Vol.2021-MUS-131 No.13
Vol.2021-SLP-137 No.13
2021/6/18

多重解像度深層分析を用いた楽音分離の実験的評価

中村 友彦[1,a)]　猿渡 洋[1,b)]

概要：時間領域音源分離は時間周波数領域を介さず直接観測音響信号を処理し分離音を出力する技術であり，end-to-end 型の深層ニューラルネットワーク（deep neural network: DNN）を用いた手法が注目を集めている．Wave-U-Net は時間領域音源分離のための DNN の 1 つであり，特徴量を畳み込み層と非線形関数で処理しつつ繰り返しダウンサンプリングした後，入力と同一の時間解像度まで同様に繰り返しアップサンプリングを行う構造を持つ．しかし，Wave-U-Net のダウンサンプリング層はデシメーションにより実装されているため，特徴量領域でエイリアシングが起こるだけでなく，音源分離に有用な情報を含みうる特徴量の一部を破棄してしまう．これらの問題を同時に解決するため，我々は以前多重解像度解析と U-Net 構造の類似性に着眼し，離散ウェーブレット変換（discrete wavelet transform: DWT）に基づくダウンサンプリング層（DWT 層）を用いた DNN ベース音源分離手法（多重解像度深層分析）を提案した．さらに，DWT 層に用いるウェーブレット基底関数を DNN と同時に学習できるよう拡張した．本稿では，これらの多重解像度深層分析とその拡張に対して詳細な検討を行うため，従来法との様々なモデルサイズでの比較を行う．この実験により，複数のモデルサイズにおいて多重解像度深層分析が従来法よりも高い分離性能を達成することを確認した．また，主観評価実験により，多重解像度深層分析が聴感上においても従来法より有意に高い分離性能を持つことを示した．

1. 序論

音源分離は観測音響信号から各音源信号を分離する技術であり，音楽音響信号加工，音声認識など様々なアプリケーションの前処理として利用できる．大量の学習データが利用できる場合，深層ニューラルネットワーク（deep neural network: DNN）を用いた教師あり音源分離が高い性能を示している [1]．

DNN を用いた音源分離は，時間周波数領域アプローチ，時間領域アプローチに大別される．時間周波数領域アプローチでは，DNN に振幅やパワースペクトログラムを入力して，各音源に対する時間周波数マスクを推定する [2–8]．このアプローチでは振幅やパワースペクトログラム領域での分離結果しか得られないため，分離結果を時間信号に変換するには適切な位相を付与する必要があり，典型的には観測信号の位相が用いられる．しかし，通常複素スペクトログラムは時間信号の冗長な表現となるため，時間信号から得られる複素スペクトログラムは同一次元の複素数ベクトル空間の特定の部分空間にしか存在しない [9]．そのため，観測信号の位相を振幅，パワースペクトログラム領域での分離結果と結合したものに対応する時間信号は，必ずしも存在するとは限らない．分離音の位相を同時に推定する方法 [10] や複素マスクを推定する手法 [11] も提案されているものの，分離結果に対応する時間信号が存在することは保証されない．一方，時間領域アプローチでは，観測音響信号を直接 DNN に入力し時間周波数領域を介さずに分離音の音響信号を出力する [12–17]．このアプローチでは，時間周波数領域アプローチでの位相に関する問題を回避することができるため，活発に研究が進められている．

Wave-U-Net は，時間領域アプローチの代表的な DNN である [15]．この DNN はエンコーダ，デコーダからなる U-Net 構造を持つ．エンコーダは，特徴量を畳み込み層と leaky rectified linear unit（ReLU）により処理しつつ，間引きを用いたダウンサンプリング（downsampling: DS）層により特徴量の時間解像度を半分にすることを繰り返す．本稿では，この DS 層をデシメーション層と呼ぶ．デコーダは，スキップコネクションによりエンコーダの各デシメーション層に入力される特徴量も参照しつつ，畳み込み層，leaky ReLU により処理しながら線形補間により信号と同一の時間解像度となるまで繰り返しアップサンプリング（upsampling: US）する．

Wave-U-Net の DS 構造は畳み込み層の受容野を指数的に広げ音源の長期の依存関係を捉えやすくするものの，以

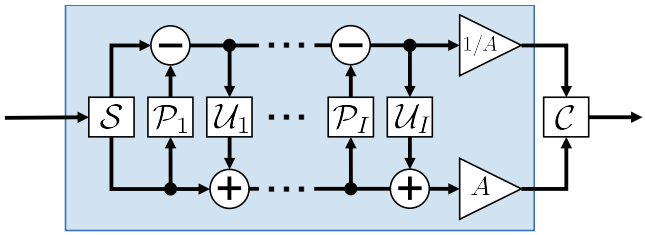
[1] 東京大学
Hongo, Bunkyo, Tokyo 113–8654, Japan
[a)] tomohiko-nakamura@g.ecc.u-tokyo.ac.jp
[b)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

前我々は信号処理の観点から見直すことでデシメーション層に内在する 2 つの問題を発見した [18,19]．信号処理の観点から DNN を再解釈すると，DNN と特徴量はそれぞれ縦続接続された非線形システムとそれらの流れる信号とみなせる．デシメーション層は単純な間引きにより実装されているため，特徴量領域でエイリアシングを引き起こす．音響認識タスク [20] や画像識別 [21] において，特徴量領域のエイリアシングにより性能が低下することが報告されており，当該エイリアシングは音源分離においても性能低下要因となりうる．これに対し画像分野では，アンチエイリアシングフィルタを DS 層の前に導入することで，画像識別性能が向上し，対角シフトした入力画像に関しても頑健に分類できることが報告されている [22]．

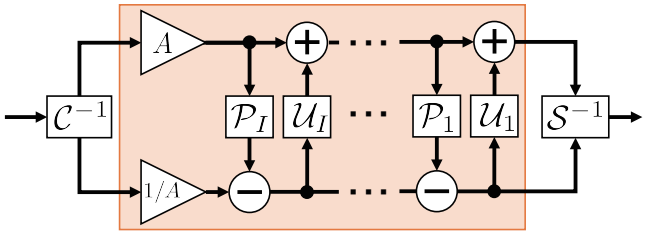
しかし，アンチエイリアシングフィルタを導入したとしても，デシメーション層は特徴量の一部を破棄してしまう．破棄された特徴量の部分に分離に有用な情報が含まれていた場合は，それより上階層に当該情報が伝搬されず分離性能の低下を招きうる．また，デコーダはスキップコネクションにより DS 前の特徴量を参照できるものの，後続の畳み込み層は並進不変性をもつため，どのインデックス成分が破棄されたか否かを区別できない．そのため，破棄された部分に含まれる情報を同階層のデコーダ部分で補償できるか否かは学習に大いに依存する．

これらの問題を同時に解決するため，Wave-U-Net と多重解像度解析 [23] の構造の類似性に着眼し，離散ウェーブレット変換（discrete wavelet transform: DWT）を用いた DS 層を構築し，DWT 層を Wave-U-Net に組み込んだ時間領域音源分離手法，多重解像度深層分析（multiresolution deep layered analysis: MRDLA）を提案した [18,19]．DWT は有限長のインパルス応答をもつ（finite impulse response: FIR）ローパスフィルタとハイパスフィルタからなる 2 チャネルフィルタバンクであり，アンチエイリアシングフィルタを備える．また，DWT で出力される 2 つのサブバンド信号から逆 DWT を用いることで入力信号を完全再構成できるため，DS により特徴量の情報が欠落しない．そのため，DWT 層を用いることでデシメーション層の 2 つの問題を同時に解決できる．DWT ではどのウェーブレットを選択するかによりローパス，ハイパスフィルタの周波数特性が変わるが，様々なウェーブレットでも多重解像度深層分析は頑健に動作することを実験的に確認した [24]．

さらに，我々はウェーブレット基底関数を DNN と同時学習できるように DWT 層拡張した [25]．しかし，単純にウェーブレット基底関数に対応するフィルタを学習できるよう拡張するだけではアンチエイリアシングフィルタを持つことを保証できていなかった．そこで，本稿では任意の FIR フィルタからなる DWT 層に関してもアンチエイリアシングフィルタを持つことを保証するため制約を導出し，



(a) DWT 層



(b) 逆 DWT 層

図 1: DWT 層と逆 DWT 層のブロック線図．青，橙色の領域はそれぞれ DS，US に関するリフティングスキームの部分を表す．$\mathcal{C}^{-1}, \mathcal{S}^{-1}$ は，それぞれ $\mathcal{C}, \mathcal{S}$ の逆を表す．

その制約を導入したウェーブレット学習可能な DWT 層を提案する．

また，本稿では多重解像度深層分析の性能を詳細に調査するため，ウェーブレットを学習する DWT 層の構造による差異や複数のモデルサイズにおける性能を比較する．これまでは従来法として Wave-U-Net のみと比較してきたため，それ以外の時間領域音源分離手法とも客観評価，主観評価を行う．

2. 多重解像度深層分析

2.1 DWT 層

本節では DWT 層の構造について述べる．DWT 層への入力特徴量を $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K] \in \mathbb{R}^{T \times K}$ とする．ここで，K はチャネル数，T は時間長を表す．以下では T は偶数の場合のみ考えるが，T が奇数の場合は DWT 層の直前にリフレクションパディング層を導入し $\boldsymbol{x}_k$ の時間長を偶数へと変えればよい．

DWT 層は，まず入力特徴量の各チャネル $\boldsymbol{x}_k$ を信号とみなして DWT を適用し，その後各チャネルでの DWT の出力をチャネル方向に結合することで，入力に対し半分の時間解像度をもつ特徴量 $[\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_{2K}] \in \mathbb{R}^{T/2 \times 2K}$ を生成する．DWT の実装にはリフティングスキームを用いた [26]．リフティングスキームは有限長のフィルタからなる DWT と逆 DWT の効率的な計算技法であり，時間分割，予測，更新，スケーリングの 4 つのステップから成る．

時間分割ステップでは，各チャネルの特徴量 $\boldsymbol{x}_k$ を偶数インデックス成分 $\boldsymbol{x}_k^{(\mathrm{even})} \in \mathbb{R}^{T/2}$，奇数インデックス成分 $\boldsymbol{x}_k^{(\mathrm{odd})} \in \mathbb{R}^{T/2}$ に分割する．この操作を $\mathcal{S}$ で表すと，当該ステップは以下のように書ける．

$$\boldsymbol{x}_k^{(\mathrm{even})}, \boldsymbol{x}_k^{(\mathrm{odd})} = \mathcal{S}\boldsymbol{x}_k \tag{1}$$

予測ステップでは，予測作用素 $\mathcal{P}$ を用いて式 (2) のように予測誤差 $\boldsymbol{d}_k \in \mathbb{R}^{T/2}$ を計算する．

$$\boldsymbol{d}_k = \boldsymbol{x}_k^{(\mathrm{odd})} - \mathcal{P}\boldsymbol{x}_k^{(\mathrm{even})} \tag{2}$$

更新ステップでは，時間分割ステップで生じた $\boldsymbol{x}_k^{(\mathrm{even})}$ のエイリアシングを低減するため，更新作用素 $\mathcal{U}$ を用いて $\boldsymbol{x}_k^{(\mathrm{even})}$ を式 (3) のように平滑化する．

$$\boldsymbol{c}_k = \boldsymbol{x}_k^{(\mathrm{even})} - \mathcal{U}\boldsymbol{d}_k \tag{3}$$

ここで，$\boldsymbol{c}_k \in \mathbb{R}^{T/2}$ は平滑化された偶数インデックス成分を表す．スケーリングステップでは，正規化定数 A を用いて式 (4) のように $\boldsymbol{c}, \boldsymbol{d}$ をスケーリングする．

$$\tilde{\boldsymbol{c}}_k = A\boldsymbol{c}_k, \quad \tilde{\boldsymbol{d}}_k = \frac{1}{A}\boldsymbol{d}_k \tag{4}$$

得られた $\tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{d}}_k$ はそれぞれ低周波成分，高周波成分に対応する．ここで，$\mathcal{P}$,$\mathcal{U}$ として非線形関数を用いることもできるが，非線形フィルタに対する周波数応答の解析は一般には困難なため，本稿ではどちらもインパルス応答長 M の FIR フィルタとする．また，更新，予測ステップは複数繰り返してもよく，各ステップで $\mathcal{P}$,$\mathcal{U}$ も異なっても良い．そのため，以下では I ペアの更新，予測ステップを持つリフティングスキームを考え，$\mathcal{P}, \mathcal{U}, \boldsymbol{c}, \boldsymbol{d}, M$ に下付き添字 $i = 1, \cdots, I$ を付与し i 番目の更新，予測ステップに関する変数であることを表す．

図 1(a) に DWT 層のブロック線図を示す．ここで，$\mathcal{C}$ は全チャネルの $\tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{d}}_k$ をチャネル方向に結合する操作を表す．FIR フィルタは時間反転したインパルス応答と信号の相関演算と等価であるため，予測，更新作用素による演算はチャネル数 1 の畳み込み層をチャネルそれぞれに適用することで実現できる．DWT 層を構成する各ステップは全て可逆であるため，決定的な $\mathcal{P}_i$,$\mathcal{U}_i$ を用いれば完全再構成性は満たされる．また，DWT 層の逆過程を行うことで逆 DWT 層を利用した US 層（逆 DWT 層）が定義できる（図 1(b) 参照）．

2.2 ネットワーク構造

図 2 に多重解像度深層分析で用いる DNN の構造を示す．この DNN は Wave-U-Net を基に構築されており，エンコーダ，デコーダはそれぞれ L 個の DS，US ブロックからなる．ここで，$l = 1, \ldots, L$ を階層インデックス，N を音源数，$C^{(\mathrm{in})}$ を観測音響信号のチャネル数とする．l 番目の DS ブロックは，フィルタ長 $f^{(\mathrm{e})}$，チャネル数 $C^{(\mathrm{e})}l$ の 1 次元畳み込み層，leaky ReLU 非線形関数，DWT 層からなる．エンコーダとデコーダの間にはボトルネックブロックがあり，フィルタ長 $f^{(\mathrm{e})}$，チャネル数 $C^{(\mathrm{m})}$ の 1 次元畳み込み層，leaky ReLU 非線形関数からなる．l 番目の US

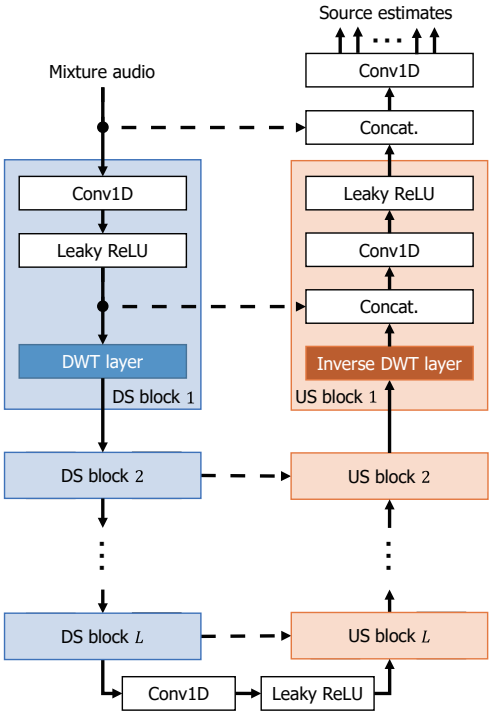


図 2: 多重解像度深層分析で用いる DNN．青，橙色の領域はそれぞれ DS，US に関する部分である．Conv1D，Leaky ReLU，Concat. は，それぞれ 1 次元畳み込み層，leaky ReLU 非線形関数，入力特徴量をチャネル方向へ結合する操作を表す．

ブロックは，まずスキップコネクションから得られた l 番目の DS ブロック内の DWT 層の入力特徴量と，逆 DWT 層によって US された $l+1$ 番目の US ブロックの出力，またはボトルネックブロックの出力をチャネル方向に結合する．その後，結合した特徴量にフィルタ長 $C^{(\mathrm{d})}$，チャネル数 $C^{(\mathrm{d})}l$ の 1 次元畳み込み層，leaky ReLU 非線形関数を適用する．デコーダから出力された特徴量は観測音響信号とチャネル方向に結合された後，フィルタ長 1，チャネル数 $NC^{(\mathrm{in})}$ の 1 次元畳み込み層により処理され，推定音源信号が出力される．推定音源信号の端にアーティファクトが生じることを避けるため，Wave-U-Net と同様に畳み込み層ではパディングを用いない．

2.3 出力層の変更

多重解像度深層分析では，出力層に関して Wave-U-Net から 2 点変更を行った．Wave-U-Net では，$N-1$ 個の分離音を推定した後 N 番目の分離音は混合音から $N-1$ 個の分離音を減算して得る．この方法は学習中であっても分離音の和が混合音と一致するものの，1 つの音源に対する分離の失敗がその他の楽器の分離音の推定に波及し，分離性能の低下に繋がりうる．これに対し，多重解像度深層分析では直接 N 個の分離音を推定する．この方法では分離音の和が混合音と一致することが保証されないものの，実験的には分離音の和と混合音の平均 2 乗誤差が十分小さく

なり分離性能も向上した．

また，Wave-U-Net は最後の畳み込み層の後に双曲線正接関数を適用するが，多重解像度深層分析では用いない．これは，4 節の実験で各曲の混合音の平均が 0，分散が 1 となるようにデータ標準化を用いるため，学習データの時間信号の値が $[-1, 1]$ の範囲に入るとは限らないからである．

3. DNN と予測，更新作用素の同時学習

3.1 予測，更新作用素とアンチエイリアシングフィルタ

リフティングスキームを構成するステップは全て可逆であるため，リフティングスキームの構造から DWT 層の完全再構成性は保証される．一方，DWT の周波数応答は $\mathcal{P}_i$,$\mathcal{U}_i$ をどう設計するかに依存する．実際，単純に $\mathcal{P}_i$,$\mathcal{U}_i$ のインパルス応答を学習パラメータとした DWT 層では完全再構成性しか保証されない [25]．この学習可能な予測，更新作用素を用いた DWT 層を trainable DWT 層（TDWT 層）と呼ぶ．本節では，アンチエイリアシングフィルタの存在を保証できる $\mathcal{P}_i$,$\mathcal{U}_i$ の設計方法を導出するため，まず DWT のフィルタとリフティングスキームの $\mathcal{P}_i$,$\mathcal{U}_i$ の z 変換領域での関係式を導出する．

DWT を構成するローパス，ハイパスフィルタの z 変換を $H_I(z), G_I(z)$，$\mathcal{P}_i$,$\mathcal{U}_i$ の z 変換を $P_i(z), U_i(z)$ とする．また，以下では簡単のためチャネルインデックス k は省略する．$\boldsymbol{x}^{(\mathrm{even})}, \boldsymbol{x}^{(\mathrm{odd})}, \tilde{\boldsymbol{c}}_i, \tilde{\boldsymbol{d}}_i$ の z 変換を，それぞれ $X^{(\mathrm{even})}(z), X^{(\mathrm{odd})}(z), \tilde{C}_i(z), \tilde{D}_i(z)$ で表す．リフティングスキームの予測，更新，スケーリングステップはそれぞれ 2×2 の行列で表現できるため，$[\tilde{C}_I(z), \tilde{D}_I(z)]^\top$ は式 (5) のように書ける [27]．

$$\begin{bmatrix} \tilde{C}_I(z) \\ \tilde{D}_I(z) \end{bmatrix} = Q_I(z) \begin{bmatrix} X^{(\mathrm{even})}(z) \\ X^{(\mathrm{odd})}(z) \end{bmatrix} \tag{5}$$

ここで，$Q_I(z)$ は以下のように定義される．

$$Q_I(z) = \underbrace{\begin{bmatrix} A & 0 \\ 0 & 1/A \end{bmatrix}}_{\text{スケーリングステップ}} \times \prod_{i=1}^{I} \left(\underbrace{\begin{bmatrix} 1 & U_i(z) \\ 0 & 1 \end{bmatrix}}_{i \text{ 番目の更新ステップ}} \underbrace{\begin{bmatrix} 1 & 0 \\ -P_i(z) & 1 \end{bmatrix}}_{i \text{ 番目の予測ステップ}} \right) \tag{6}$$

任意の FIR フィルタから構成される DWT は，式 (6) で表現できることが証明されている [27]．

一方，$H_I(z), G_I(z)$ からも同様の関係式を導出できる．$H_I(z), G_I(z)$ を偶数次の z 成分を集めたフィルタ $H_I^{(\mathrm{even})}(z), G_I^{(\mathrm{even})}(z)$ と奇数次の z 成分を集めたフィルタ $H_I^{(\mathrm{odd})}(z), G_I^{(\mathrm{odd})}(z)$ で表すと，それぞれ式 (7)，(8) となる．

$$H_I(z) = H_I^{(\mathrm{even})}(z^2) + z^{-1} H_I^{(\mathrm{odd})}(z^2) \tag{7}$$

$$G_I(z) = G_I^{(\mathrm{even})}(z^2) + z^{-1} G_I^{(\mathrm{odd})}(z^2) \tag{8}$$

この表現は $H_I(z), G_I(z)$ のポリフェーズ表現と呼ばれる [28]．$H_I(z), G_I(z)$ のポリフェーズ表現を用いて，DWT は行列形式で表現できる [27]．

$$\begin{bmatrix} \tilde{C}_I(z) \\ \tilde{D}_I(z) \end{bmatrix} = \begin{bmatrix} H_I^{(\mathrm{even})}(z) & H_I^{(\mathrm{odd})}(z) \\ G_I^{(\mathrm{even})}(z) & G_I^{(\mathrm{odd})}(z) \end{bmatrix} \begin{bmatrix} X^{(\mathrm{even})}(z) \\ X^{(\mathrm{odd})}(z) \end{bmatrix} \tag{9}$$

ここで，式 (9) の右辺の行列を $\tilde{Q}_I(z)$ と置き，式 (9) と式 (5) を比較すると，以下の式が得られる．

$$Q_I(z) = \tilde{Q}_I(z) \tag{10}$$

式 (10) の左辺は $P_i(z), U_i(z)$，右辺は $H_I(z), G_I(z)$ のみに依存するため，$H_I(z), G_I(z)$ が $P_i(z), U_i(z)$ により定まることが確認できる．

3.2 DWT 層がアンチエイリアシングフィルタを持つための制約条件

3.2.1 $I = 1$ での制約条件

本節では，式 (10) とローパス，ハイパスフィルタの定義から，DWT 層がアンチエイリアシングフィルタをもつための $P_i(z), U_i(z)$ に関する条件を導出する．$H_i(z), G_I(z)$ がローパス，ハイパスフィルタである条件は以下のように与えられる．

$$|H_I(1)| > 0, \quad H_I(-1) = 0 \tag{11}$$

$$|G_I(-1)| > 0, \quad G_I(1) = 0 \tag{12}$$

$H_I(z), G_I(z)$ のポリフェーズ表現 (7)，(8) を条件 (11)，(12) に代入することで，条件は以下のように変換される．

$$|H_I^{(\mathrm{even})}(1)| > 0, \quad H_I^{(\mathrm{even})}(1) = H_I^{(\mathrm{odd})}(1) \tag{13}$$

$$|G_I^{(\mathrm{even})}(1)| > 0, \quad G_I^{(\mathrm{even})}(1) = -G_I^{(\mathrm{odd})}(1) \tag{14}$$

以下ではまず $I = 1$ の場合を考え，その後 $I > 1$ の場合に拡張する．

t を離散時間インデックスとし，$P_i(z), U_i(z)$ のインパルス応答を $\{p_{i,t}\}_t, \{u_{i,t}\}_t$ とすると，以下の補題が成り立つ．

補題 1. $I = 1$ のとき，$H_1(z)$ と $G_1(z)$ がそれぞれローパス，ハイパスフィルタである必要十分条件は，式 (15) である．

$$\sum_t p_{1,t} = 1, \quad \sum_t u_{1,t} = \frac{1}{2} \tag{15}$$

証明は付録 A.1 を参照されたい．予測，更新作用素は畳み込み層として実装できるため，$I = 1$ の場合は対応する畳み込み層の重みを式 (15) を満たすように正規化する操作を導入すれば良い．

Vol.2021-MUS-131 No.13
Vol.2021-SLP-137 No.13
2021/6/18

Algorithm 1 WN-TDWT 層の計算

Input: $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K], \{p_{i,t}\}_{i,t}, \{u_{i,t}\}_{i,t}, A$
Output: $\{\tilde{x}_{k'}\}_{k'}$

for $k = 1$ to K **do**
  式 (1) に従い $\boldsymbol{x}_k^{(\mathrm{even})}, \boldsymbol{x}_k^{(\mathrm{odd})}$ を計算
  for $i = 1$ to I **do**
    if $i = 1$ **then**
      $\tilde{p}_{i,t} \leftarrow p_{i,t} - (\sum_\tau p_{i,\tau}/M_i - 1/M_i)$
      $\tilde{u}_{i,t} \leftarrow u_{i,t} - (\sum_\tau u_{i,\tau}/M_i - 1/2M_i)$
    else
      $\tilde{p}_{i,t} \leftarrow p_{i,t} - \sum_\tau p_{i,\tau}/M_i$
      $\tilde{u}_{i,t} \leftarrow u_{i,t} - \sum_\tau u_{i,\tau}/M_i$
    end if
    $\tilde{p}_{i,t}$ を予測作用素のインパルス応答として用いて，式 (2) に従い $\boldsymbol{d}_k$ を計算
    $\boldsymbol{x}_{(\mathrm{odd})} \leftarrow \boldsymbol{d}_k$
    $\tilde{u}_{i,t}$ を更新作用素のインパルス応答として用いて，式 (3) に従い $\boldsymbol{c}_k$ を計算
    $\boldsymbol{x}_{(\mathrm{even})} \leftarrow \boldsymbol{c}_k$
  end for
  式 (4) に従い $\tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{d}}_k$ を計算
end for
$\{\tilde{\boldsymbol{c}}_k\}_k, \{\tilde{\boldsymbol{d}}_k\}_k$ をチャネル方向に結合し $\{\tilde{x}_{k'}\}_{k'}$ を生成

3.2.2 $I > 1$ での制約条件

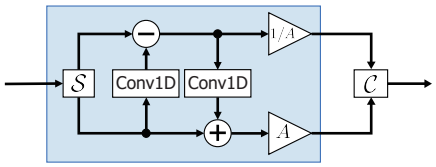
$I = 1$ の場合と同様に $I > 1$ でも必要十分条件を導出することもできるが，複数の $p_{i,t}, u_{i,t}$ 同士が複雑に関連する条件となってしまう．そのため，個々の畳み込み層に対する制約として導入することは難しく実装を困難にする．そこで，既存のウェーブレットから新たに周波数特性の異なるウェーブレットを体系的に構築できるリフティングスキームの特性を活かし，畳み込み層に対する単純な制約となるように十分条件を導出する．

$H_I(z), G_I(z)$ が条件 (13)，(14) を満たすとする．このとき以下の補題が成り立つ．

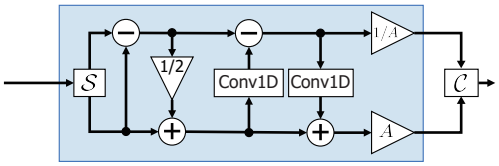
補題 2. ローパス，ハイパスフィルタ $H_I(z), G_I(z)$ をもつリフティングスキームに対し，$I+1$ 番目の予測，更新ステップをスケーリングステップの直前に導入する．この操作により得られたリフティングスキームのフィルタ $H_{I+1}(z), G_{I+1}(z)$ が，条件 (13)，(14) を満たすための十分条件は，式 (16) で与えられる．

$$\sum_t p_{I+1,t} = 0, \quad \sum_t u_{I+1,t} = 0 \tag{16}$$

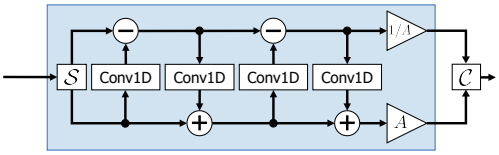
証明は付録 A.2 を参照されたい．補題 2 から，予測，更新作用素に対応する畳み込み層の重みを式 (16) に従って正規化する処理を導入すれば，ローパスフィルタ特性と完全再構成性を保証できる．補題 1，2 から得られた結果をまとめると，アンチエイリアシングフィルタと完全再構成性を保持した TDWT 層は Algorithm 1 に示すように計算できる．この DWT 層を weight-normalized trainable DWT 層（WN-TDWT 層）と呼ぶ．



(a) タイプ A



(b) タイプ B



(c) タイプ C

図 3: TDWT 層，WN-TDWT 層の構造例．表記は図 1，2 と同様である．

3.3 TDWT 層，WN-TDWT 層の構造

図 3 に，4 節の実験で用いるウェーブレットを学習可能な DWT 層の構造を示す．これらの構造は TDWT 層，WN-TDWT 層で共通とした．図 3(a) は，$I = 1$ で全ての予測，更新作用素を学習するものである．TDWT 層，WN-TDWT 層では一部の予測，更新作用素のみを事前に与え残りの作用素のみを学習してもよい．そのため，図 3(b) のように，Haar ウェーブレットに対応する予測，更新作用素を 1 段目におき，その部分は学習せずに 2 段目の予測，更新作用素のみを学習することもできる．これは，リフティングスキームの特性を活かしデータ駆動で Haar ウェーブレットを修正することに対応する．図 3(c) は，図 3(b) の 1 段目の予測，更新作用素も学習できるようにしたものである．これら 3 つの構造を区別するため，本稿ではそれぞれ順にタイプ A，B，C と呼ぶ．これらの層構造をもつ TDWT 層，WN-TDWT 層に対応する US 層も DWT 層と同様に定義でき，それぞれ逆 TDWT 層と重み正規化付き逆 WN-TDWT 層と呼ぶ．

TDWT 層，WN-TDWT 層では DWT 層に比べ，正規化処理と予測，更新作用素の学習コストがかかるものの，正規化処理については並列に実行可能なため計算時間の増加は小さい．一方，予測，更新作用素の学習にはバックプロパゲーションを用いるため，$p_{i,t}, u_{i,t}$ の勾配の計算と保持が学習中に必要となる．しかし，各作用素のパラメータ数は M_i であり，これらの DWT 層の導入によるモデルサイズの増加は他の DNN のパラメータ数に比べて非常に小さ

く，相対的に無視できる．

4. 実験的評価

4.1 実験条件

多重解像度深層分析の性能を評価するため，MUSDB18 データセット [29] を用いた楽音分離実験を行った．当該データセットは学習用データ 100 曲，テストデータ 50 曲からなり，幅広いジャンルの曲が収録されている．各曲につき，vocals，bass，drums，other の 4 つの楽器毎の音源信号とそれらの混合音が利用できる．ここで，other は曲に含まれる vocals，bass，drums 以外の楽器全てに対応する．学習データや DNN のパラメータの初期値への依存性を低減するため，学習用データ 100 曲に対し，75 曲を学習データ，25 曲を検証データとした 4 ペアを作成し，4 分割の交差検定を行った．Wave-U-Net の文献 [15] での実験条件と同じく，サンプリング周波数は 22.05 kHz としステレオ音源のまま用いた．

学習データに関して，各曲の混合音の時間波形が平均 0，分散 1 となるようにデータ標準化を行った．バッチサイズは 16 とし，バッチ生成の際には各曲から 6.68 秒（147443 サンプル）の区間をランダムに切り出した後，データ拡張として $[0.75, 1.25]$ の範囲のランダムゲインの付加，左右のチャネルのランダムな入れ替え，バッチ内の 20 %のデータに対し他曲の対応する楽器音との入れ替えを行った．これらのデータ拡張はバッチ生成時に毎回行った．ロス関数には推定分離音と正解の各音源信号との時間領域での平均二乗誤差を用い，学習率 1.0×10^{-4}，減衰率 $0.9, 0.999$ とした Adam を用いて最適化を行った．Wave-U-Net の文献 [15] での実験条件と同じく，2000 反復を 1 エポックとし過学習抑制のため 2 段階のアーリーストッピングを用いた．最初に，検証ロスが 20 エポック連続して下がらなくなるまで各モデルを学習する．その後，バッチサイズを 32，学習率を 1.0×10^{-5} に変更し，再度検証ロスが連続して 20 エポック連続して下がらなくなるまでファインチューニングを行い，最も検証ロスが小さいモデルを学習済みモデルとする．他のハイパーパラメータは，$A = \sqrt{2}, L = 12, C^{(\mathrm{m})} = 312, C^{(\mathrm{d})} = 24, f^{(\mathrm{e})} = 15, f^{(\mathrm{d})} = 5$ とした．

評価指標として，BSSEval v4 ライブラリ [1] により得られる source-to-distortion ratio（SDR）[30] を用いた．当該ライブラリは，各曲，各楽器で 1 秒毎に推定分離音の SDR を計算し，それらの中央値（トラックワイズ SDR）を得る．曲に関するトラックワイズ SDR の中央値（メディアン SDR）を評価値として出力する．計算の詳細は [1] を参照されたい．本実験では交差検証を用いるため，各データ分割でメディアン SDR を求めた後それらの平均と標準誤差を算出し評価値として用いた．

4.2 TDWT 層と WN-TDWT 層の比較

4.2.1 重みの初期化方法の影響

まず，TDWT 層，WN-TDWT 層の分離性能に関する比較を行う．TDWT 層，WN-TDWT 層を用いた多重解像度深層分析の DNN を，それぞれ TDWT モデル，WN-TDWT モデルと呼ぶ．TDWT，WN-TDWT 層の構造として図 3 に示したタイプ A，B，C の構造を用い，学習可能な予測，更新作用素のフィルタ長は 3 とした．各ネットワークで学習する作用素に対応する畳み込み層の重みは，対応する逆 DWT 層も含め全階層で共有し学習を行った．

我々が TDWT 層を提案した文献 [25] では，予測，更新作用素に関して Haar ウェーブレットと同一になるよう初期化を行った実験結果しか示していなかった．そこで，まず TDWT 層，WN-TDWT 層に対する初期化の影響を調べるため，タイプ A の TDWT 層，WN-TDWT 層に対しそれぞれ $p_{1,t}, u_{1,t}$ をランダムに初期化した場合（ランダム初期化）と Haar ウェーブレットの予測，更新作用素で初期化した場合（Haar 初期化）を比較する．ここで，エンコーダに関するハイパーパラメータは $C^{(\mathrm{e})} = 18$ とした．表 1 に，TDWT モデルと WN-TDWT モデルでのメディアン SDR の平均と標準誤差を示す．TDWT モデルに関しては，重みの初期化方法が数値安定性や分離性能を大きく影響した．ランダム初期化を用いると，学習ロスや検証ロスに関して収束はするものの突然急激に上昇することがあった．一方，Haar 初期化を用いるとそのようなロスの急激な上昇は観測されず，分離性能もランダム初期化に比べ大きく向上した．WN-TDWT モデルは，初期化によらず同程度の数値安定性や分離性能であった．この結果は，重み正規化を導入しアンチエイリアシングフィルタを保証することで，重みの初期値依存性を低減できることを示している．

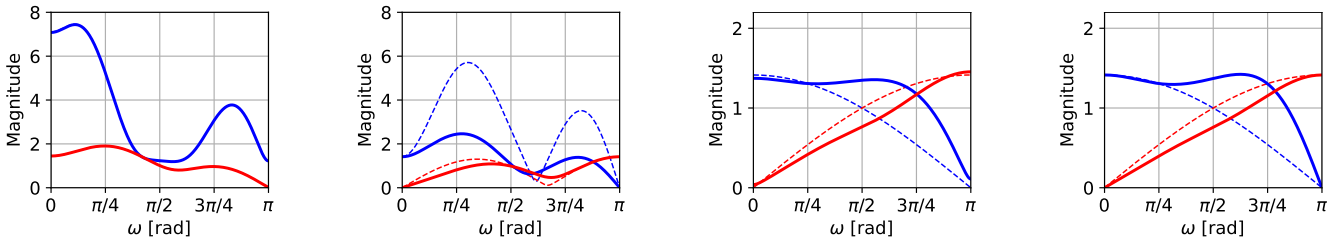
図 4 に，ランダム初期化，Haar 初期化の場合の TDWT 層，WN-TDWT 層の周波数応答を示す．ランダム初期化の TDWT 層では周波数応答が初期値から変わらなかったのに対し，WN-TDWT 層ではピーク位置はそれほど変化しなかったもののローパスフィルタのゲインが大きく変化した．Haar 初期化の場合は，TDWT 層，WN-TDWT 層ともに Haar ウェーブレットと類似しているものの，ローパスフィルタのカットオフ周波数が高周波数に寄った周波数応答が得られた．TDWT 層，WN-TDWT 層共に Haar 初期化での SDR の方がランダム初期化よりも高かったため，次節以降の実験では Haar 初期化を用いた．

4.2.2 分離性能の比較

次に，タイプ A，B，C での TDWT，WN-TDWT モデルの分離性能を比較する．両タイプ共に Haar 初期化となるように，1 段目の畳み込み層は Haar ウェーブレットの予測，更新作用素で初期化し，2 段目の畳み込み層の重みはゼロで初期化した．表 1 に示す通りタイプの差異はメディ

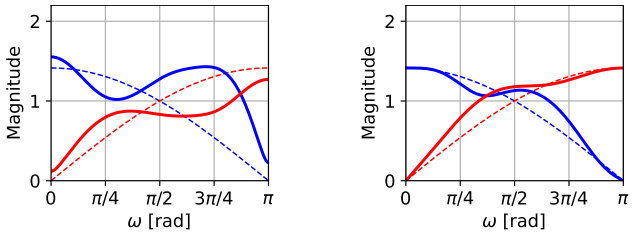
表 1: TDWT モデルと WN-TDWT モデルの SDR [dB].

Initialization	Architecture	DS layer	Instrument			
			vocals	bass	drums	other
Random	Type A	TDWT	2.68 ± 0.03	3.07 ± 0.13	3.73 ± 0.05	1.83 ± 0.07
		WN-TDWT	$\mathbf{4.62 \pm 0.12}$	$\mathbf{4.17 \pm 0.04}$	$\mathbf{5.50 \pm 0.05}$	$\mathbf{2.94 \pm 0.06}$
Haar	Type A	TDWT	4.82 ± 0.14	$\mathbf{4.47 \pm 0.15}$	5.50 ± 0.04	3.00 ± 0.06
		WN-TDWT	4.87 ± 0.11	4.44 ± 0.14	5.49 ± 0.07	3.08 ± 0.08
	Type B	TDWT	4.72 ± 0.11	4.38 ± 0.25	5.37 ± 0.13	3.07 ± 0.11
		WN-TDWT	$\mathbf{4.99 \pm 0.06}$	4.46 ± 0.13	$\mathbf{5.59 \pm 0.07}$	$\mathbf{3.17 \pm 0.04}$
	Type C	TDWT	4.82 ± 0.16	4.30 ± 0.14	5.44 ± 0.07	3.05 ± 0.06
		WN-TDWT	4.90 ± 0.08	4.36 ± 0.06	5.47 ± 0.05	3.09 ± 0.07



(a) ランダム初期化での TDWT 層　(b) ランダム初期化での WN-TDWT 層　(c) Haar 初期化での TDWT 層　(d) Haar 初期化での WN-TDWT 層

図 4: ランダム初期化，Haar 初期化でのタイプ A の構造をもつ TDWT 層，WN-TDWT 層の $H_1(z)$（青色）と $G_1(z)$（赤色）の周波数応答．実線と破線はそれぞれ初期化，学習後の周波数応答に対応する．



(a) タイプ C の TDWT 層　(b) タイプ B の WN-TDWT 層

図 5: タイプ C の TDWT 層とタイプ B の WN-TDWT 層の $H_2(z), G_2(z)$ の周波数応答．線種と色は図 4 と同様の意味を表す．

アン SDR に大きく影響しなかったが，平均的にはタイプ B の WN-TDWT モデルが最も性能が高かったため，次節以降はこのモデルを用いて比較を行う．

図 5 に，タイプ C の TDWT 層とタイプ B の WN-TDWT 層に関する周波数応答を示す．TDWT 層はローパス，ハイパス特性が保証されていないため，$H_2(z), G_2(z)$ がそれぞれ $z = -1, 1$ で厳密にゲインが 0 とならなかったものの，0 に近いゲインとなったため WN-TDWT 層と同等の性能が得られたと考えられる．一方，WN-TDWT 層は 3.2 節で導出した通り，$H_2(z), G_2(z)$ がそれぞれローパス，ハイパス特性を保持しており $z = -1, 1$ でゲインが 0 となった．

4.3 従来法との比較

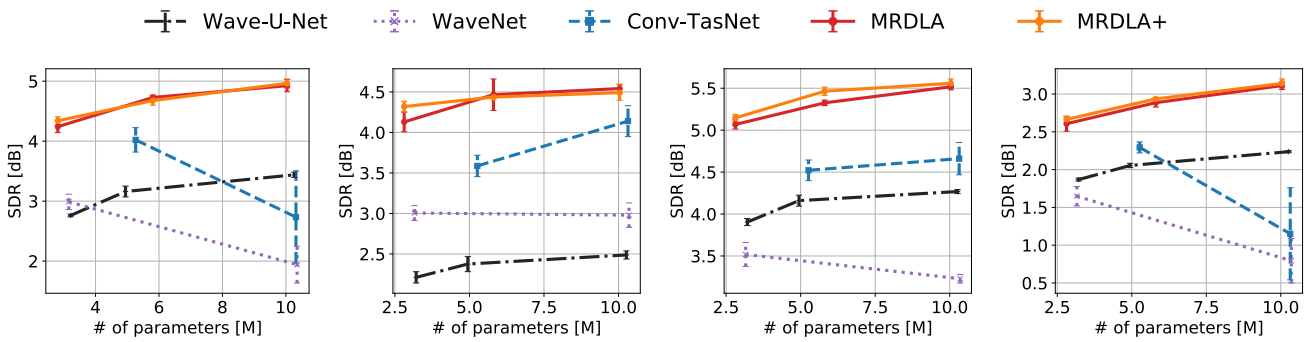
本節では，多重解像度深層分析と従来の時間領域音源分離手法を比較する．従来法として，Wave-U-Net [15] に加え，非因果的な WaveNet ベースの手法（WaveNet）[16]，Conv-TasNet [13] を用いた．全モデルは 4.1 節で述べたデータセット，データ拡張を用いて学習した．比較手法の特徴は以下の通りである．

MRDLA: 予測，更新作用素を学習しない DWT 層を用いた多重解像度深層分析である．DWT 層に用いるウェーブレットにより大きく分離性能が変化しないことを実験で確認しているため [24]，実装が簡単な Haar ウェーブレットを DWT 層のウェーブレットとして用いた．DWT 層は特徴量のチャネルサイズを 2 倍にして出力するが，デシメーション層はチャネルサイズを変化させないため，Wave-U-Net とモデルサイズを厳密に一致させることが難しい．そこで，$C^{(\mathrm{e})} = 6, 12, 18$ として複数の $C^{(\mathrm{e})}$ での性能を比較した．

MRDLA+: 上述の MRDLA の DWT 層を，タイプ B の WN-TDWT 層に置き換えたモデルである．他のハイパーパラメータは，MRDLA と同一とした．

Wave-U-Net: MRDLA の項で述べた通り，公平な比較のため原論文 [15] で用いられていた $C^{(\mathrm{e})} = 24$ だけでなく $C^{(\mathrm{e})} = 8, 12$ での性能も比較した．これらの $C^{(\mathrm{e})}$ を用いた場合，Wave-U-Net のモデルサイズは MRDLA と同程度となること確認した．

WaveNet: 原論文 [16] で提案された WaveNet はモノラ



(a) Vocals. (b) Bass. (c) Drums. (d) Other.

図 6: 様々なモデルサイズでの多重解像度深層分析と従来法の SDR [dB].

ル入力に対して設計されているため，最初の畳み込み層の入力チャネル数，最後の畳み込み層の出力チャネル数を 2 としステレオ入力を扱えるように拡張した．また，オリジナルの WaveNet のモデルサイズは $C^{(e)} = 18$ としたときの MRDLA の半分以下であったため，当該 MRDLA と同程度のモデルサイズとなるように残差ブロックのチャネル数（ [16] では k と表記）を 64 から 164 に増やした．

Conv-TasNet: WaveNet と同様にステレオ入力に対応するため，最初の畳み込み層の入力チャネル数と最後の畳み込み層の出力チャネル数を 2 とした．また，オリジナルのモデルサイズの Conv-TasNet に加え，$C^{(e)} = 18$ の MRDLA と同程度のモデルサイズとなるようにパラメータを増やした Conv-TasNet も用いた．そのために，マスク推定部のボトルネック層のチャネル数（ [13] では B と表記）を 2 倍にした．他のモデルと学習条件を揃えるため，検証ロスが連続して 20 エポック下がるか否かを停止条件としてアーリーストッピングを適用した．Conv-TasNet のロス関数はスケール不変であるため，得られる分離音のスケールは正解の音源信号のスケールと大きく異なる場合がある．この違いがあるものの，SDR はスケール依存であるため評価を行うと SDR が非常に小さい値となってしまった．そこで，このスケールの違いを補正するため，テストデータの各曲に関し，分離音の和と観測信号の平均二乗誤差を最小にするような時不変な音源毎のゲインを求め，分離音をスケーリングした．

他の実験条件は，各手法の原論文に記載されているものを用いた．

図 6 に，全モデルのメディアン SDR の平均を示す．ここで，エラーバーはメディアン SDR のデータ分割に関する標準誤差を表す．WaveNet，Conv-TasNet ではパラメータ数を増やすと平均的に SDR が低下したものの，Wave-U-Net，MRDLA，MRDLA+では増加した．また，従来法の中では，オリジナルのモデルサイズの Conv-TasNet の SDR が最も高かった．全ての楽器において，MRDLA と MRDLA+がより小さいモデルサイズで従来法よりも高い分離性能を達成した．MRDLA に比べ MRDLA+は vocals，bass で同程度，drums，other で高い SDR を示しており，ウェーブレットを学習することにより分離性能が多少向上することを確認した．一方，Wave-U-Net から MRDLA への SDR の上昇量と比べると，MRDLA から MRLDA+への SDR の上昇量は小さかった．この結果は，ウェーブレットを学習するよりもアンチエイリアシングフィルタと完全再構成性を同時に保証することの方が分離性能に関してより重要であることを示している．

MRDLA+と Wave-U-Net，MRDLA+と Conv-TasNet のペアそれぞれに対し，データ分割，楽器毎にトラックワイズ SDR に関する Wilcoxon の符号順位検定を行った．各モデルで平均メディアン SDR が最も大きいものを用いた．いずれのペアに関しても p 値が 1.0×10^{-3} よりも十分小さく，MRDLA+が Wave-U-Net と Conv-TasNet よりも統計的に有意に高い分離性能を持つことを確認した．

4.4 主観評価による比較

最後に，Wave-U-Net，Conv-TasNet，MRDLA+の分離品質に関して 12 人の被験者により主観評価実験を行った．この実験でも，各モデルで平均メディアン SDR が最も大きいものを用いた．実験は 2 回に分割して行われ，Wave-U-Net と MRDLA+，Conv-TasNet と MRLDA+に関してそれぞれプリファレンステストを行った．被験者に提示する音響信号として，MUSDB18 データセットのテストデータ 50 曲から 10 曲をランダムに選び 40 秒から 50 秒の区間を用いた．ここで，各曲で 4 つのデータ分割の 1 つから選択し，対応するデータ分割で学習されたモデルから得られた分離音を用いた．プリファレンステストでも同一のデータ分割，曲，区間を用いた．被験者には，参照音として混合音と各音源信号を提示し，各手法名については知らせずに各曲，各楽器に関し分離音をランダムな順番で提示した．また，対象楽器音の残留音も評価しやすいように，分離音に加え混合音から当該楽器の分離音を除いたマイナスワン音も提示した．これらの音は被験者が自由に何回で

表 2: Wave-U-Net と多重解像度深層分析のプリファレンステストの結果.

Instrument	Preference score [%]		p-value
	Wave-U-Net	MRDLA+	
vocals	29.17	**70.83**	$< 10^{-5}$
bass	6.67	**93.33**	$< 10^{-20}$
drums	19.17	**80.83**	$< 10^{-10}$
other	26.67	**73.33**	$< 10^{-6}$

表 3: Conv-TasNet と多重解像度深層分析のプリファレンステストの結果.

Instrument	Preference score [%]		p-value
	Conv-TasNet	MRDLA+	
vocals	13.33	**86.67**	$< 10^{-15}$
bass	13.33	**86.67**	$< 10^{-15}$
drums	10.83	**89.17**	$< 10^{-17}$
other	22.50	**77.50**	$< 10^{-8}$

も聴取できるようにした．被験者にこれらの音を提示した後，分離音の音質，分離音の歪み，分離音に残留した干渉音の自然性，対象楽器音の残留度合いを踏まえ，総合的により高い分離品質だと思うものを選択するように指示した．

表 2，3 にそれぞれ Wave-U-Net と MRDLA+，Conv-TasNet と MRDLA+のプリファレンステスト結果を示す．全楽器で MRDLA+の方が選択されており，Pearson のカイ二乗検定を用いて計算した p 値はいずれも 1.0×10^{-3} よりも十分小さかった．これらの結果は，聴感上でも MRDLA の分離品質が Conv-TasNet と Wave-U-Net に比べ全楽器で統計的に有意に高いことを示している．実際に分離結果を聴取してみると，Conv-TasNet の分離音には特に bass，drums で高周波のアーティファクトが含まれており，対象楽器音がマイナスワン音に残留している頻度が高かった．Wave-U-Net の分離音はアーティファクトが少ないものの，干渉音が他の手法に比べると非常に残留していた．一方，MRDLA+の分離音は低周波から高周波まで対象楽器音を含んでおり，他の手法の分離音に比べると音が明瞭であった．WaveNet は平均メディアン SDR が最も低かったため主観評価には含めなかったが，分離音を聴取してみると音がぶつ切れになっておりアーティファクトも多く含まれていた．主観評価実験に用いた音の一部は `http://tomohikonakamura.github.io/Tomohiko-Nakamura/demo/MRDLA/index.html` で聴取できる．

5. 結論

本稿では，多重解像度深層分析に関して詳細な客観評価および主観評価を行った．また，TDWT 層がアンチエイリアシングフィルタをもつための予測，更新作用素に対する制約条件を導出し，WN-TDWT 層を提案した．楽音分離実験により，TDWT 層に制約条件を導入しアンチエイリアシングフィルタの存在を保証することで予測，更新作用素の初期値依存性を低減でき，DWT 層を用いる場合に比べ分離性能が向上することを確認した．また，ウェーブレットを学習する拡張に比べ，アンチエイリアシングフィルタと完全再構成性を保証することによる分離性能の向上の方が大きく，周波数応答に比べ DWT 層の構造の方が音源分離により重要であることを示した．さらに，多重解像度深層分析は従来法に比べ小さいモデルサイズでより高い分離性能を達成することを確認し，主観評価実験により多重解像度深層分析の分離品質は従来法に比べ統計的に有意に高いことを示した．

謝辞 本研究は JSPS-CAS 二国間交流事業 JPJSBP120197203，JSPS 科研費科研費 JP19H01116，JP20K19818，カワイサウンド技術・音楽振興財団の助成を受けたものである．

付　　録

A.1 補題 1 の証明

$I = 1$ の場合，式 (6) は以下のように書ける．

$$Q_1(z) = \begin{bmatrix} A(1 - P_1(z)U_1(z)) & AU_1(z) \\ P_1(z)/A & 1/A \end{bmatrix} \tag{A.1}$$

式 (A.1) を式 (10) に代入し整理すると，以下の式が導出できる．

$$H_1^{(\mathrm{even})}(z) = A(1 - P_1(z)U_1(z)) \tag{A.2}$$
$$H_1^{(\mathrm{odd})}(z) = AU_1(z) \tag{A.3}$$
$$G_1^{(\mathrm{even})}(z) = \frac{P_1(z)}{A} \tag{A.4}$$
$$G_1^{(\mathrm{odd})}(z) = \frac{1}{A} \tag{A.5}$$

これらの式を式 (13)，(14) と比較すると，$P_1(1), U_1(1)$ に対する以下の条件式が得られる．

$$P_1(1) = 1, \quad U_1(1) = \frac{1}{2} \tag{A.6}$$

式 (A.6) を $p_{1,t}, u_{1,t}$ に対する条件へと変換することで，式 (15) が得られる．

A.2 補題 2 の証明

$H_I(z), G_I(z)$ は条件 (13)，(14) を満たすとする．$I + 1$ 番目の予測，更新ステップをスケーリングステップの直前に挿入すると，得られるリフティングスキームの特性は $Q_I(z)$ に関する以下の再帰式で表現できる．

$$Q_{I+1}(z) = \begin{bmatrix} A & 0 \\ 0 & 1/A \end{bmatrix} \begin{bmatrix} 1 & U_{I+1}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -P_{I+1}(z) & 1 \end{bmatrix} \times \begin{bmatrix} A & 0 \\ 0 & 1/A \end{bmatrix}^{-1} Q_I(z), \quad \text{(A.7)}$$

$$= \begin{bmatrix} 1 - P_{I+1}(z)U_{I+1}(z) & A^2 U_{I+1}(z) \\ -P_{I+1}(z)/A^2 & 1 \end{bmatrix} Q_I(z) \quad \text{(A.8)}$$

式 (A.7) の最後から 2 つ目の行列はスケーリングステップの逆を表し，$Q_I(z)$ に含まれるスケーリングステップを打ち消すため挿入した．式 (10) で表される $H_I(z), G_I(z)$ と $P_i(z), U_i(z)$ の関係を用いて，式 (A.8) を $H_I^{(\text{even})}(z), H_I^{(\text{odd})}(z), G_I^{(\text{even})}(z), G_I^{(\text{odd})}(z)$ に関する再帰式へと変換すると，

$$H_{I+1}^{(\text{even})}(z) = H_I^{(\text{even})}(z)(1 - P_{I+1}(z)U_{I+1}(z)) + A^2 G_I^{(\text{even})}(z) U_{I+1}(z) \quad \text{(A.9)}$$

$$H_{I+1}^{(\text{odd})}(z) = H_I^{(\text{odd})}(z)(1 - P_{I+1}(z)U_{I+1}(z)) + A^2 G_I^{(\text{odd})}(z) U_{I+1}(z) \quad \text{(A.10)}$$

$$G_{I+1}^{(\text{even})}(z) = G_I^{(\text{even})}(z) - \frac{P_{I+1}(z) H_I^{(\text{even})}(z)}{A^2} \quad \text{(A.11)}$$

$$G_{I+1}^{(\text{odd})}(z) = G_I^{(\text{odd})}(z) - \frac{P_{I+1}(z) H_I^{(\text{odd})}(z)}{A^2} \quad \text{(A.12)}$$

となる．ここで，条件 (13)，(14) から $H_I^{(\text{even})}(1) = H_I^{(\text{odd})}(1), G_I^{(\text{even})}(1) = -G_I^{(\text{odd})}(1)$ が成り立つため，式 (A.10)，(A.12) はそれぞれ以下のように変換できる．

$$H_{I+1}^{(\text{odd})}(1) = H_I^{(\text{even})}(1)(1 - P_{I+1}(1)U_{I+1}(1)) - A^2 G_I^{(\text{even})}(1) U_{I+1}(1) \quad \text{(A.13)}$$

$$G_{I+1}^{(\text{odd})}(1) = -G_I^{(\text{even})}(1) - \frac{P_{I+1}(1) H_I^{(\text{even})}(1)}{A^2} \quad \text{(A.14)}$$

$z = 1$ での式 (A.9)，(A.11) をそれぞれ式 (A.13)，(A.14) と比較すると，$H_{I+1}(z), G_{I+1}(z)$ に関する以下の条件が得られる．

$$P_{I+1}(1) = 0, \quad U_{I+1}(1) = 0 \quad \text{(A.15)}$$

式 (A.15) を $p_{I+1,t}, u_{I+1,t}$ に関する条件へと変換することで，条件 (16) が得られる．

参考文献

[1] Stöter, F., Liutkus, A. and Ito, N.: The 2018 signal separation evaluation campaign, *Proc. International Conference on Latent Variable Analysis and Signal Separation*, pp. 293–305 (2018).

[2] Hershey, J., Chen, Z., Le Roux, J. and Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 31–35 (2016).

[3] Jansson, A., Humphrey, E., Montecchio, N., Bittner, R. M., Kumar, A. and Weyde, T.: Singing Voice Separation with Deep U-Net Convolutional Networks, *Proc. International Society for Music Information Retrieval Conference* (2017).

[4] Takahashi, N. and Mitsufuji, Y.: Multi-scale multi-band densenets for audio source separation, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 21–25 (2017).

[5] Takahashi, N., Goswami, N. and Mitsufuji, Y.: Mmdenselstm: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation, *Proc. International Workshop on Acoustic Signal Enhancement*, pp. 106–110 (2018).

[6] Hennequin, R., Khlif, A., Voituret, F. and Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models, *Journal of Open Source Software*, Vol. 5, No. 50, p. 2154 (2020).

[7] Takahashi, N. and Mitsufuji, Y.: Densely Connected Multidilated Convolutional Networks for Dense Prediction Tasks, *ArXiv*, No. 2011.11844 (2020).

[8] Pandey, A. and Wang, D.: Dense CNN With Self-Attention for Time-Domain Speech Enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 1270–1279 (2021).

[9] Le Roux, J., Ono, N. and Sagayama, S.: Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction, *Proc. Workshop on Statistical and Perceptual Audition*, pp. 23–28 (2008).

[10] Le Roux, J., Wichern, G., Watanabe, S., Sarroff, A. and Hershey, J. R.: Phasebook and Friends: Leveraging Discrete Representations for Source Separation, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, No. 2, pp. 370–382 (2019).

[11] Koizumi, Y., Yatabe, K., Delcroix, M., Masuyama, Y. and Takeuchi, D.: Speech Enhancement Using Self-Adaptation and Multi-Head Self-Attention, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 181–185 (2020).

[12] Venkataramani, S., Casebeer, J. and Smaragdis, P.: End-to-end source separation with adaptive front-ends, *Proc. Asilomar Conference on Signals, Systems, and Computers*, pp. 684–688 (2018).

[13] Luo, Y. and Mesgarani, N.: Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 8, pp. 1256–1266 (2019).

[14] Kavalerov, I., Wisdom, S., Erdogan, H., Patton, B., Wilson, K., Le Roux, J. and Hershey, J.: Universal Sound Separation, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2019).

[15] Stoller, D., Ewert, S. and Dixon, S.: Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation, *Proc. International Society for Music Information Retrieval Conference*, pp. 334–340 (2018).

[16] Lluís, F., Pons, J. and Serra, X.: End-to-End Music Source Separation: Is it Possible in the Waveform Domain?, *Proc. INTERSPEECH*, pp. 4619–4623 (2019).

[17] Samuel, D., Ganeshan, A. and Naradowsky, J.: Meta-Learning Extractors for Music Source Separation, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 816–820 (2020).

[18] Nakamura, T. and Saruwatari, H.: Time-Domain Audio

Source Separation Based on Wave-U-Net Combined with Discrete Wavelet Transform, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 386–390 (2020).

[19] Nakamura, T., Kozuka, S. and Saruwatari, H.: Time-domain Audio Source Separation With Neural Networks Based on Multiresolution Analysis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021). to appear.

[20] Gong, Y. and Poellabauer, C.: Impact of Aliasing on Deep CNN-Based End-to-End Acoustic Models, *Proc. INTERSPEECH*, pp. 2698–2702 (2018).

[21] Zeiler, M. D. and Fergus, R.: Visualizing and Understanding Convolutional Networks, *Proc. European Conference on Computer Vision*, pp. 818–833 (2014).

[22] Zhang, R.: Making Convolutional Networks Shift-Invariant Again, *Proc. International Conference on Machine Learning*, Vol. 97, pp. 7324–7334 (2019).

[23] Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, pp. 674–693 (1989).

[24] Kozuka, S., Nakamura, T. and Saruwatari, H.: Investigation on Wavelet Basis Function of DNN-based Time Domain Audio Source Separation Inspired by Multiresolution Analysis, *Proc. International Congress and Exposition on Noise Control Engineering*, pp. 4013–4022 (2020).

[25] 小塚 詩穂里，中村 友彦，猿渡洋：ニューラルネットワークとウェーブレット基底関数の同時学習に基づく多重解像度深層分析を用いた時間領域音源分離，電子情報通信学会技術研究報告，Vol. 119, No. 439, pp. 279–284 (2020).

[26] Sweldens, W.: The lifting scheme: A custom-design construction of biorthogonal wavelets, *Applied and Computational Harmonic Analysis*, Vol. 3, No. 2, pp. 186–200 (1996).

[27] Daubechies, I. and Sweldens, W.: Factoring wavelet transforms into lifting steps, *Journal of Fourier Analysis and Applications*, Vol. 4, No. 3, pp. 247–269 (1998).

[28] Abbate, A., DeCusatis, C. and Das, P. K.: *Wavelets and Subbands: Fundamentals and Applications*, Birkhäuser, Boston, MA (2002).

[29] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. and Bittner, R.: The MUSDB18 corpus for music separation (2017).

[30] Vincent, E., Gribonval, R. and Fevotte, C.: Performance measurement in blind audio source separation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1462–1469 (2006).