令和2年度研究助成【サウンド技術振興部門】より

多重解像度深層分析に基づくEnd-to-End音源分離のためのウェーブレット基底関数の自動設計

東京大学大学院情報理工学系研究科
特任助教　中村　友彦

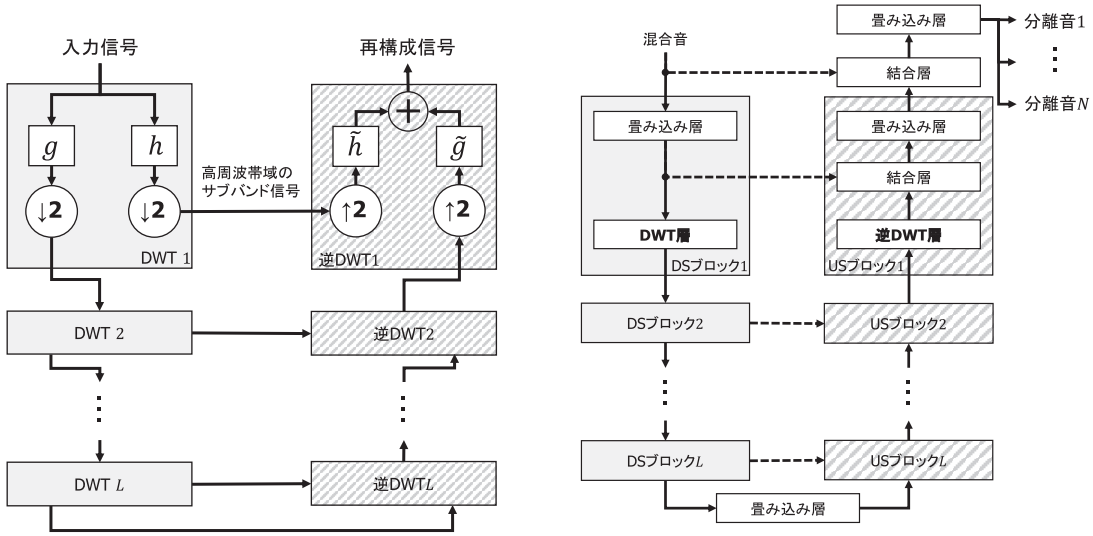
1．はじめに

人間は、複雑な音環境において選択的に音を聴き分けることができる。この能力を計算機で実現する試みが、混合音を各音源信号に分離する音源分離である。この技術は、音声認識や音楽音響信号から自動で譜面を書き起こす自動採譜、音響イベントを検出する音響イベント検知など音メディアに関する様々なシステムに利用できる。例えば､自動採譜においては､いつ、どの楽器の、どの音高が鳴っているかを音楽音響信号から認識する必要があるが、楽器や音高毎に分離することでこれらの認識の一助となる。また、楽器や音高ごとに分離した音をユーザが自由にリミックスするなど音を加工するシステムにも利用できる[1),2)]。

近年、音源分離では深層ニューラルネットワーク（deep neural network; DNN）を用いた手法が有望な結果を示している[3)]。DNNを用いた手法の中でも、観測音響信号をスペクトログラムへと変換することなく直接時間領域で処理するend-to-endアプローチが盛んに研究されている[4)-10)]。本稿では、筆者が提案したend-to-end音源分離手法である多重解像度深層分析（multiresolution deep layered analysis; MRDLA）を紹介する[10)]。その拡張について考察し課題を明らかにする。

2．多重解像度深層分析

MRDLAは、end-to-end音源分離手法の１つであるWave-U-Net[4)]をベースとする手法である。Wave-U-Netは、繰り返しダウンサンプリングを行うエンコーダと繰り返しアップサンプリングを行うデコーダからなるU-Net構造を持つ。ダウンサンプリングを繰り返すことによって、時間領域での信号の長期の依存性を捉えることができる。これに対し、筆者は信号処理の観点からこの構造を見直すことで、Wave-U-Netのダウンサンプリングには２つの問題点があることを発見した。（ⅰ）Wave-U-Netではダウンサンプリングがローパスフィルタを伴わない間引きによって実装されており、特徴量領域でエイリアシングが生じうる。（ⅱ）また、アンチエイリアシングフィルタを導入したとしても、間引きによって一部の特徴量が欠落する。そのため、欠落部分に含まれうる分離に必要な情報はそのままでは次の層に伝播しない。U-Net構造は、エンコーダの間引き前の特徴量をデコーダに渡すスキップコネクションをもち、デコーダ側で間引きによる欠落部分も含ん



(a) 多重解像度解析、合成　　(b) MRDLAのDNN

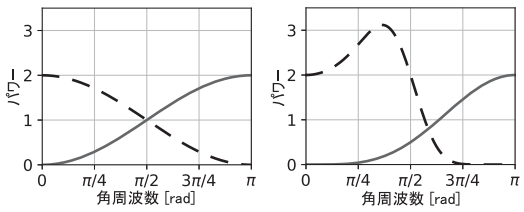
図1　L階層の多重解像度解析とMRDLAのDNNネットワーク構造の概要図。$\tilde{g}$、$\tilde{h}$ は、それぞれ分析ローパス、ハイパスフィルタ g、h に対応する合成ローパス、ハイパスフィルタを表す。DS、USはそれぞれダウンサンプリング、アップサンプリングを表す。

だ特徴量が得られる。しかし、デコーダではスキップコネクションから得られた特徴量を畳み込み層と要素毎の非線形関数によって処理するため、間引きによる欠落部分と非欠落部分を区別せずに処理してしまう。その結果、欠落部分の情報を補償できるか否かは学習に大いに依存する。

これらの問題を根本的に解決するため、ダウンサンプリングに離散ウェーブレット変換（discrete wavelet transform: DWT）を用いた新たなダウンサンプリング層が提案された[10]。当該層は、Wave-U-Net のダウンサンプリングとアップサンプリングを繰り返す構造が多重解像度解析の構造と類似していることから着想を得たものである。多重解像度解析は、DWTを用いて、入力信号を半分の時間解像度をもつサブバンド信号に繰り返しダウンサンプリングする信号解析手法である（図1(a)参照）。DWTは、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とみなせ、アンチエイリアシングフィルタを自然に含む。また、ローパスフィルタとハイパスフィルタを適切に選ぶとDWTは可逆な変換となり、DWTの逆変換（逆DWT）を用いて各レベルでのサブバンド信号から入力信号を再構成できる。すなわち、DWTは完全再構成性をもつ。そのため、DWTを用いることで上述の2つの問題を同時に解決することができる。図1(b)に示すように、MRDLAのDNNは、Wave-U-Net にDWT層と、逆DWTをアップサンプリングに用いた層（逆DWT層）を導入した構造をもつ。当該図では省略したが、各畳み込み層の直後に非線形関数が挿入されている。

3．ウェーブレット基底関数とDNNの同時学習

DWT層では、ローパス、ハイパスフィルタの周波数応答を決定するウェーブレット基底関数を自由に選択することができる。例えば、Haar基底、Deslauriers-Deubuc（DD）基底の周波数応答は図２に示す通りである。ウェーブレット基底関数は多重解像度解析の分析性能を左右するものであり、筆者はHaarウェーブレットをはじめとする既存の基底を用いたMRDLAの音源分離性能を報告してきた[10),11)]。しかし、これらは必ずしも音源分離のために設計されたものではないため、適切に基底を設計できればさらに性能が向上する可能性がある。そのような基底を人手で設計することも考えられるが、DWT層は非線形なDNNの中に組み込まれているため、基底の変更が音源分離性能にどのような影響を及ぼすかを把握するのは難しい。また、DNNのネットワーク構造や音源によっても最適な基底が変わる可能性がある。そのため、音源分離のための基底を自動で設計できることが望ましい。



（a）Haar 基底　　（b）４次のDD基底

図２　DWTのローパス（破線）、ハイパスフィルタ（実線）の周波数応答。

そこで、end-to-endアプローチの思想に則り、ウェーブレット基底関数をDNNと同時に学習することを考える。この方針においての課題は、いかに２節で提起した問題（ⅰ）、（ⅱ）を解決しつつウェーブレット基底関数とDNNを同時に学習するかである。既存の基底はアンチエイリアシングフィルタを含み、完全再構成性も満たすように設計されているため、基底を学習しない場合は問題とならなかった。一方、例えば乱数で基底を決めた場合は、DWT層の２つのフィルタがローパス、ハイパスフィルタになるとは限らず、完全再構成性も必ずしも満たさない。学習中の基底の値は様々な値を取りうるため、アンチエイリアシングフィルタを含み、完全再構成性をもつ条件を満たすような制約を設けつつ学習する必要がある。これらの制約はウェーブレット設計と密接に関係しているため、柔軟なウェーブレット設計技法（例えば、リフティングスキーム[12)]）が制約の導出に有用であろうと考え現在研究を進めている。

4．まとめ

本稿では、多重解像度解析に着想を得た

end-to-end DNN音源分離手法であるMRDLAを紹介した。また、DWT層のウェーブレット基底関数をDNNと同時に学習する拡張方法について議論し、その課題を明らかにした。本稿では1次元の信号に対するウェーブレットを扱ったが、ウェーブレット解析は球面や多様体、グラフに対しても拡張されている[13),14)]。一方、ソーシャルネットワークサービスなどでの交友関係や化合物など、グラフで表現されたデータに対するDNNも活発に研究されている[15)]。これらを鑑みると、ウェーブレットとDNNを融合する方針は音響信号処理以外の他の分野においても一考する価値があるかもしれない。

参考文献

1）T. Nakamura, H. Kameoka, K. Yoshii, and M. Goto, “Timbre replacement of harmonic and drum components for music audio signals,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7470 − 7474, May 2014.

2）T. Nakamura and H. Kameoka, “Harmonic-temporal factor decomposition for unsupervised monaural separation of harmonic sounds,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020.（in press）

3）F. Stöter, A. Liutkus, and N. Ito, “The 2018 signal separation evaluation campaign,” Proceedings of International Conference on Latent Variable Analysis and Signal Separation, pp. 293 − 305, July 2018.

4）D. Stoller, S. Ewert, and S. Dixon, “Wave-U-Net: A multi-scale neural network for end-to-end audio source separation,” Proceedings of International Society for Music Information Retrieval Conference, pp. 334 − 340, Sept. 2018.

5）S. Venkataramani, J. Casebeer, and P. Smaragdis, “End-to-end source separation with adaptive frontends,” Proceedings of Asilomar Conference on Signals, Systems, and Computers, pp. 684 − 688, Oct. 2018.

6）O. Slizovskaia, L. Kim, G. Haro, and E. Gomez, “End-to-end sound source separation conditioned on instrument labels,” Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 306 − 310, May 2019.

7）Y. Luo and N. Mesgarani, “Conv-Tas Net: 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, May 2019.

8) F. Lluís, J. Pons, and X. Serra, "End-to-end music source separation: Is it possible in the waveform domain?," Proceedings of INTERSPEECH, pp. 4619 – 4623, Sept. 2019.

9) I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. Hershey, "Universal sound separation," Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 2019.

10) T. Nakamura and H. Saruwatari, "Time-domain audio source separation based on wave-u-net combined with discrete wavelet transform," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 386 – 390, May 2020.

11) S. Kozuka, T. Nakamura, and H. Saruwatari, "Investigation on wavelet basis function of DNN-based time domain audio source separation inspired by multiresolution analysis," Proceedings of International Congress and Exposition on Noise Control Engineering, Aug. 2020.

12) W. Sweldens, "The lifting scheme: A custom-design construction of biorthogonal wavelets," Applied and Computational Harmonic Analysis, vol. 3, no. 2, pp. 186 – 200, April 1996.

13) S.T. Ali, J.-P. Antoine, and J.-P. Gazeau, "Wavelets on manifolds," Coherent States, Wavelets, and Their Generalizations, pp. 457 – 493, Springer, 2014.

14) M. Mehra, "Wavelets on arbitrary manifolds," Wavelets Theory and Its Applications, pp. 77 – 93, Springer, 2018.

15) X. Dong, D. Thanou, L. Toni, M. Bronstein, and P. Frossard, "Graph signal processing for machine learning: A review and new perspectives," IEEE Signal Processing Magazine, vol. 37, no. 6, pp. 117 – 127, 2020.