DUAL FREQUENCY- AND BLOCK-PERMUTATION ALIGNMENT FOR DEEP LEARNING BASED BLOCK-ONLINE BLIND SOURCE SEPARATION

Lukas Drude[1,2]*, Takuya Higuchi*[2]*, Keisuke Kinoshita*[2]*, Tomohiro Nakatani*[2]*, Reinhold Haeb-Umbach*[1]

[1] Paderborn University, Department of Communications Engineering, Paderborn, Germany
{drude, haeb}@nt.upb.de

[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
{higuchi.takuya, kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

Deep attractor networks (DANs) are a recently introduced method to blindly separate sources from spectral features of a monaural recording using bidirectional long short-term memory networks (BLSTMs). Due to the nature of BLSTMs, this is inherently not online-ready and resorting to operating on blocks yields a block permutation problem in that the index of each speaker may change between blocks. We here propose the joint modeling of spatial and spectral features to solve the block permutation problem and generalize DANs to multi-channel meeting recordings: The DAN acts as a spectral feature extractor for a subsequent model-based clustering approach. We first analyze different joint models in batch-processing scenarios and finally propose a block-online blind source separation algorithm. The efficacy of the proposed models is demonstrated on reverberant mixtures corrupted by real recordings of multi-channel background noise. We demonstrate that both the proposed batch-processing and the proposed block-online system outperform (a) a spatial-only model with a state-of-the-art frequency permutation solver and (b) a spectral-only model with an oracle block permutation solver in terms of signal to distortion ratio (SDR) gains.

Index Terms **—** blind source separation, deep learning, multi-channel, block-online

1. INTRODUCTION

Acoustic blind source separation is a challenging problem, which has been studied for many years [1]. The aim is to develop a system that extracts the individual source signals from speakers talking concurrently. In a multi-channel setup the blind source separation problem can be addressed with spatial clustering approaches [2, 3, 4].

In contrast, when only a single channel is available, separation has to rely on spectral features. One way is to employ dictionary-based approaches modeling individual speaker characteristics, i.e. non-negative matrix factorization (NMF) [5]. More recently, deep clustering (DC), a deep learning based single-channel source separation system was published, which is able to generalize to unseen speakers and does not assume the number of test speakers to be known at training time [6]. This is possible, since the network encodes the input spectrogram into embedding vectors that can then be clustered subsequently. DANs simplified the training recipe further and enables training with a signal reconstruction loss [7].

This work was performed while Lukas Drude was an intern at NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan.

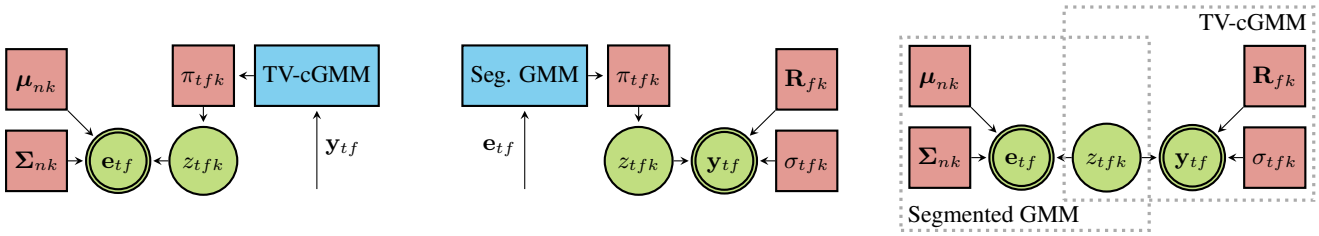
However, there are few systems that utilize both spatial and spectral features. One example uses 2D-HMMs [8] to combine spectral features with spatial observation models. A Gaussian mixture model (GMM)-based spectral model is combined with spatial features in [9]. In [10] spectral features are modeled by an NMF, while spatial features are modeled by a full rank covariance model. More recently, [11] proposed the integration of a deep neural network (DNN) based mask estimator and a complex angular-central Gaussian mixture model (cACGMM) to extract a single source. In [12] a DNN refines the source estimate in each expectation maximization (EM) iteration. In [13] we proposed modeling spectral features with a DC model and spatial features with a time-variant complex Gaussian mixture model (TV-cGMM).

Most source separation studies focus on completely overlapping speech [6, 7] and a duration of exactly one utterance. In this contribution we focus on longer mixtures with possible speech pauses to extend DC and DAN based systems to more realistic meeting scenarios. As a side effect, the produced masks can be used for speaker diarization. To be able to continuously separate an observed signal, block-online or online processing is required [14, 15, 16]. However, since both DC [6] and DANs [7] rely on BLSTMs, a generalization to online-processing is challenging. Even if we resort to block-online processing, the encoding network is optimized to separate a single mixture. It is not guaranteed that the topology of the embedding space will remain the same on a subsequent block of a possibly longer meeting. A block permutation problem arises whereby the speaker index may be permuted from block to block. Although this problem can also be addressed with speaker identification techniques, we here wish to demonstrate how to use spatial cues to solve the block permutation problem.

Short time Fourier transform (STFT) based spatial clustering models operate on each frequency independently. This yields the frequency permutation problem [3]. To jointly solve both problems, we formulate an integrated probabilistic graphical model to leverage spectral features to address the frequency permutation problem and spatial features to deal with the block permutation problem.

2. RELATION TO PRIOR WORK

In [13] an integration between DC [6] and a TV-cGMM has been proposed for an offline setup for reverberant but noise-free scenarios. In contrast, we here generalize the setup to (a) noisy scenarios and (b) block-online processing. The segment permutation problem did not occur in [13], since the whole model operated on a single block. The weighting between spatial and spectral observation model is avoided here.



(a) Fixed spatial prior model (Sec. 6.1). The posterior mask of a TV-cGMM is used as a prior for a segmented GMM.

(b) Fixed spectral prior model (Sec. 6.2). The posterior mask of a segmented GMM is used as a prior for a TV-cGMM.

(c) Full update model (Sec. 6.3). All parameters are estimated in a joint EM framework. Dashed boxes indicate conventional models.

Fig. 1: Overview of integrated batch processing models. Preprocessing blocks are shown in blue. Green circles denote random variables of the probabilistic graphical model, where the observed random variables are doubly circled. Red boxes indicate parameters to be estimated during EM iterations.

3. SIGNAL MODEL

A convolutive mixture of K independent source signals s_{tfk}, captured by D sensors is approximated in the STFT domain:

$$\mathbf{y}_{tf} = \sum_k \mathbf{h}_{fk}\, s_{tfk} + \mathbf{n}_{tf} = \sum_k \mathbf{x}_{tfk} + \mathbf{n}_{tf}, \tag{1}$$

where $\mathbf{y}_{tf}$, $\mathbf{h}_{fk}$ $\mathbf{n}_{tf}$, and $\mathbf{x}_{tfk}$ are the D-dimensional observed signal vector, the unknown acoustic transfer function vector of source k, the noise vector, and the source image at the sensors, respectively. Furthermore, t and f specify the time frame index and the frequency bin index, respectively. Since speech signals are sparse in the STFT domain, we may assume that a time frequency slot is occupied either by a single source and noise or by noise only.

4. DEEP ATTRACTOR NETWORK

DANs [7] encode a single channel mixture spectrogram to embedding vectors $\mathbf{e}_{tf}$ for each time frequency bin. The intention is to train the network such that embeddings belonging to the same source move closer together and embeddings belonging to different sources move further apart in the embedding space. In particular, a signal reconstruction loss for training the DAN is used although we later use the embeddings for clustering, which is not part of the cost function. During testing k-means is employed to cluster the embedding vectors. The motivation behind using DANs instead of DC is that the loss function can be evaluated faster during training and allow end-to-end training for future research.

5. CONVENTIONAL BATCH MODELS

In this section we present a spatial model operating on the entire mixture and a spectral model operating on each block independently. They form the basis of the models in Sec. 6. The motivation to analyze a spectral model on segments but a spatial model on the complete mixture is that there are many online or block-online formulations of spatial models, but DC and DANs are not yet generalized to the online case.

5.1. Time-variant cGMM for spatial features

The TV-cGMM is a very competitive spatial clustering approach [17]. Its efficacy was shown in the CHiME 3 and CHiME 4 challenges, where it was used in both winning systems.

It is related to a complex Gaussian mixture model but differs in that a time dependent variance parameter σ_{tfk} is factored out of the time-independent spatial correlation matrix $\mathbf{R}_{fk}$. The observation model is then given by $\mathcal{N}_{\mathbb{C}}(\mathbf{y}_{tf}; \mathbf{0}, \sigma_{tfk}\mathbf{R}_{fk})$.

The updates in the M-step with $\Gamma_{fk} = \sum_t \gamma_{tfk}^{\text{spatial}}$ are:

$$\sigma_{tfk} = \frac{1}{D}\mathbf{y}_{tf}^{\mathsf{H}}\mathbf{R}_{fk}^{-1}\mathbf{y}_{tf}, \quad \mathbf{R}_{fk} = \frac{1}{\Gamma_{fk}}\sum_t \gamma_{tfk}^{\text{spatial}}\frac{\mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathsf{H}}}{\sigma_{tfk}}. \tag{2}$$

The posterior mask is obtained with the following E-step:

$$\gamma_{tfk}^{\text{spatial}} = \frac{p(\boldsymbol{y}_{tf}; \sigma_{tfk}, \mathbf{R}_{fk})}{\sum_{k'} p(\boldsymbol{y}_{tf}; \sigma_{tfk'}, \mathbf{R}_{fk'})}. \tag{3}$$

Since the model just captures the features of a single frequency, the result is affected by the frequency permutation problem [3]. The effect of optional mixture weights is analyzed in Sec. 9.

5.2. Segmented spectral GMM for spectral features

Instead of k-means, as in the original DAN formulation [7], we employ a GMM for a more statistically sound formulation and better integration into joint models.

The segmented spectral GMM models the embedding vectors of each block n independently: $\mathcal{N}(\mathbf{e}_{tf}; \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk})$. The parameter updates are given by the M-step:

$$\boldsymbol{\mu}_{nk} = \frac{1}{\Gamma_{nk}}\sum_{t\in\mathcal{T}_n, f} \gamma_{tfk}^{\text{spectral}}\mathbf{e}_{tf}, \quad \Gamma_{nk} = \sum_{t\in\mathcal{T}_n, f} \gamma_{tfk}^{\text{spectral}}, \tag{4}$$

$$\boldsymbol{\Sigma}_{nk} = \frac{1}{\Gamma_{nk}}\sum_{t\in\mathcal{T}_n, f} \gamma_{tfk}^{\text{spectral}}(\mathbf{e}_{tf} - \boldsymbol{\mu}_{nk})(\mathbf{e}_{tf} - \boldsymbol{\mu}_{nk})^{\mathsf{T}}, \tag{5}$$

where $\mathcal{T}_n$ are all time frame indices belonging to block n. In practice, we use scaled identity covariance matrices to avoid singularities. The E-step to obtain the posterior masks is given as follows:

$$\gamma_{tfk}^{\text{spectral}} = \frac{p(\mathbf{e}_{tf}; \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk})}{\sum_k p(\mathbf{e}_{tf}; \boldsymbol{\mu}_{nk'}, \boldsymbol{\Sigma}_{nk'})}, \tag{6}$$

with $t \in \mathcal{T}_n$. Since the result is independently obtained for each block, it is affected by the block permutation problem. The effect of optional mixture weights is analyzed in Sec. 9.

6. PROPOSED INTEGRATED BATCH MODELS

This section proposes several ways to jointly solve the block permutation and frequency permutation problems described in the previous sections. All of the proposed methods fuse spectral and spatial information in different ways. The methods introduced here assume a batch-mode setting. A block-online processing approach will be introduced in the next section based on the findings in this section.

6.1. Fixed spatial prior integration model

One possible way of solving the frequency permutation problem is to use the (possibly permuted) posterior mask from Sec. 5.1 as a time dependent but fixed prior ($\pi_{tfk} \coloneqq \gamma_{tfk}^{\text{spatial}}$). Then, the spectral parameters (Eqs. (4) and (5)) and the possible permutation of the fixed prior can be updated as illustrated in Fig. 1a. The E-step is modified to account for the fact that the spatial prior is still affected by frequency permutation: The expected complete data log-likelihood is maximized to simultaneously obtain the posterior mask and the alignment Π_f:

$$\Pi_f = \underset{\Pi}{\operatorname{argmax}} \left\{ \sum_{t,k} \frac{A_{tfk}(\Pi)}{\sum_{k'} A_{tfk'}(\Pi)} \cdot \ln A_{tfk}(\Pi) \right\}, \tag{7}$$

$$\text{with } A_{tfk}(\Pi) = \pi_{tf,\Pi(k)}\, p(\mathbf{e}_{tf}; \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}). \tag{8}$$

Again, $t \in \mathcal{T}_n$. A greedy search as in [3] can be used instead of an exhaustive search. Although the segment permutation problem is not specifically addressed here, due to the fact that the spatial models are shared across all segments, the class dependent mean and covariance parameters of the spectral model are aligned automatically. After the optimal permutation Π_f is found, the new posterior mask γ_{tfk} is given by the fraction in Eq. (7).

6.2. Fixed spectral prior integration model

Here, the possibly block-permuted posterior mask of the segmented spectral GMM of Sec. 5.2 is used as a fixed prior $\pi_{tfk} \coloneqq \gamma_{tfk}^{\text{spectral}}$ as illustrated in Fig. 1b. The fixed prior therefore relates the models of different frequencies, leading to an aligned joint solution. This idea generalizes permutation free clustering [18], where a time dependent (but frequency independent) mixture weight is used to link models of different frequencies together.

By analogy with the previous section, the E-step with permutation alignment yields the alignment Π_n:

$$\Pi_n = \underset{\Pi}{\operatorname{argmax}} \left\{ \sum_{t \in \mathcal{T}_n, f, k} \frac{B_{tfk}(\Pi)}{\sum_{k'} B_{tfk'}(\Pi)} \cdot \ln B_{tfk}(\Pi) \right\}, \tag{9}$$

$$\text{with } B_{tfk}(\Pi) = \pi_{tf,\Pi(k)}\, p(\mathbf{y}_{tf}; \sigma_{tfk}, \mathbf{R}_{fk}). \tag{10}$$

After convergence, the fraction in Eq. (9) is the posterior mask γ_{tfk}.

6.3. Full update model

Although the aforementioned models both address the dual permutation problems, it can be expected that updating all spatial and spectral model parameters should yield even better source separation performance. The joint model is illustrated in Fig. 1c.

In the full update model, the M-step consists of Eqs. (2), (4) and (5). The E-step with permutation alignment is then given as follows. It is similar to Eq. (7), but includes the spectral conditional probability densities instead of the fixed priors. Again the posterior mask γ_{tfk} can then be identified as the fraction in Eq. (11) after the optimal permutation Π_f has been found ($t \in \mathcal{T}_n$):

$$\Pi_f = \underset{\Pi}{\operatorname{argmax}} \left\{ \sum_{t,k} \frac{C_{tfk}(\Pi)}{\sum_{k'} C_{tfk'}(\Pi)} \cdot \ln C_{tfk}(\Pi) \right\}, \tag{11}$$

$$\text{with } C_{tfk}(\Pi) = p(\mathbf{e}_{tf}; \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) \cdot p(\mathbf{y}_{tf}; \sigma_{tf,\Pi(k)}, \mathbf{R}_{f,\Pi(k)}) \tag{12}$$

Algorithm 1 Block-processing

1: Split into N blocks and run model of Sec. 6.3 on the first block.
2: Apply beamforming to the first block as in [13].
3: **for** n from 1 to N **do**
4: Forget all parameters but $\mathbf{R}_{n-1,fk}$ and $\boldsymbol{\Phi}_{n-1,fk}$.
5: Initialize σ_{tfk} with Eq. (2) using $\mathbf{R}_{n-1,fk}$.
6: Initialize γ_{tfk} with Eq. (3).
7: **while** not converged **do**
8: Obtain $\boldsymbol{\mu}_{nk}$ and $\boldsymbol{\Sigma}_{nk}$ with Eqs. (4) and (5).
9: Incremental update for $\boldsymbol{R}_{nfk}$ with Eq. (13).
10: Variance σ_{tfk} update with Eq. (2).
11: PE-step from Eq. (11) yields γ_{tfk} and Π_f.
12: Obtain spatial covariance matrices with Eq. (14).
13: Apply beamforming on current block as in [13].

7. BLOCK-ONLINE MODEL

To formulate a block-online model as in Alg. 1, we generalize the full update model of the previous section. Summary statistics are obtained with Eqs. (2), (4), (5), and $\Gamma_{nfk} = \sum_{t \in \mathcal{T}_n} \gamma_{tfk}$. Since the spectral statistics are only valid in one block, we drop the information after processing a block. We only carry over the spatial characteristics $\mathbf{R}_{fk}$ to automatically solve the block permutation problem of the next block with $\Gamma_{nfk}^{\text{total}} = \Gamma_{n-1,fk}^{\text{total}} + \Gamma_{nfk}$:

$$\mathbf{R}_{nfk} = \frac{\Gamma_{n-1,fk}^{\text{total}}}{\Gamma_{nfk}^{\text{total}}} \mathbf{R}_{n-1,fk} + \frac{1}{\Gamma_{nfk}^{\text{total}}} \sum_{t \in \mathcal{T}_n} \frac{\gamma_{tfk}}{\sigma_{tfk}} \mathbf{y}_{tf} \mathbf{y}_{tf}^{\mathsf{H}}. \tag{13}$$

Defining the recursion like this will yield a model that never forgets past information. This is meaningful if the acoustic scene is expected to be constant as in [19]. If desired, an additional forgetting factor can be applied to allow the model to adapt to slowly changing geometric settings.

8. BEAMFORMING

The generalized eigenvalue (GEV) beamformer has proven to be robust with respect to numerical instabilities and yields great improvements in terms of both signal to noise ratio (SNR) gain and word error rate (WER) reduction, while often outperforming the frequently used minimum variance distortionless response (MVDR) beamformer [20, 21]. Within this work, we employ the GEV beamformer as a way of separating concurrent target speakers as in [13]. For block-online processing one beamforming vector per block and frequency is obtained. The spatial covariance matrices are obtained by a smoothed update similar to Eq. (13):

$$\boldsymbol{\Phi}_{nfk}^{\text{target}} = \frac{\Gamma_{n-1,fk}^{\text{total}}}{\Gamma_{nfk}^{\text{total}}} \boldsymbol{\Phi}_{n-1,fk}^{\text{target}} + \frac{1}{\Gamma_{nfk}^{\text{total}}} \sum_{t \in \mathcal{T}_n} \gamma_{tfk} \mathbf{y}_{tf} \mathbf{y}_{tf}^{\mathsf{H}}. \tag{14}$$

The eigenvalue decomposition can then be performed on each block.

9. EVALUATION

The DAN was trained on clean WSJ [22] mixtures according to file lists provided by [6] with a sampling frequency of $8\,\mathrm{kHz}$ and an STFT size and shift of 512 and 128, respectively. In contrast to the training recipe in [7], we achieved best results with a tanh nonlinearity after the last layer. We merged the forward and backward BLSTM streams by concatenation and applied sequence normalization [23]. To test the proposed integrated models, we generated 1000 mixtures of two speakers per setting. The settings differ in terms of active speech ratio, such that the overlap as defined by a crude voice activity detector varies. An active speech ratio of $100\,\%$ means that all speakers speak concurrently without pauses. The utterances were reverberated with generated room impulse responses [24] and random reverberation times in the range of $150 - 200\,\mathrm{ms}$. We used the CHiME 3 array geometry (planar, minimum distance $10\,\mathrm{cm}$) with an unconstrained random array rotation in space. Realistic CHiME 3 background noise with matching spatial characteristics was added with a per speaker input SNR in the range of $8 - 12\,\mathrm{dB}$, where τ is the sample index in time domain and d is the sensor index:

$$\mathrm{SDR}_{\mathrm{in}} = 10\,\mathrm{dB}\log_{10}\left(\frac{\sum\limits_{\tau,k,d}|s_{kd}(\tau)|^2}{\sum\limits_{\tau,k',d}|s_{k',d}(\tau)|^2 + \sum\limits_{\tau,d}|n_d(\tau)|^2}\right). \quad (15)$$

Similarly, the output SDR were calculated intrusively [25] in time domain, where the beamforming vector for speaker k is applied to the images $\mathbf{x}_{tfk}$, $\sum_{k'}\mathbf{x}_{tfk'}$ with $k' \neq k$, and $\mathbf{n}_{tf}$, separately. We refrained from using BSSEval [26] since due to the simulation setup all the images were available and no additional estimation was necessary. All the proposed models address additive noise by introducing an additional noise class.

9.1. Evaluation of batch models

Tbl. 1 shows batch-processing results with different speaker overlaps. The TV-cGMM with frequency permutation alignment as in [3] serves as a competitive baseline (denoted by "Prior: None"). The performance of the model improves, when a frequency dependent class affiliation prior is used instead of no prior as in Eq. (3) (denoted by "Prior: fk"). The complementary baseline is a segmented GMM on deep attractor embeddings with oracle block permutation alignment. The performance is very similar with and without an additional prior.

Both the fixed spatial and fixed spectral prior models outperform the baseline in most cases. Furthermore, they resolve the frequency permutation and block permutation, such that the frequency permutation alignment of [3] and oracle block permutation alignment can be avoided altogether.

Finally, the full update model outperforms all other models both with and without an additional prior. It is also noteworthy that the performance degrades less severely with a lower active speech ratio than the TV-cGMM in our setup. An additional gain of approximately $0.2\,\mathrm{dB}$ can be obtained by carefully tuning the influence of each observation model as in [13]. In this work we refrained from any heuristic tuning to keep the model concise.

Table 1: SDR gain for all batch-mode models. Baseline methods are shown in gray. All segments contain 400 frames ($\approx 3.2\,\mathrm{s}$).

Active Speech Ratio:		73 %	60 %	47 %
Average Duration:		16.9 s	27.0 s	36.4 s
Model	Prior	SDR gain/dB		
TV-cGMM+[3]	None	16.9	15.5	13.3
	fk	17.3	16.6	15.4
DAN+GMM+Oracle Align	None	15.8	15.8	15.5
	nk	15.7	15.6	15.4
DAN+Fixed Spatial Prior	tfk	17.3	16.7	15.1
DAN+Fixed Spectral Prior	tfk	17.1	17.0	16.1
DAN+Full Update Model	None	**17.9**	**17.4**	**16.3**
	fk	17.8	**17.4**	**16.3**

9.2. Evaluation of block-online models

Tbl. 2 shows the block-online models on the same three datasets. Again, each model is evaluated with and without an additional prior. In the block-online case the full update model outperforms the other models by an even higher margin (up to $1.5\,\mathrm{dB}$) although the lower active speech ratio scenario shows the limitations of both the TV-cGMM and the full update model. Specifically, the full update model works better, when no additional prior is used. The drop in performance compared with the batch results is greater with a low speech overlap since the first block does not always contain all speakers.

10. CONCLUSION

In this contribution we demonstrated how spatial and spectral features can be integrated to solve both a frequency permutation problem and a block permutation problem. We introduced the segmented batch models as a necessity to generalize deep learning based single-channel batch models to multi-channel block-online processing. Integrated block-online processing outperforms both state of the art baseline approaches and is able to solve the block permutation problem. Future research will leverage speaker identification features to overcome the block permutation problem as regards moving speakers and evaluate its effectiveness in terms of automatic speech recognition (ASR) performance.

Table 2: Block-online results in SDR gain. Baseline methods are shown in gray. The first block consists of 400 frames ($\approx 3.2\,\mathrm{s}$), all consecutive blocks have 200 frames ($\approx 1.6\,\mathrm{s}$).

Active Speech Ratio:		73 %	60 %	47 %
Average Duration:		16.9 s	27.0 s	36.4 s
Model	Prior	SDR gain/dB		
TV-cGMM+[3]	None	15.7	14.3	11.6
	fk	16.1	15.0	12.5
DAN+GMM+Oracle Align	None	13.7	14.0	13.7
	nk	13.9	13.8	13.5
DAN+Full Update Model	None	**17.4**	**16.5**	**14.0**
	fk	17.1	16.2	13.4

11. REFERENCES

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

[2] 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.

[3] 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.

[4] S. Makino, T. Lee, and H. Sawada, *Blind speech separation*, vol. 615, Springer, 2007.

[5] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[6] J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[7] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," *arXiv preprint arXiv:1611.08930*, 2016.

[8] D. Tran Vu and R. Haeb-Umbach, "Blind speech separation exploiting temporal and spectral correlations using 2D-HMMs," in *IEEE European Signal Processing Conference (EUSIPCO)*, sep 2013, pp. 1–5.

[9] T. Nakatani, M. Souden, S. Araki, T. Yoshioka, T. Hori, and A. Ogawa, "Coupling beamforming with spatial and spectral feature based spectral enhancement and its application to meeting recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7249–7253.

[10] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *IEEE International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[11] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[12] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[13] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for BSS with deep clustering embeddings," in *Interspeech*, 2017.

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

[15] 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.

[16] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, pp. 355–368. Springer, 1998.

[17] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[18] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3238–3242.

[19] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5210–5214.

[20] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[21] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, 2017.

[22] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[23] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 444–451.

[24] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[25] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 241–244.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.