SPEAKER LOCALIZATION AND SEPARATION USING INCREMENTAL DISTRIBUTED EXPECTATION-MAXIMIZATION

Yuval Dorfan, Dani Cherkassky and Sharon Gannot

Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel
dorfany@gmail.com; dani.cherkassky@gmail.com; sharon.gannot@biu.ac.il

ABSTRACT

A network of microphone pairs is utilized for the joint task of localizing and separating multiple concurrent speakers. The recently presented incremental distributed expectation-maximization (IDEM) is addressing the first task, namely detection and localization. Here we extend this algorithm to address the second task, namely blindly separating the speech sources. We show that the proposed algorithm, denoted distributed algorithm for localization and separation (DALAS), is capable of separating speakers in reverberant enclosure without a priori information on their number and locations. In the first stage of the proposed algorithm, the IDEM algorithm is applied for blindly detecting the active sources and to estimate their locations. In the second stage, the location estimates are utilized for selecting the most useful node of microphones for the subsequent separation stage. Separation is finally obtained by utilizing the hidden variables of the IDEM algorithm to construct masks for each source in the relevant node.

Index Terms— Wireless acoustic sensor network; blind source separation; incremental estimate-maximize

1. INTRODUCTION

Blind source separation (BSS) is an unsupervised technique for recovering the underling sources from a set of their mixtures. In acoustic applications [1], as the *cocktail party* problem [2,3], the sources (speakers) are typically mixed in a convolutive manner, and the respective source separation task is referred to as convolutive BSS. The convolutive BSS problem is much more challenging compared with the instantaneous BSS, since the separation filters might have thousands of coefficients in a typical room environment. The instantaneous BSS framework can still be used for convolutive mixture separation in the frequency domain [4]. However, once the frequency domain approach is used, the inherent scaling and permutation ambiguity of BSS methods [5] will be independently encountered in each frequency band. This ambiguity has to be resolved for obtaining meaningful separation.

Additional difficulty may arise in audio applications in the under-determined case (i.e. the number of sources is greater than the number of sensors). In this case, linear separation scheme will fail to separate the sources. However, this scenario is still trackable, if the sources have a sparse representation [6]. Audio sources (such as speech and music) are often attributed by a sparse representation in the short-time Fourier transform (STFT) domain [7]. The sparseness of the speech in the time-frequency (T-F) domain attracted significant attention in the speech processing community in general, and in source separation community in particular [8–10]. Source separation in the T-F domain is achieved by clustering the T-F bins into groups, each group associated with one source signal. The clustering usually relies on features such as time difference of arrival (TDOA) [7, 11, 12].

We establish our BSS scheme on a 2-D (or a 3-D) localization procedure. The distributed localization algorithm provides reliable information regarding the sources in the room, even when the node signals are contaminated by moderate level of reverberation and interference. A modified version of the incremental distributed expectation-maximization (IDEM) algorithm [13] efficiently converges to the *global* maximum likelihood (ML) of the localization problem. The IDEM belongs to the distributed expectation maximization (DEM) group of algorithms, term coined by Nowak [14]. The IDEM is implemented over a *directed-ring*. It combines a new definition of hidden variables incorporated into the incremental EM (IEM) framework [15–17]. In this paper[1] we propose to utilize the reliable localization estimates for separating the sources, by first selecting the best receiving node per source and then applying the corresponding spectral masks deduced from the hidden variables of the IDEM algorithm. Using the source location estimates, masked signals at the same node are time-aligned and averaged to obtain the final separated signals.

The reminder of this paper is organized as follows. A brief description of the IDEM is given in Section 2. In Section 3 we present the DALAS for speaker separation in reverberant environments. Section 4 is dedicated to simulation results. Conclusions are drawn in Section 5.

[1]We would like to thank Dr. Gershon Hazan for his dedicated assistance and support.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. THE IDEM ALGORITHM

In [13] the first version of IDEM has been presented for localization. This section presents a modified version, which refers to a new set of hidden variables. First, a reminder of the localization problem is given. Then, a simplified definition of the hidden variables is presented. These hidden variables, which are actually a byproduct of the localization procedure, will be used in the derivation of the BSS algorithm. The algorithm is implemented over a directed-ring consisting of M nodes. Each node of the network consists of a microphone pair, a CPU and a communication unit. The global parameters are transmitted around the ring from one node to another. Each node updates its local hidden and transmits the maximization result to the next node.

2.1. maximum likelihood for localization - A reminder

The localization procedure starts with a pair-wise relative phase ratio (PRP) extraction:

$$\phi_m(t,k) \triangleq \frac{z_m^2(t,k)|z_m^1(t,k)|}{z_m^1(t,k)|z_m^2(t,k)|}, \forall m = 1,\ldots,M, \quad (1)$$

where $z_m^r(t,k)$ is the STFT of the rth microphone signal ($r = 1,2$) at the mth node. The time and frequency indices are $t = 1,\ldots,T$ and $k = 0,\ldots,K-1$, respectively.

These PRPs are induced by the TDOA $\tau_m(\mathbf{p}) \triangleq \frac{||\mathbf{p}-\mathbf{p}_m^2||-||\mathbf{p}-\mathbf{p}_m^1||}{c}$ of an acoustic source located in location $\mathbf{p}$, where $\mathbf{p}_m^1$ and $\mathbf{p}_m^2$ are the microphones locations of pair m, $||\cdot||$ denotes the Euclidian norm, and c is the sound velocity.

We model the PRPs using a Gaussian mixture model (GMM):

$$\phi_m(t,k) \sim \sum_{\mathbf{p}} \psi_{\mathbf{p}} \mathcal{N}^c\left(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma^2\right), \quad (2)$$

where $\psi_{\mathbf{p}}$ is the probability that the speaker emitting in bin (t,k) is located at position $\mathbf{p}$. $\mathcal{N}^c(\cdot;\cdot,\cdot)$ denotes the complex Gaussian probability density function (p.d.f.) with variance σ^2. The variance value is fixed and chosen empirically. The mean of each Gaussian can be calculated in advance on a grid of all possible locations:

$$\tilde{\phi}_m^k(\mathbf{p}) \triangleq \exp\left(-j\frac{2\pi k \tau_m(\mathbf{p})}{KT_s}\right), \forall \mathbf{p} \in \mathcal{P}, \quad (3)$$

where T_s denotes the sampling period and $\mathcal{P}$ being the set of all possible locations.

The joint p.d.f. of the PRP readings, assuming independency is given by:

$$f(\mathbf{\Phi} = \boldsymbol{\phi}; \boldsymbol{\psi}) = \prod_{m,t,k} \sum_{\mathbf{p}} \psi_{\mathbf{p}} \mathcal{N}^c\left(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma^2\right), \quad (4)$$

where $\boldsymbol{\psi} = \text{vec}_{\mathbf{p}}(\psi_{\mathbf{p}})$ and $\boldsymbol{\phi} = \text{vec}_{m,t,k}(\phi_m(t,k))$.

The ML estimate of the source locations is:

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\text{argmax}}[\log f(\mathbf{\Phi} = \boldsymbol{\phi}; \boldsymbol{\psi}) \quad \text{s.t.} \sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p}} = 1 \text{ and } 0 < \psi_{\mathbf{p}} < 1]. \quad (5)$$

2.2. Local hidden variables

The hidden variables of the IDEM algorithm are *local*, defined as the per-node association of each time-frequency bin with a source at position $\mathbf{p}$, and are denoted $y_m(t,k,\mathbf{p})$. The main modification from [13] is the direct dependency of the local hidden variables on the source position, rather than the associated TDOA. Let $\mathbf{y} = \text{vec}_{m,t,k,\mathbf{p}}(y_m(t,k,\mathbf{p}))$ be the vectorial notation of the hidden variables. The p.d.f. of $\mathbf{y}$ is given by:

$$f(\mathbf{Y} = \mathbf{y}; \boldsymbol{\psi}) = \prod_{m,t,k} \sum_{\mathbf{p}} \psi_{\mathbf{p}} y_m(t,k,\mathbf{p}). \quad (6)$$

Given the hidden variables, the p.d.f. of the observations is:

$$f(\mathbf{\Phi} = \boldsymbol{\phi}|\mathbf{y}; \boldsymbol{\psi}) = \prod_{m,t,k} \sum_{\mathbf{p}} y_m(t,k,\mathbf{p}) \times \mathcal{N}^c\left(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma^2\right). \quad (7)$$

The p.d.f. of the *complete data* can be deduced from (6)-(7):

$$f(\mathbf{\Phi} = \boldsymbol{\phi}, \mathbf{Y} = \mathbf{y}; \boldsymbol{\psi}) = \prod_{m,t,k} \sum_{\mathbf{p}} \psi_{\mathbf{p}} y_m(t,k,\mathbf{p}) \times \mathcal{N}^c\left(\phi_m(t,k), \tilde{\phi}_m^k(\mathbf{p}), \sigma^2\right). \quad (8)$$

These hidden variables, besides their role in the localization algorithm, can be utilized to construct spectral masks, as discussed in the next section.

2.3. The modified IDEM algorithm

The IDEM algorithm is based on the partial (also denoted IEM) procedure [15]. It consists of partial *E-step* updates at each node followed by an *M-step*. In this way, components of the hidden vector are estimated incrementally with the most updated values of the parameters. The IDEM is updating the hidden variables node-by-node as new data is available.

The IEM is advantageous to the conventional expectation-maximization (EM) algorithm in two important aspects, namely significantly improved convergence speed [15, 17], and weaker dependency on initialization [16, 17]. The proposed distributed method is implemented over a directed ring topology. The resulting algorithm is capable of detecting the number of active sources (including the detection of no activity) and their locations.

Algorithm 1: IDEM localization (1st stage).

input $z_m^1(t,k)$, $z_m^2(t,k); \forall m$.
Calculate $\phi_m(t,k)$ using (1).
set $\tilde{\phi}_m^k(\mathbf{p})$ using (3).
init $\hat{\psi}_{\mathbf{p}}^{(-M)}, \ldots, \hat{\psi}_{\mathbf{p}}^{(-1)}$ to uniform p.d.f..
Calculate $\upsilon_m^{(-M+1)}(t,k,\mathbf{p}), \ldots, \upsilon_m^{(0)}(t,k,\mathbf{p})$ using (10).
Calculate their mean: $\hat{\psi}_{\mathbf{p}}^{(0)} = \frac{\sum_{m,t,k} \upsilon_m^{(m-M)}(t,k,\mathbf{p})}{M \cdot T \cdot K}$.
for $\ell = 1$ **to** L **do**
  for $m = 1$ **to** M **do**
    $i = (\ell-1)M + m$ (partial iteration index).
    E-step
    Calculate $\upsilon_m^{(i)}(t,k,\mathbf{p})$ using (10).
    M-step
    Calculate
    $\hat{\psi}_{\mathbf{p}}^{(i)} = \hat{\psi}_{\mathbf{p}}^{(i-1)} + \frac{\sum_{t,k} \upsilon_m^{(i)}(t,k,\mathbf{p}) - \upsilon_m^{(i-M)}(t,k,\mathbf{p})}{M \cdot T \cdot K}$.
  end
end
output $\hat{\psi}_{\mathbf{p}}^{(LM)}, \upsilon_m^{(LM)}(t,k,\mathbf{p})$.

The *E-step* can be stated as:

$$Q(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}^{(i-1)}) \triangleq E\left\{\log\left(f(\boldsymbol{\Phi}=\boldsymbol{\phi}, \boldsymbol{Y}=\mathbf{y};\boldsymbol{\psi})\right)|\boldsymbol{\phi};\hat{\boldsymbol{\psi}}^{(i-1)}\right\} \tag{9}$$

$$= \sum_{m,t,k,\mathbf{p}} E\left\{y_m(t,k,\mathbf{p})|\phi_m(t,k);\hat{\boldsymbol{\psi}}^{(i-1)}\right\} \cdot \left[\log\psi_{\mathbf{p}} + \log\mathcal{N}^c(\phi_m(t,k);\tilde{\phi}_m^k(\mathbf{p}),\sigma^2)\right],$$

which in our case, simplifies to the calculation of:

$$\upsilon_m^{(i)}(t,k,\mathbf{p}) \triangleq E\left\{y_m(t,k,\mathbf{p})|\phi_m(t,k);\hat{\boldsymbol{\psi}}^{(i-1)}\right\} \tag{10}$$
$$= \frac{\hat{\psi}_{\mathbf{p}}^{(i-1)}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\mathbf{p}),\sigma^2\right)}{\sum_{\mathbf{p}'}\hat{\psi}_{\mathbf{p}'}^{(i-1)}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\mathbf{p}'),\sigma^2\right)}.$$

The IDEM applies a local (partial) E-step, followed by a global M-step (implemented incrementally), as summarized in Algorithm 1.

3. DISTRIBUTED ALGORITHM FOR LOCALIZATION AND SEPARATION (DALAS)

This section presents the distributed algorithm for localization and separation (DALAS). First, we deal with the detection problem, i.e. determining the number of active sources. Then, we describe the utilization of the location information to choose the best node for extracting each source. For each source we construct a spectral mask. This mask and the global localization information are utilized for node level filtering.

3.1. Number and locations of active sources

Unlike many approaches for BSS, we do not assume any a priori knowledge regarding the existence of sources and their number. The first parameter, estimated from the IDEM outputs, is the number of active sources in the enclosure.

For that, we first obtain $I(\mathbf{p})$, a location binary map, obtained by applying a threshold to the GMM weights

$$I(\mathbf{p}) = \begin{cases} 1, & \hat{\psi}_{\mathbf{p}}^{(L\cdot M)} > \text{Thr} \\ 0, & \text{else} \end{cases}, \tag{11}$$

where Thr is an empirically tuned threshold.

The number of active sources is estimated by aggregating the number of active locations:

$$\hat{N} = \sum_{\mathbf{p}\in\mathcal{P}} I(\mathbf{p}). \tag{12}$$

The location estimates are denoted $\hat{\mathbf{p}}^n, n = 1, \ldots, \hat{N}$.

3.2. Best node for each source

The next task of DALAS is to choose (for each speaker) the best receiving node. Selecting the best node is a cumbersome task. Here we proposed a simple solution:

$$m_0(n) = \underset{m(n),r}{\text{argmin}}[||\hat{\mathbf{p}}^n - \mathbf{p}_m^r||], \tag{13}$$

which selects the node with the closest microphone. This microphone is assumed to have the best input signal to noise ratio (SNR). This mechanism does not require high communication bandwidth (BW) and computational complexity. The detailed protocol for choosing the best node is out of the scope of this contribution.

3.3. Spectral masks

Source separation is obtained by utilizing spectral masking. Soft masking reduces the *musical noise* phenomenon. We propose to apply the following combination of soft and hard mask by thresholding the estimated local indicators:

$$C_m^n(t,k) = \begin{cases} 1, & \upsilon_m^{(L\cdot M)}(t,k,\hat{\mathbf{p}}^n) > T_H \\ 0, & \upsilon_m^{(L\cdot M)}(t,k,\hat{\mathbf{p}}^n) < T_L \\ \upsilon_m^{(L\cdot M)}(t,k,\hat{\mathbf{p}}^n), & \text{otherwise} \end{cases}, \tag{14}$$

where T_H and T_L are the high and the low thresholds, respectively. Their selection is a tradeoff between decreasing interference power and maintaining desired spectral contents.

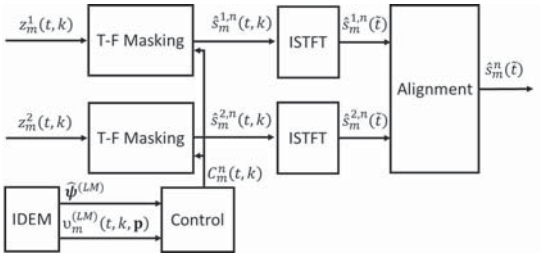

Fig. 1: Filtering at the mth microphone pair.

3.4. Node level filtering

The masking, summarized in a block-diagram depicted in Fig. 1, is applied at the best node of each source. The node signals $z_m^r(t,k)$ are masked by multiplying their STFT by the mask in (14), a value in the range $[0,1]$. Although applied locally, the location estimates are using the global information (through the control mechanism), hence potentially improving the credibility of the mask.

The masked signals, $\hat{s}_m^{r,n}(t,k)$, for speaker n at microphones $r=1,2$ of node m are transformed back to the time domain using inverse short-time Fourier transform (ISTFT) and then aligned (sub-sample delay applied) and averaged:

$$\hat{s}_m^n(t) = \frac{1}{2}\sum_{r=1}^{2} \hat{s}_m^{r,n}(t-\hat{t}_m^{r,n}), \tag{15}$$

where $\hat{t}_m^{r,n} = \frac{||\hat{\mathbf{p}}^n - \mathbf{p}_m^r||}{c}$ is the delay between the location of source n and the microphones of node m, and t is the time index. This alignment can be either implemented in the time-domain or in the frequency-domain. The result of (15) is a signal with an enhanced direct path.

3.5. Summary of DALAS

The stages of the proposed DALAS flow can now be summarized as described in Algorithm 2. The signals received by the microphones are used to execute the first stage of the algorithm, namely the IDEM algorithm. The number of sources and their locations are the outcomes of the first stage. The network chooses for each source the best node, which will be responsible for extracting its signals from the received mixture. Spectral masks (14) are applied in the STFT domain. The masked signals are aligned and averaged to improve SNR and to reduce artifacts.

4. SIMULATION STUDY AND PERFORMANCE MEASURES

4.1. Simulation setup

To evaluate the performance of the algorithms, we have simulated the following scenario. Twelve pairs (nodes) of omnidirectional microphones were located around the room at

Algorithm 2: The DALAS flow.

input $z_m^1(t,k)$, $z_m^2(t,k)$; $\forall m$.
1st stage
Execute IDEM to produce $\hat{\psi}_{\mathbf{p}}^{(L\cdot M)}$, $\upsilon_m^{(L\cdot M)}(t,k,\mathbf{p})$.
2nd stage
Estimate $\hat{N}$ using (12).
for $n=1$ **to** $\hat{N}$ **do**
  Choose the best node, $m_0(n)$ using (13).
  for $r=1$ **to** 2 **do**
    Apply masking using (14).
    Apply ISTFT.
  end
  Align masked signals and average using (15).
  output $\hat{s}_m^n(t)$.
end

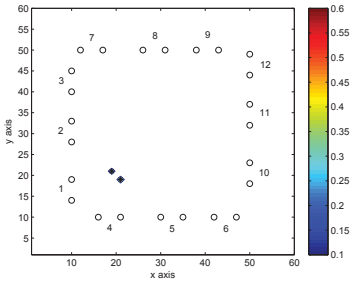

Fig. 2: Network constellation and Localization results for close distant.

the same height of 150 cm. The distance between the microphones of each pair was 50 cm as used in [13]. The dimensions of the simulated room were $6\times6\times4$ m, with low reverberation level of $T_{60}=150$ msec. Two sources, randomly located in the room at the same height as the sensors, uttered speech signals of 4 sec.

The following parameters, influencing the algorithm's performance, were empirically chosen to $\sigma^2=0.04$, Thr $=0.1$ and $L=4$. Since the localization results are quite sharp, the dependency on Thr is rather weak. The values of the masking thresholds were: $T_L=0.2$ and $T_H=0.35$. The STFT used a rectangular window and 75% overlap. The filtering part of the DALAS is examined by signal to interference ratio (SIR) measure for the input and the output signals: $\text{SIR}^n = 10\log\left(\frac{E^n}{\sum_{\tilde{n}\neq n}E^{\tilde{n}}}\right)$ dB, where E^n is the nth source power.

4.2. Results

Two examples of two concurrent sources, a man and a woman, are presented. In the first scenario the distance between the sources is rather large. In this case, based on the localization results, the algorithm selected the 1st node as the best node for extracting the woman and the 11th node as the

	Distant		Close	
	Man	Woman	Man	Woman
SIR_i	9	14	-1	4
SIR_o	19	17	19	20

Table 1: Separation measures for two sources in two cases

best node for extracting the man. This selection results in high input SIR and hence potentially improves the output separation quality. The measures are summarized in Table 1.

The second example is more challenging. The sources are very close to each other and located in the bottom left corner of the room. The localization results are depicted in Fig. 2. In this case, the algorithm selected the 4th node as the best node for separating the man and the 1st node as the best node for separating the woman. In this case the input SIR is low, since the sources are very close to each other. However, the algorithm is still capable of separating the sources and improving the SIRs significantly as evident from Table 1.

5. CONCLUSIONS

In this paper we presented the DALAS, a new separation algorithm based on the IDEM concept for localization. The distributed architecture is utilized to efficiently carry out various tasks. First, the number of active sources and their locations are estimated. Based on the location estimates, the best node for separation is selected. A byproduct of the IDEM, namely its hidden variables, is used to construct spectral masks. The estimated source locations also enable coherent averaging of the node's signals received by different microphones. Encouraging simulation results demonstrate the potential of the algorithm to blindly separate sources even in close proximity.

REFERENCES

[1] E. Vincent, C. Févotte, R. Gribonval, L. Benaroya, X. Rodet, A. Röbel, E. Le Carpentier, and F. Bimbot, "A tentative typology of audio source separation tasks," in *4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003, pp. 715–720.

[2] E.C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[4] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 139–142.

[5] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[6] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

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

[8] P.D. O'Grady and B.A. Pearlmutter, "The LOST algorithm: finding lines and separating speech mixtures," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, pp. 784296, 2008.

[9] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[10] N. Roman, D. Wang, and G.J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[11] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[12] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1900–1912, 2011.

[13] Y. Dorfan, G. Hazan, and S. Gannot, "Multiple acoustic sources localization using distributed expectation-maximization algorithm," in *the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 72–76.

[14] R.D. Nowak, "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2245–2253, 2003.

[15] R. Neal and G.E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. 1998, pp. 355–368, Kluwer Academic Publishers.

[16] P. Liang and D. Klein, "Online EM for unsupervised models," in *Proceedings of Human Language Technologies*, Stroudsburg, PA, USA, 2009, NAACL '09, pp. 611–619, Association for Computational Linguistics.

[17] M.-A. Sato, "Fast learning of on-line EM algorithm," *Rapport Technique, ATR Human Information Processing Research Laboratories*, 1999.