7L-06

LSTMによる音楽音響信号の修復法の提案 – 周波数フィルタ導入による学習データ量削減の検討 –

谷口亮輔[1], 小島諒介[1], 干場功太郎[1], 中臺一博[1,2],
1 東京工業大学　2 （株）ホンダ・リサーチ・インスティチュート・ジャパン

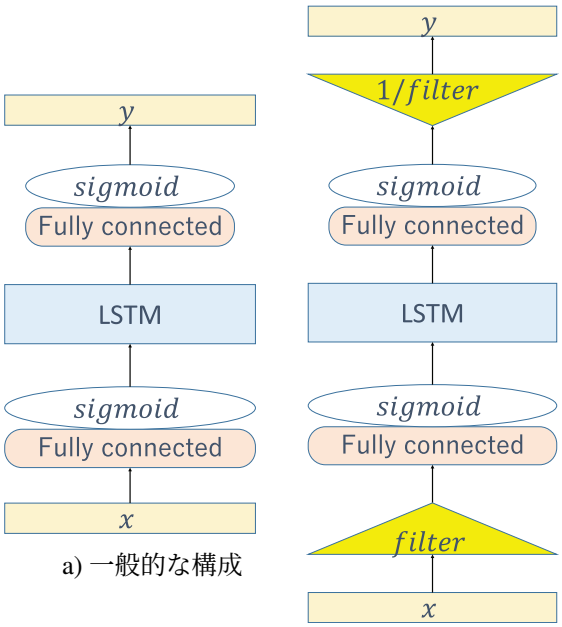
1 はじめに

本稿では，深層学習の一手法であるLSTM (Long Short-Term Memory) [1] を用いた音楽音響信号修復について報告する．一般に，深層学習では性能の高いモデルを学習するために大量のデータが必要である．実際に音楽音響信号修復に深層学習を用いると，学習データが少ない場合，情報が比較的スパースである高域の修復性能が劣化するという問題が発生する．この問題を解決するため，入力信号に対して，周波数フィルタを用いることにより，周波数方向に重みをかける手法を提案する．提案手法をフィルタを用いない一般的なLSTMを用いた方法と比較し，その有効性を示す．

2 提案法

LSTMは，再帰構造を持ち，時間的な情報を考慮に入れることが可能なRNN (Recurrent Neural Network) の改良型である．内部セルや各種ゲートの構造によってRNNでは難しい長期依存情報に対応できる．

Fig. 1 (a) は，連続する過去数フレーム分の情報から，次のフレームを予測する回帰問題としてLSTMを用いた一般的な音楽音響信号修復ニューラルネットワークである．LSTM層に加えて，入力層，全結合層，全結合層，出力層からなっている．一方，Fig. 1 (b) は，提案する音楽音響信号修復ニューラルネットワークの構成であり，Fig. 1 (a) の入出力にそれぞれフィルタ層，逆フィルタ層を追加している． 提案するニューラルネットワークでは，フィルタ層をニューラルネットワークの内部に入れることで，モデル学習時にフィルタ係数の学習が可能でなる．ただし，学習を行うのは，入力層側のフィルタのみであり，出力層側の逆フィルタの各要素は，入力フィルタ係数の単純な逆数をとるものとした． なお，いずれのニューラルネットワークも，振幅スペクトルを入出力として用いる．



a) 一般的な構成　　b) フィルタ内包型構成 (提案法)

Fig. 1　音楽音響信号修復用のニューラルネットワーク

3 評価実験

音楽データ (サンプリングレート16 kHz) に対し，フレーム長512サンプル，シフト長128サンプル，ハミング窓を窓関数として用いたSTFT (Short-Time Fourier transform) を行うことで振幅スペクトルを得る．各楽曲での最大振幅値を用いて正規化を行う．

ニューラルネットワーク学習の際には，連続する3フレーム分 ($t \sim t+2$ フレーム) の振幅スペクトルを入力し，次のフレーム ($t+3$ フレーム) の振幅スペクトルを予測するよう学習を行う．

評価の際には，欠損のない振幅スペクトル列を学習したニューラルネットワークに入力し，出力を時間方向に並べた振幅スペクトル列と，入力に用いた元の振幅スペクトル列との比較を行う．つまり，欠損のないデータが3フレーム連続して続き，その後の4フレーム目が欠損している信号に対する修復タスクを評価していることに相当する．

学習には，ジャズ楽曲25曲を用い，このうち (i) 6曲を用いた場合と (ii) 全25曲を用いた場合の2種類につ

Restoration of musical audio signal using LSTM
– Reduction of the amount of training data with frequency filtering –
Ryosuke Taniguchi[1], Ryosuke Kojima[1], Kotaro Hoshiba[1], Kazuhiro Nakadai[1,2]
1 Tokyo Institute of Technology
2 Honda Research Institute Japan Co., Ltd.

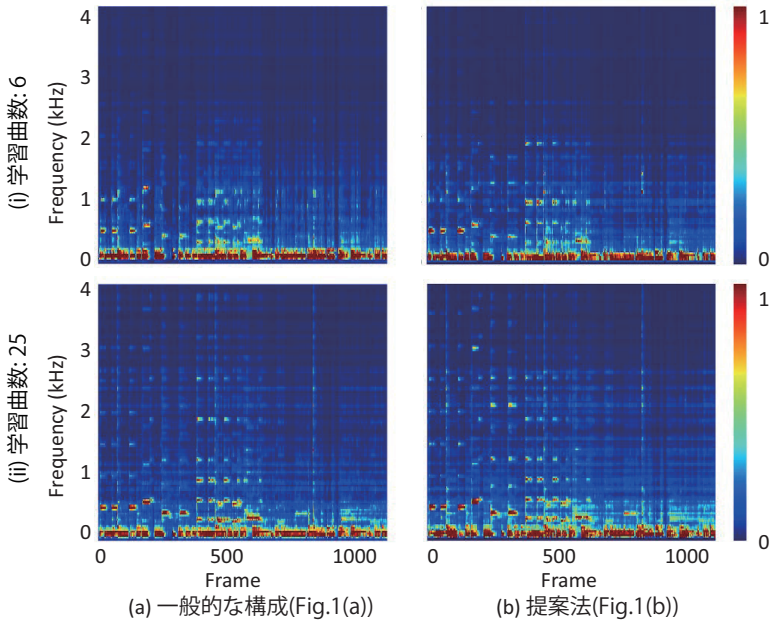


Fig. 2　出力結果

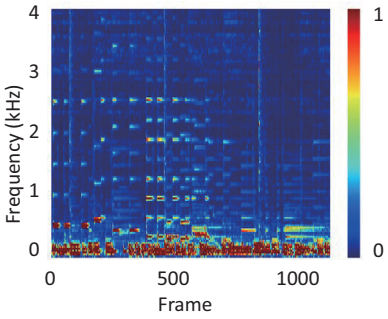


Fig. 3　正解 (入力信号)

Table 1　出力結果と正解との誤差

学習曲数	周波数フィルタ		削減率
	なし	あり	(%)
6 曲	584	549	6.0
25 曲	472	444	5.9

(誤差の小数点以下は切捨て)

いて Fig. 1 (a)，(b) それぞれについて学習し，計 4 種類のニューラルネットワークを構築した．LSTM の各ノードの初期値は，−0.2 ~ 0.2 でランダムに与えた．また，Fig. 1 (b) におけるフィルタ層の初期値には，人間の聴覚を元にした周波数重み付けである A 特性 [2] を用いた．学習率は Adam (Adaptive Moment Estimation) [3] を用いて最適化を行った．

また，評価では学習とは別のジャズ楽曲 1 曲の一部（1003 フレーム，約 8 秒分）に対して予測を行った．

4　比較結果

Fig. 2 に結果として得られた振幅スペクトログラム (周波数 0 kHz ~ 4 kHz) を示す．また，Fig. 3 に入力した振幅スペクトログラム（正解データ）を示す．

学習数が 6 曲と少ない場合には，2 kHz 周辺において，提案法 (Fig. 2 (b) (i)) が，比較手法 (Fig. 2 (a) (i)) と比べて，より精度よく信号修復ができていることがわかる．しかし，Fig. 3 と比較すれば，より高域の周波数帯域では，いずれの場合も信号修復性能が劣化していることがわかる．

一方，25 曲すべてを学習に用いた場合には，いずれの場合も高域の信号修復の性能が向上していることがわかる．特に，提案法 (Fig. 2 (b) (ii)) は，2.5 kHz~3 kHz でも修復が可能となっていることがわかる．

次に，予測結果を各フレームにおいて正解と周波数方向に平均二乗誤差をとり，時間方向に加算した結果を Table 1 に示す．学習で用いた曲数によらず，提案法は，フィルタを用いることにより 6.0 % 程度誤差が改善していることがわかる．

以上より，数曲から数十曲という少ないデータ量で学習を行った場合に，提案手法は，低域での修復性能を維持したまま高域でも精度良い修復が可能であることを示すことができた．

5　まとめ

本研究では，深層学習を用いた音楽音響信号修復を行う場合に，情報がスパースである高域の修復性能を向上させるため，周波数フィルタを用いる方法を提案した．フィルタを用いない通常の LSTM を用いたモデルと修復性能の比較実験を行った．学習で用いるデータ量が少ない場合でも周波数フィルタを用いることで，低域での修復性能を維持したまま高域まで修復が可能であることが示された．深層学習で用いる学習データ削減が期待できる．

謝辞　本研究は, JSPS 科研費 24220006, 16H02884, 16K00294 および，JST ImPACT タフロボティクスチャレンジの助成をうけた．

参考文献

[1] F. A. Gers *et al.*, “Learning to Forget: Continual Prediction with LSTM”, NEURAL COMPUTATION, Vol.12, No.10, pp.2451-2471, 2000.

[2] 西山他, “音響振動工学”, コロナ社, 1979.

[3] Kingma, D. P. *et al.*, “ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION”, International Conference on Learning Representations, pp.1-13, 2015.