1-R-35

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）

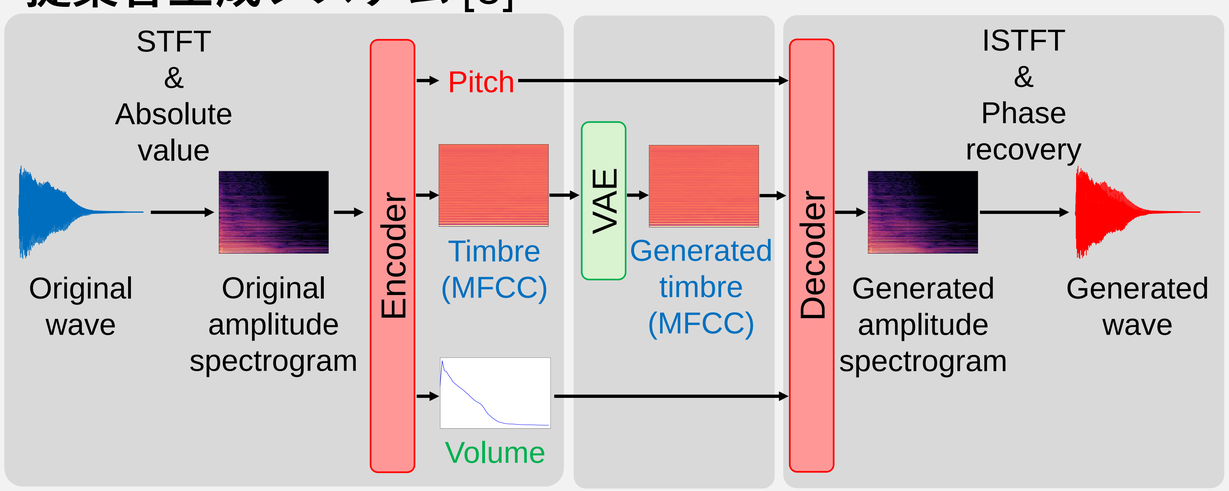
1. 研究背景

変分自己符号化器（VAE）による音色変換・生成

・VAEによる音色・音高の分離表現学習 [1]
・知覚的メトリクス正規化付きVAE [2]

音の三要素である音高・音色・音量が**混在**

提案音生成システム [3]



1. 入力音響信号から音高・音色・音量を抽出
2. 音色のみVAEで学習し，音色を生成
3. 生成された音色と音高・音量から音響信号を出力

提案音生成システムの問題

音高・音色・音量から振幅スペクトログラムを予測する**デコーダ**が必要

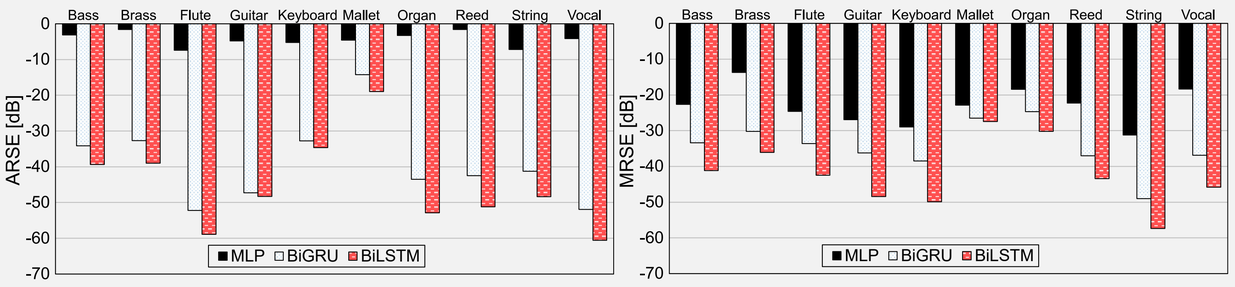
DNNを用いたデコーダにより達成

2. DNNデコーダ

振幅スペクトログラム予測DNNデコーダ

・入力：音色（メル周波数ケプストラム係数：MFCC）
　　音量（時間フレームごとの振幅の総和）
　　音高（音高別に学習したDNNデコーダの選択に使用）
・出力：振幅スペクトログラム
・アーキテクチャ
　- 多層パーセプトロン（MLP）
　- 双方向再帰型ニューラルネットワーク
　　・長・短期記憶（BiLSTM）
　　・ゲート付き回帰型ユニット（BiGRU）

客観評価結果



・評価指標
　- 振幅相対二乗誤差（ARSE）
　- MFCC相対二乗誤差（MRSE）
・すべての楽器で両指標においてBiLSTMが最も高評価
・打楽器に属するMalletは評価が低い→予測が困難

本発表の主題

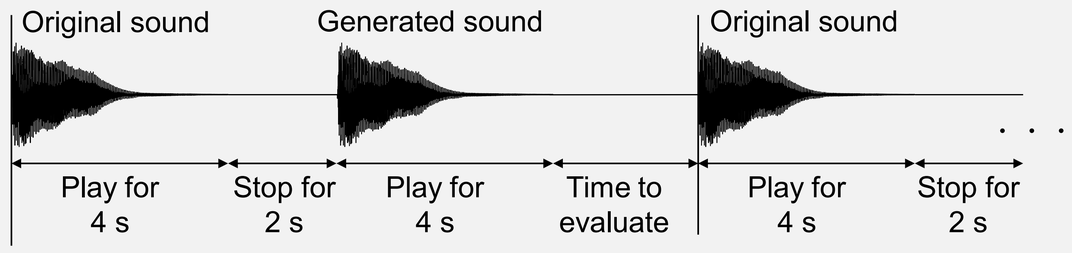
音楽は**人間の感性**と関連するもの

人間が実際に聞いたうえで**主観的に評価**する必要がある

3. 主観評価実験方法

Degradation mean opinion score（DMOS）

・本来の楽器音→予測した楽器音の順に再生し，どの程度劣化して聞こえるかを評価



MATLABで作成した実験用アプリケーション

実験手順

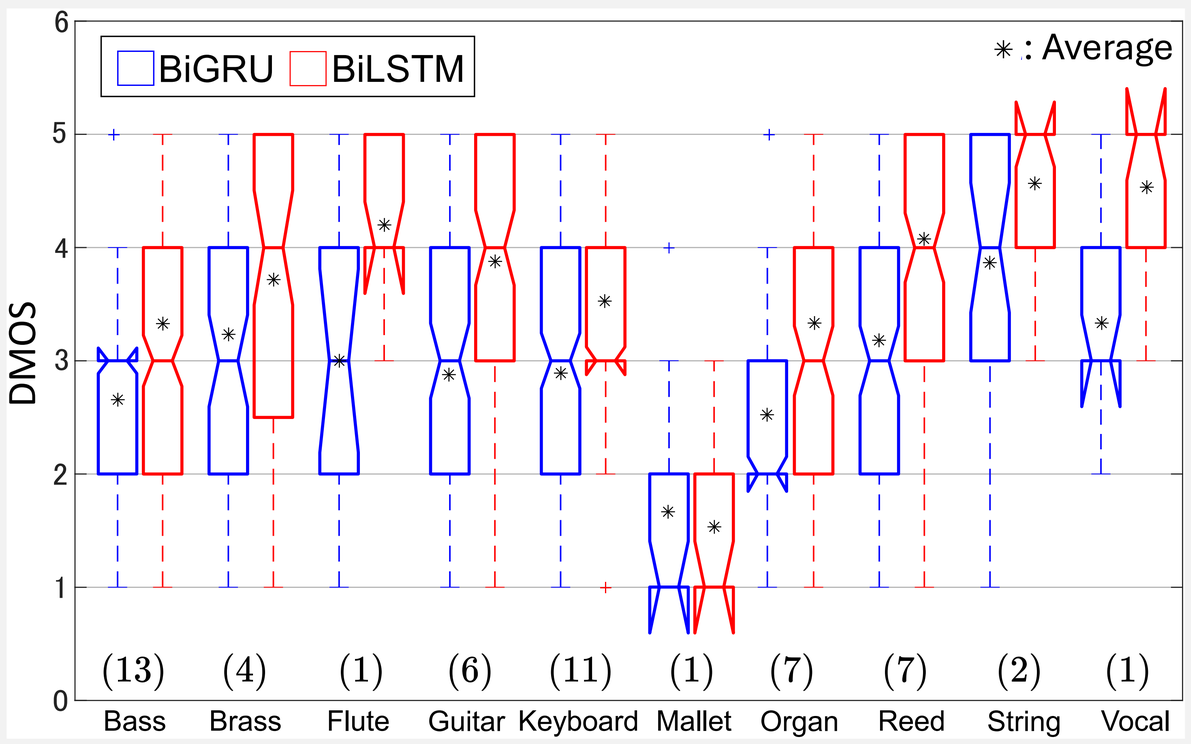
1. インストラクションを読む
2. ユーザーネームを入力
3. 再生ボタンを押下し，音源を視聴
4. 評価を行い，提出
5. 3と4を繰り返す
6. 問題数の半分で休憩

4. 実験

実験条件

・データセット：Nsynthに含まれる１つの音高のみ
　　BiGRU及びBiLSTMの予測結果を対象
・被験者　　：10~40代の男女15名

実験結果



・Malletを除いてBiLSTMの方が評価の平均が高い
・Malletは客観評価と同様にどちらも評価が低い
　- 原因：Malletの特徴的な調波構造

・ノッチが被っていない楽器（Brass, Guitar, Organ, Reed, String, Vocal）
　- 有意水準95%で真の中央値が異なる
　- BiLSTMが予測精度が高い

・主観評価の結果が客観評価の結果と類似
　- 客観評価指標は人間の感性にあった評価が行えている

被験者からのヒアリング

「短音の楽器音は本来の音の後にうなりのようなものがあり，全く異なる音に聞こえた」
→実際にBassに含まれる短い音は評価平均が低い

参考文献

[1] P. Esling, A. Chemla-RomenuSantos, and A. Bitton, “Generative timbre space: regularizing variational autoencoders with perceptual metrics,” *in Proc. DAFX*, 2018.
[2] 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.
[3] S. Kawaguchi and D. Kitamura, “Amplitude spectrogram prediction from mel-frequency cepstrum coefficients using deep neural network,” *Journal of Signal Proceeding*, vol. 27, no. 6, pp. 207－211, 2023.