深層学習に基づく間引きインジケータ付き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），中村友彦，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．

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，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，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．そこで著者らは，DNN の予測が内挿となるように周波数ビンを間引く分割手法を新たに提案した [7]．この手法では，Fig. 2 に示すような方法で周波数領域 MASS を適用する周波数ビンと DNN で予測する周波数ビンを分割する．これらの分割手法を前述の音源分離フレームワーク [5, 6] に適用し，音源分離性能を評価したところ，Fig. 2 (b) のように不等間隔に間引くことで，高精度な音源分離を達成することが可能となった．以上の知見から，前段の音源分離処理に有効な周波数ビンを上手く選択することで，高精度な DNN の予測，延いては高品質な音源分離が可能になると考えられる．

そこで，本稿では任意の周波数ビンの間引き方（分割手法）に対応可能な音源分離フレームワークを提案する．本手法では，混合信号のスペクトログラムに対し，パワーが大きい帯域が音源分離に有効な周波数ビンであるとみなして分割する．この時，DNN は間引かれた周波数ビンの分離信号成分を予測しなければならないため，どこを間引いたかを表すインジケータを新たに設け，DNN の入力情報として用いた．

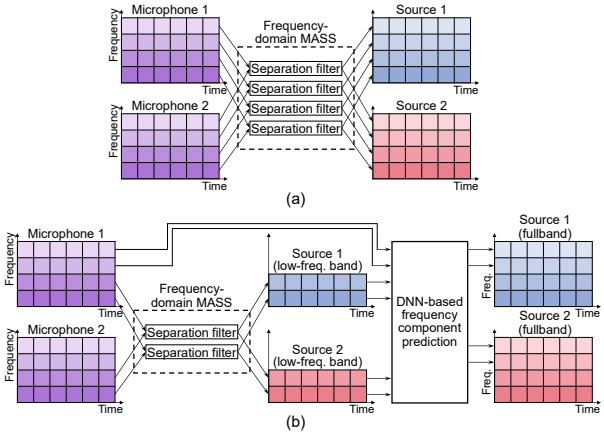

Fig. 1　(a) Full-band frequency-domain MASS and (b) proposed MASS frameworks using DNN-based frequency component prediction.

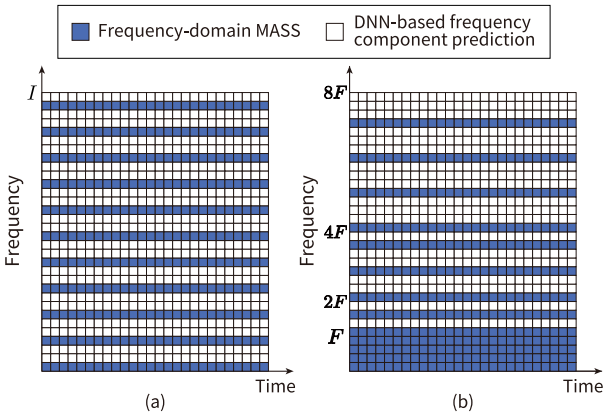

Fig. 2　Frequency decimations: (a) even-interval and (b) uneven-interval decimations.

2　従来手法：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．

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

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

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

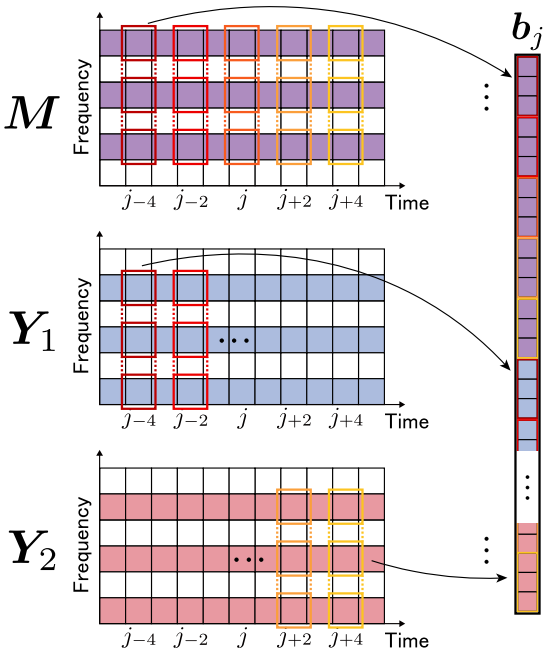

Fig. 3 Input vector of DNN, where $N = 2$, $I = 6$, and $J = 11$.

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} 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）[7]．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．

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. 3 に示す．但し，$\mathbb{F} = \{i | i = 1, 3, 5\}$ 及び $\mathbb{F}' = \{i | i = 2, 4, 6\}$ として等間隔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

本稿で用いる DNN の構造は文献 [7] と同様である．DNN で予測される各音源のソフトマスクは，周波数毎の音源に関する和が 1 となる必要があるため，周波数毎の Softmax 関数を適用する．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

[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 をハーフオーバーラップ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．

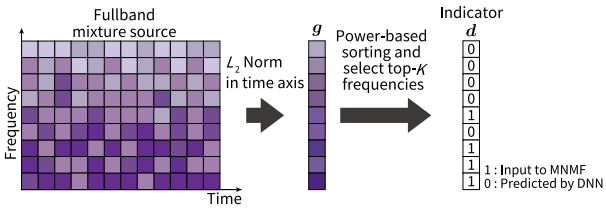

Fig. 4 Calculation of decimation indicator based on frequency-wise observed power values.

の非間引き周波数成分の正解ベクトルは次式となる．

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

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

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が最小となるように学習される．

3 提案手法

3.1 観測パワーに基づく周波数ビンの選択手法

Fig. 1 (b) の前段処理において，音源分離処理に有効な周波数ビンを選択することで，高精度な音源分離が期待できる．しかし，音源毎にヒューリスティックに選択を行うのは現実的ではない．そこで，混合信号の振幅スペクトログラムに対し，時間方向の L_2 ノルムを取った際にパワーの大きい周波数ビンが，高精度な音源分離に有効であると仮定する．Fig. 4 のように混合信号の各周波数ビンに対して，以下のような時間方向の L_2 ノルムを算出したベクトルを $\boldsymbol{g}$ とする．

$$\boldsymbol{g} = \left(\|\tilde{\boldsymbol{m}}_1\|, \|\tilde{\boldsymbol{m}}_2\|, \cdots, \|\tilde{\boldsymbol{m}}_I\|\right)^T \in \mathbb{R}_{\geq 0}^I \quad (13)$$

$$\tilde{\boldsymbol{m}}_i = (m_{i1}, m_{i2}, \cdots, m_{iJ})^{\mathrm{T}} \in \mathbb{C}^J \quad (14)$$

ベクトル $\boldsymbol{g}$ は周波数ビン毎の観測パワー値に対応するため，$\boldsymbol{g}$ の要素を降順にソートし，上位 K の要素の周波数ビンインデクスを $\mathbb{F}$ の要素と定義することで，パワーの大きい周波数ビンだけに周波数領域 MASS を適用できる．上位 K 以外の周波数ビンインデクスは $\mathbb{F}'$ の要素（即ち，間引かれる周波数ビン）となる．Fig. 4 は間引き率 1/2 の場合（$K = \lfloor I/2 \rfloor$）を示している．ここで，$\lfloor \cdot \rfloor$ は床関数である．

3.2 インジケータを用いたモデル学習

提案手法では，前節の方法に基づいて間引かれる周波数ビンが決まるため，DNN が分離信号を予測する帯域は観測信号毎に異なる．そこで，間引いた周波数ビンを明示的に示すインジケータを作成し，新たな DNN の入力に用いる．このインジケータを $\boldsymbol{d} \in \{0,1\}^I$ とし，音源分離処理を行う帯域（$\mathbb{F}$ の要素）を 1，DNN によって予測を行う帯域（$\mathbb{F}'$ の要素）を 0 とする．

$$\boldsymbol{d} = [d_1, d_2, \cdots, d_I]^T \quad (15)$$

$$d_i = \begin{cases} 1 & (\text{if } i \in \mathbb{F}) \\ 0 & (\text{otherwise}) \end{cases} \quad (16)$$

このベクトル $\boldsymbol{d}$ を式（7）に結合し，これを DNN の入力とする．

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

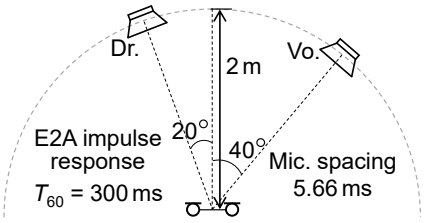

Fig. 5 Impulse responses used in experiment.

4 実験

4.1 条件

本実験では，Table 1 の信号に，Fig. 5 に示す RWCP [10] 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フレームワーク [7]，及び時間方向のパワーを基に分割し間引きインジケータを用いる提案フレームワークの 5 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, 7] と同様である．MNMF の基底数は 30 本とした．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）[9] 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

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，$L = 20, 40, \cdots$ 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 1/2 に削減されており，等間隔，不等間隔，パワーに基づく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 ID1〜3 に対しては，パワーに基づく分割手法が他の分割手法に比べ高精度の音源分離を達成している．これは，パワーの大きい周波数帯域を前段の音源分離処理に入力す

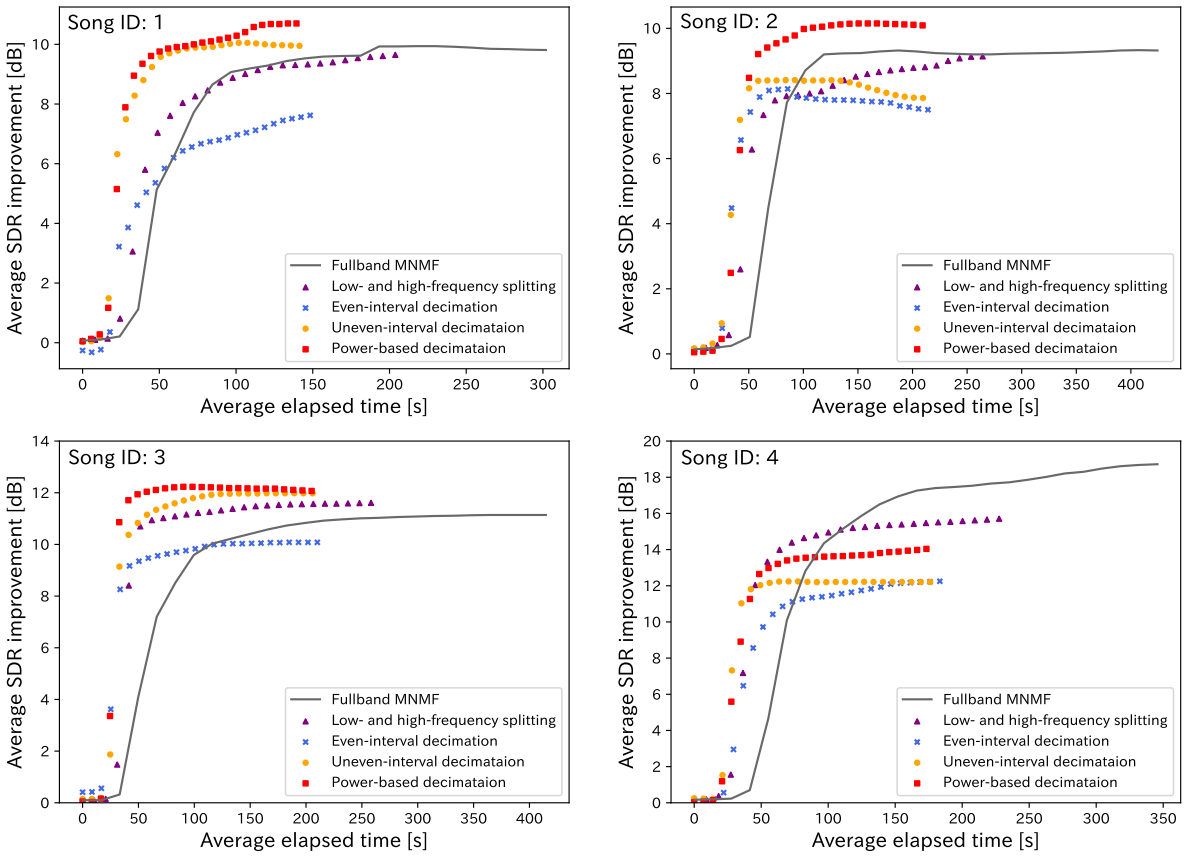

Fig. 6　Average SDR improvements and their elapsed times for each song.

ることで，高精度な分離信号が得られ，DNN の予測精度向上に寄与していると考えられる．また，Song ID4 について，低周波帯域と高周波帯域に分割する手法及びパワーに基づく間引き手法は，他の間引き手法と比べ性能が高いことから，Song ID4 は低周波帯域に信号のパワーが集中していると考えられる．従って，低周波帯域をより多く間引く等間隔及び不等間隔間引き手法の性能が低下したと推測される．

5　おわりに

本稿では，音源分離に有効な周波数ビンを任意に選択し分割する手法を提案し，従来の音源分離フレームワークに適用した．実験結果より，従来の各分割手法と比較して分離性能が向上したことが明らかとなった．従って，パワーの大きい周波数帯域が前段処理である，周波数領域 MASS に有効であるという裏付けとなる．このように，音源分離フレームワークの性能が向上するような分割の指標を調査する事が今後の課題である．

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．

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] 渡辺瑠伊, 北村大地, 猿渡洋, 高橋祐, 近藤多伸, “深層学習に基づく周波数帯域補間手法による音源分離処理の高速化” **日本音響学会** *2021* **年春季研究発表会講演論文集**, pp. 213–216, 2021.

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

[9] 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.

[10] 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.