卒業論文

独立深層学習行列分析に基づく拡散性雑音下教師有りリアルタイム多チャネル音声抽出

03-230625 仲西優貴

指導教員 猿渡洋 教授

2025 年 1 月

東京大学工学部計数工学科システム情報コース

概要

本論文では，deep neural network (DNN) に基づく音源モデルを導入した拡散性雑音下でのリアルタイム音声抽出フレームワークを提案する．音声抽出とは，目的音声と雑音が混合した観測信号から目的音声を抽出する技術である．拡散性雑音はあらゆる環境に存在するため，拡散性雑音下での音声抽出技術は音声認識システムや補聴器システムに応用することが可能である．このような応用を行う場合には，音声抽出はリアルタイムに行われる必要がある．

本論文は拡散性雑音下におけるリアルタイム音声抽出において DNN で音源モデルを推定するフレームワークを提案する．本フレームワークはランク制約付き空間共分散行列推定法 (RCSCME) のリアルタイム動作を行う．RCSCME は独立低ランク行列分析 (ILRMA) と呼ばれる手法とランク制約のある空間共分散行列の推定手法を組み合わせた，拡散性雑音下での高い音声抽出性能を持つ手法である．しかし ILRMA には連続的なスペクトルを持つ音声に対する音源の分離性能が限定的であるという問題がある．そこで，音声抽出で用いる ILRMA は分離行列を求めることができれば他の手法で代替可能であることを考慮し，音声抽出の前段として独立深層学習行列分析 (IDLMA) を導入し，DNN による音源モデルの推定を行い音源を分離するリアルタイム音声抽出を提案する．また，単チャネル DNN により抽出された音声から音声の雑音のみの区間を推定し，得られた雑音の空間共分散行列を RCSCME の事前分布としてリアルタイム音声抽出に用いることを提案する．提案手法が従来の ILRMA を前段としたリアルタイム音声抽出よりも音声抽出性能において優れることを実験的に示す．

目次

第 1 章

序論

1.1 本論文の背景

音声抽出とは, 目的となる音声と雑音が混合した観測信号から目的となる音声を抽出する技術である. マイクロホンなどで音を捉える際には目的音声と雑音の混合音を捉えているため, 目的音声を何らかの形で利用するためには混合音から目的音声を抽出する必要がある. 音声抽出技術を活用することで, 例えば会話などにおいて背景雑音を抑制したクリアな声をユーザに提示することができる[1], [2] . 音声認識システムの入力音声に対して音声抽出技術を適用して雑音を抑制することで, より精度の高い音声認識を期待することができる[3].
caption: [Application of speech extraction]) 次に音源分離および音声抽出手法について述べる. 音源分離とは, 複数の音源から到来した音が混合した観測信号からそれぞれの混合前の信号を分離する技術である. したがって目的とする音声を抽出するための音声抽出に音源分離技術を応用することができる. 音源分離問題は, 観測に用いるマイクロホン数が単数 (単チャネル) か複数 (多チャネル) か, および学習を事前情報無しで行う (ブラインド) か事前情報ありで行うかの観点で分類することができる. 単チャネルの場合音響的特徴のみしか用いることができない一方で, 多チャネルの場合は空間的特徴を利用して音源分離を行うことが可能である.

ブラインドでない分離手法としては, Wiener フィルタ[4] やビームフォーマ[5]–[7] がある. Wiener フィルタは最小平均二乗誤差規範により, 目的音源および雑音源のパワースペクトログラムを用いて目的音源のパワースペクトログラムを推定する. ビームフォーマでは, マイクロホンアレイの位置関係や目的音源の到来方向の情報を用いて目的音源を推定する. ただしこれらの手法は目的音源や雑音源の音響的情報や空間的情報を必要とし, 十分でなければ推定精度が低下する恐れがある.

音源信号や収音環境について全て未知であるとする手法はブラインド音源分離 (blind source separation: BSS) と呼ばれる[8]. BSS には事前情報が必要ないため, 様々な状況で利用することが可能である. 単チャネルの BSS の手法としては, 音源のパワースペクトログラムの持つ特徴をモデル化することで分離を行う非負値行列因子分解 (nonnegative matrix factorization: NMF) [9] が提案されている [10] . 一方で多チャネルの場合には空間的な情報を用いてより高精度に分

離をすることができる．多チャネルの BSS の手法には，時間領域における瞬時混合を仮定した独立成分分析 (independent component analysis: ICA) [11]–[13] がある．しかしこの手法は畳み込み混合された信号に直接適用することができない．そこで短時間 Fourier 変換 (short-time Fourier transform: STFT) で時間周波数領域に変換したときに残響長が STFT の窓長よりも十分短ければ周波数ごとの瞬時混合とみなせることを用いて, 畳み込み混合信号に ICA を適用する手法である周波数領域独立成分分析 (frequency-domain ICA: FDICA) [14]–[16] が提案されている. FDICA は各周波数ビンごとの統計的独立性のみに基づくため, 各出力信号の大きさが定まらない問題 (スケール問題) や周波数ビンの順番が定まらない問題 (パーミュテーション問題) が生じる．スケール問題については，複数のマイクロホンから基準となるマイクロホンを選択し，各分離信号の総和がそのマイクロホンの観測信号に一致するようスケールを定める projection back (PB) 法 [17] が提案されている．パーミュテーションを解決して音源分離を行うための手法として，FDICA を改良した独立ベクトル分析 (IVA) [18]–[20] が提案されている．さらに IVA の音源のパワースペクトログラムを NMF で表現する独立低ランク行列分析 (independent low-rank matrix analysis: ILRMA) [21], [22] により，さらに高精度の音源分離が達成されている．これらの手法は分離音の歪みを抑制して分離することが可能であるが, 背景雑音として全方位から到来する拡散性雑音を完全に除去することは不可能であり, 分離された目的音に雑音成分が残留してしまう [23] ．複数音源の混合音から 1 つの独立な成分を抽出する独立ベクトル抽出 (independent vector extraction: IVE) [24] が提案されているが，モデル自体は FDICA と同様に点音源仮定に基づいているため，拡散性を有する雑音下での分離性能は限定的である．また，FDICA や IVA，ILRMA などの線形時不変なフィルタでの分離を前段処理として実行し，その出力に対し Wiener フィルタやスペクトル減算 [25] 等の単チャネル手法を適用して雑音を抑制する手法が数多く提案されている [26]–[35] が，これらの手法は厳密には統計的枠組みに基づいておらず精度は限定的である [36].

その他の多チャネル BSS 手法としては，各音源の空間特徴を表す空間共分散行列 (spatial covariance matrix: SCM) を用いるフルランク SCM モデルが提案されている [37]．フルランク SCM モデルは線形時不変な分離フィルタを用いる分離手法の空間モデルと比較して自由度が高く，拡散性の音源を適切にモデル化できると考えられる．また，各音源のパワースペクトログラムを NMF によりモデル化した多チャネル NMF (multichannel NMF: MNMF) [38], [39] や，モデル近似により MNMF のパラメータ数を減らして計算コストを削減した FastNMF [40], [41] が提案されているが, 計算コストが高いことや初期値に頑健でないこと, 分離音が歪んでしまうことなどの問題点がある[21] .

拡散性雑音下での音声抽出において, 拡散性雑音をモデル化し高い音声抽出性能を持つブラインド音声抽出手法として，ランク制約付き空間共分散行列推定法 (rank-constrained SCM estimation: RCSCME) [42] が提案されている．この手法の前段処理として ILRMA などの音源間の独立性に基づく線形時不変分離フィルタによる音源分離を行う．拡散性雑音下で M 個のマイクロホンを用いてこの前段処理を実行すると, 拡散性雑音のみからなる $M-1$ 個の分離音と, 目的音声と拡散性雑音からなる 1 個の分離音が得られる [34] ．目的音声の含まれない $M-1$ 個の分

離音から，雑音 SCM のランク $M-1$ 成分を推定できる．また，点音源である目的音声を正確に打ち消す $M-1$ 個の雑音の分離フィルタから, 目的音声のステアリングベクトルを高精度に推定することができる．さらに SCM の残りランク 1 成分と目的音声や雑音に関する残りのパラメータを推定し, これらを多チャネル Wiener フィルタに適用する．RCSCME は MNMF や FastNMF よりも計算コストが低く，初期値に頑健であり，高い音声抽出性能を実現している．さらに RCSCME は高速に動作させることが可能であるため，前段処理として ILRMA を用いて後段で RCSCME を実行するリアルタイム音声抽出システム [43] が提案され，リアルタイムに高い精度の音声抽出が実現することが確認されている．

BSS は事前情報を一切未知とする一方で, 近年は膨大な音のデータベースを活用して音源分離を行う教師あり音源分離の枠組みが注目されている. 音のデータベースとは観測信号に含まれる音源と同じ種類の異なる音源を集めたものである．例えば，ピアノのデータベースには様々な楽曲を演奏したピアノの音が収録される. このような蓄積されたデータベースを利用して音源信号の周波数構造のような特徴を学習することできる．この枠組みにおいて deep neural network (DNN) は単チャネル分離 [44]–[47] と多チャネル分離 [48]–[51] の両方で高い性能を達成している．DNN では十分な録音データ数が確保されている場合，その時間周波数構造を効果的にモデル化することができる．一方で空間モデルについては．部屋の形状や観測マイクロホンの位置，音源位置の微小な変化に多大な影響を受けるため，DNN により汎用的な空間モデルを学習することは困難である．したがって，学習済み DNN モデルによる音源モデル推論とブラインドな空間モデル推定を組み合わせるのは合理的である. フルランク SCM の枠組みに DNN による音源モデルを導入した多チャネル音源分離の枠組み [44] が提案されているが, 空間共分散の推定に大きな計算コストがかかる上に空間パラメータの推定は困難であり精度が十分でない[21]．より空間モデルのパラメータが少なく軽量で安定した手法として，独立深層学習行列分析 (independent deeply learned matrix analysis: IDLMA) [52] が提案されている．IDLMA は ICA による分離行列のブラインド推定と DNN による音源モデルの教師あり推論を組み合わせた分離手法であり，ILRMA の NMF 音源モデルよりも柔軟な音源モデルであるため NMF でモデル化が困難な非低ランクな音源信号も適切にモデル化することが期待される．IDLMA は元々音楽分離に提案された手法であるが，音声強調に IDLMA を利用すること [53] も提案され，雑音環境での音声強調において優れた音声強調性能を示すことが確認されている．

1.2 本論文の目的

1.3 本論文の構成

第 2 章では，従来手法として ILRMA，RCSCME，リアルタイム音声抽出，および独立深層学習行列分析を用いた音声抽出について述べる．第 3 章では，拡散性雑音下での独立深層学習行列分

析を用いたリアルタイム音声抽出のフレームワークを提案し，単チャネル音声抽出 DNN の出力に基づく雑音の空間共分散行列の事前分布の RCSCME への導入について述べる．第 4 章では，音声抽出性能の実験的評価により提案手法の有効性を確認する．第5章では本論文の結論を述べる．

第 2 章

従来手法

2.1 はじめに

本章では，提案手法について説明する上で必要となる従来手法を説明する．2.2 節では点音源から到来する音に対するブラインド音源分離手法である ILRMA について説明する．2.3 節では拡散性雑音下での音声強調手法である RCSCME について説明する．続いて 2.4 節では，ILRMA を用いたリアルタイム音声抽出フレームワークの先行研究について説明する．2.5 節では，RCSCME の前段として独立深層学習行列分析を用いた音声強調手法について説明する．最後に 2.6 節で本章のまとめを述べる．

2.2 ILRMA

本節では，音源は点音源であり残響時間が STFT の窓長よりも十分に長いことを仮定する。

2.3 RCSCME

2.4 拡散性雑音下でのリアルタイム音声抽出

2.5 独立深層学習行列分析に基づく音声抽出

2.6 まとめ

第 3 章

独立深層学習行列分析に基づくリアルタイム音声抽出

3.1 はじめに

3.2 DNN を用いた音源モデル推定

3.3 目的音声欠損に対する補完処理

3.4 雑音自己教師あり空間共分散行列の推定

3.5 雑音自己教師あり空間共分散行列を事前分布とするリアルタイムRCSCME

3.6 まとめ

第 4 章

実験的評価

4.1 はじめに

4.2 実験条件

4.3 各パラメータを変化させた時の音声抽出性能の変化

4.4 各提案手法の音声抽出性能の比較

4.5 まとめ

第５章

結論

謝辞

参考文献

[1] F. Mustière, M. Bouchard, H. Najaf-Zadeh, R. Pichevar, L. Thibault, and H. Saruwatari, “Design of multichannel frequency domain statistical-based enhancement systems preserving spatial cues via spectral distances minimization,” Signal Processing, vol. 93, pp. 321–325, 2013.

[2] M. Une, Y. Kubo, N. Takamune, D. Kitamura, H. Saruwatari, and S. Makino, “Evaluation of multichannel hearing aid system by rank-constrained spatial covariance matrix estimation,” in Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1874–1879, 2019.

[3] 高橋祐, 猿渡洋, and 鹿野清宏, “独立成分分析を導入した空間的サブトラクションア レーによるハンズフリー音声認識システムの開発,” 電子情報通信学会論文誌 D, vol. 93, no. 3, pp. 312–325, 2010.

[4] N. Wiener, “Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications,” The MIT press, 1964.

[5] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” Proceedings of the IEEE, vol. 57, no. 8, pp. 1408–1418, 1969.

[6] O. Frost, “An algorithm for linearly constrained adaptive array processing,” Proceedings of the IEEE, vol. 60, no. 8, pp. 926–935, 1982.

[7] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming,” IEEE Transactions on Antennas and Propagation, vol. 30, no. 1, pp. 27–34, 1982.

[8] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, “A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF,” APSIPA Transactions on Signal and Information Processing, vol. 8, no. e12, pp. 1–14, 2019.

[9] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[10] P. Smaragdis, B. Raj, and M. Shashanka, “Supervised and semi-supervised separation of sounds from single-channel mixtures,” in Proceedings of Independent Component Analysis and Signal Separation (ICA), pp. 414–421, 2007.

[11] P. Comon, “Independent component analysis, a new concept?,” Signal Processing, vol. 36, no. 3, pp. 287–314, 1994.

[12] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” Neural Computation, vol. 7, no. 6, pp. 1129– 1159, 1995.

[13] S. Amari A. Cichocki and H. Yang, “A new learning algorithm for blind signal separation,” Advances in Neural Information Processing Systems, 1995.

[14] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” Neurocomputing, vol. 22, no. 1-3, pp. 21–34, 1998.

[15] S. Ikeda and N. Murata, “A method of ica in time frequency domain,” in Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation (ICA), pp. 365–371, 1999.

[16] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, “Blind source separation based on a fast-convergence algorithm combining ICA and beamforming,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 2, pp. 666–67, 2006.

[17] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” Neurocomputing, vol. 41, no. 1-4, pp. 1–24, 2001.

[18] A. Hiroe, “Solution of permutation problem in frequency domain ica, using multivariate probability density functions,” in Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation (ICA), pp. 601–608, 2006.

[19] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp. 70–79, 2007.

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

[21] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 9, pp. 1626–1641, 2016.

[22] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation with independent low-rank matrix analysis,” Audio source separation, pp. 125–155, 2018.

[23] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, “Equivalence between frequency-domain blind source separation and frequency- domain adaptive beamforming for convolutive mixtures,” EURASIP Journal on Advances in Signal Processing, vol. 2003, no. 11, pp. 1–10, 2003.

[24] Z. Koldovský and P. Tichavský, “Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence,” IEEE Transactions on Signal Processing, vol. 67, no. 4, pp. 1050–1064, 2019.

[25] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113–120, 1979.

[26] H.-Y. Kim, F. Asano, Y. Suzuki, and T. Sone, “Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer,” IEICE Trans- actions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 79, no. 12, pp. 2151–2158, 1996.

[27] M. Mizumachi and M. Akagi, “Noise reduction by paired-microphones using spectral subtraction,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 1001–1004, 1998.

[28] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, “Speech enhancement using nonlinear microphone array based on noise adaptive complementary beamforming,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 83, no. 5, pp. 866–876, 2000.

[29] J. Meyer and K. Simmer, “Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction,” in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 1167–1170, 1997.

[30] I. A. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence,” IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 709– 716, 2003.

[31] D. Kolossa and R. Orglmeister, “Nonlinear postprocessing for blind speech separation,” in Proceedings of Independent Component Analysis and Blind Signal Separation (ICA), pp. 832–839, 2004.

[32] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Blind extraction of dominant target sources using ica and time-frequency masking,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2165–2173, 2006.

[33] Y. Mori et al., “Blind separation of acoustic signals combining SIMO- model-based independent component analysis and binary masking,” EURASIP Jour- nal on Advances in Signal Processing, vol. 2006, pp. 1–17, 2006.

[34] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, “Blind spatial subtraction array for speech enhancement in noisy environment,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 4, pp. 650–664, 2009.

[35] R. Miyazaki, H. Saruwatari, R. Wakisaka, K. Shikano, and T. Takatani, “Theoretical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction,” in Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), pp. 19–24, 2011.

[36] 猿渡洋, “最近の音声処理に用いられるマイクロホンアレー技術,” 日本音響学会誌, vol. 66, no. 10, pp. 521–526, 2010.

[37] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1830–1840, 2010.

[38] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 550–563, 2010.

[39] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Multichannel extensions of nonnegative matrix factorization with complex-valued data,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 5, pp. 971–982, 2013.

[40] N. Ito and T. Nakatani, “Fastmnmf: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization,” in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 371–375, 2019.

[41] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, “Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices,” in Proceedings of European Signal Processing Conference (EUSIPCO), pp. 1–5, 2019.

[42] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, “Blind speech extraction based on rank-constrained spatial covariance matrix estimation with multivariate generalized gaussian distribution,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1948–1963, 2020.

[43] Y. Ishikawa, K. Konaka, T. Nakamura, N. Takamune, and H. Saruwatari, “Real-time Speech Extraction Using Spatially Regularized Independent Low-rank Matrix Analysis and Rank-constrained Spatial Covariance Matrix Estimation,” in Proceedings of HSCMA, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.12477

[44] E. M. Grais, M. U. Sen, and H. Erdogan, “Deep neural networks for single channel source separation,” in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3734–3738, 2014.

[45] S. Uhlich, F. Giron, and Y. Mitsufuji, “Deep neural network based instrument extraction from music,” in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2135–2139, 2015.

[46] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep U-Net convolutional networks,” in Proceedings of International Society for Music Information Retrieval (ISMIR), 2017.

[47] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[48] A. 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.

[49] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, “Exploring multi-channel features for denoising-autoencoder-based speech enhancement,” in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 116–120, 2015.

[50] Y.-H. Tu, J. Du, L. Sun, and C.-H. Lee, “LSTM-based iterative mask estimation and post-processing for multi-channel speech enhancement,” in Proceedings of Asia- Pacific Signal and Information Processing Association Annual Summit and Confer- ence (APSIPA ASC), 2017.

[51] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, “Deep learning based speech beamforming,” in Proceedings of International Confer- ence on Acoustics, Speech and Signal Processing (ICASSP), pp. 5389–5393, 2018.

[52] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, and S. Takamichi, “Independent Deeply Learned Matrix Analysis for Determined Audio Source Separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 10, pp. 1601-1615, 2019.

[53] S. Misawa, N. Takamune, T. Nakamura, D. Kitamura, H. Saruwatari, and M. Une, “Speech Enhancement by Noise Self-Supervised Rank-Constrained Spatial Covari-

ance Matrix Estimation via Independent Deeply Learned Matrix Analysis," in Proceedings of Asia- Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021.