音響特徴量からのDNNスペクトログラム予測の主観評価*

☆川口翔也（農工大），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（DNN）に基づく音色変換は様々なものが研究されている．例えば，DNNの中でも変分自己符号化器（VAE）と呼ばれる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を用いた楽器音の解析や生成が提案されている [1]．Fig. 1にVAEに数字の手書き画像を適用したものを示す．Fig. 1のように，潜在空間と呼ばれる空間にそれぞれの数字の集合を見ることができる．さらに空間上では，「7」と「9」の集合の間に相当する乱数を入力することで「7」と「9」の中間の手書き数字が出力される．

文献 [2] では，このような各集合の相対関係の学習を，楽器音の音色集合に対して適用することで，VAEを用いた新しい音色変換アルゴリズムの構築を目指した．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，「ギターとピアノの中間の音色」等の聞いたことのない音色を持つ音響信号を作成できる．さらに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．

2　提案音生成システム

2.1　提案音生成システムの詳細及び現状

文献 [2] にて提案した提案音生成システムの全体概要をFig. 2に示す．入力となる音響信号に対してエンコーダを通し，音高，音色，及び音量の3つの特徴量を抽出する．さらに抽出された音色のみをVAEに入力し，VAEの出力として音色を得る．こうして得られた音高，VAEで生成された音色，及び音量の3つの特徴量をデコーダに入力し，新しい音響信号を生成・出力する．

提案音生成システムを実現には，前述の3つの特徴量から振幅スペクトログラムの予測が必要となるが，実現可能かは不明であった．そこで，文献 [2] では提案音生成システムにおいて必要な部分システムとして3つの特徴量から振幅スペクトログラムを予測するデコーダを構築した．特に，多層パーセプトロン（MLP），ゲート付き回帰型ユニット（GRU）を用いた双方向再帰型ニューラルネットワーク（BiGRU），及び長・短期記憶（LSTM）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（BiLSTM）の3種類を比較し，どのようなネットワーク構造が高精度な

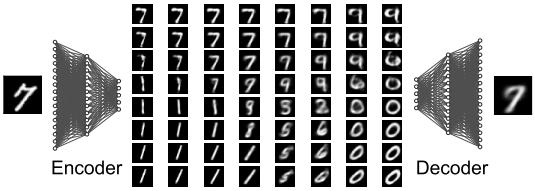

Fig. 1　Latent space of VAE trained with images of handwritten numbers.

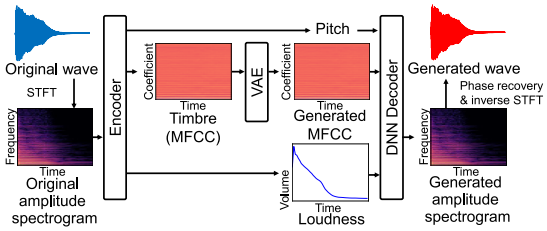

Fig. 2　Process flow of proposed timbre conversion system [2].

振幅スペクトログラムの予測に効果的かを客観的評価を用いて評価を行い，BiLSTM及びBiGRUをデコーダとして用いることで高精度な予測が可能であるという結果を得た．

2.2　本稿の目的

文献 [2] では，音色の予測精度を客観評価指標によって評価した．しかしながら，提案音生成システムの目的は，音楽という人間の感性と関連する芸術につながるものである．従って，提案音生成システムで予測した音響信号の音質及び精度は，実際に人間が聞いた上で主観的に評価されるべきであると考える．そこで，本稿では，提案音生成システムで予測した音響信号が入力の音響信号にどの程度近いかを被験者による主観評価する実験を実施し，性能について考察を行う．

3　主観評価実験

3.1　実験手法

本稿では，予測精度の主観評価方法としてdegradation mean opinion score（DMOS）法 [3] を用いる．DMOS法はFig. 3に示すように，先に入力の音響信号を被験者に提示し，次に予測した音響信号を提示することで，予測した音響信号が入力の音響信号からどの程度劣化したかを評価させる方法である．各音響信号は，Fig. 4に示す項目及び評点に基づいて評価される．

今回の主観評価実験は，Fig. 4に示すMATLABで

*Subjective evaluation of spectrogram prediction from acoustic features based on deep neural network. By Shoya KAWAGUCHI (Tokyo University of Agriculture and Technology) and Daichi KITAMURA (NIT Kagawa).

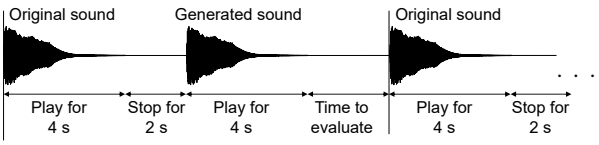

Fig. 3　Process flow of DMOS.

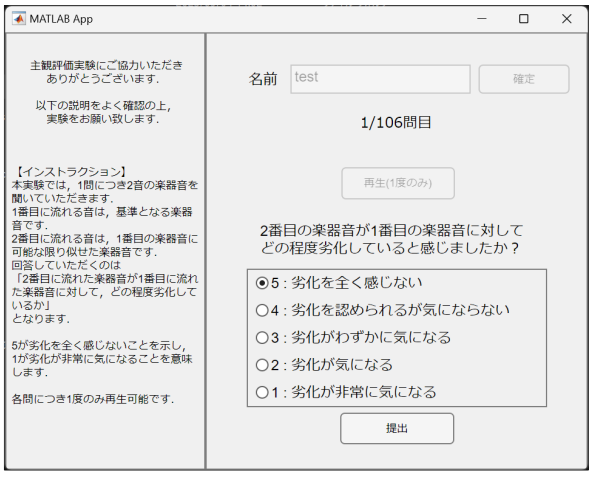

Fig. 4　Screen of MATLAB-based GUI application used in subjective test.

作成した GUI アプリケーションを用いて行った．以下に実験手順を示す．

(a) MATLAB の GUI アプリケーションを実行する
(b) Fig. 4 に示す画面が表示され，被験者はインストラクションを読む
(c) テキストボックスにユーザーネームを入力し，確定ボタンを押下する．
(d) 被験者は再生ボタンを押下し，音源を視聴する
(e) 視聴後に評価を行い，提出ボタンを押下する
(f) 手順 4 及び手順 5 を指定の問題数繰り返す
(g) 問題数の半分が終了時点で休憩を取る
(h) ウィンドウを閉じて終了する

3.2　実験条件

本稿で評価対象とする音響信号は，文献 [2] の客観評価において高い精度で予測できていると評価された BiGRU 及び BiLSTM を用いたデコーダで予測した音響信号である．実験では，音色のみを比較するために全ての音響信号の音高は統一した．被験者は，10 代から 40 代の男女 15 名を対象とした．

3.3　結果

被験者による主観評価実験の結果を Fig. 5 に示す．水平線は中央値，箱は四分位範囲，アスタリスクは平均値，ひげは中央値から四分位数までの 1.5 倍の範囲，箱のノッチは有意水準 95%で真の中央値となりうる範囲をそれぞれ示す．図中の括弧書きされた数字は各楽器のテストデータ数を示している．DMOS 値は値が大きい程，高精度に振幅スペクトログラムを予測できていることを示しており，評価の平均値で比較すると Mallet の除いた全ての楽器において，BiLSTM

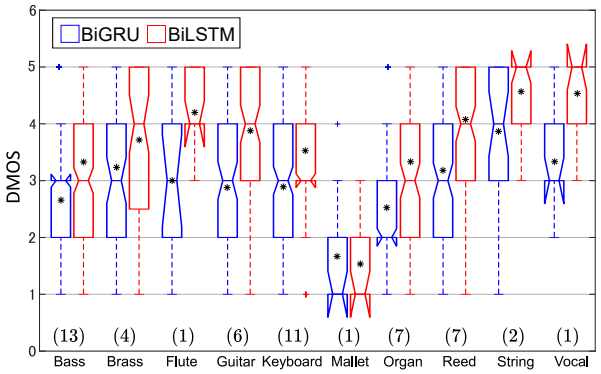

Fig. 5　Results of subjective test with DMOS for each instrument by BiGRU and BiLSTM.

の予測精度が高いことがわかる．Mallet については，評価の平均値は BiGRU の方が高いがどちらにおいても低い評価となった．これは Mallet が打楽器に分類され，特殊な調波構造であることが原因であると考えられる．実験後，被験者にヒアリングを行ったところ，「短音の音響信号は，本来の音の後にうなりのようなものがあり，全く異なる楽器音に聞こえた」という意見があり，実際に短音の音響信号の DMOS 値は非常に低い結果となった．さらに，ノッチが被っていない Brass，Flute，Guitar，Organ，Reed，String，Vocal については，有意水準 95%で真の中央値が異なることがわかった．

4　まとめ

本稿では，文献 [2] で作成した音高，音色，及び音量の 3 つの音響特徴量から振幅スペクトログラムの予測を行う DNN デコーダの主観評価実験を行った．実験結果より，BiLSTM をデコーダとして用いることで，人間が聞いたうえで違和感のない楽器音の振幅スペクトログラムを予測できていることを確認した．加えて，文献 [2] の客観評価結果と同様の結果が得られたことから，文献 [2] で用いた客観評価指標は人間の感性にあった評価が行えていたことをわかった．

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 23K24908 の助成を受けた．

参考文献

[1] 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.
[2] S. Kawaguchi and D. Kitamura, “Amplitude spectrogram prediction from mel-frequency cepstrum coefficients using deep neural networks,” *Journal of Signal Processing*, vol. 27, no. 6, pp. 207–211, 2023.
[3] ITU-T Recommendation P.800, “Methods for subjective determination of transmision quality”, 1996.
[4] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, “Neural audio synthesis of musical notes with WaveNet autoencoders,” *in Proc. ICML*, pp. 1068–1077, 2017.