E-004

DNN を利用した音源モデルが IDLMA の性能に与える影響の調査・検討

衛藤　吉彦 †　吉村　宏紀 †,††　松村　寿枝 †††　清水　忠昭 †,††　西山　正志 †,††
岩井　儀雄 †,††

1　はじめに

われわれの生活には様々な音が存在している．それらから任意の音だけを分離・抽出する技術が求められている．例えば，人の音声を認識するアシスタントアプリケーションやスマートスピーカーでは話者だけを抽出する必要がある．自動採譜を行うアプリケーションでは複数鳴っている楽器音をそれぞれ抽出する必要がある．このように複数の音がなっている状況下で任意の音だけを取り出す技術のことを音源分離という．　音源分離は様々な分野に細分化されるが，本稿では音源位置や混合系が未知の条件で，観測された音源のみから混合前の真の音源を推定するブラインド音源分離（Blind Source Separation: BSS）を取り扱う．

BSS は真の音源から収録マイクまでの空間情報を推定し分離を行う手法と，音源情報によって真の音源の推定し分離を行う手法で二つに大別される．空間情報とは音源位置，録音機器の位置，録音空間の広さ，形状，残響のことであり，これらをモデル化したものを空間モデルと呼ぶ．音源情報とは対象となる真の音源のことであり，その特徴である音色や強さ，長さなどの要素をモデル化したものを音源モデルと呼ぶ．

空間モデルと音源モデル，両方を融合させた ILRMA(Independent Low-Rank Matrix Analysis: ILRMA) [1] がある．この手法は従来の手法よりも高精度な分離を可能にしている．また，近年ではディープニューラルネットワーク (Deep Neural Networks: DNN) を用いた音源分離手法として ILRMA と DNN による音源モデル推定を組み合わせた手法である独立深層学習行列分析 (Independent Deeply Learned Matrix Analysis: IDLMA) [2] が提案されている．これは分離対象音源の学習データが得られるという条件下で，「女性音声」や「ギター」等の分離対象と同じ音源属性を事前に学習しておくことで高精度な分離を可能にする手法である．これによって各音源の様々な混合比での分離を行うことができる．

しかし，IDLMA の DNN の学習の際に利用している音楽データは 50 曲分のみである．また，音源 1 フレームに対しては 1 パターンでの音量倍率でしか学習されていない．さらに，隠れ層の層数 4 層，各層のユニット数 1024 で固定し，窓長を変化させて比較のみを行っている．本研究では新たに別パターンでの音量倍率での学習をし，更に隠れ層とユニット数を変化させて学習させる．

これにより，本稿では従来の IDLMA で比較が行われていなかった学習データ数による精度の比較，および層数・ユニット数での精度の比較を行い，IDLMA での分離に最適な DNN を調査・検討することを目的とする．

2　関連手法

2.1　空間モデルと音源モデルの両方を用いた手法

2.1.1　ILRMA

複数マイクがある場合に各マイクでの音源の音量差や空間モデルを推定し分離する IVA と音源モデルを推定し分離する NMF を組み合わせた手法として独立低ランク行列分析 (Independent Low-Rank Matrix Analysis: ILRMA) [1] が提案されている．空間モデルの推定には IVA [3] を用いているため ILRMA は優決定条件の手法となる．IVA の音源の生成モデルがベクトルであったのに対し，ILRMA では低ランク行列へと拡張したことによって，効率的な最適化とより高精度な音源分離を達成している．ILRMA は分離行列と音源モデルを反復更新することで分離を進めていく．

2.1.2　IDLMA

ILRMA では音源モデルの推定を NMF で行っていたが，これを DNN で行うように変えた手法として IDLMA [2] が提案された．IDLMA は ILRMA と同様に複素ガウス分布に基づいて音源モデル及び空間モデルを推定する．このとき混合音源から n 番目の音源の振幅 (モデルパワースペクトログラム) を推定する DNN をあらかじめ学習させておく．例えばボーカルとベースが混ざった混合音源がある場合を仮定すると，混合音源からボーカルを強調して出力する DNN とベースを強調して出力する DNN の二つを学習することで分離を行う．各音源の DNN は，分離対象となる音源データとその他の音源信号の学習データを任意の比率で混合した信号の振幅値を入力とし，混合前の対象音源の音源信号の振幅値を予想するように学習する．このように n 個の DNN を学習させることで適切な音源モデルを構築することができ，より音源モデル及び分離行列の推定に活用できる．

ILRMA と大まかな流れは同じだが音源モデルの推定を

† 鳥取大学大学院持続性社会創成科学研究科
Graduate School of Sustinability Science, Tottori University
†† 鳥取大学工学部附属クロス情報科学研究センター
Cross-Informatics Research Center, Tottori University
††† 奈良高専
National Instuitute od Technology(Kosen), Nara College

NMF から DNN での推定に変更している．IDLMA は対象の音源のみを強調する DNN を教師あり音源モデルとして活用しているが，空間モデルは依然としてブラインドに推定できるため，汎化性能の高いアルゴリズムになっている．

3　本研究の狙い

IDLMA で検討されている DNN が隠れ層 4 層，隠れ層のユニット数が 1024 で隠れ層と出力層のユニットに ReLu を用いた全結合型 NN のみになっており，データ数も音楽データセット 50 曲分を学習しているのみである．STFT の窓長を変更しての比較は行われているが，学習データ数，DNN の層数，ユニット数を変更しての比較を行われていない．そこで本研究では IDLMA の DNN の学習データ数，層数，ユニット数を変更して比較を行い，音源分離の精度へ与える影響を調査することを狙いとする．

まず，真の音源に各フレームおきに一様乱数をかけ，混合音源を作り n 番目の音源を学習する DNN_n へ入力する．これを b 回繰り替えす．これによって様々な混合比で学習することができ，b 倍のデータ数による学習が可能になる．b 倍のデータを学習した DNN を $\mathrm{DNN}_{b,n}$ とする．$\mathrm{DNN}_{b,n}$ を適用した IDLMA の出力の平均 SDR を取り比較を行う．次に DNN の隠れ層を $(c = 3, 4, 5, 6, 7)$ 層 に変更した DNN を用い比較を行い，最後にユニット数を $(e = 512, 1024, 2048)$ へと変更し同様に比較を行う．学習データを増やすことで汎化性を高め様々な混合音源へと対応できると考える．また，先述の隠れ層 4 層，隠れ層のユニット数 1024 よりも複雑な DNN によって更に SDR を高めることが出来ると考える．

4　評価実験

4.1　データセットと評価指標

学習及び評価に使うデータセットは SiSEC2016 [4] の音楽データセット DSD100 の 100 曲分のボーカル (Vo)，ベース (Ba) の 2 音源のデータを利用した．DSD100 中の Dev データ 50 曲を DNN 音源モデルの学習データとし，Test データ 50 曲を評価データとした．DNN 入力時はサンプリング定数 8000，窓長 4096 ms で SFTF を行った．

音源分離性能の客観評価尺度には Signal-to-DistortionRatio (SDR) [5] [6] を用いた．SDR は信号対歪み比とも呼ばれ，分離音源のクオリティを示す音源分離分野で使われる指標である．SDR の値が高いほど，分離性能を良いことを表している．

図 1，2，3 では 14dB の少し上に黒いラインがある．これは参考文献である [2] と同一条件の学習データ数 $b = 1$ 倍，隠れ層数 $c = 4$，ユニット数 $e = 1024$, 音源数 $n = 2$ で学習を行った DNN を用いた IDLMA の SDR である．このラインと比較することで従来手法から本実験でどの程度改善されたかを示す．学習データはフレーム毎で混合倍率を変化させ，様々な混合音源に対応させた．評価時は DSD100 の Test データ 50 曲を全曲で同じ混合倍率を使って混合音源を作成した．これを IDLMA に入力し，最終的に出力された分離音源の SDR

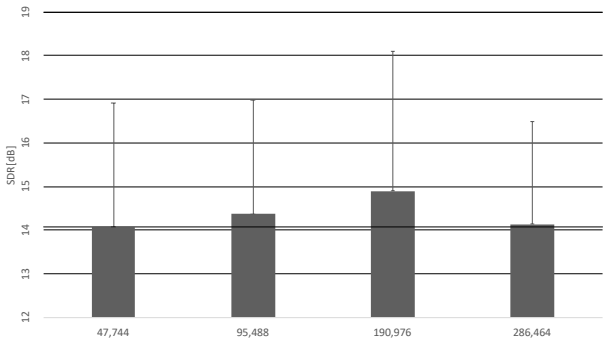

図 1: 学習データ別　100 epoch 時の分離音源の平均 SDR

上位 5 曲の 100 epoch 時での平均 SDR を示した．

4.2　学習データ数を増やして学習した IDLMA の比較

まず，IDLMA の DNN の性能が学習データ数にどの程度影響を受けるか調査するために学習データを増やして実験を行った．

この実験では隠れ層数 $c = 4$, 各隠れ層のユニット数 $e = 1024$, 音源数 $n = 2$ で固定し，学習データ数 $b = (1, 2, 4)$ 倍で変更し比較した．また $(b = 1, 2, 4)$ としたとき，フレーム数は $(J = 47744, 95488, 190976, 286, 464)$ フレームとなる．

100 epoch での平均 SDR を示した図 1 に示した．図 1 を見ると学習データ数 $J = 190,976$ のときが最も良い SDR を示していることが分かった．しかし，学習データ数によってエラーバーの大きな変動はなく，安定した SDR を得ることはできなかった．以降の実験では学習データ数は全て $J = 190,976$ で行った．学習データ数が b=6 倍となった時，SDR が大きく低下し，過学習が発生したと思われる．これは同一の学習データの音量倍率の変更を繰り返した結果に起こったものと思われる．新たに別の学習データを増やせば過学習を避け，更に学習をすることができると考えられる．

4.3　隠れ層数を変更して学習した IDLMA の比較

次に DNN の隠れ層を変更し学習，IDLMA へ実装して実験を行った．パラメーターは学習データ数 $b = 4$ 倍 ($J = 190976$ フレーム), 各隠れ層のユニット数 $e = 1024$, 音源数 $n = 2$ は固定, 隠れ層数 $(c = 3, 4, 5, 6)$ と変更し比較した．

図 2 は更新 100epoch での平均 SDR を示したものである．図 2 より僅かながら 5 層より 6 層が最も良い平均 SDR を示すことが分かった．なお標準偏差は $(c = 3, 4, 5, 6, 7)$ に対して $(2.757, 3.188, 2.505, 2.847, 5.2571)$ となり，$c = 5$ のときが最も安定していることが分かった．3 層から 6 層まで層数が増えるごとに SDR が上昇し続けたが，7 層の DNN では過度に音源モデルを学習してしまった結果，汎化性の低下を招いてしまっていると考えられる．この結果を受けて次の実験では隠れ層を 6 層で固定した．

4.4　ユニット数を変更して学習した IDLMA の比較

最後に DNN の各隠れ層のユニット数を変更し，IDLMA へ実装して実験を行う．学習データ，評価データにおいては先の

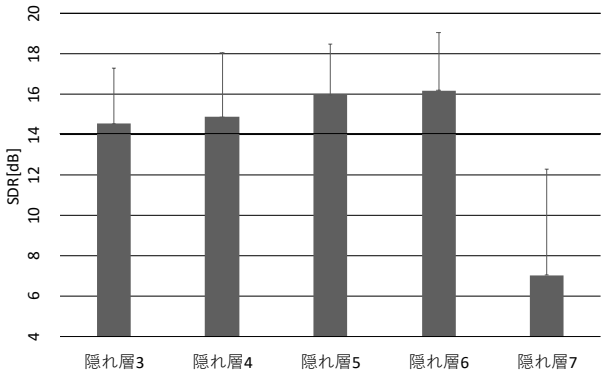

図 2: DNN の層数別　100 epoch 時の分離音源の平均 SDR

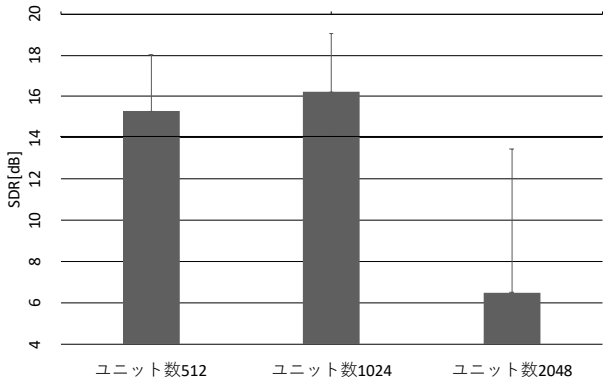

図 3: DNN のユニット数別　100 epoch 時の分離音源の平均 SDR

実験と同様である．パラメーターは学習データ数 $b = 4$ ($J = 190976$ フレーム) 倍, 各音源数 $n = 2$, 隠れ層数 $c = 6$ に固定した．また，隠れ層のユニット数 ($e = 512, 1024, 2048$) に変更して比較した．

図 3 は 100 epoch での平均 SDR を示したものである．図 3 よりユニット数 $e = 1024$ の際が最も良い SDR を示すことが分かった

これまでの結果より IDLMA に適した全結合型 NN は学習データ数 $b = 4$ ($J = 190976$ フレーム) 倍, 隠れ層数 $c = 6$, 各隠れ層のユニット数 $e = 1024$ となった．e=2048 の SDR が低下してしまった結果は 4.2 と同様に過度に音源モデルを学習してしまったからと考えられる．

5　まとめ

本稿では従来の IDLMA では行われていなかった DNN の学習データ数，層数，ユニット数を変更して IDLMA の精度比較を行った．その結果，DNN が全結合型 NN の場合，学習データ数 $b = 4$ ($J = 190976$) 倍, 隠れ層数 $c = 6$, 各隠れ層のユニット数 $e = 1024$ が最も高精度であり，学習データと隠れ層を増やすほど精度が向上する傾向が見られた．また，図 3 で確認できるように最も良い結果である学習データ数 $b = 4$ ($J = 190976$ フレーム) 倍, 隠れ層数 $c = 6$, 各隠れ層のユニット数 $e = 1024$ 時と，従来手法である学習データ数 $b = 1$ 倍，隠れ層数 $c = 4$，ユニット数 $e = 1024$ 時を比較すると，約 2dB 程度の改善しかなされていない．根本的に全結合型 DNN では精度の限界を迎えている可能性があるので，今後の課題として全結合型 NN 以外の DNN による IDLMA の精度検証などが挙げられる．

謝辞　本研究は JSPS 科研費 20K11886 の助成を受けたものである．

参考文献

[1] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Determined blind source separation with independent low-rank matrix analysis. In *Audio source separation*, pp. 125–155. Springer, 2018.

[2] Shinichi Mogami, Hayato Sumino, Daichi Kitamura, Norihiro Takamune, Shinnosuke Takamichi, Hiroshi Saruwatari, and Nobutaka Ono. Independent deeply learned matrix analysis for multichannel audio source separation. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1557–1561. IEEE, 2018.

[3] Nobutaka 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. IEEE, 2011.

[4] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave. The 2016 signal separation evaluation campaign. In Petr Tichavský, Massoud Babaie-Zadeh, Olivier J.J. Michel, and Nadège Thirion-Moreau, editors, *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, pp. 323–332, Cham, 2017. Springer International Publishing.

[5] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, Vol. 14, No. 4, pp. 1462–1469, 2006.

[6] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.