一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-9 (11/25)

Quad-directional LSTMを用いた音楽音響信号修復法の提案
Musical Audio Signal Restoration using Quad-directional LSTM

谷口亮輔 *1, 干場功太郎 *1, 中臺一博 *1,2
Ryosuke TANIGUCHI*1, Kotaro HOSHIBA*1, Kazuhiro NAKADAI*1,2
東京工業大学 工学院 システム制御系 *1
（株）ホンダ・リサーチ・インスティチュート・ジャパン *2
Tokyo Institute of Technology*1, Honda Reserch Institute Japan*2
{taniguchi, hoshiba, nakadai}@ra.sc.e.titech.ac.jp

Abstract

本稿ではLSTM（Long Short-Term Memory）を用いた音楽音響信号の修復法を提案する．実際にLSTMを適用した場合，情報が比較的スパースである高域の学習が十分でなくなり，修復性能が劣化してしまう．この問題に対し，我々は，入力信号に対して高域を強調するような周波数フィルタを用いて，その解決を試みた．また，この手法の拡張として，順方向の時系列情報だけではなく，逆方向の時系列情報も考慮したBLSTM（Bi-directional LSTM）を用いる方法を提案した．今回，そのさらなる拡張として，時間方向のみではなく，周波数方向の系列情報も考慮することが可能なQDLSTM（Quad-directional LSTM）を用いることを提案し，評価を行う．その結果，時間方向BLSTMのみと比較してより高音域での修復性能が向上することを確認した．

1 はじめに

近年，深層学習が多くの分野に活用され，その有用性が示されている．本稿では，その一手法であるLSTM（Long Short-term memory）を音楽音響信号修復に適用することを検討する．一般的に，深層学習を用いて性能の高いモデルを学習するためには，大量のデータが必要である．実際の音楽音響信号修復にLSTMを適用した場合，学習データ量が少ないと情報が比較的スパースである高域の学習が十分でなくなり，修復性能が劣化してしまう．この問題に対し，これまで我々は，入力信号に対して高域を強調するような周波数フィルタを用いて，その解決を試みた[1]．また，その拡張として，順方向の時系列のみではなく，逆方向の時系列情報も考慮したBLSTM（Bi-directional LSTM）を用いることを提案した[2]．本稿では時間方向のみではなく，周波数方向のBLSTMを構成し，その両方を用いるQDLSTM（Quad-directional LSTM）を用いることを提案する．提案手法では，時間方向のみでは難しい周波数方向の系列を考慮することが可能であるため，より高音域までの詳細な修復が可能になると考えられる．

2 音楽音響信号修復に用いる深層学習

はじめに，音楽音響信号修復に用いる深層学習手法について説明する．

2.1 Recurrent Newral Network（RNN）

RNNは，音楽音響信号のような系列データを扱うのに適した深層学習手法である[3]．特徴としては，Recurrentの名前の通り，系列方向に対して再帰構造を持つことにある．系列番号tにおける，RNNの入力を$\boldsymbol{x}_t$，出力y_tの関係式は以下のように表せる．

$$\boldsymbol{y}_t = f(\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{x}_t + \boldsymbol{W}_{rec}\boldsymbol{y}_{t-1} + \boldsymbol{b}) \quad (1)$$

ここで，fは活性化関数と呼ばれる非線形関数であり，$\boldsymbol{W}_{\boldsymbol{x}}$は入力に対する重み，$\boldsymbol{W}_{rec}$は以前の出力に対する重み，$\boldsymbol{b}$はバイアスを表す．この構造により，RNNは現在と過去の情報を合わせて考慮に入れた出力をすることが可能となっている．しかし，単純な再帰構造のみであるため，データが長くなると，学習時に用いる誤差の伝播の際，消失や発散を起こしてしまう．これを解決するため，RNNの改良型として，LSTMが提案されている[4]．

2.2 Long Short-Term Memory (LSTM)

RNNでは難しい長期依存関係のデータを扱うため，LSTMは，内部に情報を保持する機能，および各種入出力にゲートを持った構造となっている．実際のLSTMの実装には，複数のバージョンが存在するが，今回は1999年に提案されたGersらによるForget Gate (忘却ゲート) をもつ構造のLSTM[4]を用いた．LSTMの構造は図1のようになっており，式(2)~(4)として記述することができる．

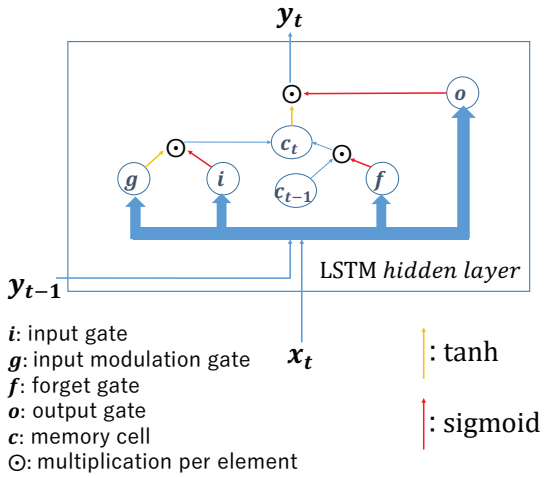


図 1: LSTM の構造

$$\begin{pmatrix} \boldsymbol{i} \\ \boldsymbol{f} \\ \boldsymbol{o} \\ \boldsymbol{g} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{T}_{2n,4n} \begin{pmatrix} x_t \\ \boldsymbol{y}_{t-1} \end{pmatrix} \tag{2}$$

$$\boldsymbol{c}_t = \boldsymbol{f} \odot \boldsymbol{c}_{t-1} + \boldsymbol{i} \odot \boldsymbol{g} \tag{3}$$

$$\boldsymbol{y}_t = \boldsymbol{o} \odot \tanh(\boldsymbol{c}_t) \tag{4}$$

$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{i}, \boldsymbol{f}, \boldsymbol{o}, \boldsymbol{g}, \boldsymbol{c} \in \Re^n$
σ はシグモイド関数，$\odot$ は要素ごとの積をあらわす．
$\boldsymbol{T}_{2n,4n}$ は $2n$ 次元ベクトルから $4n$ 次元ベクトルを作る写像．

$\boldsymbol{i}$ のインプットゲートは入力を表し，$\boldsymbol{g}$ の入力判断ゲートによって調節されることで直近のデータをどれだけ用いるかが判断できる．$\boldsymbol{c}$ は記憶領域のメモリセルであり，忘却判断ゲート $\boldsymbol{f}$ を通して次の時間での依存関係を表すために用いられる．出力判断ゲート $\boldsymbol{o}$ は次の層へ渡す信号の量を調節する．それぞれの判断ゲートは，直前，現在のメモリ状態，隣接する層からの入力値を用いて，その時々に情報を用いるかという判断のみを行う．そのため，判断ゲートの学習では誤差の消失，異常な増加に対する影響を受けにくいため，長期の依存関係を扱うことができる．

2.3 Bi-directional LSTM (BLSTM)

単一の LSTM のみでは，順方向のみの系列しか考慮できない．そこに，逆順の系列を入力とすることで，逆方向の系列を考慮することができる LSTM を合わせて用いることにより，正負両方向（Bi-directional）での系列を考慮できるようにしたものが BLSTM[5] である．BLSTM は単純に二つの LSTM を組み合わせたものであるため，順方向の処理，逆方向の処理二つの LSTM を用いる．つまり，入力が N 個の系列 x の場合，順方向 LSTM に対しての入力は $\{x_1, x_2, x_3, \cdots, x_N\}$ の順になる．一方，逆方向 LSTM には $\{x_N, x_{N-1}, x_{N-2}, \cdots, x_1\}$ の順で入力をする．その後，順方向，逆方向それぞれの出力 $\boldsymbol{y}_F$ と $\boldsymbol{y}_B$ とを統合する．これらを式で表すと，

$$\boldsymbol{y}(t) = \boldsymbol{y}_F(t) \oplus \boldsymbol{y}_B(N - t + 1) \tag{5}$$

となる．順方向と逆方向それぞれの出力系列の位置を合わせるために，二つの LSTM の出力を $\oplus$ によって統合する．

3 システム構成

3.1 LSTM を用いた音楽音響信号修復モデル

LSTM を用いた場合の音楽音響信号修復モデルを述べる．構成としては，図 2 a) に示すように，入力層，線型結合（全結合）層，LSTM，線型結合層，出力層からなる．入出力としては音楽音響信号に STFT（Short-TimeFourier transform）をかけて得られる各フレームの振幅スペクトルを用いる．音楽音響信号修復は，過去数フレームの情報を入力として，次の 1 フレームを予測するという回帰問題として扱う．このモデル関数を f_{LSTM} と表記すると，時刻 t での出力 y_t は，時刻 1 から t までの入力 x_1^{t-1} より，以下のように定義できる．

$$\boldsymbol{y}_t = f_{LSTM}(\boldsymbol{x}_1^{t-1}) \tag{6}$$

3.2 フィルタ内包型 LSTM

LSTM を用いた音楽音響信号修復モデルを図 2 b) に示す．構成としては，図 2 a) の音楽音響信号修復モデルに対し，入力の直後にフィルタ層を，出力の直前に逆フィルタ層を挿入した形である．この内，逆フィルタ層の係数は，必ず，フィルタ層での逆数を取るようフィルタ層と逆フィルタ層の係数を一体で更新するように設計した．周波数フィルタがネットワークに内包された構造になっているため，最初に初期値の設定は必要なものの，それ以降は，LSTM の学習と同時に学習できるため，データから最適なフィルタを学習することが可能である．フィルタを W_{filt} とし，逆フィルタを W^*_{filt} と表すと，

$$\boldsymbol{y}_t = \boldsymbol{W}^*_{filt} f_{LSTM}(\boldsymbol{W}_{filt} \boldsymbol{x}_1^{t-1}) \tag{7}$$
$$(\text{ただし}, \boldsymbol{W}_{filt} \boldsymbol{W}^*_{filt} = (1, 1, 1, 1, 1, \cdots)^T)$$

となる．

3.3 フィルタ内包型 BLSTM

フィルタ内包型 LSTM を Bi-directional LSTM に適用した場合について述べる．先に述べたフィルタ内包型 LSTM の LSTM を Bi-directional LSTM に拡張した音楽音響信号修復モデルを図 2 c) に示す．線形結合層，LSTM 層，線形結合層の三層をひとまとめとして，順方向を F-LSTM，逆方向を B-LSTM と呼ぶことにする．F-LSTM に対しては，入力 $\boldsymbol{x}$ は $\{x_1, x_2, x_3, \cdots, x_N\}$ の順になる．一方，B-LSTM には $\{x_N, x_{N-1}, x_{N-2}, \cdots, x_1\}$ の順で入力をす

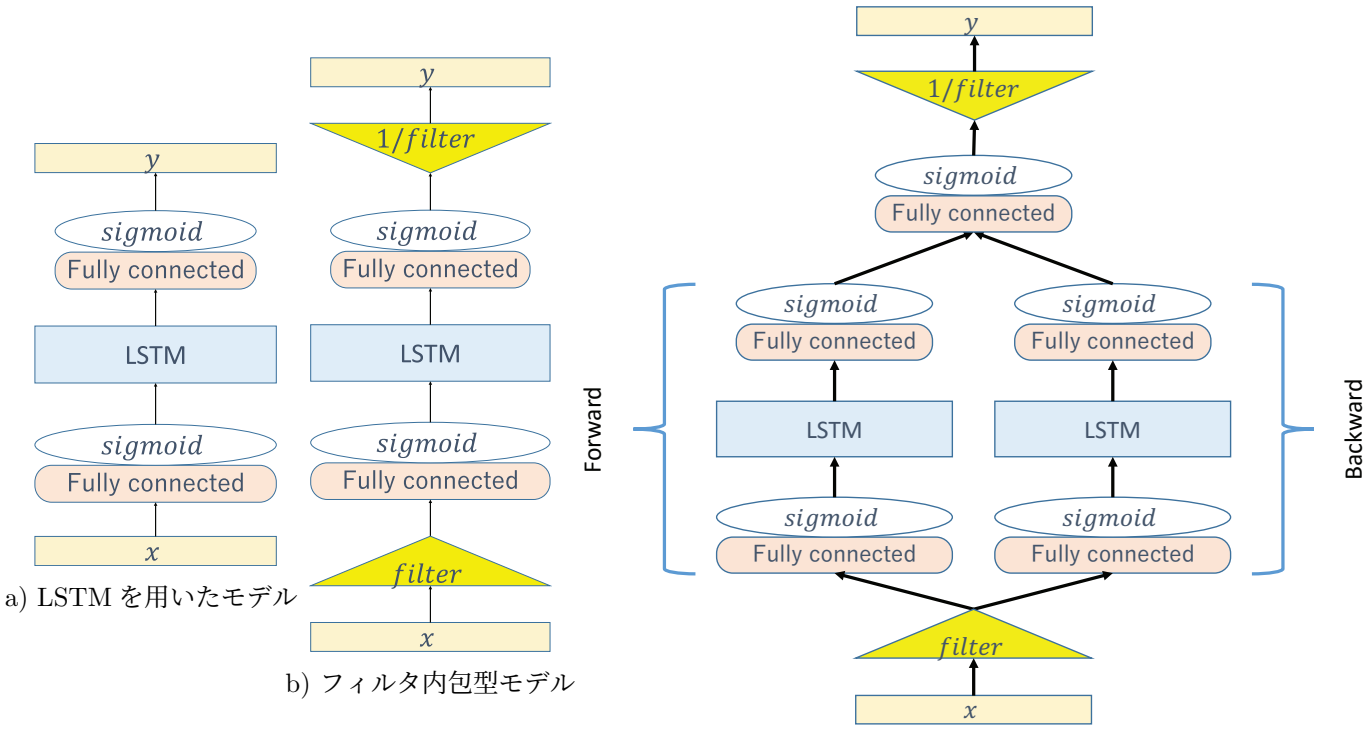


図 2: 音楽音響信号修復用のニューラルネットワーク

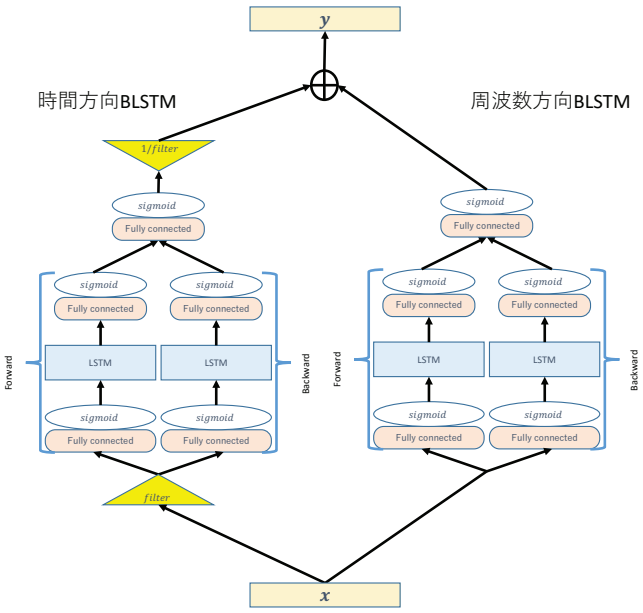


図 3: 提案モデル

る．その後，全結合層によりそれぞれからの出力をどのように合わせるかを判断する．入力 $\boldsymbol{x}$ が N 個の系列であるとし，全結合層を W_l とすると，

$$\boldsymbol{y}_t = \boldsymbol{W}^*_{filt}\boldsymbol{W}_l \begin{pmatrix} f_{F-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_1^{t-1}) \\ f_{B-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_N^{t+1}) \end{pmatrix} \quad (8)$$

となる．

3.4 Quad-directional LSTM

時間方向のみの BLSTM では，倍音など周波数方向での系列情報が保持されず，高音域での修復性能が依然として十分ではない．その問題を解決するため，本稿で拡張した点である Quad-directional LSTM について述べる．音楽音響信号は，時間方向に関して関係性をもつ系列信号であるが，倍音などから周波数方向にも関係性を持つ．そのため，STFT をかけたスペクトログラムにおいては，周波数方向にも系列データであるとみなすことができる．この周波数方向について LSTM を合わせて用いることで，時間方向，周波数方向の二次元での系列を考慮できるようにしたものが QDLSTM である [6]．本稿で用いたモデルは図 3 となる．時間方向に関しては，先のフィルタ内包型 BLSTM で用いたものを利用するが，周波数方向ではフィルタを用いない BLSTM を構成した．これら二つを独立に学習させ，その出力を足し合わせるという構成である．時間方向の系列 N 個，周波数方向の系列 F 個の信号に対して，時間方向 BLSTM の出力 $\boldsymbol{y}^{time}$，周波数方向 BLSTM の出力 $\boldsymbol{y}^{freq}$ とし，t, f をそれぞれ時間，周波数のインデックスとすると，式は以下のようになる．

$$\boldsymbol{y}_t^{time} = \boldsymbol{W}^*_{filt}\boldsymbol{W}_l^{time} \begin{pmatrix} f^{time}_{F-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_1^{t-1}) \\ f^{time}_{B-LSTM}(\boldsymbol{W}_{filt}\boldsymbol{x}_N^{t+1}) \end{pmatrix} \quad (9)$$

$$\boldsymbol{y}_f^{freq} = \boldsymbol{W}_l^{freq} \begin{pmatrix} f^{freq}_{F-LSTM}(\boldsymbol{x}_1^{f-1}) \\ f^{freq}_{B-LSTM}(\boldsymbol{x}_F^{f+1}) \end{pmatrix} \quad (10)$$

Algorithm 1 時間方向 BLSTM の学習

for $t = 1, 2, \ldots, l - 35$ **do**
  $x = W_{filt}\, data\{t, t+1, \ldots, t+35\}$
  $true = data\{t+3, t+4, \ldots, t+32\}$
  for $n = 1, 2, \ldots, 30$ **do**
    $y^*_{for}[n] = f^{time}_{F-LSTM}\{x^{n+2}_n\}$
    $y^*_{back}[31-n] = f^{time}_{B-LSTM}\{x^{35-n}_{37-n}\}$
  end for
  for $m = 1, 2, \ldots, 30$ **do**
    $y[m] = W^*_{filt} W^{time}_l \begin{pmatrix} y^*_{for}[m] \\ y^*_{back}[m] \end{pmatrix}$
    compute $loss$ of $y[m]$ and $true[m]$
    $loss_{sum} += loss$
  end for
  propagate $loss_{sum}$ backwards
end for

Algorithm 2 周波数方向 BLSTM の学習

for $t = 1, 2, \ldots, l - 11$ **do**
  $x = data\{t, t+1, t+2\} \& data\{t+9, t+10, t+11\}$
  $true = data\{t+3, t+4, t+5, t+6, t+7, t+8\}$
  for $f = 1, 2, \ldots, 257$ **do**
    $y^*_{for}[t, f] = f^{freq}_{F-LSTM}\{x_{t,f}\}$
    $y^*_{back}[t, 258-f] = f^{freq}_{B-LSTM}\{x_{t,258-f}\}$
  end for
  for $m = 1, 2, \ldots, 257$ **do**
    $y[t, m] = W^{freq}_l \begin{pmatrix} y^*_{for}[t, m] \\ y^*_{back}[t, m] \end{pmatrix}$
    compute $loss$ of $y[t]$ and $true[t]$
    $loss_{sum} += loss$
  end for
  propagate $loss_{sum}$ backwards
end for

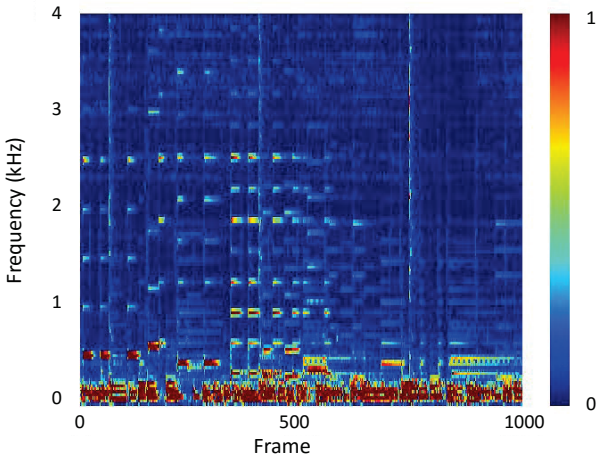
$$\boldsymbol{y}_{t,f} = \boldsymbol{y}^{freq}_{t,f} \oplus \boldsymbol{y}^{time}_{t,f} \tag{11}$$

4 評価実験

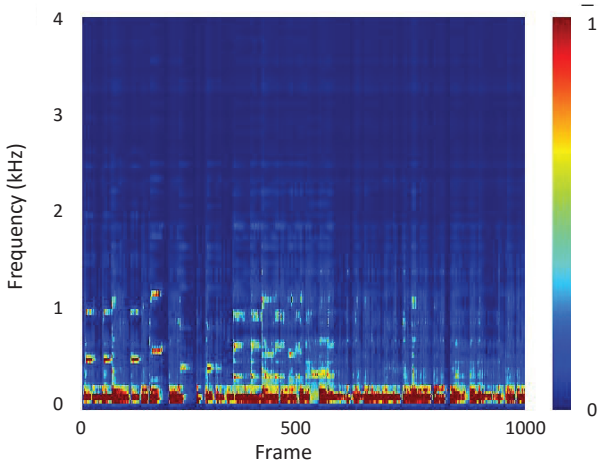
楽曲データ (サンプリングレート 16 kHz) に対し，フレーム長 512 サンプル，シフト長 128 サンプル，ハミング窓を窓関数として用いた STFT を行うことで振幅スペクトルを得る．各楽曲での最大振幅値を用いて正規化を行う．ニューラルネットワーク学習の際には，連続する 3 フレーム分 ($t \sim t+2$ フレーム) の振幅スペクトルを入力し，次のフレーム ($t+3$ フレーム) の振幅スペクトルを予測するよう学習を行う．評価の際には，欠損のない振幅スペクトル列を学習したニューラルネットワークに入力し，出力を時間方向に並べた振幅スペクトル列と，入力に用いた元の振幅スペクトル列との比較を行う．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 6 曲を用い，図 2 a), b), c)，図 3 それぞれについて学習した 3 種類のニューラルネットワークを構築した．最小化する損失関数には MSE（Mean Square Error）を用いた．また，図 2 b), c)，図 3 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 [7] を用いた．学習率は Adam (Adaptive Moment Estimation) [8] 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（1006 フレーム，約 8 秒分）に対して予測を行い，予測結果の内 1000 フレームを比較に用いた．順方向のみの深層学習モデルの場合，学習をするためには単純に t を 1 ずつ増加させ，そのたびごとの出力と正解との誤差を算出し，逆伝播させることが可能である．しかし，BLSTM の場合，順方向からの出力と逆方向からの出力との時間を合わせるために，ある程度の長さの範囲を指定する必要がある．そのため，本稿では，BLSTM の学習は，Algorithm1，Algorithm2 にしたがうものとした．ここで，l は学習に用いる音楽データのフレーム長である．また，QDLSTM における時間方向，周波数方向それぞれの BLSTM からの出力の統合には，二つの出力の要素の平均をとるという方法で行った．

5 比較結果

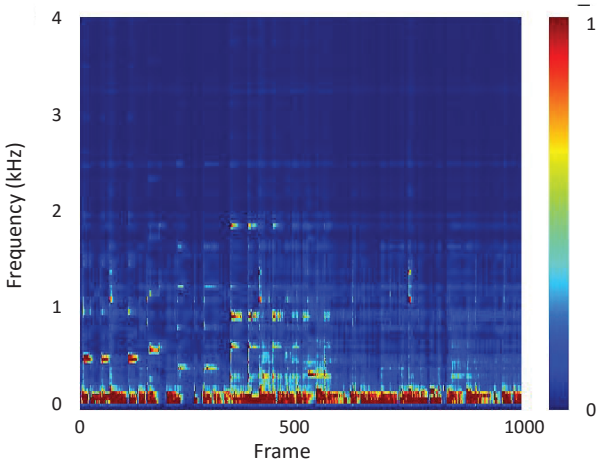
各モデルからの出力結果として得られた振幅スペクトログラム（0∼4 kHz）を図 4 b), c), d),e),f) に示す．また，図 4 a) に正解データの振幅スペクトログラムを示す．フィルタを内包した場合，通常の LSTM モデルではできない 2 kHz 付近の修復が可能となっている．また，通常の LSTM モデルでは 1 kHz 部分に大きく出ている誤差がフィルタ内包型では改善されている．しかし，音の輪郭（オンセット・オフセット）の関係が不明瞭なままとなっている．一方時間方向 Bi-directional LSTM の拡張では，2.5 kHz 付近の修復が可能となっている部分があり，また，音の輪郭部分が明瞭となっている．しかし，3 kHz 以上の高音域に関しては修復性能が十分でなく，倍音成分をうまく出力できていない．周波数方向 Bi-directional LSTM モデルでは，時間方向の BLSTM では修復ができていない 3 kHz 付近での修復が可能となり，より高音域までの周波数方向の関係を処理することが可能となっている．しかし，時間方向での関係性は保たれておらず，音信号のつながりが不明瞭となっている．これは，スペクトログラムに逆 STFT をかけ，音楽データとして再構成した際に顕著に感じらる．ピアノなどの倍音は修復されているため，音色としては豊かなものとなっている．しかし，各発音や残響音に違和感がのこり，結果として聞き心地の悪いものとなってい



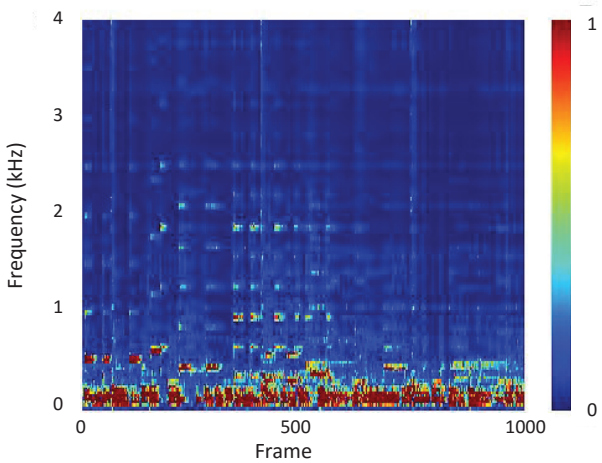
a) 正解データ



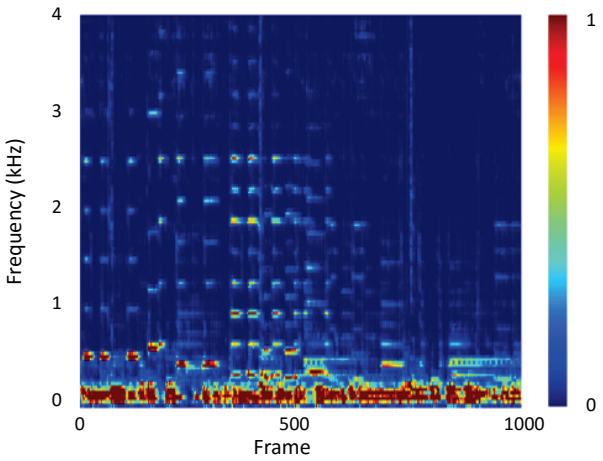
b) LSTM を用いたモデル



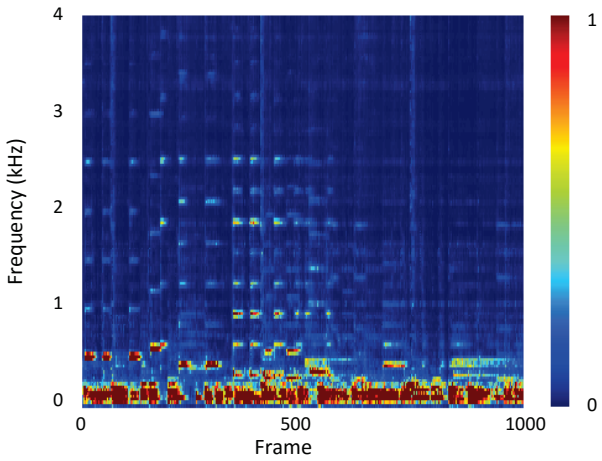
c) フィルタ内包型モデル



d) フィルタ内包型時間方向 BLSTM モデル



e) 周波数方向 BLSTM モデル



f) Quad-directional LSTM

図 4: 各モデルによる予測結果と正解データ

る．これらに対して，本稿での拡張である QDLSTM として再構成したものが図 4 f) である．時間方向 BLSTM の音の輪郭の明瞭さ，残響音関係の修復に加え，周波数方向 BLSTM の 2.5 kHz 以上での修復性能が向上している．音

表 1: 各モデル出力の SDR

モデル	SDR
LSTM	5.00
フィルタ内包型 LSTM	5.26
フィルタ内包型時間方向 BLSTM	8.77
周波数方向 BLSTM	6.51
QDLSTM	8.67

楽データとして再構成した場合，時間方向 BLSTM の際に感じられたピアノの倍音部分の表現の弱さが改善されており，より生楽器のピアノ音色として自然なものとなっている．それに加え，周波数方向 BLSTM の結果で感じられた発音間の不自然さが軽減され，自然な演奏として感じられる．しかし，依然として 3 kHz 付近から 4 kHz 付近にあるシンバルの倍音成分や，クラッシュシンバル，ライドシンバルのレガートなどの細かな残響による時間関係の修復はできておらず，ノイズのように感じられるものとなっている．

表 1 は，式 (12) によって算出した SDR（signal-to-distortion ratio）である．SDR は，修復音全体における正解との差と，正解との比をとったものであり，修復結果の歪の指標となる．s は正解，y は出力であり，数値が大きいほど歪が少ないことを示す．

$$SDR = 10\log_{10}\frac{\sum s^2}{\sum (s-y)^2} \tag{12}$$

周波数方向 BLSTM は，スペクトログラム上では最も高周波まで修復が可能であるが，聴覚上でも感じられた通り歪の大きい結果となっている．一方時間方向 BLSTM では歪が最も少ない結果となっている．QDLSTM は，SDR の値では時間方向 BLSTM に対してやや劣ったものとなっているが，周波数方向 BLSTM の高音域修復性能を持ったまま，歪を抑えられるということがわかる．

6 考察

周波数方向 BLSTM において，修復フレーム 6 フレームに対して前 3 フレーム，後ろ 3 フレームからの修復のため，時間関係の修復が十分にされていない．そのため，周波数方向での強い関係をもつ倍音成分に対しては系列情報として認識されるため，ピアノなどの倍音が強く出ているものには対応ができているのだと考えられる．しかし，残響に重なって新しく発音された場合などでは，残響音の信号は強度が弱いためうまく認識がされず，新しく発音された部分の倍音のみが修復されている．結果として，時間方向での系列がうまく処理されず，音楽として不自然な修復結果となっていると考えられる．また，QDLSTM として時間方向，周波数方向の二つの BLSTM を統合した場合，今回の統合方法では単純に各要素の平均を取るという方法をとっている．そのため，周波数方向 BLSTM には出ていない残響部分は時間方向 BLSTM の出力によって補完され，その逆に時間方向 BLSTM では出ていない高音域の倍音が周波数方向 BLSTM によって補完されている．これにより，フロント楽器であるため信号が強く出ているピアノや，低音域であるベースなどは音色が豊かになり，聞き心地のよい修復ができていると考えられる．しかし，シンバルレガートの部分など，残響音によって表現をされている奏法などでは，その弱い信号をどちらの BLSTM においても修復がされていないため，ノイズのような信号のままとして残されている．これらにより，今回用いた統合方法での QDLSTM では，聴覚的に自然なピアノ，ベースなど音階楽器と，ノイズのように感じられるシンバルなど打楽器とで乖離したような修復結果となっていると考えられる．また，単純な平均を取っている統合方法では，片方の BLSTM では出ているがもう一方では出ていない信号部分などが薄まってしまい，結果として正解から離れてしまうという部分も確認できる．これらの問題に対して，音階楽器と打楽器とでの特徴を深層学習に対して与える方法を検討する必要がある．また，二つの BLSTM の統合を，周波数ビンごと，時間ごとで比率を変えることができれえば，より詳細な修復が可能になると考えられる．

7 終わりに

本稿では，深層学習の一手法である LSTM を用いた音楽音響信号修復法として，周波数フィルタを内包するモデルを提案し，その拡張として時間方向，周波数方向の二つの Bi-directional LSTM を用いた Quad-directional LSTM に適用した．提案手法は，過去の手法に対してより高い修復性能を示すことを，修復タスクを想定した評価実験によって示した．

謝辞 本研究は、JSPS 科研費 16H02884, 16K00294, 17K00365 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] 谷口他, “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”, 第 79 回情報処理学会全国大会, 2017

[2] 谷口他, “Bi-directional LSTM を用いた音楽音響信号修復法の提案”, 第 35 回 日本ロボット学会学術講演会 , 2017

[3] Jerey L Elman. “Finding structure in time.” Cognitive science, 14(2):179211, 1990.

[4] F.A.Gers *et al.*,“Learning to forget: Continual prediction with LSTM.” ICANN ’99, 1999 p. 850 855

[5] A. Graves, J. Schmidhuber. “Framewise Phoneme Classification with Bidirectional LSTM Networks.” IJCNN 2005, Montreal, Canada, pp. 2047-2052.

[6] A. Graves, S. Fernndez, “J. Schmidhuber. Multi-Dimensional Recurrent Neural Networks.” ICANN 2007, Porto, Portugal, pp. 549-558.

[7] 西山他, “音響振動工学”, コロナ社, 1979.

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