研究概要報告書 【サウンド技術振興部門】

研究題目	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	報告書作成者	中村 友彦
研究従事者	中村 友彦		
研究目的	人間は，様々な音が混在する環境下でも周囲の状況や起きている事象を把握できる．これは，混合音中の特定の音を選択的に聴き分ける能力を備えているからである．この選択的な聴取能力を計算機によって実現する試みは音源分離と呼ばれ，様々な音アプリケーションの基盤技術である．高性能な音源分離が実現できれば，介護や監視システムで必要となる音響イベント検知技術（いつどんな音イベントが鳴っているかを検知）の性能向上に寄与するだけでなく，ユーザが自由に音楽をリミキシングできるシステムの実現などエンターテインメントにも寄与しうる．音源分離は長年研究されているものの，様々な環境で高性能に動作させることは未だチャレンジングな問題である． 近年，深層学習を用いた音源分離手法が高い分離性能を示しており，特に End-to-End アプローチに基づく深層学習を用いた音源分離（end-to-end 音源分離）が注目を集めている．従来の音源分離手法は短時間 Fourier 変換などを用いて得られた振幅やパワースペクトログラムに対して処理を行っていたが，End-to-end 音源分離は入力信号を直接処理し時間波形領域で直接分離音を出力する．そのため，従来の音源分離手法では別々に扱っていたスペクトログラムの位相を分離に直接活用できる． 我々は，そのような end-to-end 音源分離手法の 1 つとして以前多重解像度深層分析と名付けた音源分離手法を提案した．当該手法は，Wave-U-Net と呼ばれる深層ニューラルネットワーク（DNN）の構造が信号処理で提案された多重解像度解析に類似していることにヒントを得て構築されたものである．このアナロジーから，Wave-U-Net 内で行われるダウンサンプリングの際に，特徴量領域でエイリアシングや情報の一部欠落が起こりうることを発見した．また，それらの問題を解決するため，アンチエイリアシングフィルタと完全再構成を備えた離散ウェーブレット変換（DWT）を用いたダウンサンプリング層（DWT 層）を開発した（説明書図 1 参照）． DWT においてどのウェーブレット基底関数を選ぶかは重要である．例えば，多重解像度解析においてはウェーブレットの選択により DWT の周波数特性が変わるため，分析性能がウェーブレットに左右される．これに対し，DWT 層では Haar ウェーブレットなど様々な既存のウェーブレットを使用できる．しかし，それらは必ずしも音源分離用に設計されておらず，多重解像度深層分析の性能を制限している可能性がある．また，実験的に人手でウェーブレット基底関数を設計することもできるが，音源やネットワーク構造により適切なウェーブレットが変わる可能性が高く，学習コストの高い DNN においては現実的でない． そこで本研究では，ウェーブレット基底関数を DNN と同時に学習できるように拡張した学習可能 DWT 層を提案した．具体的には，ウェーブレット基底関数と DNN の同時学習法を提案し，それを用いてどのようなウェーブレット基底関数が得られるか，分離性能にどう影響するかを調査した．		

研究概要報告書【サウンド技術振興部門】

研究内容	本研究では，学習可能 DWT 層を提案し，学習されたウェーブレット基底関数や分離性能への影響を実験により調査する． **1. ウェーブレット基底関数を DNN と同時に学習できる DWT 層** DWT 層は，リフティングスキームと呼ばれる DWT の効率的な計算技法を用いて実装される（説明書図 1 参照）．リフティングスキームは，時間分割・予測・更新・スケーリングステップからなる．時間分割ステップでは，入力信号を偶数・奇数インデックス成分に分割する．予測ステップでは，予測作用素を用いて偶数インデックス成分から奇数インデックスを予測しその残差を出力する．更新ステップでは，更新作用素を用いて予測残差を用いて偶数インデックス成分を平滑化する．最後に，スケーリングステップでは，予測残差と平滑化された成分を定数によりスケーリングすることで，入力信号の高周波・低周波成分に対応するサブバンド信号を得る． リフティングスキームでは作用素と更新作用素を定めると，対応するウェーブレット基底関数を用いた DWT が実現できる．そのため，予測作用素と更新作用素を DNN の標準的な学習手法である誤差逆伝播法の枠組みで学習できればよい．ここで，これらの作用素として有限のインパルス応答長のフィルタ（FIR フィルタ）を用いれば，DNN の標準的な構成要素である畳み込み層を用いて実装でき，DNN と同時に学習もできる．また，任意の有限長のウェーブレット基底関数をもつ DWT はリフティングスキームの形で表現できることが証明されており，様々なウェーブレットを表現できる．そのため，これらの作用素による演算を畳み込み層に置き換えた DWT 層として，学習可能 DWT 層を検討する． **2. 学習可能 DWT 層のウェーブレット基底関数と分離性能に関する実験的解析** 実験により，ウェーブレット基底関数を DNN と同時に学習する効果に関し調査を行った．まず，予測・更新作用素に対応する畳み込み層の初期値依存性について調査した．これは，リフティングスキームでは完全再構成性が各ステップの可逆性により保証されるものの，アンチエイリアシングフィルタ，すなわちローパスフィルタを持つか否かは予測・更新作用素に依存するためである．ランダムに初期化する場合と，Haar ウェーブレットの予測・更新作用素と一致するよう初期化する場合（Haar 初期化）を比較した． 次に，構造による分離性能の差異を調査した．比較には 3 種類の構造の学習可能 DWT 層を用いた（説明書図 2 参照）．タイプ A は 1 つの予測・更新作用素を用いる場合，タイプ B は Haar ウェーブレットを修正する形で予測・更新作用素を学習する場合，タイプ C は 2 つの予測・更新作用素を同時に学習する場合に対応する．各作用素の初期化手法として，上述の実験で分離性能が高かったものを採用した．また，学習後の学習可能 DWT 層のローパス・ハイパスフィルタの周波数特性を，学習前後や他のウェーブレットでの周波数特性と比較した．

研究概要報告書 【サウンド技術振興部門】

研究のポイント	● 我々は end-to-end 音源分離手法の 1 つとして多重解像度深層分析を提案してきた. 本研究では, 多重解像度深層分析で用いられる DWT 層 のウェーブレット基底関数を DNN と同時に学習する方法を検討した. ● DWT においてウェーブレットは周波数特性を定める重要な役割をもつ. DWT 層では Haar ウェーブレットなど様々な既存のウェーブレットを利用できるものの, それらは音源分離用に設計されておらず, 多重解像度深層分析の性能を制限している可能性があった. ● 本研究では, リフティングスキームの予測・更新作用素を畳み込み層により表現することで, ウェーブレット基底関数を DNN と同時に誤差逆伝播法を用いて学習できるように DWT 層を拡張した. ● 実験により, 適切に初期値を設定することでローパス・ハイパスフィルタに近い周波数特性となるウェーブレット基底関数が得られることを確認した. また, 学習可能 DWT 層を用いることで, わずかに分離性能が向上することを確認した.
研究結果	● 学習可能 DWT 層を用いた多重解像度深層分析を用いた楽音分離実験により, 予測・更新作用素に対応する畳み込み層のパラメータ初期化方法が数値安定性や分離性能に影響することを発見した(説明書表 1 参照). ランダム初期化の場合, 急激に学習・検証ロスの値が上昇することがあった. 一方, Haar 初期化では数値的に安定して学習でき, 分離性能もランダム初期化に比べ向上した. ● 学習可能 DWT 層の周波数特性を確認したところ, ランダム初期化では学習前後で周波数特性が変化しなかっただけでなく, ローパス・ハイパスフィルタに類似した周波数特性は得られなかった. また, 学習前でもゲインの高い周波数帯域があり, 数値的不安定性の一因であることが示唆された. Haar 初期化では学習後にもローパス・ハイパスフィルタに近い周波数特性が得られ, 学習可能 DWT においては初期化が重要であることを確認した. ● 一方, 学習可能 DWT 層の構造は分離性能に大きく影響しなかったが, Haar ウェーブレットによる DWT 層の代わりに学習可能 DWT 層を用いることで, わずかに分離性能が向上した(説明書図 4 参照). 特にパラメータ数が小さい場合, タイプ A や B の構造がベースやドラムの分離に有効であることを確認した. ● 本研究結果を含む論文が, IEEE/ACM Transactions on Audio, Speech, and Signal Processing に採録された [1]. [1] Tomohiko Nakamura, Shihori Kozuka, and Hiroshi Saruwatari, “Time-Domain Audio Source Separation with Neural Networks Based on Multiresolution Analysis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1687–1701, Apr. 2021.
今後の課題	● 本研究期間では分離対象として楽音を用いたが, 提案した枠組みは話声や環境音の分離にも適用できる. これらの分離対象においても学習可能 DWT 層がどの程度有効であるかを検証することも重要である. ● ランダム初期化を用いた実験により, 周波数特性と DNN の学習の数値的不安定性の関連性が示唆された. これを突き詰めていけば, 信号処理の古典的概念と深層学習の数値的安定性を関連付けて議論できる可能性がある.

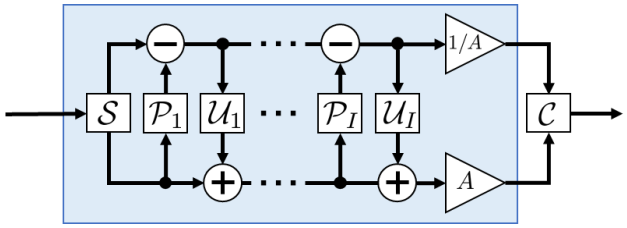
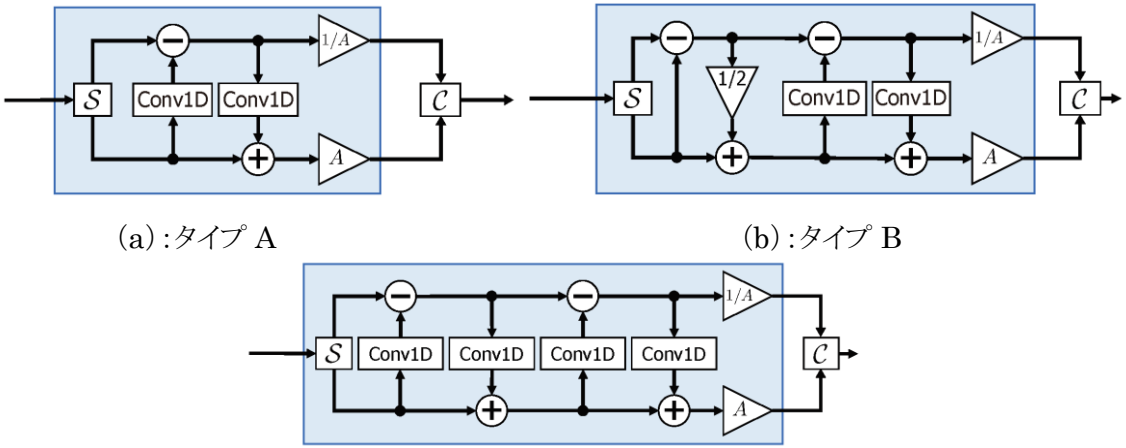


図 1: I個の予測・更新作用素を持つ DWT 層. ここで, $\mathcal{S}$は入力信号の偶数・奇数インデックス成分への分割操作,$\mathcal{P}_i, \mathcal{U}_i$ $(i = 1, \ldots, I)$はそれぞれ予測・更新作用素, $\mathcal{C}$はチャネル方向への結合操作を表す.



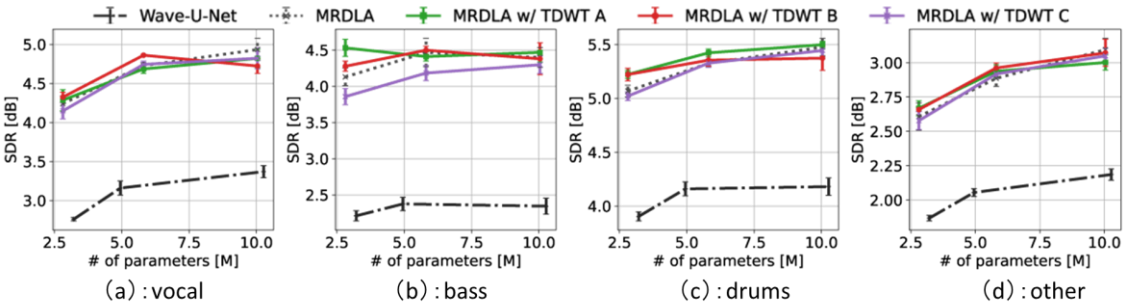
(a) : タイプ A　　(b) : タイプ B

(c) : タイプ C

図 2: 実験に用いた学習可能 DWT 層の構造. Conv1D は1次元の畳み込み層を表す.

表 1: 様々なパラメータ初期化と構造による signal-to-distortion ratio (SDR) [dB]

Initialization	Architecture	Instrument			
		vocals	bass	drums	other
Random	Type A	2.68 ± 0.03	3.07 ± 0.13	3.73 ± 0.05	1.83 ± 0.07
Haar	Type A	$\mathbf{4.82 \pm 0.14}$	$\mathbf{4.47 \pm 0.15}$	$\mathbf{5.50 \pm 0.04}$	$\mathbf{3.00 \pm 0.06}$
Haar	Type B	4.72 ± 0.11	4.38 ± 0.25	5.37 ± 0.13	3.07 ± 0.11
	Type C	4.82 ± 0.16	4.30 ± 0.14	5.44 ± 0.07	3.05 ± 0.06



(a) : vocal　(b) : bass　(c) : drums　(d) : other

図 4: Wave-U-Net と, DWT 層または学習可能 DWT 層を用いた多重解像度深層分析の SDR. MRDLA では Haar ウェーブレットによる DWT 層を用い, MRDLA w/ TDWT A, B, C ではそれぞれタイプ A, B, C の学習可能 DWT 層を用いた.

(注:フローチャート図, ブロック図, 構成図, 写真, データ表, グラフ等 研究内容の補足説明にご使用下さい。)