深層学習に基づく周波数帯域補間手法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

多チャネル音源分離（multichannel audio source separation: MASS）とは，複数のマイクロフォンによって得られる観測信号から，混合前の信号を推定する技術である．有名な周波数領域 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 では，混合系を音源と周波数毎の空間共分散行列でモデル化し，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）でモデル化している．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推定している．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が必要である．

一方で，深層学習（deep neural networks: DNN）は音響信号処理においても一般的となり，単一チャネル音源分離 [2, 3] や音源分離を目的とした音響帯域拡張 [4] といった様々な課題解決に利用されている．また著者らは，Fig. 1(b) に示すような，DNN に基づく音響帯域拡張によって音源分離処理を高速化するフレームワーク [5, 6] を提案している．この手法では，信号を低周波帯域と高周波帯域の二つに分けることを考える．前段では，低周波帯域に周波数領域 MASS が適用され，各分離信号が推定される．後段では，得られた低周波帯域の分離信号及び高周波帯域の混合信号の二つを用いて，DNN が混合前の音源の高周波帯域を予測する．DNN の予測の計算コストが周波数領域 MASS の計算コストよりも十分小さい場合，本手法によって全体の計算コストを削減できる．

後段の DNN における周波数成分の予測は，分離信号の帯域拡張問題に対応するため，(混合信号の高周波帯域を用いてはいるが）分離信号の周波数成分の外挿に等しく，比較的難しい推論が要求される．しかし，Fig. 1(b) のフレームワークは，低周波帯域と高周波帯域の分割に限らず，任意の周波数ビンの組み合わせを，周波数領域 MASS（前段）と DNN に基づく周波数成分予測（後段）で分担できる．従って，後段の DNN に基づく周波数成分予測が内挿（補間）となるように間引く周波数ビンを選択する方が，より高精度な推論が可能であると考えられる．

本稿では，上記のアイデアに基づき，Fig. 1(b) の提案フレームワークにおける周波数ビンの間引き方法を二種類提案する．これらの手法が，音源分離フレームワークに与える影響を実験により明らかにする．

2 提案手法：音源分離フレームワーク

2.1 定式化

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

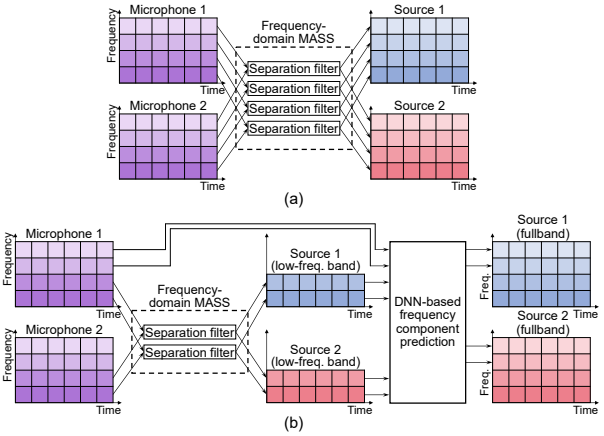

Fig. 1 (a) Full-band frequency-domain MASS and (b) proposed frameworks using DNN-based frequency component prediction.

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

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

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，及びマイクロフォン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．また，観測された多チャネルの混合信号は $\boldsymbol{x}_{ij} = \sum_n \boldsymbol{s}_{ijn}$ と仮定する．

2.2 DNN に基づく周波数成分予測

リファレンスチャネル（m_{ref} とする）の音源信号のスペクトログラムを $\boldsymbol{Y}_n = \boldsymbol{S}_{nm_{\mathrm{ref}}}$ とし，混合信号のスペクトログラムを $\boldsymbol{M} = \boldsymbol{X}_{m_{\mathrm{ref}}}$ と定義する．今，前段の周波数領域 MASS に入力される周波数ビンの集合を $\mathbb{F} \subset \{i | i = 1, 2, \cdots, I\}$ とし，逆に間引く周波数ビンの集合を $\mathbb{F}' \subset \{i | i = 1, 2, \cdots, I\}$ と定義する（$\mathbb{F}'$ は $\mathbb{F}$ の補集合となる）．さらに，$\mathbb{F}$ 及び $\mathbb{F}'$ の要素のインデクスをそれぞれ $f = 1, 2, \cdots, |\mathbb{F}|$ 及び $f' = 1, 2, \cdots, |\mathbb{F}'|$ とする．これらの集合を用いて，混合信号 $\boldsymbol{M}$ の全周波数ビンの内，前段の周波数領域 MASS に入力する周波数ビンだけをまとめた行列を $\boldsymbol{M}^{(\mathrm{P})} \in \mathbb{C}^{|\mathbb{F}| \times J}$ と定義する．さらに，前段の周波数領域 MASS には入力しない（間引く）周波数ビンだけをまとめた行列を $\boldsymbol{M}^{(\mathrm{Q})} \in \mathbb{C}^{|\mathbb{F}'| \times J}$ と定義する．同様に，音源信号 $\boldsymbol{Y}_n$ も $\boldsymbol{Y}_n^{(\mathrm{P})} \in \mathbb{C}^{|\mathbb{F}| \times J}$ 及び $\boldsymbol{Y}_n^{(\mathrm{Q})} \in \mathbb{C}^{|\mathbb{F}'| \times J}$ に分割される．文献 [5, 6] で提案した Fig. 1(b) に示すような，低周波帯域と高周波帯域に分割する間引き方法は $\mathbb{F} = \{i | i = 1, 2, \cdots, \lfloor I/2 \rfloor\}$

*Fast audio source separation based on deep-neural-network-based frequency component interpolation. By Rui WATANABE, Daichi KITAMURA (NIT Kagawa), Hiroshi SARUWATARI (UTokyo), Yu TAKAHASHI, and Kazunobu KONDO (Yamaha).

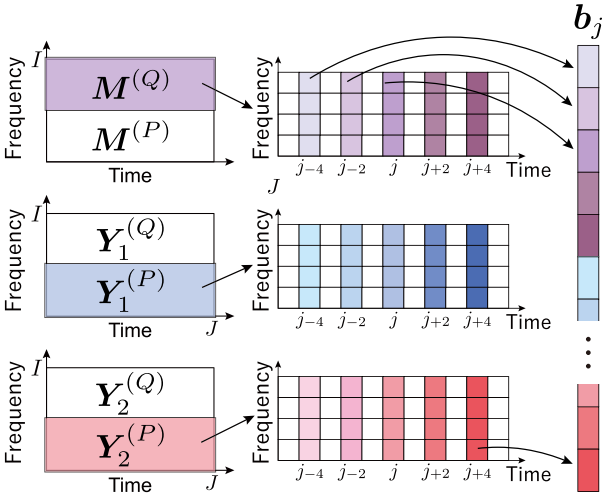

Fig. 2　Input vector of DNN, where $N = 2$.

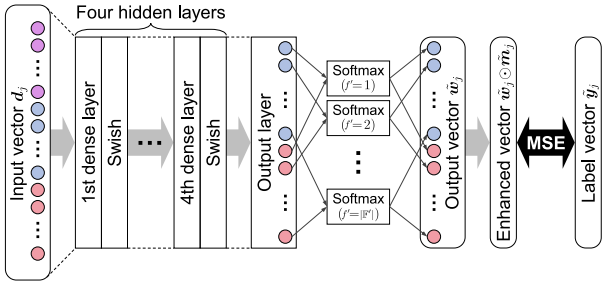

Fig. 3　DNN architecture, where $N = 2$.

及び $\mathbb{F}' = \{i | i = \lfloor I/2 \rfloor + 1, \lfloor I/2 \rfloor + 2, \cdots, I\}$ と表される．ここで，$\lfloor \cdot \rfloor$ は床関数である．

DNN は，集合 $\mathbb{F}$ に属する周波数ビンをまとめた全音源信号の行列 $\boldsymbol{Y}_1^{(\mathrm{P})}, \boldsymbol{Y}_2^{(\mathrm{P})}, \cdots, \boldsymbol{Y}_N^{(\mathrm{P})}$ と集合 $\mathbb{F}'$ に属する周波数ビンをまとめた混合信号の行列 $\boldsymbol{M}^{(\mathrm{Q})}$ を入力とし，集合 $\mathbb{F}'$ に属する周波数ビンをまとめた全音源信号の行列 $\boldsymbol{Y}_1^{(\mathrm{Q})}, \boldsymbol{Y}_2^{(\mathrm{Q})}, \cdots, \boldsymbol{Y}_N^{(\mathrm{Q})}$（前段の周波数領域 MASS で分離せずに間引いた成分）を予測する．より具体的には，DNN は $\boldsymbol{M}^{(\mathrm{Q})}$ から $\boldsymbol{Y}_n^{(\mathrm{Q})}$ を得るようなソフトマスクを予測し出力する．

$N = 2$ における DNN モデルの入力ベクトルを Fig. 2 に示す．但し，$\mathbb{F} = \{i | i = 1, 2, \cdots, \lfloor I/2 \rfloor\}$ 及び $\mathbb{F}' = \{i | i = \lfloor I/2 \rfloor + 1, \lfloor I/2 \rfloor + 2, \cdots, I\}$ として低周波帯域と高周波帯域に分割した場合の図を示している．混合信号の間引き周波数成分行列 $\boldsymbol{M}^{(\mathrm{Q})}$ 及び各音源の非間引き周波数成分行列 $\boldsymbol{Y}_n^{(\mathrm{P})}$ の時間フレーム j におけるベクトルはそれぞれ次のように表される．

$$\boldsymbol{m}_j^{(\mathrm{Q})} = \left(m_{1j}^{(\mathrm{Q})}, \cdots, m_{f'j}^{(\mathrm{Q})}, \cdots, m_{|\mathbb{F}'|j}^{(\mathrm{Q})}\right)^{\mathrm{T}} \in \mathbb{C}^{|\mathbb{F}'|} \tag{3}$$

$$\boldsymbol{y}_{nj}^{(\mathrm{P})} = \left(y_{n1j}^{(\mathrm{P})}, \cdots, y_{nfj}^{(\mathrm{P})}, \cdots y_{n|\mathbb{F}|j}^{(\mathrm{P})}\right)^{\mathrm{T}} \in \mathbb{C}^{|\mathbb{F}|} \tag{4}$$

ここで，$m_{f'j}^{(\mathrm{Q})}$ 及び $y_{nfj}^{(\mathrm{P})}$ はそれぞれ $\boldsymbol{M}^{(\mathrm{Q})}$ 及び $\boldsymbol{Y}_n^{(\mathrm{P})}$ の要素である．DNN が各音源の時間 j における非間引き周波数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 (3) 及び (4) 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．

$$\overline{\boldsymbol{m}}_j^{(\mathrm{Q})} = \left(\boldsymbol{m}_{j-2c}^{(\mathrm{Q})\mathrm{T}}, \cdots, \boldsymbol{m}_j^{(\mathrm{Q})\mathrm{T}}, \cdots, \boldsymbol{m}_{j+2c}^{(\mathrm{Q})\mathrm{T}}\right)^{\mathrm{T}} \tag{5}$$

$$\overline{\boldsymbol{y}}_{nj}^{(\mathrm{P})} = \left(\boldsymbol{y}_{n(j-2c)}^{(\mathrm{P})\mathrm{T}}, \cdots, \boldsymbol{y}_{nj}^{(\mathrm{P})\mathrm{T}}, \cdots, \boldsymbol{y}_{n(j+2c)}^{(\mathrm{P})\mathrm{T}}\right)^{\mathrm{T}} \tag{6}$$

$$\boldsymbol{b}_j = \left(\overline{\boldsymbol{m}}_j^{(\mathrm{Q})\mathrm{T}}, \overline{\boldsymbol{y}}_{1j}^{(\mathrm{P})\mathrm{T}}, \cdots, \overline{\boldsymbol{y}}_{nj}^{(\mathrm{P})\mathrm{T}}, \cdots, \overline{\boldsymbol{y}}_{Nj}^{(\mathrm{P})\mathrm{T}}\right) \in \mathbb{C}^{(2C+1)(|\mathbb{F}'|+N|\mathbb{F}|)} \tag{7}$$

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．この $\boldsymbol{b}_j$ の各要素の振幅値を取ったベクトル $|\boldsymbol{b}_j|$ が DNN の入力ベクトルとなる．ここで，ベクトルや行列に対する絶対値記号 $|\cdot|$ は，要素毎の振幅値を取ったベクトルや行列である．

2.3　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 [7] 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．

n 番目の音源信号の間引き周波数成分行列の振幅値 $|\boldsymbol{Y}_n^{(\mathrm{Q})}|$ を推定するためのソフトマスクを $\boldsymbol{W}_n \in \mathbb{R}_{[0,1]}^{|\mathbb{F}'| \times J}$ とし，次式のように表す．

$$|\boldsymbol{Y}_n^{(\mathrm{Q})}| \approx \boldsymbol{W}_n \odot |\boldsymbol{M}^{(\mathrm{Q})}| \tag{8}$$

ここで，$\odot$ は要素毎の積を表す．DNN の出力ベクトル $\tilde{\boldsymbol{w}}_j$ は，全音源のソフトマスク $\boldsymbol{W}_1, \cdots, \boldsymbol{W}_N$ の時間フレーム j におけるベクトルを連結したものである．

$$\tilde{\boldsymbol{w}}_j = \left(\boldsymbol{w}_{1j}^{\mathrm{T}}, \cdots, \boldsymbol{w}_{nj}^{\mathrm{T}}, \cdots, \boldsymbol{w}_{Nj}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}_{[0,1]}^{N|\mathbb{F}'|} \tag{9}$$

$$\boldsymbol{w}_{nj} = \left(w_{n1j}, \cdots, w_{nf'j}, \cdots, w_{n|\mathbb{F}'|j}\right)^{\mathrm{T}} \in \mathbb{R}_{[0,1]}^{|\mathbb{F}'|} \tag{10}$$

ここで，$w_{nf'j}$ は $\boldsymbol{W}_n$ の要素であり，出力層の Softmax 関数により，$\sum_n w_{nf'j} = 1 \ \forall \ f', j$ である．全音源の非間引き周波数成分の正解ベクトルは次式となる．

$$|\tilde{\boldsymbol{y}}_j| = \left(|\boldsymbol{y}_{1j}^{(\mathrm{Q})}|^{\mathrm{T}}, \cdots, |\boldsymbol{y}_{nj}^{(\mathrm{Q})}|^{\mathrm{T}}, \cdots, |\boldsymbol{y}_{Nj}^{(\mathrm{Q})}|^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{N|\mathbb{F}'|} \tag{11}$$

$$\boldsymbol{y}_{nj}^{(\mathrm{Q})} = \left(y_{n1j}^{(\mathrm{Q})}, \cdots, y_{nf'j}^{(\mathrm{Q})}, \cdots, y_{n|\mathbb{F}'|j}^{(\mathrm{Q})}\right)^{\mathrm{T}} \in \mathbb{C}^{|\mathbb{F}'|} \tag{12}$$

ここで，$y_{nf'j}^{(\mathrm{Q})}$ は $\boldsymbol{Y}_n^{(\mathrm{Q})}$ 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

[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．

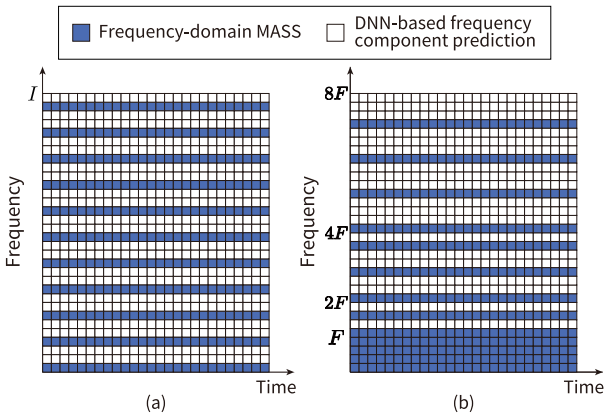

Fig. 4　Proposed frequency decimations: (a) even-interval with $D = 3$ and (b) uneven-interval decimations.

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

最小となるように学習される．

$$\mathrm{MSE}(|\tilde{\boldsymbol{y}}_j|, \tilde{\boldsymbol{w}}_j \odot |\tilde{\boldsymbol{m}}_j|) = \frac{1}{N|\mathbb{F}'|}\left\| |\tilde{\boldsymbol{y}}_j| - \tilde{\boldsymbol{w}}_j \odot |\tilde{\boldsymbol{m}}_j| \right\|_2^2 \tag{13}$$

$$\tilde{\boldsymbol{m}}_j = \left(\underbrace{\boldsymbol{m}_j^{(\mathrm{Q})^{\mathrm{T}}}, \cdots, \boldsymbol{m}_j^{(\mathrm{Q})^{\mathrm{T}}}}_{N} \right) \in \mathbb{C}^{N|\mathbb{F}'|} \tag{14}$$

3　DNN の予測が内挿となる間引き方法

文献 [5, 6] では，低周波帯域と高周波帯域に分割するような間引き方法で周波数領域 MASS に係る計算コストを削減していた．DNN に基づく周波数成分予測の精度を向上するために，本稿では DNN の予測が内挿となる新しい間引き方法を提案する．

3.1　等間隔間引き手法

等間隔に周波数ビンを間引く手法を，Fig. 4(a) に示す．これは，周波数領域 MASS に入力される周波数ビンの集合を $\mathbb{F} = \{i | i = 1, D+1, 2D+1, 3D+1, \cdots\}$ と定義することと等価である（$\mathbb{F}'$ は $\mathbb{F}$ の補集合とする）．ここで，D は等間隔間引きにおける間引き率の逆数である．即ち，間引き後の周波数ビン数と間引き前の周波数ビン数の関係は $|\mathbb{F}| = I/D$ となる．このように等間隔に周波数ビンを間引くことで，後段の間引かれた周波数の分離信号成分の予測は内挿となり，従来の高周波帯域予測（外挿）よりも高精度になることが期待される．

3.2　不等間隔間引き手法

前節の間引き手法では，全ての周波数において等間隔の間引き処理を行っている．しかしながら，人間の聴覚の認知は周波数に対して対数的に変化することから，低周波帯域はほとんど間引かず，高周波帯域ほど間引き数を増やす方が妥当である可能性がある．そこで，Fig. 4(b) のように，不等間隔に周波数ビンを間引く手法を提案する．

本手法では，基準となる周波数ビン F を定め，$0 \sim F$ の区間では，全周波数ビンを周波数領域 MASS に入力し，$F \sim 2F$ の区間では，周波数ビンを一つ飛ばしで間引く．更に，$2F \sim 4F$ の区間は二つ飛ばし，$4F \sim 8F$ の区間では三つ飛ばしのように，高周波帯域になるにつれ間引き間隔を増やしていくことで，不等間隔の間引きを実現する．

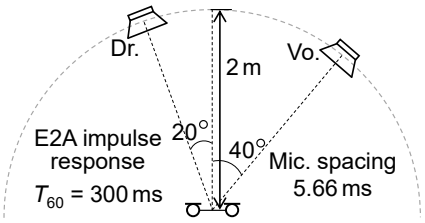

Fig. 5　Impulse responses used in experiment.

4　実験

4.1　条件

本実験では，Table 1 の信号に，Fig. 5 に示す RWCP [9] 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．比較手法として，全周波数ビンで MNMF を適用する手法（フルバンド MNMF），低周波帯域及び高周波帯域に分割（但し $\mathbb{F} = \{i | i = 1, 2, \cdots, \lfloor I/2 \rfloor\}$）する提案フレームワーク [5, 6]，等間隔間引き及び不等間隔間引きを行う提案フレームワークの 4 手法を用いた．提案フレームワークの DNN の学習データは文献 [5, 6] と同様である．MNMF の基底数については，全手法において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 30 本に設定した．MNMF の空間相関行列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，分離の度合いと音質の両方を示す source-to-distortion ratio（SDR）[8] の改善量を用いた．MNMF の計算には，Intel Core i7 8700 CPU，DNN の予測には NVIDIA GeForce GTX1660Ti GPU を用いた．

4.2　結果

Table 1 で示した 4 曲に対し，フルバンド MNMF，低周波帯域と高周波帯域に分割する従来手法，等間隔間引き手法，及び不等間隔間引き手法で実験を行った．各手法において，異なる乱数値を用いて 5 回実験を行った際の平均 SDR 改善量及び平均処理時間を Fig. 6 に示す．なお，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，$L = 10, 20, \cdots$ 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. 6 の提案手法は L の設定毎に SDR 値をプロットしている．低周波帯域と高周波帯域に分割する手法では，周波数領域 MASS に入力する行列の周波数ビン数が全体の

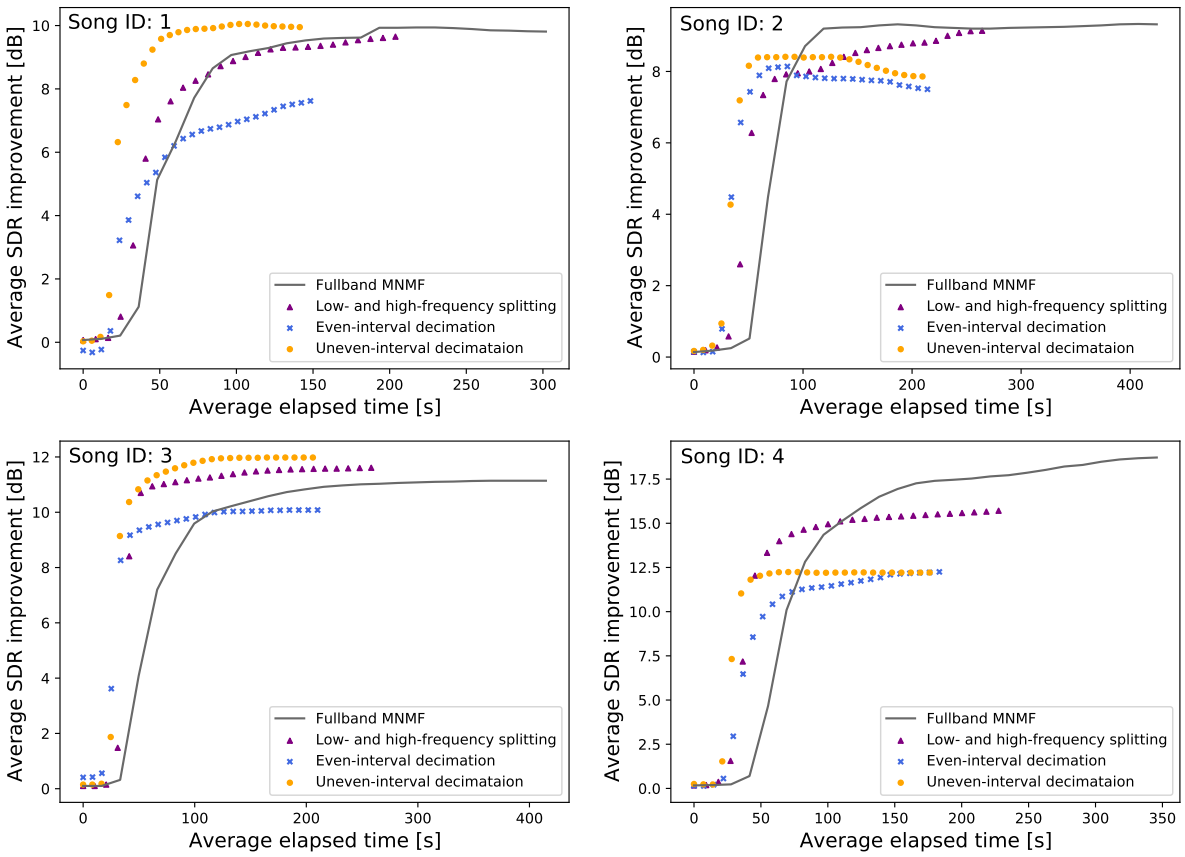

Fig. 6　Average SDR improvements and their elapsed times for each song.

1/2 に削減されており，等間隔及び不等間隔間引き手法ではおよそ 1/3 に削減されている．従って，計算時間もこれらの比率に応じて短縮されることを期待している．

結果を見ると，多少のオーバーヘッドが生じることから，提案フレームワークの手法の計算時間はフルバンド MNMF の 1/2 倍又は 1/3 倍までは削減されなかったが，それでもより短い時間で高い SDR 改善量を実現していることが分かる．Song ID1 及び Song ID3 の結果では，低周波帯域と高周波帯域に分割する手法と比較し，不等間隔間引き手法は高精度な音源分離が達成され，これは提案フレームワークの DNN の予測が内挿となった効果と予想される．しかし，等間隔間引き手法では性能改善が得られておらず，上記の要因は内挿か外挿かではなく，間引き方法の違いによって生じた可能性も残る．

5　おわりに

本稿では，周波数ビンを間引くことで全体の計算コストを削減する音源分離フレームワークにおいて，等間隔及び不等間隔の周波数間引き方法を新たに提案した．実験結果より，不等間隔間引きを行う手法は，低周波帯域及び高周波帯域に分割する手法とおおよそ同程度の性能を保ちながら，計算コストに削減できることが分かった．

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．

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] 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.

[3] 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.

[4] 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.

[5] 渡辺瑠伊, 北村大地, 猿渡洋, 高橋祐, 近藤多伸, “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” **日本音響学会** *2020* **年秋季研究発表会講演論文集**, pp. 131–134, 2020.

[6] R. Watanabe, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, “DNN-based frequency component prediction for frequency-domain audio source separation,” *Proc. EUSIPCO*, pp. 805–809, 2020.

[7] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” *arXiv preprint*, arXiv:1710.05941, 2017.

[8] 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.

[9] 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.