音源分離のための深層学習に基づく音響帯域拡張*

☆渡辺瑠伊, 北村大地 (香川高専)

1 はじめに

ブラインド音源分離 (blind source separation: BSS) とは，複数の音源が混合した観測信号から，混合系に関する事前情報を用いることなく混合前の音源信号を推定する技術である．BSS の手法として，独立ベクトル分析 (independent vector analysis: IVA) [1] や独立低ランク行列分析 (independent low-rank matrix analysis: ILRMA) [2, 3], 多チャネル非負値行列因子分解 (multichannel nonnegative matrix factorization: MNMF) [4] などが提案されている．これらの手法はいずれも，混合信号を短時間フーリエ変換（short-time Fourier transform: STFT）して得られる時間周波数行列（スペクトログラム）に対して，周波数毎の時不変混合系または時不変分離系を推定する技術である．比較的高品質な BSS が可能である反面，パラメータ推定には逆行列演算等を含む反復計算が必要であり，実際の応用においては計算時間が問題になることが多い．

一方，音響帯域拡張 [5–7] とは，対象となる音源の低周波帯域から高周波帯域を復元するような技術である．近年では深層学習 (deep neural networks: DNN) に基づく手法 [8, 9] が数多く提案されている．DNN に基づく手法では通常，帯域拡張の対象となる音響信号のスペクトログラム (あるいはその振幅やパワー) を DNN に入力し，帯域拡張された音響信号のスペクトログラムを予測するようなモデルを学習する．音響帯域拡張は，ナイキスト周波数以上の成分を外挿することとなるため，高精度な推定は通常難しい問題である．

本稿では，より効率的な音源分離を目的とし，先に挙げた周波数領域の BSS 手法と DNN に基づく音響帯域拡張を組み合わせる新しい音源分離フレームワークを提案する．また，実験を通して，提案手法の妥当性について調査する．

2 提案手法

2.1 動機

本稿で提案する音源分離フレームワークの概要を Fig. 1 に示す．提案手法では，前段として，混合信号の低周波帯域のみに BSS を適用し，分離信号の低周波帯域を推定する．次に，後段として，DNN に基づく音響帯域拡張によって，分離信号の高周波帯域を予測し，分離信号の全周波数成分を得る．この時，後段の DNN に基づく音響帯域拡張では，既存の音響帯域拡張とは異なり，混合信号の高周波帯域を追加の入力情報として用いることができる．この情報は音響帯域拡張の外挿問題に対して強力な手がかりであり，これによって分離信号の高周波帯域の高精度な予測が実現できる．あるいは，後段の DNN は，混合信号の高周波帯域の音源分離器として解釈することもできるが，その場合にも，あらかじめ分離されている低周波帯域の情報が強力な手がかりとなる．結果として，BSS で分離する周波数成分を減らすことができ，また DNN による予測を専用のエッジコンピューティングデバイス等で高速化できるならば，全体的な計算時間の短縮ができる．また，マイクロホンアレイを用いる BSS では，空間エイリアシングにより高周波数帯域において分離が困難な帯域が存在する問題があるが，提案フレームワークではこれを解決できる可能性がある．

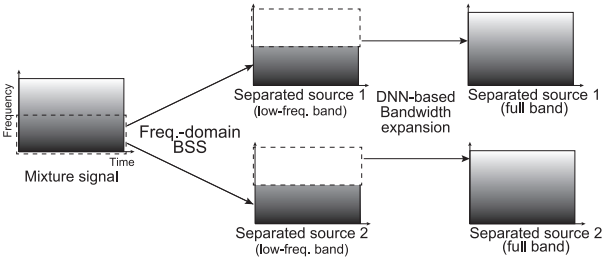

Fig. 1 Proposed framework.

なお，本稿では音源数が 2 個の状況のみを対象とするが，次節以降で説明する DNN の学習法を拡張することで，任意の音源数にも適用できる．

2.2 定式化

本稿では，混合信号と 2 つの音源信号の計 3 種類の信号を取り扱う．混合音源及び 2 つの音源信号に対して STFT を適用し，要素毎に絶対値をとった振幅スペクトログラムをそれぞれ $\boldsymbol{M} \in \mathbb{R}_{\geq 0}^{I \times J}$，$\boldsymbol{S}_1 \in \mathbb{R}_{\geq 0}^{I \times J}$，及び $\boldsymbol{S}_2 \in \mathbb{R}_{\geq 0}^{I \times J}$ とおく．この時，$i = 1, ..., I$ 及び $j = 1, ..., J$ はそれぞれ全周波帯域の周波数ビン及び時間フレームのインデクスである．また，高周波帯域と低周波帯域の境界となる周波数ビンインデクスを I'（但し $1 \leq I' < I$）と定義し，DNN は $I'+1 \sim I$ の範囲の周波数ビンにおける分離信号（高周波帯域）$\boldsymbol{S}_1^{(\mathrm{H})} \in \mathbb{R}_{\geq 0}^{(I-I') \times J}$, $\boldsymbol{S}_2^{(\mathrm{H})} \in \mathbb{R}_{\geq 0}^{(I-I') \times J}$ を推定する．また，混合信号の低周波帯域及び高周波帯域をそれぞれ $\boldsymbol{M}^{(\mathrm{L})} \in \mathbb{R}_{\geq 0}^{I' \times J}$ 及び $\boldsymbol{M}^{(\mathrm{H})} \in \mathbb{R}_{\geq 0}^{(I-I') \times J}$ とおき，2 つの音源信号の低周波帯域を $\boldsymbol{S}_n^{(\mathrm{L})} \in \mathbb{R}_{\geq 0}^{I' \times J}$ とおく．ここで，$n = 1, 2$ は音源のインデクスである．従って，$\boldsymbol{M}^{(\mathrm{L})}$ と $\boldsymbol{M}^{(\mathrm{H})}$ の 2 つの行列を行方向に結合すると，$\boldsymbol{M}$ に一致する．$\boldsymbol{S}_n^{(\mathrm{L})}$ と $\boldsymbol{S}_n^{(\mathrm{H})}$ についても同様である．

2.3 DNN の入力情報

提案手法の DNN は，分離信号の高周波帯域の予測を時間フレーム毎に行うモデルである．DNN の入力ベクトル及び予測ベクトルを Fig. 2 に示す．まず，時間フレーム j における混合信号の高周波帯域及び分離信号の低周波帯域を次式で表す．

$$\boldsymbol{m}_j^{(\mathrm{H})} = \left(m_{1j}^{(\mathrm{H})}, m_{2j}^{(\mathrm{H})}, \cdots, m_{(I-I')j}^{(\mathrm{H})}\right)^{\mathrm{T}} \quad (1)$$

$$\boldsymbol{s}_{n_j}^{(\mathrm{L})} = \left(s_{n_1 j}^{(\mathrm{L})}, s_{n_2 j}^{(\mathrm{L})}, \cdots, s_{n_{I'} j}^{(\mathrm{L})}\right)^{\mathrm{T}} \quad (2)$$

*Bandwidth expansion based on deep neural networks for audio source separation by Rui Watanabe (NIT Kagawa) and Daichi Kitamura (NIT Kagawa).

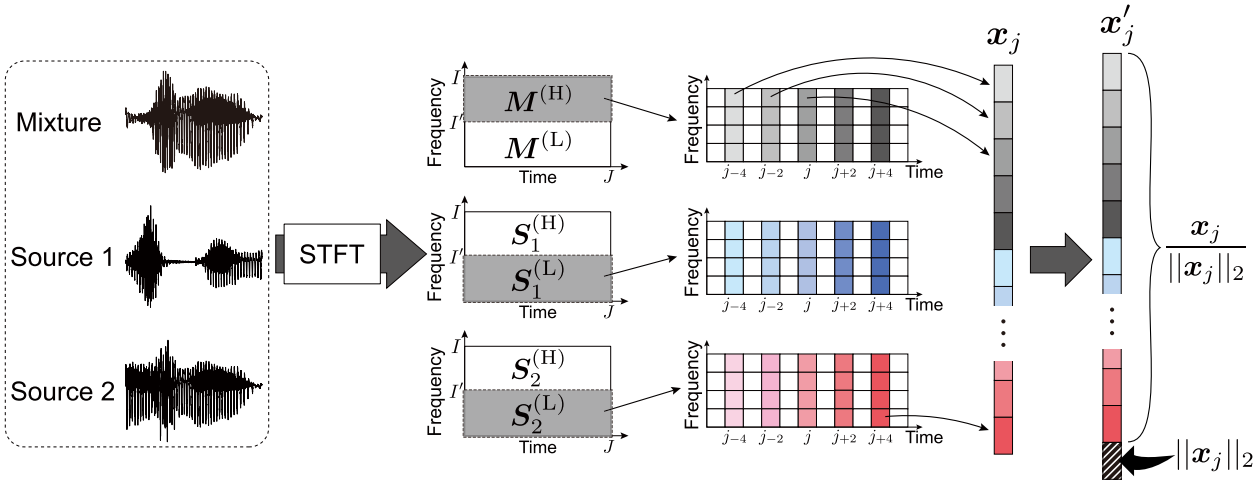

Fig. 2　Pre-processing of input vector for DNN.

ここで，$m_{ij}^{(\mathrm{H})}$ は $\boldsymbol{M}^{(\mathrm{H})}$ の ij 要素であり，$s_{n_{ij}}^{(\mathrm{L})}$ は，$\boldsymbol{S}_n^{(\mathrm{L})}$ の ij 要素を表す．即ち，$\boldsymbol{m}_j^{(\mathrm{H})}$ 及び $\boldsymbol{s}_{n_j}^{(\mathrm{L})}$ はそれぞれ $\boldsymbol{M}^{(\mathrm{H})}$ 及び $\boldsymbol{S}_n^{(\mathrm{L})}$ の時間フレーム j 近傍の周波数ベクトルを結合したベクトルである．時間フレーム j に対する高周波帯域を予測する際，DNN の入力として与える情報は $j, j\pm2, j\pm4$ のように j 近傍の時間フレームの列ベクトル (式 (1) 及び式 (2)) を結合したベクトルとする[1]．これを $\boldsymbol{x}_j$ とおくと，次式のように構成される．

$$\boldsymbol{x}_j = \Big(\boldsymbol{m}_{j-4}^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{m}_{j-2}^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{m}_{j}^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{m}_{j+2}^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{m}_{j+4}^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{s}_{1_{j-4}}^{(\mathrm{L})^{\mathrm{T}}}, \cdots, \boldsymbol{s}_{1_{j+4}}^{(\mathrm{L})^{\mathrm{T}}}, \boldsymbol{s}_{2_{j-4}}^{(\mathrm{L})^{\mathrm{T}}}, \cdots, \boldsymbol{s}_{2_{j+4}}^{(\mathrm{L})^{\mathrm{T}}}\Big)^{\mathrm{T}} \quad (3)$$

なお，DNN の学習を容易にするため，$\boldsymbol{x}_j$ に対し L_2 ノルムが 1 となる正規化を施す．このとき，$\boldsymbol{x}_j$ の大きさ（音量）の情報を保持するため，正規化係数のスカラー値を 1 次元として新たに加えたベクトルを $\boldsymbol{x}'_j$ とし，これが DNN の入力ベクトルとなる．具体的には次式のようになる．

$$\boldsymbol{x}'_j = \left(\frac{1}{\|\boldsymbol{x}_j\|_2}\boldsymbol{x}_j^{\mathrm{T}}, \|\boldsymbol{x}_j\|_2^{\mathrm{T}}\right)^{\mathrm{T}} \quad (4)$$

ここで，$\|\cdot\|_2$ は L_2 ノルムを表す．

2.4　DNN の学習法

DNN の中間層（隠れ層）は，全結合層 4 層で構成され，各隠れ層及び出力層の活性化関数には rectified liner unit（ReLU）を用いた．また，各隠れ層の次元数は全て出力層と同じ次元数とした．

DNN が出力する各音源信号の高周波帯域の予測を $\hat{\boldsymbol{S}}_n^{(\mathrm{H})} \in \mathbb{R}_{\geq 0}^{(I-I')\times J}$ とすると，時間フレーム j に対応する出力層の予測ベクトルは，$\hat{\boldsymbol{S}}_1^{(\mathrm{H})}, \hat{\boldsymbol{S}}_2^{(\mathrm{H})}$ の列ベクトル $\hat{\boldsymbol{s}}_{1_j}^{(\mathrm{H})}$ 及び $\hat{\boldsymbol{s}}_{2_j}^{(\mathrm{H})}$ を結合したものが出力される．これ

[1]ここで近傍の時間フレームを 1 つ置きに採用している理由は，STFT において短時間のシフト幅を窓長の半分に設定した場合を想定しているためであり，実際には STFT の条件に応じて選択の余地がある．

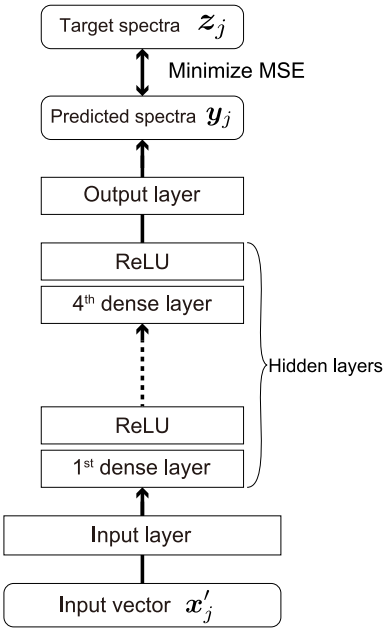

Fig. 3　DNN architecture.

を $\boldsymbol{y}_j$ とすると次のように表せる．

$$\boldsymbol{y}_j = \left(\hat{\boldsymbol{s}}_{1_j}^{(\mathrm{H})^{\mathrm{T}}}, \hat{\boldsymbol{s}}_{2_j}^{(\mathrm{H})^{\mathrm{T}}}\right)^{\mathrm{T}} \quad (5)$$

$$\hat{\boldsymbol{s}}_{n_j}^{(\mathrm{H})} = \left(\hat{s}_{n_{1j}}^{(\mathrm{H})}, \hat{s}_{n_{2j}}^{(\mathrm{H})}, \cdots, \hat{s}_{n_{(I-I')j}}^{(\mathrm{H})}\right)^{\mathrm{T}} \quad (6)$$

ここで，$\hat{s}_{n_{ij}}^{(\mathrm{H})}$ は $\hat{\boldsymbol{S}}_n^{(\mathrm{H})}$ の ij 要素を表す．

また，予測ベクトル $\hat{\boldsymbol{s}}_{1_j}^{(\mathrm{H})}$ 及び $\hat{\boldsymbol{s}}_{2_j}^{(\mathrm{H})}$ に対する正解ラベルはそれぞれ $\boldsymbol{s}_{1_j}^{(\mathrm{H})}$，$\boldsymbol{s}_{2_j}^{(\mathrm{H})}$ であり，これらを結合したベクトルを $\boldsymbol{z}_j$ とすると，次式で表される．

$$\boldsymbol{z}_j = \left(\boldsymbol{s}_{1_j}^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{s}_{2_j}^{(\mathrm{H})^{\mathrm{T}}}\right)^{\mathrm{T}} \quad (7)$$

従って DNN は，次式で示す $\boldsymbol{y}_j$ と $\boldsymbol{z}_j$ 間の平均二乗誤差（mean squared error: MSE）を最小化するように学習する．

$$\mathrm{MSE}(\boldsymbol{y}_j, \boldsymbol{z}_j) = \frac{1}{2(I-I')}\|\boldsymbol{z}_j - \boldsymbol{y}_j\|_2^2 \quad (8)$$

Table 1　STFT conditions

FFT length	64 ms
Shift length	32 ms
Window function	Hamming window

Table 2　Number of units for each layer in DNN

Input layer	3846 units
All hidden layers	514 units
Output layer	514 units

3　実験

3.1　実験条件

本実験では，提案するフレームワークにおいて，混合信号の高周波帯域の入力情報がどの程度帯域拡張の精度を向上させるかについて実験的に確認する．即ち，前段の BSS から完全に分離された音源信号（低周波帯域のみ）が得られたという仮定の下で，後段の DNN に基づく音響帯域拡張の精度について，混合信号の高周波帯域を用いる場合と用いない場合の差を確認する．

実験では，データセットに SiSEC2016 [10] の音楽信号データセット DSD100 を使用し，ドラム (Dr.) とボーカル (Vo.) の 2 音源を用いた．100 曲の音源から，DSD100 の dev データにある no. 51〜100 の 50 曲と test データにある no. 26〜50 の 25 曲の計 75 曲を DNN の学習データとし，残りの test データにある no. 1〜25 の 25 曲を評価データとした．また，学習用音源 75 曲のうち，ランダムに選んだ 15 曲を学習時の検証データとして用い，学習の early stopping を行う目安とした．さらに，全ての音源をサンプリング周波数 16 kHz へリサンプルし，無音時間が少ない区間として適当な 60 秒を全曲切り取った信号を学習及びテストデータとした．なお，学習データは Fig. 4 に示すようにスペクトログラムを 75 曲全て時間フレーム方向に統合して作成した．

本実験では，低周波帯域を 0〜4 kHz，高周波帯域を 4〜8 kHz とした．従って，境界となるインデクスは $I' = \mathrm{floor}(I/2)$ となる．ここで，floor(·) は小数点以下切り捨てを表す．

DNN の最適化法には Adam [11] を用いハイパーパラメータはそれぞれ $\varepsilon = 1.0 \times 10^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ 及び学習率 $\eta = 0.001$ とした．その他の学習パラメータについては，ミニバッチサイズを 1024，エポック数を 1000 として誤差逆伝搬学習を行った．また，過学習防止のため early stopping を設け，50 エポックの内に評価データに対する精度向上が見られなかった場合学習を打ち切るように設定した．その他の実験条件は Tables 1 及び 2 に示す．

DNN によって予測された分離信号の高周波帯域のスペクトログラム $\hat{\boldsymbol{S}}_n^{(\mathrm{H})}$ は，DNN の入力に用いた低周波帯域のスペクトログラム $\boldsymbol{S}_n^{(\mathrm{L})}$ と結合することで，全周波帯域の分離信号のスペクトログラムを得られる．但し，本稿の DNN は振幅スペクトルだけを予測し，位相スペクトルは推定しないため，分離信号の高

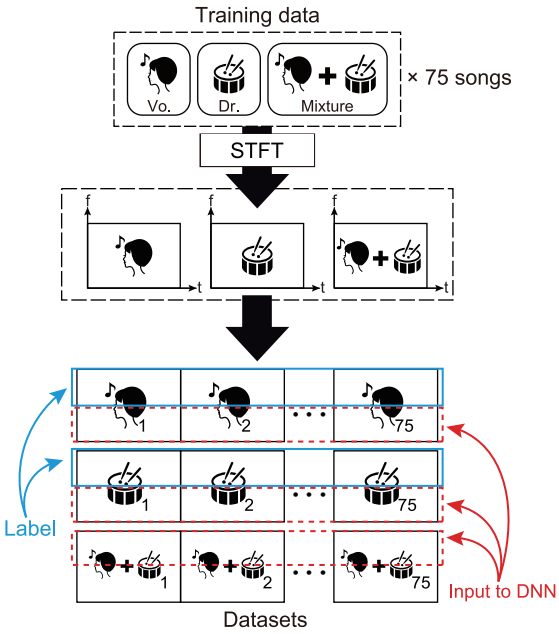

Fig. 4　Training datasets of vocals, drums, and their mixture.

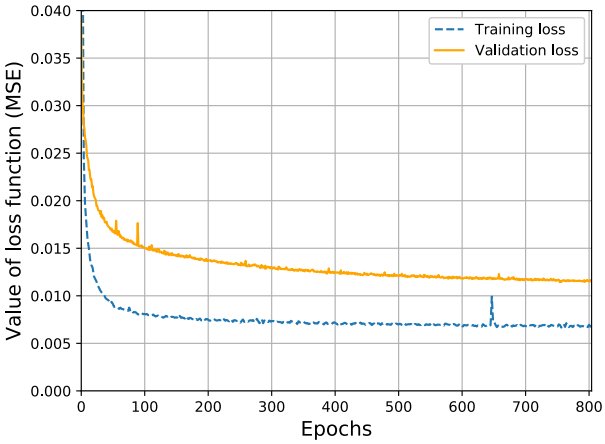

Fig. 5　Convergence behaviors of training and validation loss values.

周波帯域のスペクトログラム $\hat{\boldsymbol{S}}_n^{(\mathrm{H})}$ には混合信号の高周波帯域のスペクトログラム $\boldsymbol{M}^{(\mathrm{H})}$ の位相をそのまま用いた．このようにして作成された分離信号は逆 STFT を経て時間領域に変換される．

客観評価尺度には，音響帯域拡張の前後の sources-to-artifacts ratio (SAR) [12] を用いた．SAR は信号処理によって生じた歪みの少なさを表したものであり，高周波帯域が失われたことによる歪みをどの程度復元できるかという尺度として用いている．

最後に，本実験では，入力データに混合信号の高周波帯域 $\boldsymbol{M}^{(\mathrm{H})}$ を用いない場合の DNN による高周波帯域の予測結果を比較する．このとき，DNN の構造や学習データ，その他の条件等は全て統一し，純粋に「混合信号の高周波帯域 $\boldsymbol{M}^{(\mathrm{H})}$ が，分離信号の帯域拡張の精度をどの程度向上させるか」を比較している．即ち，Fig. 1 に示す提案フレームワークの妥当性の一部を検証する実験である．

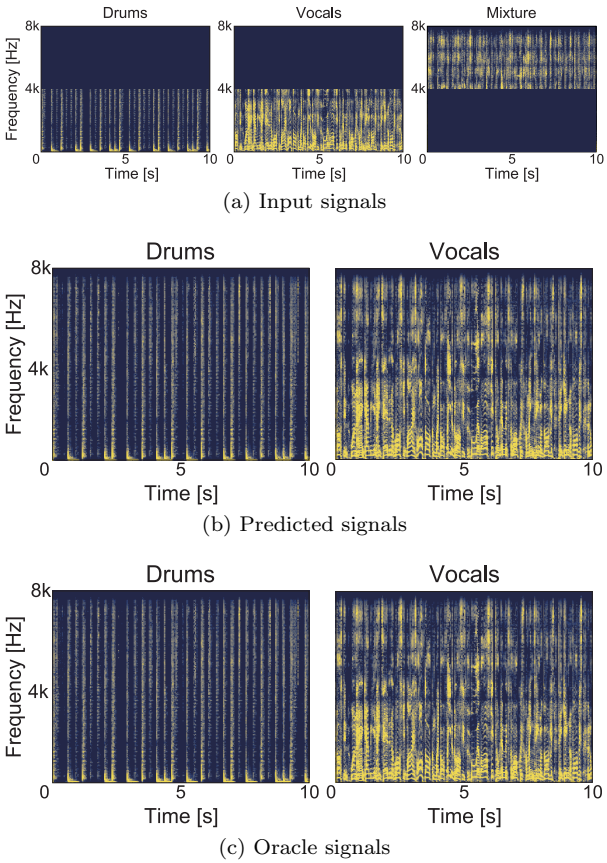

Fig. 6 Spectrograms of (a) input, (b) predicted, and (c) oracle signals.

3.2 結果

3.1 節の条件で学習させた提案手法の DNN の MSE の収束結果を Fig. 5 に示す．エポックを重ねるごとに学習データ及び評価データに対する MSE が減少していることから DNN は適切に学習出来ていることが分かる．また，early stopping により，804 エポックの時点で学習は打ち切られた．

Fig. 6 は，提案手法の音響帯域拡張の一例として，入力データ，提案手法の DNN の予測結果から復元した全周波帯域の分離信号，及び正解の分離信号のスペクトログラムをそれぞれ示している．Fig. 6 より，ドラム及びボーカル音源のそれぞれの特徴を DNN が学習し，高周波帯域を高精度で復元していることが分かる．また，test データ no. 1〜25 の平均 SAR 値及び平均 SAR 改善値を Table 3 に示す．ここで，Table 3 (b) 及び (c) の違いは，入力データに混合信号の高周波帯域を与えるか否かである．これらの値からも，Dr. と Vo. の両音源において，高周波帯域の復元によって SAR が大きく改善していることが分かる．特に，混合信号の高周波帯域を DNN に与えない場合と提案手法の差は大きく，Fig. 1 に示す提案フレームワークの妥当性の一部が実験的に示されている．

4 まとめ

本稿では，音源分離のコスト削減を目的とした DNN に基づく音響帯域拡張を提案し，混合信号の高周波帯域と各分離信号の低周波帯域の入力情報から，分離信号の高周波帯域を高精度に予測できることを実験的に示した．提案フレームワークの妥当性をさらに確認するために，提案フレームワークが，全周波帯域に BSS を適用する場合よりも (a) 短い計算時間となるか，(b) 分離精度は同程度以上となるか，の 2 点について今後さらに確認する必要がある．

Table 3 Average SAR values and their improvements

		Dr.	Vo.
(a) Narrow-band signal		22.1 dB	20.5 dB
(b) Full-band signal predicted by using only narrow-band sources		22.1 dB	20.6 dB
(c) Full-band signal predicted by using both narrow-band sources and high-frequency mixture		23.6 dB	23.7 dB
Improvement	(b) - (a)	0.0 dB	0.1 dB
	(c) - (a)	1.5 dB	3.2 dB

謝辞 本研究の一部は JSPS 科研費 19K20306 及び NVIDIA GPU Grant の助成を受けたものである．

参考文献

[1] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2006.

[2] 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 Trans. ASLP*, vol. 24, no. 9, pp. 1622–1637, 2016.

[3] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, pp. 125–155. 2018.

[4] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[5] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. MPCA*, 2002.

[6] P. Smaragdis and B. Raj, "Example-driven bandwidth expansion," in *Proc. WASPAA*, 2007, pp. 135–138.

[7] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *Proc. ICASSP*, 2009, pp. 145–148.

[8] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *Proc. ISCA*, 2015.

[9] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *ICASSP*, 2015, pp. 4395–4399.

[10] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*. Springer, 2017, pp. 323–332.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.