深層学習に基づく音響帯域拡張による音源分離処理の高速化*

☆渡辺瑠伊，北村大地（香川高専），猿渡洋（東大），高橋祐，近藤多伸（ヤマハ）

1 はじめに

音源分離（audio source separation: ASS）とは，観測された混合信号から特定の信号を推定する技術である．スマートスピーカなどの代表的なオーディオデバイスにはマイクアレイが搭載されており，多チャネル ASS（multichannel ASS: MASS）が適用されている．有名な MASS 手法の一つに多チャネル非負値行列因子分解（multichannel nonnegative matrix factorization: MNMF）[1] がある．MNMF は，混合系を音源と周波数毎の空間共分散行列（spatial covariance matrix: SCM）[2] でモデル化し，更に各音源の時間周波数構造を非負値行列因子分解（nonnegative matrix factorization: NMF）[3] でモデル化する．そして，推定された空間モデルと音源モデルを用いて周波数毎の分離フィルタを計算する．MNMF は，事前情報無しで高品質な音源分離が可能であるが，パラメータの推定に膨大な計算コストを必要とする．

近年は深層学習（deep neural networks: DNN）が様々な課題解決に利用されており，DNN に基づく単一チャネル ASS の研究が盛んである [4,5]．多チャネル信号に対しても，DNN に基づく MASS [6] が提案されているが，ネットワーク構造が大きくなりがちであるため，製品への実装コストが高い問題がある．

本稿では，各音源の周波数成分を予測する DNN [7] を既存の周波数領域 MASS のポスト処理を用いた低コストかつ効率的な MASS フレームワークを提案する．このフレームワークは，Fig. 1 に示すように，(a) 狭帯域（限られた数）の周波数に周波数領域 MASS を適用する処理及び (b) スモールサイズの DNN を用いて MASS を適用しなかった周波数の音源成分を予測する処理の二つで構成される．(a) の処理で MASS を適用する周波数のビン数を減らすことができるため，(b) の処理の DNN の予測が効率的であれば，音源分離全体の計算コストを削減でき，エッジコンピューティングデバイス等への実装が可能となる．本稿では，ASS の一例として MNMF [1] のみを扱うが，提案フレームワークでは，周波数毎に音源を分離する ASS（例えば [8,9] 等）であれば，いかなる手法にも適用可能である．

2 周波数領域 MASS

2.1 定式化

N 及び M をそれぞれ音源数及びマイクロホン数とすると，短時間フーリエ変換（short-time Fourier transform: STFT）で得られる多チャネル信号及び混合信号の複素成分は次のように表される．

$$\boldsymbol{s}_{ijn} = (s_{ijn1}, \cdots, s_{ijnm}, \cdots, s_{ijnM})^{\mathrm{T}} \in \mathbb{C}^N \quad (1)$$

$$\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijm}, \cdots, x_{ijN})^{\mathrm{T}} \in \mathbb{C}^M \quad (2)$$

ここで，$i = 1, 2, \ldots, I$，$j = 1, 2, \ldots, J$，$n = 1, 2, \ldots, N$，及び $m = 1, 2, \ldots, M$ はそれぞれ，周波数ビン，時間フレーム，音源の数，及びマイク数のインデクスである．また，式 (1) 及び (2) のスペクトログラムをそれぞれ，$\boldsymbol{S}_{nm} \in \mathbb{C}^{I \times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I \times J}$ と定義する．多チャネル信号 $\boldsymbol{s}_{ijn}$ はソースイメージと呼ばれる．また，観測された多チャネルの混合信号はソースイメージの総和 $\boldsymbol{x}_{ij} = \sum_n \boldsymbol{s}_{ijn}$ であると仮定する．

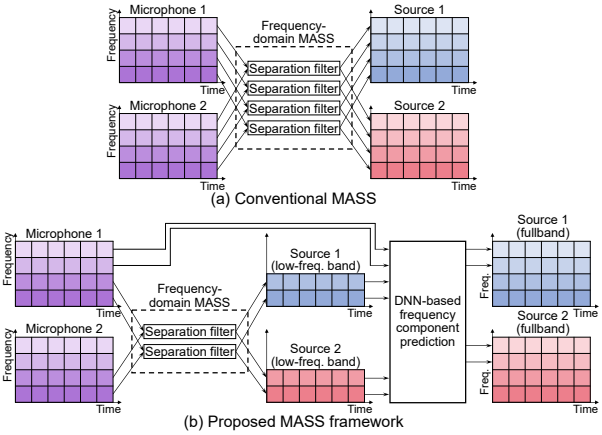

Fig. 1 Frameworks of (a) conventional and (b) proposed frequency-domain MASS.

2.2 MNMF

MNMF [1] では，ソースイメージに対して次のような生成モデルを仮定している．

$$\boldsymbol{s}_{ijn} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{ijn}\boldsymbol{H}_{in}) \quad (3)$$

$$\sigma_{ijn} = \sum_k z_{kn} t_{ik} v_{kj} \quad (4)$$

ここで，$\mathcal{N}(\boldsymbol{0}, \sigma_{ijn}\boldsymbol{H}_{in})$ は，共分散行列 $\sigma_{ijn}\boldsymbol{H}_{in}$ を持つ平均ゼロの多変量複素ガウス分布である．また，$t_{ik} \geq 0$ 及び $v_{kj} \geq 0$ は，NMF のパラメータ（基底とアクティベーション）であり，$k = 1, 2, \ldots, K$ は NMF の基底のインデクス，$z_{kn} \in \mathbb{R}_{[0,1]}$ は基底 K を N 個の音源にクラスタリングする潜在変数である．共分散行列が時変成分 σ_{ijn} と時不変行列 $\boldsymbol{H}_{in} \in \mathbb{C}^{M \times M}$ で定義され，前者は NMF による音源モデル及び後者は SCM [2] と呼ばれる空間モデルに対応する．これらのパラメータを推定することで，多チャネルの音源分離が可能となる．各パラメータは最尤推定で求められるが，$\boldsymbol{H}_{in}$ の更新式は，次式のように各時間周波数スロットにおいてサイズ $\boldsymbol{M}$ の行列の逆行列演算を含むため，膨大な計算コストが必要である．

$$\boldsymbol{H}_{in} \leftarrow \boldsymbol{R}_{in}^{-1/2} \left(\boldsymbol{R}_{in}^{1/2} \boldsymbol{A}_{in} \boldsymbol{R}_{in}^{1/2} \right)^{1/2} \boldsymbol{R}_{in}^{-1/2} \quad (5)$$

$$\boldsymbol{R}_{in} = \sum_{j,k} z_{kn} t_{ik} v_{kj} \boldsymbol{D}_{ij}^{-1} \quad (6)$$

$$\boldsymbol{D}_{ij} = \sum_n \sigma_{ijn} \boldsymbol{H}_{in}, \quad (7)$$

$$\boldsymbol{A}_{in} = \boldsymbol{H}_{in} \left(\sum_j \sigma_{ijn} \boldsymbol{D}_{ij}^{-1} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{H}} \boldsymbol{D}_{ij}^{-1} \right) \boldsymbol{H}_{in} \quad (8)$$

*Fast audio source separation using bandwidth expansion based on deep neural networks. By Rui WATANABE, Daichi KITAMURA (NIT Kagawa), Hiroshi SARUWATARI (UTokyo), Yu TAKAHASHI, and Kazunobu KONDO (Yamaha).

推定されたパラメータを用いて，次式のマルチチャネル Wiener フィルタを構成することで，分離されたソースイメージ $\hat{\boldsymbol{s}}_{ijn}$ が得られる．

$$\hat{\boldsymbol{s}}_{ijn} = (\sigma_{ijn}\boldsymbol{H}_{in})\,\boldsymbol{D}_{ij}^{-1}\boldsymbol{x}_{ij} \tag{9}$$

3 提案手法

3.1 動機

MNMF を含む周波数領域 MASS は比較的高精度な音源分離が可能であるが，アルゴリズムの複雑さから製品への実装コストが高い．また大規模な DNN は，メモリに制限のある小型デバイスでのエッジコンピューティングには適していない．

この問題を解決するために，周波数領域 MASS とスモールサイズの DNN を組み合わせた新たなフレームワークを提案する．このフレームワークは，Fig. 1 に示すように，限られた数の周波数（低周波帯域）にのみ MASS を適用し，他の周波数（高周波帯域）の分離信号は DNN で予測する．このとき，DNN の入力データは (a) 狭帯域の分離信号及び (b) その他の周波数の混合信号であり，この両者からその他の周波数帯域の分離信号を予測する．(a) だけでなく (b) も入力データとして予測に利用するため，小規模な DNN であっても，高精度な予測が可能となる．

簡単のため，本稿では 2 音源の場合（$N=2$）のみを扱うが，提案フレームワークは後述の DNN モデルを拡張するだけで，3 音源以上にも適用できる．また，Fig. 1 のように低周波帯域と高周波帯域を分割する例だけを扱うが，他の分割方法にも一般化できる．

3.2 DNN の入力情報

$\boldsymbol{Y}_1 \in \mathbb{R}_{\geq 0}^{I\times J}$ 及び $\boldsymbol{Y}_2 \in \mathbb{R}_{\geq 0}^{I\times J}$ を 2 音源の振幅スペクトログラム，$\boldsymbol{M} \in \mathbb{R}_{\geq 0}^{I\times J}$ を混合信号の振幅スペクトログラムとする．音源信号や観測信号は多チャネルである場合，$\boldsymbol{Y}_1$，$\boldsymbol{Y}_2$，及び $\boldsymbol{M}$ は基準マイクロホンを用いて $\boldsymbol{Y}_1 = \mathrm{abs}(\boldsymbol{S}_{11})$，$\boldsymbol{Y}_2 = \mathrm{abs}(\boldsymbol{S}_{21})$，及び $\boldsymbol{M} = \mathrm{abs}(\boldsymbol{X}_1)$ と定める．ここで，$\mathrm{abs}(\cdot)$ は要素毎の絶対値を表す．高周波帯域と低周波帯域に分割した際の境界の周波数を $i = I'$ とすると，Fig. 1 のように，$\boldsymbol{Y}_n$ は低周波帯域 $\boldsymbol{Y}_n^{(\mathrm{L})} \in \mathbb{R}_{\geq 0}^{I'\times J}$ 及び高周波帯域 $\boldsymbol{Y}_n^{(\mathrm{H})} \in \mathbb{R}_{\geq 0}^{(I-I')\times J}$ に分けられる．同様に，$\boldsymbol{M}$ の低周波帯域と高周波帯域は $\boldsymbol{M}^{(\mathrm{L})} \in \mathbb{R}_{\geq 0}^{I'\times J}$ 及び $\boldsymbol{M}^{(\mathrm{H})} \in \mathbb{R}_{\geq 0}^{(I-I')\times J}$ となる．

DNN は，2 音源の低周波帯域 $\boldsymbol{Y}_1^{(\mathrm{L})}$ 及び $\boldsymbol{Y}_2^{(\mathrm{L})}$ と混合信号の高周波帯域 $\boldsymbol{M}^{(\mathrm{H})}$ から，2 音源の高周波帯域成分（MASS で分離しなかった成分）を予測する．正確には，DNN は $\boldsymbol{M}^{(\mathrm{H})}$ から $\boldsymbol{Y}_1^{(\mathrm{H})}$ 及び $\boldsymbol{Y}_2^{(\mathrm{H})}$ を得るような 2 音源分のソフトマスクを出力する．

DNN モデルの入力ベクトルを Fig. 2 に示す．時間フレーム j における混合信号の高周波帯域及び 2 音源の低周波帯域のベクトルは次のように表される．

$$\boldsymbol{m}_j^{(\mathrm{H})} = \left(m_{1j}^{(\mathrm{H})}, m_{2j}^{(\mathrm{H})}, \cdots, m_{(I-I')j}^{(\mathrm{H})}\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{I-I'} \tag{10}$$

$$\boldsymbol{y}_{nj}^{(\mathrm{L})} = \left(y_{n1j}^{(\mathrm{L})}, y_{n2j}^{(\mathrm{L})}, \cdots, y_{nI'j}^{(\mathrm{L})}\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{I'} \tag{11}$$

ここで，$m_{ij}^{(\mathrm{H})}$ 及び $y_{nij}^{(\mathrm{L})}$ はそれぞれ $\boldsymbol{M}^{(\mathrm{H})}$ 及び $\boldsymbol{Y}_n^{(\mathrm{L})}$ の ij 要素である．DNN が各音源の時間 j における高周波帯域成分を予測する場合，j 周辺の成分も重要である．そこで，式 (10) 及び (11) の隣接時間フレームを連結し[1]，以下のようなベクトルを定義する．

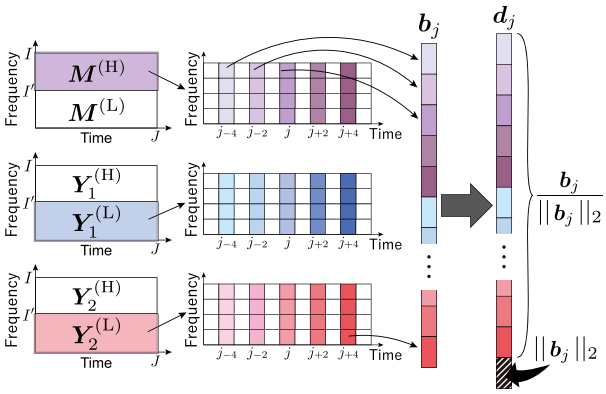

Fig. 2 Input vector of DNN.

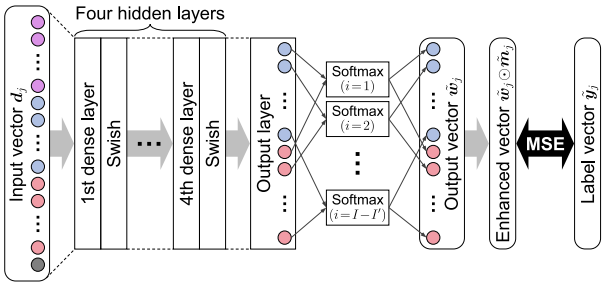

Fig. 3 DNN architecture.

$$\overline{\boldsymbol{m}}_j^{(\mathrm{H})} = \left({\boldsymbol{m}_{j-2c}^{(\mathrm{H})}}^{\mathrm{T}}, \cdots, {\boldsymbol{m}_{j}^{(\mathrm{H})}}^{\mathrm{T}}, \cdots, {\boldsymbol{m}_{j+2c}^{(\mathrm{H})}}^{\mathrm{T}}\right)^{\mathrm{T}} \tag{12}$$

$$\overline{\boldsymbol{y}}_{nj}^{(\mathrm{L})} = \left({\boldsymbol{y}_{n(j-2c)}^{(\mathrm{L})}}^{\mathrm{T}}, \cdots, {\boldsymbol{y}_{nj}^{(\mathrm{L})}}^{\mathrm{T}}, \cdots, {\boldsymbol{y}_{n(j+2c)}^{(\mathrm{L})}}^{\mathrm{T}}\right)^{\mathrm{T}} \tag{13}$$

$$\boldsymbol{b}_j = \left({\overline{\boldsymbol{m}}_j^{(\mathrm{H})}}^{\mathrm{T}}, {\overline{\boldsymbol{y}}_{1j}^{(\mathrm{L})}}^{\mathrm{T}}, {\overline{\boldsymbol{y}}_{2j}^{(\mathrm{L})}}^{\mathrm{T}}\right) \in \mathbb{R}_{\geq 0}^{(2C+1)(I+I')} \tag{14}$$

ここで，$c = 0, 1, \ldots, C$ は，隣接時間フレームのインデクスである．さらに，DNN の学習を安定化するために，$\boldsymbol{b}_j$ を次式のように正規化する．このとき，正規化係数も連結して入力ベクトル $\boldsymbol{d}_j$ が構成される．

$$\boldsymbol{d}_j = \left(\frac{1}{\|\boldsymbol{b}_j\|_2}\boldsymbol{b}_j^{\mathrm{T}}, \|\boldsymbol{b}_j\|_2\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{(2C+1)(I+I')+1} \tag{15}$$

ここで，$\|\cdot\|_2$ は L_2 ノルムを表す．

3.3 DNN の出力と学習

DNN の構造を Fig. 3 に示す．4 層の隠れ層は全て全結合層であり，出力層と同じ次元数となっている．また，隠れ層の活性化関数として，Swish [10] を使用している．出力層では，各音源のソフトマスクの和が周波数毎に 1 となる必要があるため，周波数毎に Softmax 関数が適用される．

高周波帯域の n 番目の音源を得るためのソフトマスクを $\boldsymbol{W}_n \in \mathbb{R}_{[0,1]}^{(I-I')\times J}$ とし，次式のように表す．

$$\boldsymbol{Y}_n^{(\mathrm{H})} \approx \boldsymbol{W}_n \odot \boldsymbol{M}^{(\mathrm{H})} \tag{16}$$

[1] 時間フレームを $j-2, j, j+2$ のようにスキップするのは，STFT をハーフシフトで行うことにより隣接する時間フレームに冗長成分が含まれるためである．

Table 1　Song names of dry source in test dataset

Song ID	Song name	Signal length [s]
1	dev1_bearlin-roads	14.0
2	dev2_another_dreamer-the_ones_we_love	25.0
3	dev2_fort_minor-remember_the_name	24.0
4	dev2_ultimate_nz_tour	18.0

ここで，$\odot$ は要素ごとの積を表す．DNN の出力ベクトル $\tilde{\boldsymbol{w}}_j$ は，時間フレーム j における各音源のソフトマスク $\boldsymbol{W}_1$ 及び $\boldsymbol{W}_2$ を連結したものである．

$$\tilde{\boldsymbol{w}}_j = \left(\boldsymbol{w}_{1j}^{\mathrm{T}}, \boldsymbol{w}_{2j}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}_{[0,1]}^{2(I-I')} \tag{17}$$

$$\boldsymbol{w}_{nj} = \left(w_{n1j}, w_{n2j}, \cdots, w_{n(I-I')j}\right)^{\mathrm{T}} \in \mathbb{R}_{[0,1]}^{I-I'} \tag{18}$$

ここで，w_{nij} は $\boldsymbol{W}_n$ の要素であり，出力層の Softmax 関数により，$\sum_n w_{nij} = 1 \ \forall\, i, j$ である．2 音源の高周波帯域の正解ベクトルは次式となる．

$$\tilde{\boldsymbol{y}}_j = \left(\boldsymbol{y}_{1j}^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{y}_{2j}^{(\mathrm{H})^{\mathrm{T}}}\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{2(I-I')} \tag{19}$$

$$\boldsymbol{y}_{nj}^{(\mathrm{H})} = \left(y_{n1j}^{(\mathrm{H})}, y_{n2j}^{(\mathrm{H})}, \cdots, y_{n(I-I')j}^{(\mathrm{H})}\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{I-I'} \tag{20}$$

ここで，$y_{nij}^{(\mathrm{H})}$ は $\boldsymbol{Y}_n^{(\mathrm{H})}$ の要素である．DNN モデルは，次式の平均二乗誤差（mean squared error: MSE）が最小となるように学習される．

$$\mathrm{MSE}(\tilde{\boldsymbol{y}}_j, \tilde{\boldsymbol{w}}_j \odot \tilde{\boldsymbol{m}}_j) = \frac{1}{2(I-I')} \|\tilde{\boldsymbol{y}}_j - \tilde{\boldsymbol{w}}_j \odot \tilde{\boldsymbol{m}}_j\|_2^2 \tag{21}$$

$$\tilde{\boldsymbol{m}}_j = \left(\boldsymbol{m}_j^{(\mathrm{H})^{\mathrm{T}}}, \boldsymbol{m}_j^{(\mathrm{H})^{\mathrm{T}}}\right) \in \mathbb{R}_{\geq 0}^{2(I-I')} \tag{22}$$

3.4　分離信号の時間領域への再構成

DNN では，分離信号の低周波帯域 $\boldsymbol{Y}_1^{(\mathrm{L})}$ 及び $\boldsymbol{Y}_2^{(\mathrm{L})}$，及び混合信号の高周波帯域 $\boldsymbol{M}^{(\mathrm{H})}$ を入力とし，ソフトマスク $\tilde{\boldsymbol{w}}_j$ を予測する．推定された高周波成分は 式 (16) で求められる．得られた $\boldsymbol{Y}_n^{(\mathrm{L})}$ 及び $\boldsymbol{Y}_n^{(\mathrm{H})}$ を周波数方向に連結することで，フルバンドの振幅スペクトログラム $\boldsymbol{Y}_n$ が得られる．

$\boldsymbol{Y}_n$ を時間信号に戻す際に，$\boldsymbol{Y}_n^{(\mathrm{H})}$ の位相を復元する必要がある．提案手法では，混合信号 $\boldsymbol{M}^{(\mathrm{H})}$ の位相を $\boldsymbol{Y}_1^{(\mathrm{H})}$ 及び $\boldsymbol{Y}_2^{(\mathrm{H})}$ に付与して逆 STFT を適用する．

4　実験と結果

提案フレームワークの妥当性を確認するため，(a) DNN に基づく周波数成分の予測性能比較，及び (b) ドラム (Dr.) とボーカル (Vo.) の混合音源を用いた MASS の 2 つの実験を行った．

4.1　周波数成分の予測性能の比較実験

4.1.1　実験条件

まず提案フレームワークの妥当性を実証する予備実験として，DNN に基づく音源毎の周波数成分の予測性能が，混合信号の高周波帯域 $\boldsymbol{M}^{(\mathrm{H})}$ の有無によってどの程度変化するかを評価した．即ち，$\boldsymbol{M}^{(\mathrm{H})}$ を入力する DNN と入力しない DNN の 2 つを用意した．混合信号を入力しない DNN は単純に音源信号の低周波帯域 $\boldsymbol{P}^{(\mathrm{L})} \in \mathbb{R}_{\geq 0}^{I' \times J}$ のみを用いて高周波帯域 $\boldsymbol{P}^{(\mathrm{H})} \in \mathbb{R}_{\geq 0}^{(I-I') \times J}$ を予測するモデル [11] である．このとき，$\boldsymbol{P}^{(\mathrm{H})}$ 及び $\boldsymbol{P}^{(\mathrm{L})}$ は混合前の音源信号の振幅スペクトログラムである．また，入力ベクトル $\boldsymbol{q}_j$ と正解ベクトル $\boldsymbol{p}^{(\mathrm{H})}$ はそれぞれ次のようになる．

$$\boldsymbol{q}_j = \left(\frac{1}{\|\overline{\boldsymbol{p}}_j\|_2} \overline{\boldsymbol{p}}_j^{\mathrm{T}}, \|\overline{\boldsymbol{p}}_j\|_2\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{(2C+1)I'+1} \tag{23}$$

$$\overline{\boldsymbol{p}}_j^{(\mathrm{L})} = \left(\boldsymbol{p}_{(j-2c)}^{(\mathrm{L})^{\mathrm{T}}}, \cdots, \boldsymbol{p}_j^{(\mathrm{L})^{\mathrm{T}}}, \cdots, \boldsymbol{p}_{(j+2c)}^{(\mathrm{L})^{\mathrm{T}}}\right)^{\mathrm{T}} \tag{24}$$

$$\boldsymbol{p}_j^{(\mathrm{L})} = \left(p_{1j}^{(\mathrm{L})}, p_{2j}^{(\mathrm{L})}, \cdots, p_{I'j}^{(\mathrm{L})}\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{I'} \tag{25}$$

$$\boldsymbol{p}_j^{(\mathrm{H})} = \left(p_{1j}^{(\mathrm{H})}, p_{2j}^{(\mathrm{H})}, \cdots, p_{(I-I')j}^{(\mathrm{H})}\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{I-I'} \tag{26}$$

ここで，$p_{ij}^{(\mathrm{L})}$ 及び $p_{ij}^{(\mathrm{H})}$ はそれぞれ $\boldsymbol{P}^{(\mathrm{L})}$ 及び $\boldsymbol{P}^{(\mathrm{H})}$ の要素である．$\boldsymbol{P}^{(\mathrm{H})}$ の位相復元には Griffin–Lim（GL）アルゴリズム [12] を用いた．混合信号を入力しない DNN では，入力は 1 音源のみとなるため，Dr. 及び Vo. に対してそれぞれ専用の DNN を用意した．一方，混合信号を入力する DNN は，$\boldsymbol{Y}_1^{(\mathrm{L})}$，$\boldsymbol{Y}_2^{(\mathrm{L})}$ 及び $\boldsymbol{M}^{(\mathrm{H})}$ を入力として，ソフトマスクを予測し，$\boldsymbol{Y}_1^{(\mathrm{H})}$ 及び $\boldsymbol{Y}_2^{(\mathrm{H})}$ を求める．なお，条件を揃えるため，この実験では混合信号を入力する DNN も $\boldsymbol{Y}_n^{(\mathrm{H})}$ の位相を GL アルゴリズムで復元する．さらに本実験では，$\boldsymbol{Y}_1^{(\mathrm{L})}$ 及び $\boldsymbol{Y}_2^{(\mathrm{L})}$ に，完全分離音源を用いている．

DNN の学習には，SiSEC2016 [13] の MUS の Dr. 及び Vo. を 100 曲分使用した．STFT におけるハミング窓長は 128 ms，シフト長は 64 ms とした．また，$I = 1025$（8 kHz）とした．入力ベクトル及び正解ベクトルとして，混合信号を入力しない DNN では $\boldsymbol{q}_j$ 及び $\boldsymbol{p}_j^{(\mathrm{H})}$，混合信号を入力する DNN では，$\boldsymbol{d}_j$ 及び $\tilde{\boldsymbol{y}}_j$ を作成し，混合信号 $\boldsymbol{M}$ は 2 音源のドライソースの時間領域和とした．高周波帯域と低周波帯域の境界周波数を $I' = 512$（4 kHz）とし，隣接時間フレーム数は $C = 2$ とした．DNN の最適化法には Adam を使用し，そのパラメータは $\varepsilon = 1.0 \times 10^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, 及び $\eta = 0.001$ とした．また，エポック数を 1000，ミニバッチサイズを 128 に設定した．

Dr. 及び Vo. のテストデータとして，SiSEC2011 [14] 内の Table 1 に示す 4 曲を使用した．再構成された時間信号について，人工的な歪みを評価するために，sources-to-artifacts ratio (SAR) [15] を使用した．

4.1.2　結果

各 DNN モデルで予測された信号の SAR を Table 2 に示す．この結果より，混合信号を入力する DNN の性能が，混合信号を入力しない DNN の性能よりも優れている．これは，混合信号の高周波帯域を DNN に入力する提案フレームワークの妥当性を示している．

4.2　音楽信号の混合信号を用いた MASS 実験

4.2.1　条件

本実験では， Table 1 の信号に，Fig. 4 に示す RWCP [16] の E2A インパルス応答を畳み込むことで，2 チャネル（$N = M = 2$）で観測した Dr. 及び Vo. の混合信号を生成した．MNMF では，音源分離性能が高くなる基底数をあらかじめ実験的に調べ，最適であった $K = 13$ に設定した．SCM の初期値は単位行列とし，他の MNMF パラメータは乱数で初期化した．音源の分離性能の指標として，分離の度合いと

Table 2　SARs of predicted fullband source signal

Song ID	DNN w/o mixture	DNN w/ mixture
1	Dr.: 21.1 dB	Dr.: **28.0** dB
	Vo.: 21.8 dB	Vo.: **31.5** dB
2	Dr.: **22.0** dB	Dr.: 21.8 dB
	Vo.: 12.7 dB	Vo.: **19.6** dB
3	Dr.: 15.0 dB	Dr.: **20.4** dB
	Vo.: 11.2 dB	Vo.: **18.5** dB
4	Dr.: 11.0 dB	Dr.: **18.2** dB
	Vo.: 10.4 dB	Vo.: **15.3** dB

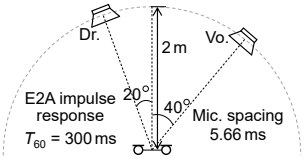

Fig. 4　Impulse responses used in experiment.

音質の両方を示す source-to-distortion ratio（SDR）[15] を用いた．MNMF の計算には，AMD Ryzen7 2700X CPU，DNN の予測には NVIDIA GeForce RTX2080Ti GPU を用いた．

4.2.2　結果

両手法において，異なる乱数値を用いて 10 回実験を行った際の，全周波帯域の MNMF 及び提案フレームワークの平均 SDR と処理時間の関係を Fig. 5 に示す．提案手法の処理時間には，DNN の予測にかかる時間が含まれるが，その時間は 0.1 s 以下である．また，提案手法では MNMF の反復回数（L）を，$L=10, 20, \ldots$ のように設定し，各条件において DNN の予測を行ったため， Fig. 5 の提案手法は L の設定毎に SDR 値をプロットしている．また，$L=300$ での平均 SDR も矢印で示している．これらの結果から，提案手法では，全周波帯域の MNMF と比較し計算コストを半分近く削減できることが確認できる．例として， Fig. 5 (d) では，全周波帯域の MNMF が 120 s で 13 dB に達するのに対して，提案手法では 50 s 以下で 13 dB を達成している．これは，MNMF の周波数が $I=1025$ から $I-I'=512$ へと削減されたためである．また，収束点において，提案手法は Fig. 5 (a)，(b)，及び (d) の 3 曲で全周波帯域 MNMF よりも優れていることが確認できる．これは，提案手法の DNN が 100 曲分のデータを学習したことで，高精度なソフトマスクを推定できたためと推測される．

5　まとめ

本稿では，既存の ASS 手法と DNN に基づく周波数成分予測を組み合わせた，低コストな音源分離フレームワークを提案した．本手法では，周波数を間引いて音源分離を行い，間引かれた周波数の分離音源成分を DNN で予測することで，計算コストを削減している．実験結果から，全周波帯域での MASS 手法と比較して，提案手法がより高速に同程度の音源分離を達成できることを確認した．今後の課題として，低周波帯域と高周波帯域という分割方法ではなく，音源分離に有効な周波数にのみ ASS を適用する手法に拡張すること等が挙げられる．

謝辞　本研究の一部は JSPS 科研費 19K20306，19H01116，及び NVIDIA GPU Grant の助成を受けたものである．

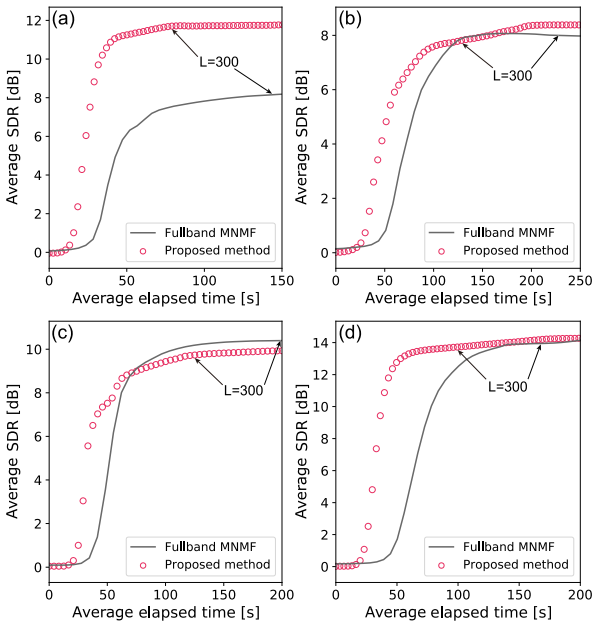

Fig. 5　Average SDRs and their elapse times.

参考文献

[1] 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.
[2] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.
[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[4] P.-S. Huang, M. Kim, M. H.-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. ASLP*, vol. 23, no. 12, pp. 2136–2147, 2015.
[5] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," *Proc. ICASSP*, pp. 31–35, 2016.
[6] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," *Proc. ICASSP*, pp. 116–120, 2015.
[7] 渡辺瑠伊, 北村大地, "音源分離のための深層学習に基づく音響帯域拡張" **日本音響学会** *2020* **年春季研究発表会講演論文集**, pp. 221–224, 2020.
[8] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, 2010.
[9] 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. 1626–1641, 2016.
[10] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint*, arXiv:1710.05941, 2017.
[11] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," *Proc. ICASSP*, pp. 4395–4399, 2015.
[12] D. Griffin and J. Lim, "Signal estimation from modified shorttime Fourier transform," *IEEE Trans. ASSP*, vol. 32, no. 2, pp. 236–243, 1984.
[13] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," *Proc. LVA/ICA*, pp. 323–332, 2012.
[14] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," *Proc. Latent Variable Analysis and Signal Separation*, pp. 414–422, 2012.
[15] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.
[16] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. LREC*, pp. 965–968, 2000.