［研究助成（A）］

ウェーブレット解析と深層学習に基づく時間領域音源分離の検討

Time-Domain Audio Source Separation Based on Wavelet Analysis and Deep Learning

2201018

研究代表者　東京大学大学院 情報理工学系研究科　特任助教　中　村　友　彦

［研究の目的］

音は人間や機械にとって主要な外界認識の手がかりの１つである。しかし，我々の身の回りには多様な音が溢れており，多くの場合様々な音事象が混ざったまま観測されてしまう。人間はこのような複雑な音環境の中でもそれぞれの音を聴き分ける能力を備えているため，周囲の状況，事象を把握できる。この能力を計算機で実現する試みが音源分離であり，監視・介護システム，人間の演奏に合わせて自動で伴奏を再生するシステムなど，人間と協働しつつ周囲の音事象の認識を行うシステムの実現には必須である。本研究では，混合音から各音源信号を分離する技術である音源分離に取り組む。

［研究の内容，成果］

1. 背景

学習のための分離対象の音響信号が事前に大量に入手できる場合には，深層ニューラルネットワーク（deep neural network：DNN）を用いた音源分離手法が高い性能を示している［1］。DNN が導入された当初は，それ以前の音源分離手法と同様に時間周波数領域で分離を行う手法が主流であったが，近年は混合音の時間波形を直接処理し時間周波数領域を介さず分離する一気通貫型 DNN が注目を集めている。

音源分離における代表的な一気通貫型 DNN の１つが Wave-U-Net である［2］。この DNN はエンコーダ，デコーダからなる U-Net 構造をもつ。エンコーダは，特徴量を畳み込み層，非線形層，時間方向に間引く層（デシメーション層）に順に適用する処理を繰り返す。デコーダは，エンコーダのデシメーション層に入力された特徴量を参照しながら，特徴量を畳み込み層と非線形層で処理しつつ繰り返しアップサンプリングを行う。

U-Net 構造で用いられる繰り返しダウンサンプリングする構造は，効率的に畳み込み層の受容野を拡大でき広い範囲のデータの関連性を捉えることに寄与するため，深層学習では広く用いられている。しかし，信号処理の観点からこの構造を見直ことで，デシメーション層をダウンサンプリングに用いる際に生じる２つの問題を発見した。そこで，本研究ではこれらの問題を解決するための手法について検討を行った。

2. 動機

2.1　デシメーション層の問題点と信号処理

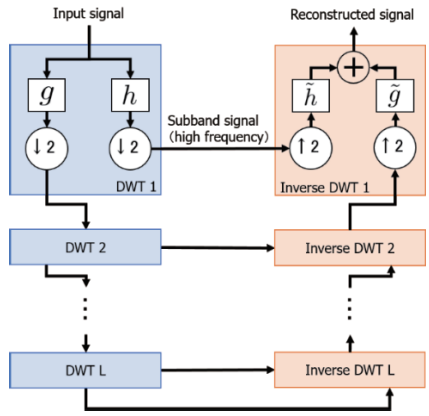
信号処理の立場から見ると，DNN の各層はフィルタに対応し，特徴量はフィルタで処理される信号とみなせる。この解釈に則れば，デシメーション層では，入力特徴量に含まれる高周波成分が低周波成分側に混入する現象（エイリアシング）が起き，それに起因する雑音が重畳

された出力が得られてしまう。特徴量領域でのエイリアシングによる性能低下は音声認識や画像認識で報告されており［3, 4］，音源分離においても性能低下を招く可能性がある。また，デシメーション層は入力特徴量の半分を破棄してしまう。そのため，その部分に含まれていた情報は，より粗い解像度の特徴量を処理する階層には到達しない。また，U-Net 構造では，エンコーダの各階層で間引きされる前の特徴量をデコーダが参照できるようスキップ接続が導入されているものの，それらの特徴量は並進不変な畳み込み層と要素毎の非線形層により処理される。並進不変性により，デシメーション層によってどのインデックスの成分が破棄されたのか否かが区別されないまま処理されてしまう。そのため，デコーダが破棄された成分に含まれる情報を補償できるか否かは学習に強く依存する。

適切に学習ができればこれら 2 つの問題の影響を緩和できる可能性があるが，必ずしもそのような学習を実行できる保証はない。これらの問題はデシメーション層の構造そのものに起因するため，本質的な解決にはダウンサンプリング構造の見直しが必要である。そこで，本研究では，エイリアシングを低減しつつ情報の欠落を防ぐ構造をもつダウンサンプリング層の構築を行った。

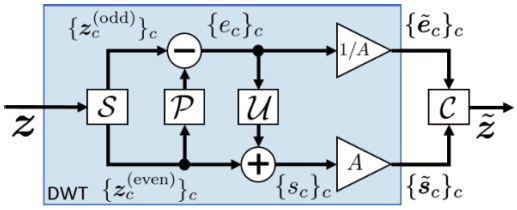
2.2　U-Net 構造とウェーブレット解析

所望のダウンサンプリング層の構築のため，U-Net とウェーブレット解析の分野で提案された多重解像度解析の構造の類似性に着眼した（図 1 参照）。多重解像度解析は，離散ウェーブレット変換（discrete wavelet transform：DWT）を繰り返し用いて信号を複数の時間解像度のサブバンド信号に分解する手法である［5］。DWT は低域，高域通過フィルタからなる 2 チャネルフィルタバンクであり，信号を半分の時間解像度を持つ低周波，高周波成分に分解する。そのため，アンチエイリアシングフィルタを備える。また，サブバンド信号に対して



g, h はそれぞれ低域，高域通過フィルタを表し，$\tilde{g}, \tilde{h}$ は対応する合成フィルタを表す。1）より引用

図 1　多重解像度解析，合成の概要図



1）より引用

図 2　DWT 層

逆 DWT 変換を繰り返し適用することで，原信号を再構成できる，すなわち完全再構成性をもつ。したがって，DWT の前後で信号の情報は保たれる。

3.　提案法

3.1　DWT 層

上述の通り DWT はアンチエイリアシングフィルタ，完全再構成性の両者を備える。そのため，ダウンサンプリング操作として DWT を用いることでデシメーション層に内在する 2 つの問題を同時に解決できる。この考えに基づき，DWT を用いたダウンサンプリング層（DWT 層）を提案した（図 2 参照）。

DWT 層での処理は 2 段階に分けられる。まず，各チャネルの特徴量を信号とみなして DWT を適用し，各チャネルにつき時間解像度が半分の 2 つのサブバンド信号を得る。その後，全チャネルのサブバンド信号をチャネル方向に結合することで，ダウンサンプルされた特徴量

表 1　従来のダウンサンプリング層と DWT 層の比較

	アンチエイリアシングフィルタ	完全再構成性
デシメーション層	無し	無し
平均プーリング層	**有り**	無し
squeezing 操作	無し	**有り**
DWT 層	**有り**	**有り**

を得る。

DWT の実装には，リフティングスキームと呼ばれる技法を用いた。この技法により DWT の全過程がインプレース演算で書けるため，推定時の計算量を低減でき，デシメーション層に比べ大きく計算量が増加することはない。また，リフティングスキームは可逆であるため，DWT 層の逆過程として逆 DWT を用いたアップサンプリング層（逆 DWT 層）も定義できる。

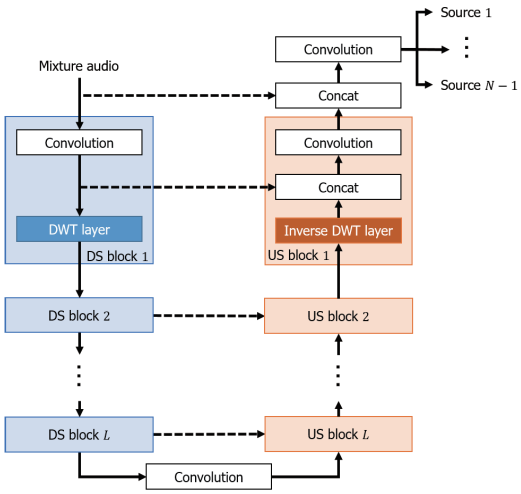
3.2　従来のダウンサンプリング層との関係

DWT 層はアンチエイリアシングフィルタ，完全再構成性どちらも備えている。一方，従来のダウンサンプリング層は，これらの性質のどちらも備えていないか，いずれか一方のみを備えているものだけである（表 1 参照）。例えば，ダウンサンプリング層としてよく用いられる平均プーリング層は，隣接する特徴量の平均を 2 以上のストライドで計算する。そのため，アンチエイリアシングフィルタとして機能する。しかし，一般に平均のみから入力された特徴量を復元することはできないため，完全再構成性はもたない。

また，正規化フローで用いられる squeezing 操作は，特徴量を偶数，奇数インデックスの成分に分解しそれらをチャネル方向に結合した特徴量を出力する。この操作は明らかに完全再構成性を満たすが，間引きいたものをそのまま結合するため，特徴量領域でエイリアシングが起こる。

3.3　多重解像度深層分析

Wave-U-Net のデシメーション層を DWT 層に置き換えた DNN を構築し，それを用いた音源分離手法（多重解像度深層分析）を提案した。図 3 に多重解像度深層分析で用いる DNN の構造を示す。ただし，畳み込み層の後の非線形関数は省略した。



1）より引用

図 3　多重解像度深層分析の DNN

エンコーダは L 個のダウンサンプリングブロックから成り，各ダウンサンプリングブロックで特徴量は畳み込み層，Leaky ReLU，DWT 層により処理される。最上階層のダウンサンプリングブロックから出力された特徴量は，畳み込み層，Leaky ReLU で処理され，デコーダに入力される。デコーダは，L 個のアップサンプリングブロックから成る。各アップサンプリングブロックは，同一階層のダウンサンプリングブロックの入力と上層のアップサンプリングブロックの出力をチャネル方向に結合し，畳み込み層，Leaky ReLU，逆 DWT 層で処理される。最後に，畳み込み層，双曲線正接関数を通して分離音が得られる。また，分離音の和が混合音と一致するように，最後の音源の分離音は，混合音からそれ以外の分離音を除算することで得る。

4.　実験的評価

4.1　実験条件

アンチエイリアシングフィルタと完全再構成性の音源分離性能に対する効果を，楽音分離実験により評価した。学習，評価データとして，

MUSDB18 データセットからそれぞれ 100 曲，50 曲を用いた。ただし，学習データのうち 25 曲をランダムに選択し検証データとして用いた。このデータセットでは，各曲に対し bass, drums, other, vocals の 4 つの楽器毎の録音とそれらの混合音が付与されている。サンプリング周波数は 22.05 kHz とし，[2] と同様ステレオのまま用いた。

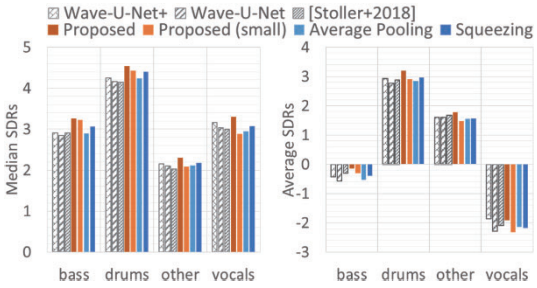
比較手法として Wave-U-Net を用いた。ここで，提案法の DWT 層のウェーブレットして Haar ウェーブレットを用いた。また，アンチエイリアシングフィルタと完全再構成の効果を別々に検証するため，デシメーション層の代わりに表 1 の平均プリーリング，squeezing 操作を用いた Wave-U-Net の変種（それぞれ Average Pooling, Squeezing）とも比較を行った。

デシメーション層は特徴量のチャネル数を変えないものの，DWT 層はチャネル数を 2 倍にするため，Wave-U-Net のデシメーション層を単純に DWT 層に置き換えるだけではパラメータ数が増加してしまう。そこで，公正な比較のため，Wave-U-Net の各畳み込み層のチャネルを倍にした変種（Wave-U-Net+），提案法の各チャネルを半分にした変種（Proposed（small））も用いた。

分離性能の指標として，BSSEval v4 ライブラリで計算される source-to-distortion ratio（SDR）を各曲で計算し，曲に関するそれらの中央値，平均値を求めた。また，初期値の違いによる性能差を低減するため，5 つのランダムシードで得られた結果を平均し，指標を算出した。他の詳しい実験条件に関しては，1）を参照されたい。

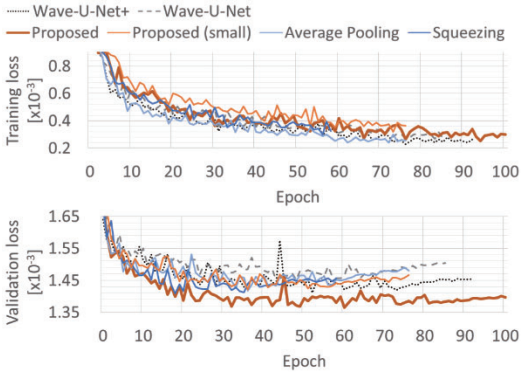
4.2　結果と考察

図 4 に各手法での音源分離性能を示す。ここで，[Stoller+2018] は [2] で報告された SDR を表す。Wave-U-Net+，Wave-U-Net に比べそれぞれ Proposed, Proposed（small）の方が半分程度のパラメータ数であるにも関わらず，



1）より引用

図 4　音源分離性能の比較



1）より引用

図 5　学習，検証ロスの時間発展

提案法の方が SDR の中央値，平均値に関しどちらも高いか同程度であった。この結果は，DWT 層の方がデシメーション層に比べ，音源分離に適していることを示している。

Average Pooling, Squeezing はアンチエイリアシングフィルタか完全再構成性のいずれかを持つダウンサンプリング層を用いているものの，いずれも提案法の性能に達しなかった。この結果は，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考慮することの重要性を示している。

図 5 に，学習，検証ロスの時間発展を示す。提案法では学習ロスは他の手法に比べ大きいものの，検証ロスは小さくなっていることを発見した。この結果は，DWT 層の導入により過学習が抑制できることを示唆している。

本実験ではウェーブレットとして Haar ウェーブレットを用いたが，他の代表的なウェーブレットを用いた場合でも Haar ウェーブレットと同等の性能が得られることを確認し

た。詳細は 2）を参照されたい。多重解像度解析ではウェーブレットの選択により分析性能が大きく変化するものの，DWT 層はウェーブレットの選択に頑健であり，アンチエイリアシングフィルタと完全再構成性を導入することがより重要であることが確認できた。

［今後の研究の方向，課題］

本研究により，ダウンサンプリング層でのアンチエイリアシングフィルタと完全再構成性の重要性が確認できた。しかし，この結果は Wave-U-Net 型の音源分離においてであり，他の DNN 構造における有効性は未調査である。また，ダウンサンプリング層は様々なタスクの DNN で頻繁に現れるため，音源分離以外のタスクにおいてどのような効果があるか検証することは重要である。さらに，提案した音源分離を前処理として用いたときに，後段のアプリケーションでどのような影響があるかも検証を行いたい。

［参 考 文 献］

［1］ F. Stöter, A. Liutkus, and N. Ito : “The 2018 signal separation evaluation campaign,” in Proceedings of International Conference on Latent Variable Analysis and Signal Separation, pp. 293–305, Jul. 2018.

［2］ D. Stoller, S. Ewert, and S. Dixon : “Wave-U-Net : A multi-scale neural network for end-to-end audio source separation,” in Proceedings of International Society for Music Information Retrieval Conference, pp. 334–340, Sep. 2018.

［3］ Y. Gong and C. Poellabauer : “Impact of aliasing on deep CNN-based end-to-end acoustic models,” in Proceedings of INTERSPEECH, pp. 2698–2702, Sep. 2018.

［4］ R. Zhang : “Making convolutional networks shift-invariant again,” in Proceedings of International Conference on Macine. Learning, vol. 97, pp. 7324–7334, Jul. 2019.

［5］ S. Mallat : “A theory for multiresolution signal decomposition : the wavelet representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674–693, Jul. 1989.

［成果の発表，論文等］

1） Tomohiko Nakamura and Hiroshi Saruwatari : “Time-domain Audio Source Separation based on Wave-U-Net Combined with Discrete Wavelet Transform,” in Proceedings of the 45th International Conference on Acoustics, Speech, and Signal Processing, pp. 386–390, May 2020.

2） Shihori Kozuka, Tomohiko Nakamura and Hiroshi Saruwatari : “Investigation on Wavelet Basis Function of DNN-based Time Domain Audio Source Separation Inspired by Multiresolution Analysis,” in Proceedings of the 49th International Congress and Exposition on Noise Control Engineering, pp. 4013–4022, Aug. 2020.