双方向LSTMによるラウドネス及びMFCCからの振幅スペクトログラム予測と評価*

☆川口翔也，北村大地（香川高専）

1　はじめに

生成モデル系の深層ニューラルネットワーク (deep neural network: DNN) に基づく楽器音信号や音色変換は様々なものが研究されている．例えば，楽器音信号の生成手法として，Differentiable digital signal processing（DDSP）[1] がある．また，潜在的な特徴量を教師無しで学習できる生成モデル系 DNN の変分自己符号化器（variational auto-encoder: VAE）を用いた楽器音解析や生成方法として，知覚的メトリクスに基づく正則化付き VAE [2] や VAE による音色及び音高の分離表現学習 [3, 4, 5] などが提案されている．DDSP は複数の正弦波やフィルタリングされたノイズで楽器音信号を合成する手法であるため，シンセサイザーのようにやや人工的な楽器音が合成される．一方，音色・音高の分離表現学習では，楽器音信号の音量に関しては明示的に分離表現されておらず，特徴量の解釈も比較的難しいと思われる．

著者らは現在，音高，音色，及び音量という楽器音信号を構成する 3 要素を，これまでよく検討された（レガシーな）特徴量としてそれぞれ抽出し，音色のみを VAE に入力する新しい音色変換アルゴリズムの構築を目指している [6]．以後，このシステムを「提案音生成システム」と呼ぶ．提案音生成システムを用いることで，複数種類の楽器音信号の音色の内挿や外挿が可能になることを期待しており，新しい芸術及び音楽の発展に寄与することができると考える．

提案音生成システムの実現には，合成音生成時に，音高，操作された音色，及び音量の 3 つの特徴量から合成音の振幅スペクトログラムを予測する必要があるが，解析的な処理での高精度な予測は困難である．そのため，文献 [6] ではまず，3 つの特徴量から DNN を用いて振幅スペクトログラムを予測できるかについて検討した．具体的には，多層パーセプトロン（multi-layer perceptron: MLP）及び再帰型ニューラルネットワーク（recurrent neural network: RNN）の 1 つであるゲート付き回帰型ユニット（gated recurrent unit: GRU）[7] を用いた双方向再帰型ニューラルネットワーク（bidirectional RNN using GRU: BiGRU）の 2 種類の DNN で，振幅スペクトログラムの予測精度を比較し，MLP と BiGRU のどちらが振幅スペクトログラムの予測に効果的かについて，実験的に調査した．文献 [6] では予測精度について具体的な評価が未実施であったため，本稿では，長・短期記憶（long-short term memory: LSTM）[8] ユニットを用いた双方向再帰型ニューラルネットワーク（bidirectional RNN using LSTM unit: BiLSTM）を新たに加えた 3 種類のネットワーク構造で振幅スペクトログラムの予測精度を客観的に比較する．

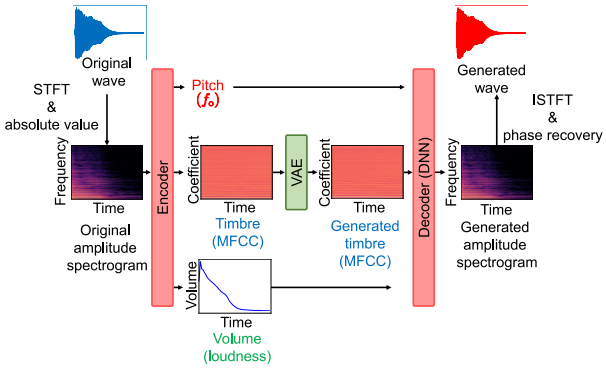

Fig. 1　Detailed process flow of the proposed sound generation system [6].

2　提案手法

2.1　提案音生成システム全体の説明

提案音生成システムの全体図を Fig. 1 に示す．まず，入力となる音響信号をエンコーダに通し，音高，音色，及び音量の 3 つの特徴量を抽出する．この 3 つの特徴量には，それぞれレガシーな尺度として，基本周波数 f_0，メル周波数ケプストラム係数（mel-frequency cepstrum coefficient: MFCC）[9]，及びラウドネスをそれぞれ用いる．ここで MFCC とは，時間周波数領域で表現された音声及び楽器音等の音色の特徴量である．音色のみの特徴量を可能な限り抽出しているため，音高や音量はあまり反映されない特徴量である．さらに抽出された MFCC のみを VAE に入力し，複数種類の楽器が埋め込まれた MFCC の潜在空間を学習する．出力部では，基本周波数，VAE で生成された MFCC，及びラウドネスの 3 つの特徴量を何らかのデコーダに入力し，振幅スペクトログラムに復元することで，合成された楽器音信号を生成する．本稿では，このデコーダに DNN を用いる．

以上が提案音生成システムの概要である．学習時は Fig. 1 に示す入力と出力（original wave と generated wave）間の損失が小さくなるように，VAE 及びデコー

*Amplitude spectrogram prediction and evaluation from MFCC and loudness using bidirectional LSTM. By Shoya KAWAGUCHI and Daichi KITAMURA (NIT Kagawa).

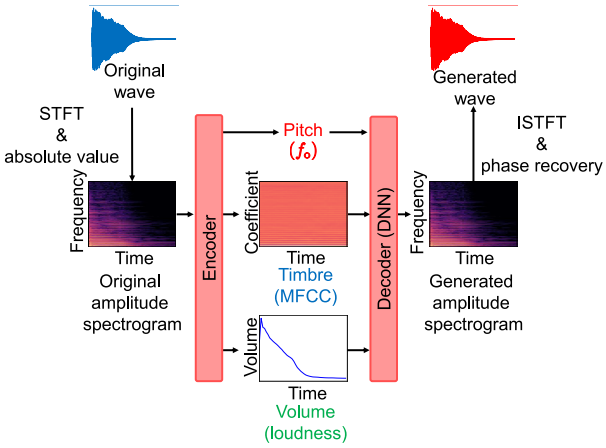

Fig. 2　Training process flow of the proposed DNN-based timbre decoder.

ダをそれぞれ学習する．学習後は，複数種類の楽器が埋め込まれた MFCC の潜在空間上で任意の値を与えることで，新しい音色を持つ楽器音信号を生成できることが期待される．この提案音生成システムは構想段階であり，文献 [6] では次節で述べるように，f_0・MFCC・ラウドネスから振幅スペクトログラムを予測する DNN デコーダの性能が簡易的に調べられただけである．

2.2　本稿で扱う問題

Fig. 1 の提案音生成システムでは，f_0，VAE で生成された MFCC，及びラウドネスの 3 つの特徴量から振幅スペクトログラムを生成する必要がある．しかしながら，MFCC は音色のみを低次元空間で表現した特徴量であるため，これらの 3 つの特徴量から解析的に振幅スペクトログラムを高精度に求めることはできない．そこで文献 [6] では，DNN に基づくデコーダを Fig. 2 に示す方法で学習し，その予測精度について簡易的に調査した．ただし，このデコーダは MFCC 及びラウドネスのみを入力とする DNN を想定している．音高の特徴量は離散的であることから，DNN の入力に与えるのではなく，各音高専用に学習した DNN を選択するために用いる．すなわち，予め学習された音高依存の DNN を複数用意し，いずれかの DNN が入力の音高により選択される．このような方式を取ることで，DNN に基づくデコーダは音高に対する汎化性能を獲得する必要がなくなり，より高精度な振幅スペクトログラムの予測が可能になると考えられる．

本稿では，文献 [6] の実験をより詳細にした結果を報告する．具体的には，デコーダとして用いる DNN に，文献 [6] で調査された MLP 型 DNN 及び BiGRU 型 DNN だけでなく，BiLSTM 型 DNN を加えた 3 種類を取り扱い，予測精度を実験的に比較する．さらに，合成された楽器音信号の MFCC の予測誤差を定量的に評価し，性能について議論する．

2.3　DNN に基づくデコーダ

提案 DNN デコーダの入力には，MFCC とラウドネスを用いる．この時のラウドネスは DDSP の文献中の方法 [1] よりも簡易的な抽出方法として，次式で計算する．

$$v_j = \sum_{i=1}^{I} |z_{ij}| \tag{1}$$

ここで，z_{ij} は音響信号に短時間 Fourier 変換（short-time Fourier transform: STFT）を適用し得られた複素スペクトログラム $\boldsymbol{Z}$ の要素であり，$i = 1, 2, \cdots, I$ 及び $j = 1, 2, \cdots, J$ はそれぞれ周波数ビンのインデクス及び時間フレームを示す．一方，MFCC はラウドネスの影響を排除するために，次式の時間フレーム毎の正規化を施したスペクトログラムから求める．

$$\hat{z}_{ij} = \frac{z_{ij}}{v_j} \tag{2}$$

即ち，式 (2) で得られる正規化済みパワースペクトログラム $|\hat{\boldsymbol{Z}}|^{\cdot 2}$ を用いて MFCC を求める．この MFCC を $\boldsymbol{C} \in \mathbb{R}^{K \times J}$ と定義する．提案 DNN デコーダの入力は，MFCC とラウドネスを次式で結合した行列とする．

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{v}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{(K+1) \times J} \tag{3}$$

ここで，$\boldsymbol{v} = [v_1, v_2, \cdots, v_J]^{\mathrm{T}}$ 及び K は MFCC の次数である．この次数は，MFCC への変換時に用いられるメルフィルタバンクと呼ばれるバンドパスフィルタのフィルタ数と同じである．

前述の 3 種類の提案 DNN デコーダのうち Bi-GRU 及び BiLSTM について詳細を説明する．BiGRU 及び BiLSTM は双方向 RNN（bidirectional RNN: BiRNN）の一種であり，MFCC の時間方向のように連続的な系列の次元を持つ入力に対して，その系列方向の再帰性を考慮した学習ができる．過去から未来の方向（順方向）の RNN の出力を過去側から順に $\boldsymbol{h}_1^{(\mathrm{forward})}, \boldsymbol{h}_2^{(\mathrm{forward})}, \cdots, \boldsymbol{h}_J^{(\mathrm{forward})}$ とし，未来から過去の方向（逆方向）の RNN の出力を未来側から順に $\boldsymbol{h}_1^{(\mathrm{backward})}, \boldsymbol{h}_2^{(\mathrm{backward})}, \cdots, \boldsymbol{h}_J^{(\mathrm{backward})}$ とする．提案 DNN デコーダでは，式 (3) の $\boldsymbol{X}$ を Fig. 3 のように入力する．具体的には，Fig. 4 に示すように，$\boldsymbol{X}$ の時間 j における入力ベクトル $\boldsymbol{x}_j$ から 4 つの GRU もしくは LSTM ユニットを通して，時刻 j における出力ベクトル $\boldsymbol{h}_j$ を出力する．この時，BiGRU 及び BiLSTM では順方向の出力ベクトル $\boldsymbol{h}_j^{(\mathrm{forward})}$ 及び逆方向の出力ベクトル $\boldsymbol{h}_{J-j}^{(\mathrm{backward})}$ が出力され，その要素毎の積を取ったベクトルを時間 j の出力ベクトル $\boldsymbol{h}_j$ として扱う．なお，Fig. 4 の GRU 及び LSTM ユニットの

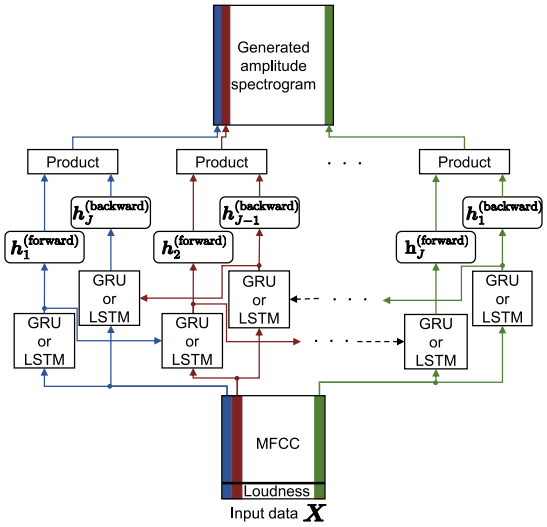

Fig. 3　Architecture of BiRNN used as the DNN decoder.

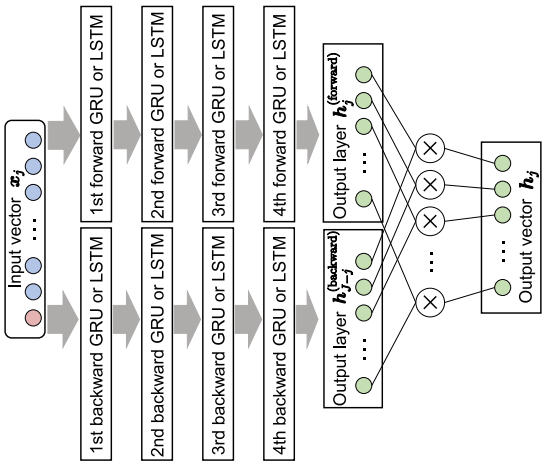

Fig. 4　Data flow in the multi-layer BiGRU at time frame j.

出力ベクトルの要素数は全て I と設定している．

提案 DNN デコーダの学習では，次式の平均二乗誤差（mean squred error: MSE）ロス又は multi-scale spectral（MSS）ロスを損失関数に用いる．

$$L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = ||\mathbf{y} - \mathbf{p}||_2^2 \tag{4}$$

$$L_{\mathrm{MSS}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||_1 + ||\log \boldsymbol{y} - \log \hat{\boldsymbol{y}}||_1 \tag{5}$$

ここで，$\boldsymbol{y} \in \mathbb{R}_{\geq 0}^{IJ}$ 及び $\hat{\boldsymbol{y}} \in \mathbb{R}_{\geq 0}^{IJ}$ はそれぞれ入力と予測の振幅スペクトログラムをベクトル化したものであり，$||\cdot||_p$ は L_p ノルムを表す．

3　振幅スペクトログラム予測実験

3.1　実験条件

本章では，提案 DNN デコーダの振幅スペクトログラムの予測精度を確認するための実験について述べる．本稿で新たに追加された BiLSTM の構成は，文献 [6] で使用した BiGRU の GRU を LSTM ユニットに入れ替えた構造であり，Figs. 3 及び 4 に示す通りである．MFCC の客観評価指標には，次式で定義される MFCC 相対二乗誤差（MFCC relative squared error: MRSE）を用いる．

$$\mathrm{MRSE} = 10 \log \frac{\sum_{j=1}^{J} \sum_{k=2}^{14} (c_{kj} - \hat{c}_{kj})^2}{\sum_{j=1}^{J} \sum_{k=2}^{14} c_{kj}^2} \,[\mathrm{dB}] \tag{6}$$

ここで，c_{kj} は DNN の入力振幅スペクトログラムの MFCC $\boldsymbol{C}$ の要素であり，$\hat{c}_{kj}$ は予測結果の振幅スペクトログラムから直接算出した MFCC である．MRSE は値が低いほど高精度であることを示す．また，式 (6) の通り，音色の情報を強く含む傾向にある MFCC の 2 から 14 次元を用いる．その他の実験条件については文献 [6] と同様とした．学習データ及びテストデータについても，文献 [6] を参照されたい．

3.2　実験結果

Figs. 5 及び 6 はそれぞれ，学習済みの MLP，BiGRU，及び BiLSTM に対してテストデータ中のピアノ及びギターの G4 音の振幅スペクトログラムを入力した際の予測結果の例を示している．具体的には，ピアノとギターの G4 音の 36 種類の音色を学習データに用いて構築した MLP 型 DNN，BiGRU 型 DNN，及び BiLSTM 型 DNN の 3 つに対して，テストデータ中のあるピアノの G4 音を入力した予測結果が Fig. 5 であり，テストデータ中のあるギターの G4 音を入力した予測結果が Fig. 6 である．いずれの結果をみても，MLP を用いた振幅スペクトログラムの予測は失敗している．このような予測失敗の傾向は学習データに対しても同様であり，MLP を用いて MFCC とラウドネスのみから振幅スペクトログラムを予測することが非常に困難であることを示唆している．一方，BiGRU 及び BiLSTM を用いた予測では，ピアノとギターの両楽器のテストデータで調波構造や時間的な推移を正確に予測できている．BiLSTM については，低周波帯域における調波構造以外の振幅の予測も行えていることがわかる．これは，BiLSTM 及び BiGRU が MFCC 及びラウドネスの時間方向の連続性を考慮しつつ学習できたことに起因している．

Fig. 7 (a) 及び (b) に示す，MRSE を評価指標として用いた客観的評価では，ピアノ及びギターの両楽器において BiLSTM が最も高精度に予測が行えていることがわかる．さらに MSE ロス及び MSS ロスについては，MSS ロスの方が高精度に予測できるモデルを実現していることがわかる．

4　おわりに

本稿では，文献 [6] で取り上げたネットワークに BiLSTM 型 DNN を加え，同条件で振幅スペクトログラムを予測する実験を行い，客観的評価指標として

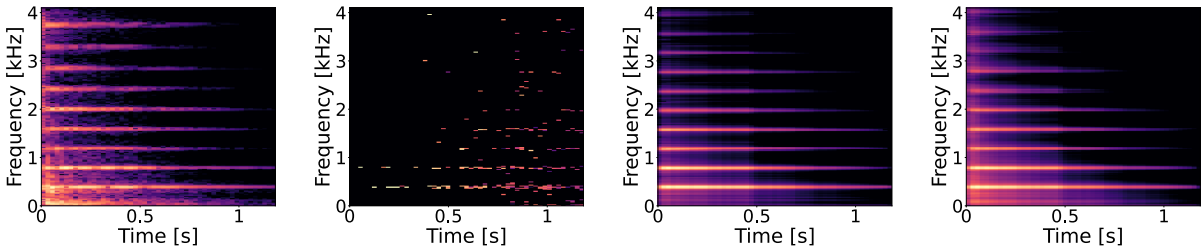

Fig. 5　Example of spectrograms of piano test data: (a) input, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

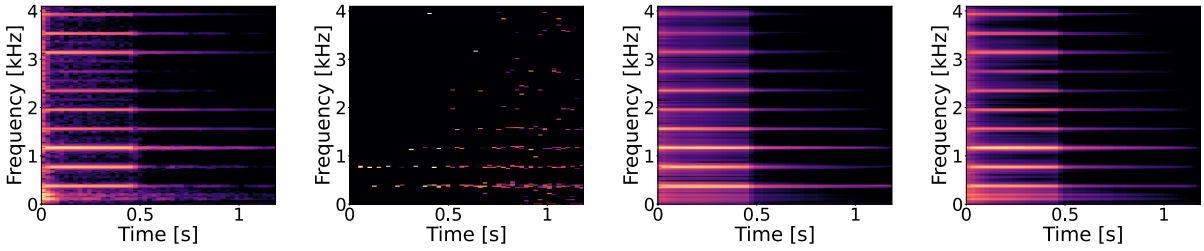

Fig. 6　Example of spectrograms of guitar test data: (a) input, (b) predicted by MLP, (c) predicted by BiGRU, and (d) predicted by BiLSTM.

MRSE を用いた評価を行った．実験結果より，BiLSTM を用いた DNN デコーダが最も高精度で振幅スペクトログラムを予測可能であることが分かった．

謝辞

本研究の一部は公益信託小野音響学研究助成基金及び JSPS 科研費 22H03652 及びの助成を受けた．

参考文献

[1] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, “Differentiable Digital Signal Processing,” *in Proc. ICLR*, 2020.

[2] P. Esling, A. Chemla-RomenuSantos, and A. Bitton, “Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics,”*in Proc. DAFX*, 2018.

[3] Y. J. Luo, K. Agres, and D. Herremans, “Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders,”*in Proc. ISMIR*, 2019.

[4] Y. J. Luo, K. W. Cheuk, T. Nakano, M. Goto, and D. Herremans, “Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds,” *in Proc. ISMIR*, pp 700–707, 2020.

[5] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii, and S. Morishima, “Pitch-timbre disentanglement of musical instrument sounds based on VAE-based metric learning,”*in Proc. ICASSP*, pp. 111–115, 2021.

[6] 川口翔也, 北村大地, “双方向 RNN による MFCC 及びラウドネスからの振幅スペクトログラム予測,” *IPSJ SIGMUS*, vol. 2022-MUS-134, no. 60, pp. 1–6.

[7] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXiv: 1406.1078, 2014.

[8] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” *IEEE Trans. Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[9] F. Zheng, G. Zhang, and Z. Song, “Comparison of different implementations of MFCC,” *J. Computer Science and Technology*, vol. 16, pp. 582–589, 2001.

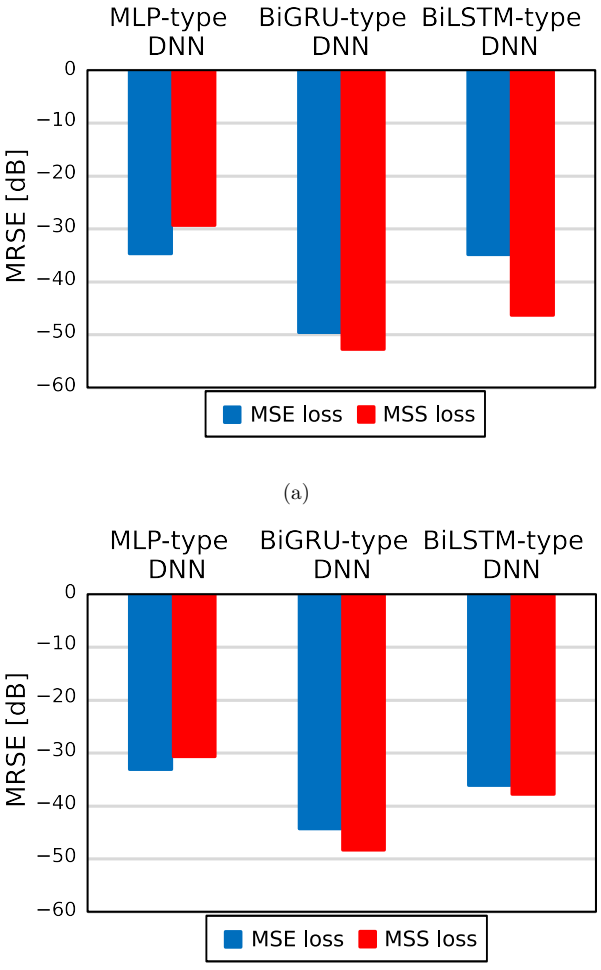

Fig. 7　Average MRSE of the predicted amplitude spectrograms for (a) piano and (b) guitar.