the-art separation and tracking algorithms assume some prior knowledge on either of the tasks, or comprise two successive algorithms for solving each task separately. In this paper, we simultaneously solved both tasks by defining a factor graph model and deriving a novel inference algorithm, based on the LBP scheme. Although in this paper we focused on the practical problem of speaker tracking and separation, the proposed inference algorithm solves a general graphical model, which consists of parallel Markov chains and multiple observations, where the associations of the observations to the Markov chains are unknown. This type of model can be applied in various other problems.

The proposed algorithm was evaluated using both simulated data and real recordings measured in our lab in natural conditions, demonstrating the capabilities of the proposed algorithm in various conditions, including different SNR and reverberation levels, source to microphone distances and DOA velocities. The proposed algorithm outperformed a reference algorithm, based on variational inference, in almost all cases and for all evaluation metrics. Moreover, we have shown that these results can be achieved with lower computation time and without relying on any prior knowledge on the initial speakers' DOAs.

We conclude that the proposed algorithm presents a new methodology for solving the challenging task of simultaneous separation and tracking, which achieves efficient high-quality performance, and can also be adopted in other domains. Future research should be focused on extending the algorithm to highly reverberant environments, unknown noise characteristics, and scenarios with unknown microphone locations or a moving array.

Appendix

A. Observation Likelihood

In this section, we simplify the conditional probability of the observations given the hidden variables (11). We first estimate the speech PSD using the maximum likelihood estimator (MLE), and then we factorize the probability and substitute the estimated PSD to obtain the final expression (13). Here we denote $b_{t,k} \equiv d_t(a_{t,k}) \in [1 \dots M]$ for simplicity.

1) Speech PSD estimation: We estimate the unknown parameters $\boldsymbol{\phi}_{\boldsymbol{s}} = \text{vec}_{t,k,m}\{\phi_{s,t,k}(m)\}$ using ML. To this end, we write the marginal distribution of the observations, by marginalizing out the hidden variables:

$$P(\mathbf{z};\boldsymbol{\phi}_{\boldsymbol{s}}) = \sum_{\mathbf{b}} \prod_{t,k} P(\mathbf{z}_{t,k}|\mathbf{b}_{t,k}) P(\mathbf{b}) \tag{40}$$

where $\mathbf{b} = \text{vec}_{t,k}\{b_{t,k}\}$ and $P(\mathbf{b})$ is the prior probability of $\mathbf{b}$ which depends on the priors $P(\mathbf{a})$ and $P(\mathbf{d})$.

The MLE for $\phi_{s,\tilde{t},\tilde{k}}(m)$ is obtained by maximizing (40) w.r.t. $\phi_{s,\tilde{t},\tilde{k}}(m)$. We first rearrange the marginal distribution by excluding the $(\tilde{t},\tilde{k})$th observation from the product and summation:

$$P(\mathbf{z};\boldsymbol{\phi}_{\boldsymbol{s}}) = \sum_{\mathbf{b}_{\tilde{t},\tilde{k}}} \left[P(\mathbf{z}_{\tilde{t},\tilde{k}}|\mathbf{b}_{\tilde{t},\tilde{k}}) \sum_{\substack{\mathbf{b}\backslash \\ \mathbf{b}_{\tilde{t},\tilde{k}}}} \prod_{\substack{t,k\backslash \\ (\tilde{t},\tilde{k})}} P(\mathbf{z}_{t,k}|\mathbf{b}_{t,k}) P(\mathbf{b}) \right]. \tag{41}$$

Substituting (11) into (41), and explicitly writing the first summation over all candidates, we have:

$$P(\mathbf{z};\boldsymbol{\phi}_{\boldsymbol{s}}) = \sum_{w=1}^{M} \mathcal{N}(\mathbf{z}_{\tilde{t},\tilde{k}}, \mathbf{0}, \Phi_{\mathbf{z},\tilde{t},\tilde{k}}(w)) \cdot C \tag{42}$$

where $C \equiv \sum_{\mathbf{b}\setminus\mathbf{b}_{\tilde{t},\tilde{k}}} \prod_{t,k\setminus(\tilde{t},\tilde{k})} P(\mathbf{z}_{t,k}|\mathbf{b}_{t,k})P(\mathbf{b})$ denotes a positive term, independent of the parameter of interest $\phi_{s,\tilde{t},\tilde{k}}(m)$. Then, taking the derivative w.r.t $\phi_{s,\tilde{t},\tilde{k}}(m)$ we get:

$$\frac{\partial P(\mathbf{z};\boldsymbol{\phi_s})}{\partial \phi_{s,\tilde{t},\tilde{k}}(m)} = \frac{\partial \mathcal{N}(\mathbf{z}_{t,k},\mathbf{0},\Phi_{\mathbf{z},t,k}(m))}{\partial \phi_{s,\tilde{t},\tilde{k}}(m)} \cdot C. \tag{43}$$

By setting this derivative to zero we get the MLE for $\phi_{s,t,k}(m)$ [35]:

$$\hat{\phi}_{s,t,k}(m) = |\hat{s}_{\mathbf{w},t,k}(m)|^2 - \phi_{v,k}(m). \tag{44}$$

where $\hat{s}_{\mathbf{w},t,k}(m)$ is the MVDR-BF output defined in (14), and $\phi_{v,k}(m)$ is the PSD of the residual noise at the output of the MVDR-BF, and is given by:

$$\phi_{v,k}(m) \equiv \frac{1}{\mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)}. \tag{45}$$

2) Likelihood factorization: We factorize the likelihood of the observation to obtain a simpler expression. We first define the a posteriori SNR of the signal impinging the array from the mth candidate position as:

$$\eta_{t,k}(m) = \frac{|\hat{s}_{\mathbf{w},t,k}(m)|^2}{\phi_{v,k}(m)} \tag{46}$$

and the a priori SNR as:

$$\zeta_{t,k}(m;\phi_{s,t,k}(m)) = \frac{\phi_{s,t,k}(m)}{\phi_{v,k}(m)}. \tag{47}$$

According to (11), the conditional distribution of a single observation given the hidden data is given by:

$$\mathcal{N}(\mathbf{z}_{t,k},\mathbf{0},\Phi_{\mathbf{z},t,k}(m)) = \frac{1}{\pi^N \det(\Phi_{\mathbf{z},t,k}(m))} \exp(-\mathbf{z}^H \left(\Phi_{\mathbf{z},t,k}(m)\right)^{-1}\mathbf{z}). \tag{48}$$

Using the definition of $\Phi_{\mathbf{z},t,k}(m)$ (12) and Sylvester's determinant theorem, the determinant can be written as:

$$\begin{aligned}&\det(\Phi_{\mathbf{z},t,k}(m))\\ &= \det(\Phi_{\mathbf{v},k}) \cdot \det(1+\phi_{s,t,k}(m)\mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m))\\ &= \det(\Phi_{\mathbf{v},k}) \cdot (1+\zeta_{t,k}(m;\phi_{s,t,k}(m))).\end{aligned}$$

In addition, using the Woodbury identity, the inversion of $\Phi_{\mathbf{z},t,k}(m)$ can be written as:

$$\Phi_{\mathbf{z},t,k}(m)^{-1} = \Phi_{\mathbf{v},k}^{-1} - \frac{\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)\mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}}{\phi_{s,t,k}(m)^{-1} + \mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)}.$$

By substituting these relations into the p.d.f., we can factorize it as following:

$$\mathcal{N}(\mathbf{z}_{t,k},\mathbf{0},\Phi_{\mathbf{z},t,k}(b_{t,k})) = T_{t,k}(b_{t,k};\phi_{s,t,k}(b_{t,k})) \cdot G_{t,k} \tag{49}$$

where $G_{t,k}$ aggregates all terms which do not depend on m:

$$G_{t,k} = \frac{1}{\pi^N \det(\Phi_{\mathbf{v},k})} \exp\left(-\mathbf{z}^H\Phi_{\mathbf{v},k}^{-1}\mathbf{z}\right) \equiv \mathcal{N}(\mathbf{z}_{t,k},\mathbf{0},\Phi_{\mathbf{v},k}) \tag{50}$$

and $T_{t,k}(m;\phi_{s,t,k}(m))$ aggregates the other terms:

$$T_{t,k}(m;\phi_{s,t,k}(b_{t,k})) = \frac{1}{1+\zeta_{t,k}(m;\phi_{s,t,k}(m))} \cdot \exp\left(\frac{\mathbf{z}^H\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)\mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}\mathbf{z}}{\phi_{s,t,k}(m)^{-1} + \mathbf{g}_k^H(m)\Phi_{\mathbf{v},k}^{-1}\mathbf{g}_k(m)}\right).$$

Using (14),(45), (46) and (47) we can write $T_{t,k}(m;\phi_{s,t,k}(m))$ in a simple way:

$$T_{t,k}(m;\phi_{s,t,k}(m)) = \frac{1}{1+\zeta_{t,k}(m;\phi_{s,t,k}(m))} \exp\left(\frac{\zeta_{t,k}(m;\phi_{s,t,k}(m))\eta_{t,k}(m)}{1+\zeta_{t,k}(m;\phi_{s,t,k}(m))}\right). \tag{51}$$

Note that $T_{t,k}(m;\phi_{s,t,k}(m))$ is the LRT, as presented in [36, Eq. (14)]. The LRT tests whether $\mathbf{z}_{t,k}$ is either associated with a speaker located in the mth candidate DOA or with noise only. Using the estimator of $\phi_{s,t,k}(m)$ we can further simplify $T_{t,k}(m;\phi_{s,t,k}(m))$. Dividing (44) by $\phi_{v,k}(m)$ and using the definitions in (46) and (47), we obtain: $\zeta_{t,k}(m;\hat{\phi}_{s,t,k}(m)) = \eta_{t,k}(m) - 1$. Finally, by substituting this relation into (51), we obtain:

$$T_{t,k}(m;\hat{\phi}_{s,t,k}(m)) = \frac{1}{\eta_{t,k}(m)} \exp\left(\eta_{t,k}(m) - 1\right). \tag{52}$$

B. Factor Graphs

In this section, we briefly review the definition of factor graphs and their inference methods based on [25], [37].

1) Definition: Let $\{x_1, x_2, ..., x_Q\}$ be a set of Q discrete-valued random variables. We consider the joint probability mass function $P(\mathbf{x}) = P(x_1, x_2, ..., x_Q)$, which is assumed to be factored into a product of functions:

$$P(\mathbf{x}) = \frac{1}{C}\prod_{u\in\mathcal{U}} f_u(\mathbf{x}_u) \tag{53}$$

where u is an index that labels the functions from a set $\mathcal{U}$, where each function $f_u(\mathbf{x}_u)$ has arguments $\mathbf{x}_u \subset \{x_1, x_2, ..., x_Q\}$. We assume that the functions $f_u(\mathbf{x}_u)$ are non-negative and finite, so that $P(\mathbf{x})$ is a well-defined probability distribution. Here, C is a normalization constant.

A factor graph is a bipartite graph that expresses the factorization structure in (53). A factor graph has a variable node (which we draw as a circle) for each variable x_i, and a factor node (which we draw as a square) for each function f_u, with an edge connecting variable node x_i to factor node u if and only if $x_i \in \mathbf{x}_u$.

2) Inference: For a given graph with given factors, one may be interested in two different goals. The first is to find the marginals of each variable, i.e. $P(x_i)\ \forall i$, and the other is to find the most probable state, i.e. $\operatorname{argmax}_{\mathbf{x}} P(\mathbf{x})$. An exact inference for factor graphs is obtained using the belief propagation (BP) algorithm. When implemented for computing the marginal p.d.f., the BP algorithm is also known as the *sum-product* algorithm, and when implemented for finding the most probable state, it is called the *max-product* algorithm. In the sum-product algorithm messages are sent

from the factors to the variables and vice-verse, using the following equations:

$$n_{i\to u}(x_i) = \prod_{c\in\mathcal{G}\{x_i\}/u} m_{c\to i}(x_i) \tag{54a}$$

$$m_{u\to i}(x_i) = \sum_{\mathbf{x}_u/x_i} f_u(\mathbf{x}_u) \prod_{j\in\mathcal{G}\{u\}/x_i} n_{j\to u}(x_j) \tag{54b}$$

where $n_{i\to u}(x_i)$ is the message from the ith variable to the uth factor, $m_{u\to i}(x_i)$ is the opposite direction message, $\mathcal{G}\{x_i\}$ is the set of neighbouring factors of x_i and $\mathcal{G}\{u\}$ is the set of neighbouring variables of u. We can then obtain the marginal probability of a particular variable x_i using:

$$P(x_i) \propto \prod_{u\in\mathcal{G}\{x_i\}} m_{u\to i}(x_i) \tag{55}$$

where the sign $\propto$ means that one should normalize this expression to obtain the final distribution. In the *max-product* algorithm summations are replaced by the `max` operator. The max-product algorithm is out of the scope of this article.

The sum-product algorithm is proved to converge to the true marginals in tree-structured graphs [38]. However, when the graph contains loops this algorithm is not proved to coverage to the true marginal. The loopy belief propagation (LBP) [31] is an extension of the BP algorithm for loopy graphs, in which messages are updated repeatedly, in an arbitrary order, until a termination condition is met. In practice, it has been observed that this algorithm often provides good estimates of the marginals.

C. Derivation of the Messages from the Observation Factors

In this section, we derive the messages from the observation factors to its neighboring variables. For general derivation, we assume here that each variable sends a message to the observations. We denote by $\delta_{t,k,j}(:)$ and $q_{t,k}(:)$ the messages from the DOA and the association variables, respectively.

1) The message from the observations to the association variables: Using (54b) the messages to $a_{t,k}$ are given by:

$$\upsilon_{t,k}(a_{t,k}) = \sum_{d_t(1)}\sum_{d_t(2)}\dots\sum_{d_t(J)} \Upsilon_{t,k}(a_{t,k}, d_t(1)\dots d_t(J)) \prod_{i=1}^{J}\delta_{t,k,i}(d_t(i)).$$

Substituting the definition of $\Upsilon_{t,k}$ (18) we obtain:

$$\upsilon_{t,k}(a_{t,k}) = \sum_{d_t(1)}\sum_{d_t(2)}\dots\sum_{d_t(J)} T_{t,k}(d_t(a_{t,k})) \prod_{i=1}^{J}\delta_{t,k,i}(d_t(i)). \tag{56}$$

Note that the expression $T_{t,k}(d_t(a_{t,k}))$ is constant for all summations except for the sum over $d_t(a_{t,k})$, hence we rearragne the summations as follows:

$$\upsilon_{t,k}(a_{t,k}) = \sum_{d_t(a_{t,k})} T_{t,k}(d_t(a_{t,k})) \sum_{d_t(:)/d_t(a_{t,k})} \prod_{i=1}^{J}\delta_{t,k,i}(d_t(i)).$$

Since each message $\delta_{t,k,i}(d_t(i))$ is influenced by only one summation, we can switch the sum and product operations:

$$\sum_{d_t(a_{t,k})} T_{t,k}(d_t(a_{t,k}))\cdot\delta_{t,k,a_{t,k}}(d_t(a_{t,k})) \prod_{i\neq a_{t,k}}\sum_{d_t(i)}\delta_{t,k,i}(d_t(i)).$$

In order to further simplify this expression, we multiply and divide it by the term $\sum_{d_t(a_{t,k})}\delta_{t,k,a_{t,k}}(d_t(a_{t,k}))$ to obtain

$$\frac{\sum_{d_t(a_{t,k})} T_{t,k}(d_t(a_{t,k}))\cdot\delta_{t,k,a_{t,k}}(d_t(a_{t,k}))}{\sum_{d_t(a_{t,k})}\delta_{t,k,a_{t,k}}(d_t(a_{t,k}))} \underbrace{\prod_{i}\sum_{d_t(i)}\delta_{t,k,i}(d_t(i))}_{\text{Const}}. \tag{57}$$

Since the messages are not normalized anyway, we can ignore the constant term, and we finally obtain:

$$\upsilon_{t,k}(a_{t,k}) \propto \frac{\sum_m T_{t,k}(m)\cdot\delta_{t,k,a_{t,k}}(m)}{\sum_m \delta_{t,k,a_{t,k}}(m)} \equiv \rho_{t,k}(a_{t,k}) \tag{58}$$

2) The messages from the observations to the DOA variables: The incoming messages are coming from $d_t(1),\dots,d_t(j-1),d_t(j+1),\dots,d_t(J)$ and $a_{t,k}$, therefore:

$$\upsilon_{t,k}(d_t(j)) = \sum_{a_{t,k}}\sum_{d_t(:)/d_t(j)} T_{t,k}(d_t(a_{t,k}))q_{t,k}(a_{t,k})\prod_{i\neq j}\delta_{t,k,i}(d_t(i))$$

where $q_{t,k}(a_{t,k})$ is uniform for the uniform distribution model (8) or defined by (34) for the MRF model (30). We split the first summation over all possible values of $a_{t,k}\in[1\dots J]$ to a sum over j and summations over all other values:

$$= \underbrace{\sum_{d_t(:)/d_t(j)} T_{t,k}(d_t(j))q_{t,k}(j)\prod_{i\neq j}\delta_{t,k,i}(d_t(i))}_{(*)}$$
$$+ \sum_{a_{t,k}\neq j} q_{t,k}(a_{t,k}) \underbrace{\sum_{d_t(:)/d_t(j)} T_{t,k}(d_t(a_{t,k}))\prod_{i\neq j}\delta_{t,k,i}(d_t(i))}_{(**)}.$$

This expression consists of two terms. In $(*)$ the term $T_{t,k}(d_t(j))q_{t,k}(j)$ depends on $d_t(j)$, hence we take it out of the summation and switch the order of the sum and product operations to obtain:

$$(*) = T_{t,k}(d_t(j))q_{t,k}(j)\prod_{i\neq j}\sum_{d_t(i)}\delta_{t,k,i}(d_t(i))$$

The term $(**)$ is same as (56) and similarly to (57) it can simplified to:

$$(**) = \rho_{t,k}(a_{t,k})\prod_{i\neq j}\sum_{d_t(i)}\delta_{t,k,i}(d_t(i)).$$

The overall message is now given by:

$$= T_{t,k}(d_t(j))q_{t,k}(j)\underbrace{\prod_{i\neq j}\sum_{d_t(i)}\delta_{t,k,i}(d_t(i))}_{\text{const}}$$
$$+ \sum_{a_{t,k}\neq j}\rho_{t,k}(a_{t,k})q_{t,k}(a_{t,k})\underbrace{\prod_{i\neq j}\sum_{d_t(i)}\delta_{t,k,i}(d_t(i))}_{\text{const}}.$$

Dividing the message by the constant $q_{t,k}(j)\prod_{i\neq j}\sum_{d_t(i)}\delta_{t,k,i}(d_t(i))$, we finally obtain:

$$m_{t,k}(d_t(j)) \propto T_{t,k}(d_t(j)) + \frac{\sum_{a_{t,k}\neq j} q_{t,k}(a_{t,k})\rho_{t,k}(a_{t,k})}{q_{t,k}(j)}. \tag{59}$$

REFERENCES

[1] J. DiBiase, H. Silverman, and M. Brandstein, *Microphone arrays : signal processing techniques and applications*. Springer Verlag, 2001, ch. Robust Localization in Reverberant Rooms, pp. 157–180.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[3] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, May 2000. [Online]. Available: http://dx.doi.org/10.1016/S0893-6080(00)00026-5

[4] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Location feature integration for clustering-based speech separation in distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 354–367, 2013.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sign. Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, 2017.

[7] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[8] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[9] S. Makino, *Audio Source Separation*. Springer, 2018.

[10] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[11] M. I. Mandel and N. Roman, "Enforcing consistency in spectral masks using markov random fields," in *European Signal Processing Conference (EUSIPCO)*, 2015, pp. 2028–2032.

[12] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 2, pp. 392–402, 2014.

[13] M. D. Titterington, "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 257–267, 1984.

[14] O. Cappé and E. Moulines, "On-line expectation–maximization algorithm for latent data models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[15] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 2233–2243, 2014.

[16] O. Schwartz, Y. Dorfan, E. A. Habets, and S. Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[17] Y. Dorfan, O. Schwartz, B. Schwartz, E. A. Habets, and S. Gannot, "Multiple DOA estimation and blind source separation using estimation-maximization," in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, 2016.

[18] O. Schwartz, Y. Dorfan, M. Taseska, E. A. Habets, and S. Gannot, "DOA estimation in noisy environment with unknown noise power using the EM algorithm," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 86–90.

[19] K. Weisberg, S. Gannot, and O. Schwartz, "An online multiple-speaker DOA tracking using the Cappé-Moulines recursive expectation-maximization algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 656–660.

[20] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 4, pp. 728–739, 2008.

[21] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3191–3195.

[22] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational em algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 8, pp. 1408–1423, 2016.

[23] M. Swartling, N. Grbic, and I. Claesson, "Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, vol. 4, 2006, pp. 833–836.

[24] A. Brendel, B. Laufer-Goldshtein, S. Gannot, R. Talmon, and W. Kellermann, "Localization of an unknown number of speakers in adverse acoustic conditions using reliability information and diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7898–7902.

[25] F. R. Kschischang, B. J. Frey, H.-A. Loeliger, *et al.*, "Factor graphs and the sum-product algorithm," *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, 2001.

[26] G. Colavolpe and G. Germi, "On the application of factor graphs and the sum-product algorithm to ISI channels," *IEEE Transactions on Communications*, vol. 53, no. 5, pp. 818–825, 2005.

[27] D. Kipnis and R. Diamant, "A factor-graph clustering approach for detection of underwater acoustic signals," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 5, pp. 702–706, 2018.

[28] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Georgia Institute of Technology, Tech. Rep., 2012.

[29] A. Cunningham, M. Paluri, and F. Dellaert, "DDF-SAM: Fully distributed SLAM using constrained factor graphs," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 3025–3030.

[30] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 467–475.

[31] B. J. Frey and D. J. MacKay, "A revolution: Belief propagation in graphs with cycles," in *Advances in neural information processing systems*, 1998, pp. 479–485.

[32] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings*, 2002, pp. 209–213.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[35] H. Ye and D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Sign. Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[36] T. Yu and J. H. Hansen, "A speech presence microphone array beamformer using model based speech presence probability estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 213–216.

[37] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on information theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[38] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.