人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-057-5 (11/20)

空間モデルを考慮した深層学習ベースの音源分離

DNN based speech source separation considering spatial model

戸上真人 [1*]

[1] LINE 株式会社
[1] LINE Corporation

Abstract: Recently, deep learning based speech source separation has been evolved rapidly. A neural network (NN) is usually learned independently of a spatial model. However, a research question remains whether the NN that is trained such as configuration is really optimal when speech source separation is performed with the spatial model. In this paper, I will introduce conventional statistical model based speech source separation and deep learning based speech source separation. After that, I will introduce four research directions which incorporate a spatial model into the NN structure.

1 はじめに

会議議事録の書き起こしシステムなどの音声認識システムを用いたシステム，及びテレビ会議システムなどの音声通話システムでは，マイクロホンに入ってくる音は様々な雑音・複数の人の話し声が混ざった音となる。このような様々な音が混ざったマイクロホン入力信号を音源毎に分離する音源分離技術に対する注目が集まっている。

これまで音源分離技術としてブラインド音源分離技術 (BSS) に関する検討が盛んになされてきた [2, 4, 7–10, 17, 24]。BSS では分離に必要なパラメータをマイク入力信号から求めるが，マイク入力信号以外に何も情報が無いとパラメータを一意に決定することが難しいため，追加の情報として音の伝搬過程および信号源に関する二つの手がかり・モデルを利用する手法が一般的に用いられている。音の伝搬過程に関するモデルは空間モデルと呼ばれ，マイク入力信号を音源と空間的なインパルス応答の線形畳み込み混合でモデル化する事が一般的である。一方，信号源に関するモデルとして音声の統計的性質（特に優ガウス性）に基づき，音源の原信号をラプラス分布 [7, 9] や時変ガウス分布 [4] でモデル化するような構成がよく用いられている。本稿ではこの信号源に関するモデルを統計モデルと呼ぶ。これら統計モデルを用いた音源分離技術はパラメータ最適化のための繰り返し計算に基づくアルゴリズムの検討 [16, 17, 25] と歩調を合わせて検討が進んでいる。

一方で，近年教師ありの音源分離手法として，深層学習に基づく音源分離方式に関する検討が広く進みつつある。例えば，ディープクラスタリング [6,23], パーミュテーション不変学習 [26, 27], ディープアトラクタネットワーク [1,11], BSS とのハイブリッド方式 [12,14,15] が提案されている。深層学習に基づく音源分離方式では，音源の周波数特性などの音源の特徴を従来の統計モデルに基づく音源モデルと比較し，より正確に捉えられる事が期待できる。加えて，音源分離のパラメータ最適化のために繰り返し計算に基づく方法を用いる必要が無いという利点も有する。こうしたことから深層学習に基づく音源分離方式の検討が飛躍的に進んでいる。特に，空間モデルを求めるための時間周波数マスクを深層学習により求める手法 [5, 6, 23, 26, 27] の検討が進んでいる。これらの手法では一般的に時間周波数マスクを学習するための教師データとして，時間周波数毎の S/N に基づく真の時間周波数マスクを定義し，その真の時間周波数マスクに近い時間周波数マスクをニューラルネットワークが出力するように学習を進める。これらの構成ではニューラルネットワーク学習時には，時間周波数マスクを用いて推定した空間モデルの精度を考慮することなく学習を行う。しかし，学習したニューラルネットワークを空間モデルを用いた音源分離に接続するとしたときには，時間周波数マスクを用いて推定した空間モデルの精度の影響を大きく受けるため，学習時においても時間周波数マスクを用いて推定した空間モデルの精度を考慮することが望ましいと考えられる。また時変ガウスモデルを用いた音源分離 [4] のように時々刻々フィルタの形状が変化する場合，時間周波数マスクと共に，時間周波数毎の音源の分散も求める必要がある。時間周波数マスクは空間モデルを推定するために用いる変数であり，一方で時間周波数毎の音源の分散は分離フィルタの形状を変化させるための変数であり，それぞれ役割が異なる。した

*連絡先：LINE 株式会社
E-mail: masahito.togami@linecorp.com

がって，空間モデルの影響を考慮し，それぞれの役割に適合した形で変数を推定することが望ましいと考えられる。

こうしたことから，著者らは，ニューラルネットワーク学習時において空間モデルの影響を考慮する方式として，次の４つの構成について検討を進めてきている。

1. 空間モデルの影響を考慮したニューラルネットワークの損失関数
2. ニューラルネットワークの構造の中に空間モデルを用いた音源分離を埋め込む方法
3. 所望音源の到来方向の情報をアトラクタとして用いて音源分離に必要なパラメータを推定するフレームワーク
4. 統計モデルに基づく音源分離法を疑似教師信号生成機として用いる教師無しニューラルネットワーク学習法

本稿では，これらの４つの方向性について紹介する。

2 空間モデルの影響を考慮したニューラルネットワークの損失関数 [18]

時間周波数領域でのマイク入力信号を $\boldsymbol{x}_{l,k}$ (l がフレームインデックス，k が周波数インデックス）とする。$\boldsymbol{x}_{l,k}$ は N_m 個の要素からなるベクトルとする。N_m はマイクロホン数である。時変ガウスモデルに基づく音源分離では複数チャネルのウィナーフィルタ (MWF)$\boldsymbol{W}_{i,l,k}$ (N_m 行 N_m 列) を使って，i 番目の音源信号 $\boldsymbol{c}_{i,l,k}$ を以下のように推定する。

$$\hat{\boldsymbol{c}}_{i,l,k} = \boldsymbol{W}_{i,l,k}\boldsymbol{x}_{l,k} \tag{1}$$

ここで MWF は

$$\boldsymbol{W}_{i,l,k} = v_{i,l,k}\boldsymbol{R}_{i,k}\left(\sum_{j=0}^{N_s-1} v_{j,l,k}\boldsymbol{R}_{j,k}\right)^{-1} \tag{2}$$

で求める事ができる。N_s は音源数，$v_{i,l,k}$ は i 番目の音源の時間周波数成分ごとの分散，$\boldsymbol{R}_{i,k}$ は空間共分散行列とする。空間共分散行列は時間周波数マスク $M_{i,l,k}$ を用いて以下のように推定される。

$$\boldsymbol{R}_{i,k} = \frac{1}{\sum_l M_{i,l,k}}\sum_l M_{i,l,k}\boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H \tag{3}$$

時間周波数マスク $M_{i,l,k}$ は音源の時間周波数成分毎の分散 $v_{i,l,k}$ と共にニューラルネットワークを介して推定される。これまで，一般的に時間周波数マスク $M_{i,l,k}$ を推定するためのニューラルネットワークの学習時には，時間周波数マスクの正解値を定義し，推定したマスクがその正解値に近づくようにニューラルネットワークを学習してきた。したがって，ニューラルネットワーク学習時には推定したマスクを用いて算出した空間モデルを通して音源分離した結果を評価してはいなかった。また，$M_{i,l,k}$ と $v_{i,l,k}$ は共に時間周波数毎の i 番目の音源の音量に関連する変数となるが，$M_{i,l,k}$ は空間共分散推定に用いられ，$v_{i,l,k}$ は時間毎のフィルタ形状を決めるために用いられるといったように，空間モデルを用いた音源分離における役割はそれぞれ異なる。したがって，$M_{i,l,k}$ と $v_{i,l,k}$ を推定するニューラルネットワークを i 番目の音源の音量を教師信号として学習することは必ずしも望ましくなく，空間モデルの影響を考慮した上で学習することが望ましいと考えられる。そこで我々は，$M_{i,l,k}$ と $v_{i,l,k}$ を推定するニューラルネットワークを音源分離信号 $\hat{\boldsymbol{c}}_{i,l,k}$ が真の音源信号 $\boldsymbol{c}_{i,l,k}$ に近づくように学習する手法を提案している [18]。提案法では，音源信号の事後確率を以下のように計算する。

$$p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k}) = \mathcal{N}(\boldsymbol{c}_{i,l,k}|\hat{\boldsymbol{c}}_{i,l,k}, \boldsymbol{V}_{i,l,k}) \tag{4}$$

ここで $\mathcal{M}$ はニューラルネットワークのパラメータとする。$\boldsymbol{V}_{i,l,k}$ は $\boldsymbol{c}_{i,l,k}$ の共分散行列であり $\hat{\boldsymbol{c}}_{i,l,k}$ と同様に $\boldsymbol{W}_{i,l,k}$，$v_{i,l,k}$ と $\boldsymbol{R}_{i,k}$ から算出することができる。$p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ 算出のためのブロック構成を図 1 に示す。ニューラルネットワークの損失関数としては，負の対数事後確率 $-\log p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ を用いる。提案する損失関数では，分散 $\boldsymbol{V}_{i,l,k}$ が一種の正則化効果を及ぼし $\hat{\boldsymbol{c}}_{i,l,k}$ が $\boldsymbol{c}_{i,l,k}$ から大きくずれている場合であっても，損失が過度に大きくなることを防ぐ効果があると期待できる。実際に二乗誤差と比較して提案する損失関数を用いて学習した結果，分離性能が向上することを確認している。また，残響除去と音源分離のためのニューラルネットワークを同時に学習する手法も提案されている [20]。

3 ニューラルネットワークの構造の中に空間モデルを用いた音源分離を埋め込む方法 [19]

空間モデルをニューラルネットワークに統合するための２つの構造として，ニューラルネットワークの構造の中に空間モデルを用いた音源分離を埋め込む方法を紹介する（図 2)。前章で紹介したようなニューラルネットワークで分離に必要なパラメータを推定しそのパラメータを使って音源分離を行うような構成の場合，図 2(a) で示すようにニューラルネットワークは順方向の計算中には空間モデルを参照しない。これに対して，空間モデルにより適合したパラメータをニューラルネッ

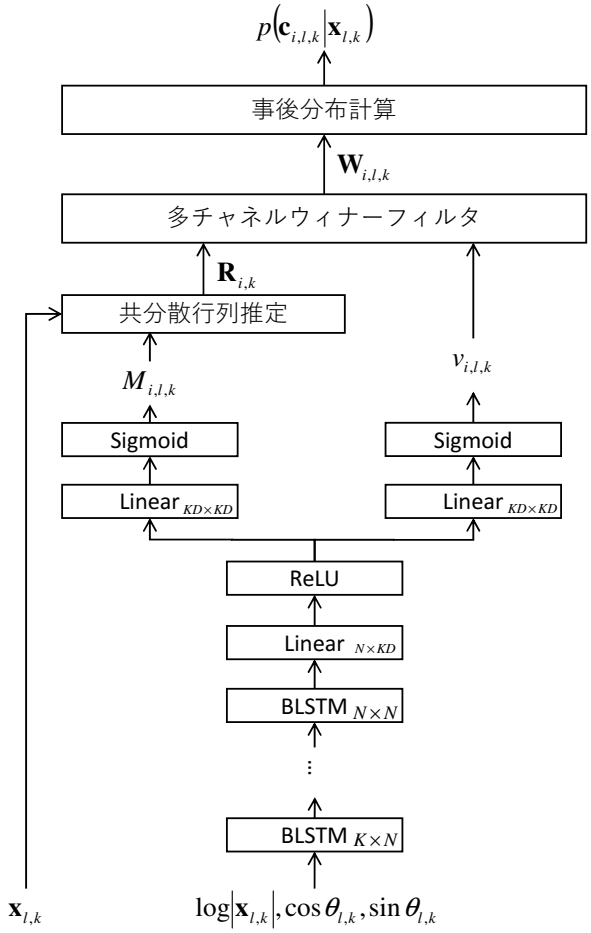

図 1: ニューラルネットワークを介した音源信号の事後確率算出ブロック

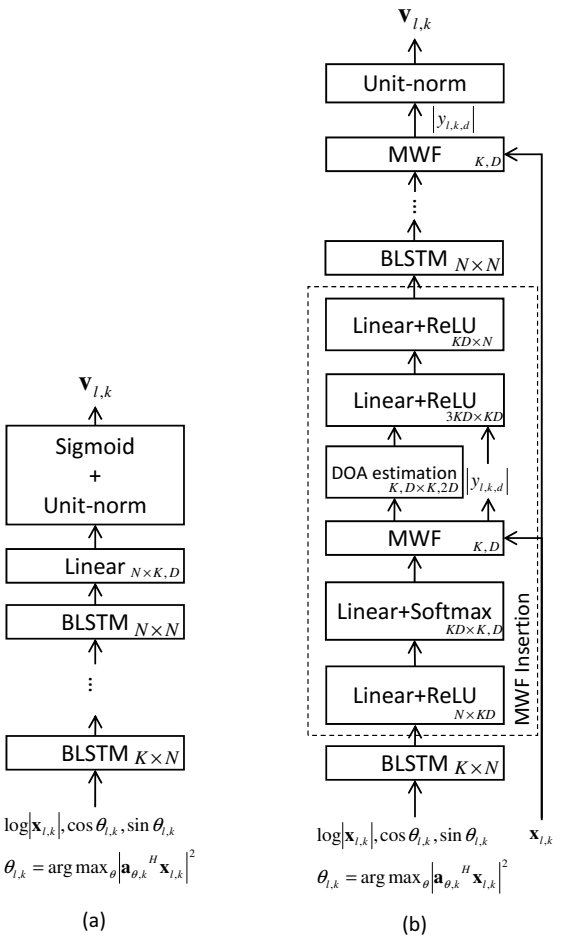

図 2: ブロック構成

トワークで推定可能となることを狙い，図 2(b) で示すように，順方向の計算時に空間モデルを参照する構成を提案する。本構成では，各 BLSTM の出力信号を時間周波数マスク $M_{i,l,k}$ に変換する。そして変換したマスクから共分散行列を構築し時不変の MWF に基づき音源分離実施する。その分離結果を次の BLSTM の入力信号として変換する。BLSTM の出力信号は時不変の MWF に変換され，BLSTM の出力信号の自由度よりも時不変の MWF の自由度が低いことから，音源分離を埋め込むことにより空間モデルに適合する形で自由度が落とすことが可能と考えられる。

ディープクラスタリング [6, 23] の構成に提案法の枠組みを適用し，音源分離性能が向上することを確認した。特に，空間モデルの適合度が向上することにより音源分離後の歪が減少することが確認された。

4 所望音源の到来方向の情報をアトラクタとして用いて音源分離に必要なパラメータを推定するフレームワーク [13]

所望音源の到来方向が分かっており，その方向の音のみを分離抽出したいというケースは多い。例えば，対話型ロボットでロボットの顔が向いている方向の音を取りたいというケースや，カメラ画像で人の顔を認識し人の方向の音だけを抽出したいというケースなどである。このような場合，音源分離した後に所望音源の到来方向から所望の分離音を選択するという構成よりも，音源分離の段階から到来方向の情報を利用した方が効率的に所望信号の情報と妨害音の情報を切り分けて推定できると推察される。そのような事から，所望音源の到来方向の情報を一種のアトラクタとして用いて，音源分離に必要なパラメータを推定するフレームワー

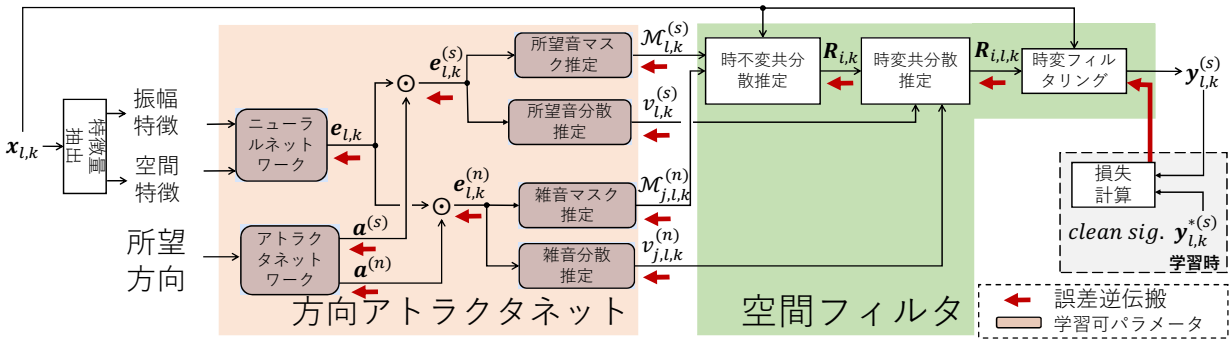

図 3: ブロック構成

クを提案した（図 3)。本構成は時変フィルタを用いた音源分離を実行するため，時間周波数マスクと共に，音源の時間周波数毎の分散を必要とする。これらのパラメータに対する教師信号を定義するのではなく，分離音に対する教師信号を定義することにより時間周波数マスクと音源の時間周波数毎の分散を推定するニューラルネットワークを統合的に学習することが可能となる。

5 統計モデルに基づく音源分離法を疑似教師信号生成機として用いる教師無し NN 学習法 [21]

深層学習を用いた音源分離で仮定する真の分離音は通常手に入らないことが多い。そのようなシーンでは，真の分離音がなくともニューラルネットワークを学習することが求められる。このようなニーズに対して，近年，教師無しのニューラルネットワーク学習法が提案されている [3,22]。これらの教師無しニューラルネットワーク学習法では，時間周波数マスクの教師信号が無くとも，各音源の時間周波数マスクを推定するためのディープクラスタリング構成のニューラルネットワークを学習することが可能な構成となっている。しかし，残響や背景雑音が存在する場合，各音源の時間周波数マスク以外にも様々な変数を推定することが必要になり，空間モデルを用いた音源分離の分離信号がより良くなるようにこれらの変数を推定することが望まれる。これに対して，真の分離音の代わりに，統計モデルに基づく音源分離法を疑似教師信号生成機として用いて，統計モデルに基づく音源分離法が出力する分離音に深層学習を用いた音源分離の分離音が近づくようにニューラルネットワークを学習する手法を提案する（図 4)。統計モデルに基づく音源分離法としては時変の MWF を用いる。共分散行列に各音源の直接音の共分散行列と共に残響と背景雑音の共分散行列も加えることにより，残響・背景雑音耐性を高める。統計モデルに基づく音源分離法の出力信号中に含まれる分離エラーに対して過度に追従することを防ぐために，Kullback Leibler Divergence (KLD) に基づく損失関数を用いる。

実験の結果，統計モデルに基づく音源分離法よりも高い分離性能を得ることが可能であることを確認している。

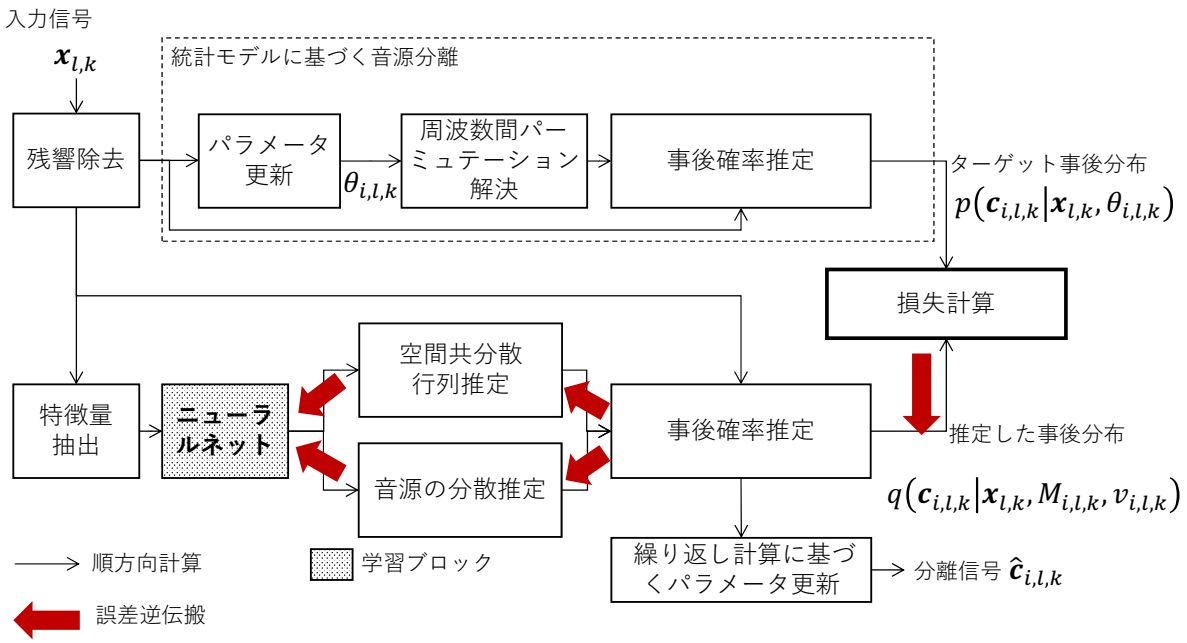

図 4: ブロック構成

6 まとめ

本稿では，空間モデルの影響を深層学習時にも考慮する 4 つの方向性を示した。深層学習に基づく方式の進展が著しいが，音源分離の空間モデルのような物理的な知識を活用可能なシステムにおいては，ニューラルネットワーク単体で全てを学習するのは望ましくなく，物理的な知識と深層学習に基づく方式をどう融合するかが今後の一つの重要な方向性と考えている。

参考文献

[1] Z. Chen, Y. Luo, and N. Mesgarani. Deep attractor network for single-microphone speaker separation. In *ICASSP 2017*, pp. 246–250, March 2017.

[2] P. Common. Independent component analysis, a new concept ? *Signal Processing*, Vol. 36, No. 3, pp. 287–314, April 1994.

[3] L. Drude, D. Hasenklever, and R. Haeb-Umbach. Unsupervised training of a deep clustering model for multichannel blind source separation. In *ICASSP 2019*, pp. 695–699, May 2019.

[4] N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio Speech Lang. Process.*, Vol. 18, No. 7, pp. 1830–1840, 2010.

[5] H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel, and J. Le Roux. Improved mvdr beamforming using single-channel mask prediction networks. In *Interspeech 2016*, pp. 1981–1985, 2016.

[6] J.R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP 2016*, pp. 31–35, 2016.

[7] A. Hiroe. Solution of permutation problem in frequency domain ica using multivariate probability density functions. In *Proceedings ICA*, pp. 601–608, Mar. 2006.

[8] N. Ito, S. Araki, and T. Nakatani. Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing. In *EUSIPCO 2016*, pp. 1153–1157, Aug 2016.

[9] T. Kim, H.T. Attias, S.-Y. Lee, and T.-W. Lee. Independent vector analysis: an extension of ica to multivariate components. In *Proceedings ICA*, pp. 165–172, Mar. 2006.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari. *Determined Blind Source separation with Independent Low-Rank Matrix Analysis*, chapter 6, pp. 125–155. Springer Publishing Company, Incorporated, 2018.

[11] Y. Luo, Z. Chen, and N. Mesgarani. Speaker-independent speech separation with deep attractor network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 4, pp. 787–796, April 2018.

[12] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono. Independent deeply learned matrix analysis for multichannel audio source separation. In *EUSIPCO 2018*, pp. 1557–1561, Sep. 2018.

[13] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi. Deep speech extraction with time-varying spatial filtering guided by desired direction attractor. In *ICASSP 2020*, 2020.

[14] A.A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.*, Vol. 24, No. 9, pp. 1652–1664, 2016.

[15] A.A. Nugraha, A. Liutkus, and E. Vincent. Deep neural network based multichannel audio source separation. In *Audio Source Separation*. Springer, March 2018.

[16] N. Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 189–192, Oct 2011.

[17] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Trans. Audio, Speech, and Language Process.*, Vol. 21, No. 5, pp. 971–982, May 2013.

[18] M. Togami. Multi-channel Itakura Saito distance minimization with deep neural network. In *ICASSP 2019*, pp. 536–540, May 2019.

[19] M. Togami. Spatial constraint on multi-channel deep clustering. In *ICASSP 2019*, pp. 531–535, May 2019.

[20] M. Togami. Joint training of deep neural networks for multi-channel dereverberation and speech source separation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3032–3036, 2020.

[21] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome. Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function. In *ICASSP 2020*, 2020.

[22] E. Tzinis, S. Venkataramani, and P. Smaragdis. Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information. In *ICASSP 2019*, pp. 81–85, May 2019.

[23] Z.Q. Wang, J. Le Roux, and J.R. Hershey. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *ICASSP 2018*, pp. 1–5, 2018.

[24] 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, July 2004.

[25] K Yoshii, R Tomioka, D Mochihashi, and M Goto. Infinite positive semidefinite tensor factorization for source separation of mixture signals. *30th International Conference on Machine Learning, ICML 2013*, pp. 1613–1621, 01 2013.

[26] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva. Multi-microphone neural speech separation for far-field multi-talker speech recognition. In *ICASSP 2018*, pp. 5739–5743, April 2018.

[27] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP 2017*, pp. 241–245, March 2017.