奏法による音程変化を表現可能な撥弦楽器音合成の検討

Synthesis of plucked string instrument sounds capable of expressing pitch changes depending on playing techniques

廣畑 和音
Hirohata Kazuto

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 本報告では，Deep Neural Network (DNN) を用いたエレクトリックギターにおける音程変化を伴う奏法を表現可能な楽器音合成方式を提案する．提案方式モデルは楽器音合成システムである Deep Performer を踏襲し，奏法および弦の情報を入力部に付与する拡張をおこなう．この拡張によって，奏法による音色や音程の変化の捕捉と合成楽器音の品質向上を試みる．評価実験の結果，提案方式ではメルスペクトログラムの推定精度が向上し，5 段階評価の MOS テストにおいても Deep Performer より高い評価が得られた．

1 はじめに

撥弦楽器であるギターのような生楽器を用いる楽曲制作には，全ての楽器の演奏を収録することが必要であり，一人で楽曲制作を完結させることは困難であった．しかし近年では，楽器音を合成可能なソフトウェア音源を使用することで，楽器演奏が未経験であっても手軽に生音を含む楽曲の制作が可能になっている．これらの音源は信号処理モデル音源，サンプリング音源，物理モデリング音源の 3 つに大きく分類され [1]，高品質な製品が様々なデベロッパから提供されている．一方でギターの音源においては，楽器特有の奏法の表現や自然な楽器音の再現が難しく，柔軟性と品質の点で更なる改善が求められている．

従来の合成方式と異なる手法として，DNN (Deep Neural Network) を利用した楽器音合成方式が提案されている．楽器音合成は一連の音符系列を入力とする音響信号生成と見なすことができる．そのため，音素系列から音声を合成する TTS （Text-to-Speech）の技術を応用した研究が数多くおこなわれており [2]，WaveNet [3] を応用したモデル [4] や，FastSpeech [5] をベースにしたモデル [6] を筆頭に，様々なモデルが提案されている [7–12]．

本報告では，DNN を用いたエレクトリックギターにおける音程変化を伴う奏法を表現可能な楽器音合成方式を検討する．提案方式モデルでは Deep Performer を踏襲し，入力部に弦と奏法のラベルを付与することで，弦による音色の違いや奏法による音程の変化が表現可能なモデルに拡張する．そして，楽譜特徴量をフレームレベルにアップサンプリングした後にエンコーダを適用するようモデルを変更する．この変更によってメルスペクトログラムのフレーム間の関係を捉え，推定精度を向上させることを期待する．

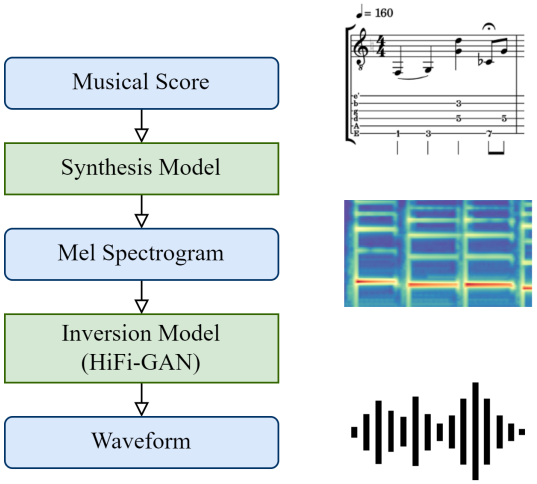

図 1: 提案方式モデルの処理の流れを示す概要図

2 先行研究

楽器音合成システムである Deep Performer [6] は，実際に楽器を演奏した際の時間的なズレを予測する Alignment Model，メルスペクトログラムを推定する Synthesis Model，メルスペクトログラムを波形に変換する Inversion Model の 3 モデルから構成されている．

Synthesis Model は Transformer [13] エンコーダデコーダモデルで構成されており，入力である楽譜特徴量を埋め込んだベクトルがエンコーダに入力される．エンコーダから出力された中間特徴量は Polyphonic Mixer でアップサンプリングされ，最終的にデコーダからメルスペクトログラムが推定される．

また Inversion Model では，ニューラルボコーダである HiFi-GAN [14] を使用している．HiFi-GAN は本来音声合成用のモデルであるが，Deep Performer においては楽器音でモデルを学習させることで楽器音合成に適用している．

3 提案方式

提案方式モデルの概要図を図 1 に示す．基本的な構造は Deep Performer の枠組みを踏襲しており，Alignment Model を除く Synthesis Model と Inver-

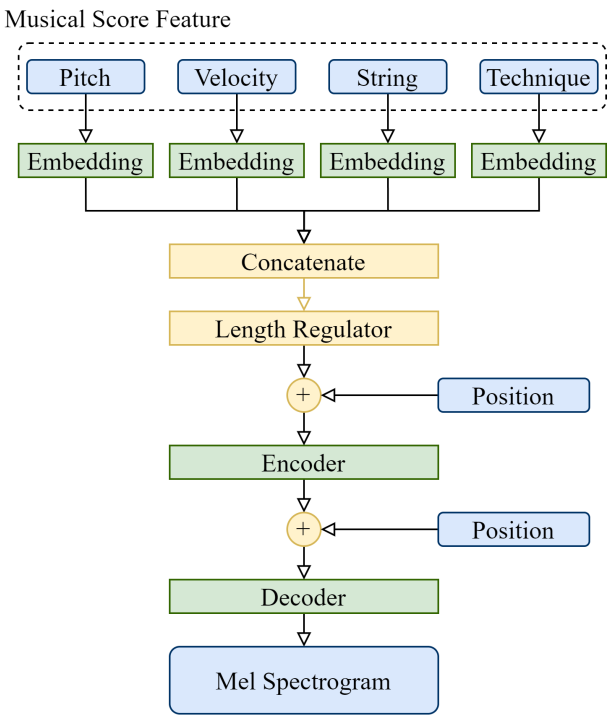

図 2: Synthesis Model のモデル図

表 1: 提案方式モデルにおいて表現可能な奏法

ラベル	奏法	説明
なし	ピッキング	弦をピックではじく
1	ベンド (UP)	音程を上げる
2	ベンド (UP-DOWN)	音程を上げて下げる
3	ビブラート	音程を揺らす

sion Model を連結させる形で構成されている．

次に Synthesis Model のモデル図を図 2 に示す．Deep Performer では，楽譜特徴量として音高，ベロシティ，オンセット，継続長の 4 つの値が使用されているが，提案方式では音高，ベロシティ，そして弦と奏法のラベルを入力として利用する．弦ラベルは 1 弦から 6 弦に対応する 1 から 6 の値が音符ごとに割り振られる．弦ラベルを用いることで，弦の違いによる音色の違いを対応付け，推定精度を向上させることを目的としている．奏法においては表 1 に示す奏法がラベルとして与えられ，各奏法特有の音程の変化を捉えることを期待している．また，Polyphonic Mixer の部分は Length Regulator とし，合成可能な楽器音を単音に限定することでタスクの難易度を落としている．そして，提案方式ではエンコーダと Length Regulator の順序を入れ替え，フレームレベルにアップサンプリングされた特徴量がエンコーダに入力される．前後の音素によって発音に変化が見られる調音結合をもつ音声信号と比べ，楽器音においては前後の音が出音に大きな影響をもたらすことは少ない．そのため，音符レベルよりもフレームレベルでエンコーディングする方が推定精度を高めるためには有効であるとの考えから，このような改良をおこなっている．この部分については 5 章の評価実験で検証する．

また，Inversion Model では，Deep Performer と同様に HiFi-GAN を使用する．学習の際には最初に Synthesis Model を学習させ，重みを固定させてから Inversion Model の学習を開始することで End-to-End での学習を実現している．

4 使用するデータセット

3 章で述べた提案方式モデルでは，楽器音とそれに対応する弦と奏法の両方の情報を含む楽譜データが必要となる．そのため，楽譜データとして Lakh MIDI Dataset [15] を使用し，その楽譜データを自作のサンプリング音源によって合成した楽器音のペアから構成されるデータセットを用意する．Lakh MIDI Dataset 上の楽譜データには弦や奏法の情報は含まれていないため，次の方法で弦と奏法の情報を付与する．

- **弦**：音符系列から運指を推定することで弦の情報を付与．音符の音高から考えられる運指を列挙し，最も指の移動距離が小さい運指から弦を割り当てる．指の移動距離 L はフレット間の距離 L_f と弦間の距離 L_s の積 $(1+L_f)(1+L_s)$ の合計によって計算される．
- **奏法**：四分音符よりも音価の長い MIDI ノートに対して，表 1 の奏法の中からランダムに選択．

サンプリング音源は，表 1 に記載している奏法の単音のサンプルを全ての弦の全フレットにおいて録音し，楽譜データを基にサンプルを繋ぎ合わせることで音を合成する．約 5 秒のフレーズが約 12 万個含まれており，合計約 180 時間分のデータセットとなっている．

サンプルの収録では，ギターに Fender Player Telecaster[*1]，オーディオインターフェースに Native Instruments Komplete Audio 2[*2]を使用し，サンプリング周波数 48 kHz，量子化ビット数 24 bit で記録した．

5 評価実験

5.1 客観評価実験

奏法ラベルによって奏法特有の音色や音程の変化を制御可能か評価するため，学習に含めていない音についてメルスペクトログラムを推定する．推定したメルスペクトログラムの概観を Ground Truth と比較し，音程の変化を表現可能であるか確認する．

表 1 の奏法のラベルを付与してメルスペクトログラムを推定した結果を図 3 に示す．(a) がベンド (UP),

[*1] https://www.fender.com/ja-JP/electric-guitars/telecaster/player-telecaster/0145212515.html
[*2] https://www.native-instruments.com/jp/products/komplete/audio-interfaces/komplete-audio-1-audio-2/

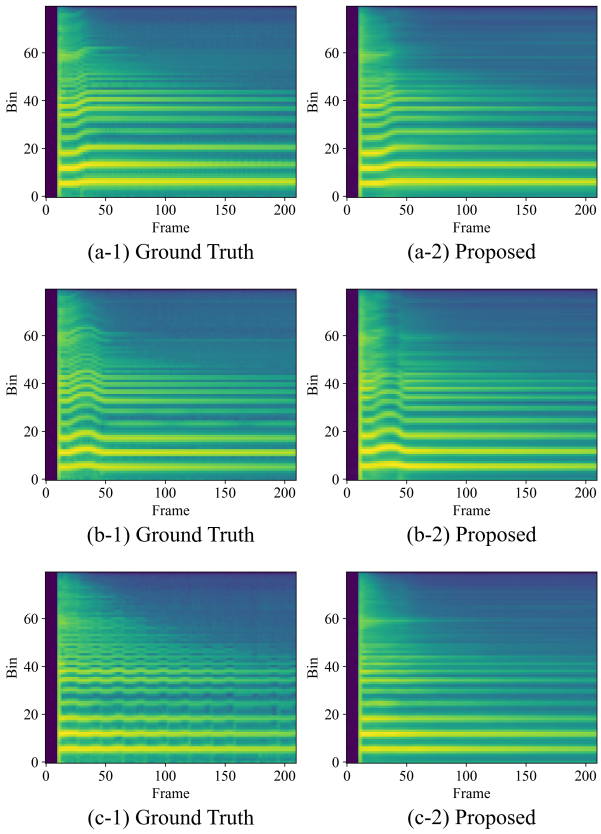

図 3: 学習外の音において奏法ラベルを付与してメルスペクトログラムを推定した結果

(b) がベンド (UP-DOWN)，(c) がビブラートとなっており，左図が Ground Truth，右図が提案方式によって推定したメルスペクトログラムである．図 3 (a)，(b) より，ベンド (UP) とベンド (UP-DOWN) では提案方式で Ground Truth のメルスペクトログラムと同様の音程の変化が確認される．これらの奏法は音程の変化量とタイミングがある程度共通であるため，ラベルのみによって十分奏法の特徴が表現できている．一方，図 3 (c) のビブラートでは，提案方式において音程の揺れが再現できていない．これはビブラートによる揺れの周期が学習データ内で一定でないためであると考えられる．このような音程の揺れを再現するには，メルスペクトログラムを基本周波数と対応付けて学習をおこなう必要がある．また合成の際には MIDI のピッチベンドの機能を用いることで，利用者の意図通りに音程の変化を制御することも可能であると考える．

次に，Synthesis Model における変更点の有効性を評価するため，推定したメルスペクトログラムの二乗平均平方根誤差 (RMSE) を比較する．以下の 3 つのモデルで比較をおこなう．

1. **Deep Performer**
2. **Proposed**
3. **Proposed (Encoder → LR)** : エンコーダの次に Length Regulator を配置したモデル

表 2: 推定されたメルスペクトログラムの RMSE

	RMSE (dB)
Deep Performer	2.649
Proposed	0.405
Proposed (Encoder → LR)	0.510

表 3: MOS テストの結果 (評価 ± 標準誤差)

	生楽器音の再現性	奏法の再現性
Ground Truth	4.13 ± 0.098	4.41 ± 0.071
Session Guitarist	3.89 ± 0.087	4.21 ± 0.078
Deep Performer	1.83 ± 0.078	1.76 ± 0.070
Proposed	2.95 ± 0.085	3.01 ± 0.092

20 個のフレーズについて RMSE を計算し，その平均をとった結果を表 2 に示す．Deep Performer と Proposed を比較すると，入力部の改良によって大幅に推定誤差が減少していることが分かる．これは弦及び奏法ラベルを付与することで，入力と出力の対応付けが明確になり推定精度が向上したと考えられる．また Proposed と Proposed (Encoder → LR) を比較すると，フレームレベルでエンコーダに入力する Proposed の形式が楽器音のメルスペクトログラム推定において有効であることが示されている．

5.2 主観評価実験

合成楽器音の品質評価として，以下の評価項目で 5 段階評価の MOS テストを実施する．

- 評価項目 A：生楽器音の再現度
- 評価項目 B：奏法の再現度

比較対象は，Ground Truth，市販のサンプリング音源[*3] (Session Guitarist)，Deep Performer，提案方式 (Proposed) の 4 つである．各音源で合成した楽器音を聴いて「非常に悪い」「悪い」「普通」「良い」「非常に良い」を 1 点から 5 点とした 5 段階の評価をしてもらい，各モデルごとの平均値 (Mean Opinion Score:MOS) をスコアとする MOS テストをおこなった．なお，本実験の参加者は 10 名である．

実験結果を表 3 に示す．どちらの評価項目においても，提案方式が Deep Performer を上回っているが，Ground Truth と比べると低い評価となった．

Ground Truth よりも評価の低くなった理由として，HiFi-GAN が楽器音合成向けでないことが挙げられる．

[*3] https://www.native-instruments.com/jp/products/komplete/guitar/session-guitarist-electric-sunburst-deluxe/

楽器音合成のために HiFi-GAN を用いる場合，次のような問題が生じる．

- 継続長の大きい音の自然性の低下
- 高域や低域での自然性の低下

この問題の解決には，歌声合成のために提案されたニューラルボコーダ [16] を楽器音合成に応用することが改善策として考えられる．

6　まとめと今後の課題

本報告では，弦と奏法の情報を用いることで，撥弦楽器特有のアーティキュレーションを表現可能な楽器音合成方式を検討した．提案方式として，メルスペクトログラムを推定する Synthesis Model とメルスペクトログラムを楽器音信号に変換する Inversion Model を統合したモデルを提案した．この方式では，Synthesis Model において弦と奏法の情報をラベルとして付与することで，入出力の明確な対応付けによる推定精度の向上と楽器音合成方式としての表現力の拡張を試みた．そして，メルスペクトログラム推定精度向上のため，フレームレベルにアップサンプリングした楽譜特徴量をエンコーダへの入力とする改良をおこなった．また，評価実験として客観評価実験および主観評価実験を実施した．客観評価実験ではメルスペクトログラムの概観および推定誤差の比較をおこない，推定精度の面で提案方式が Deep Performer を上回った．そして単純な奏法であれば，学習外の音であっても奏法ラベルによって音程の変化を制御が可能であることが確認された．主観評価実験では，生楽器音の再現性と奏法の再現性の 2 つを評価項目とした 5 段階評価の MOS テストを実施した．実験の結果，両方の項目において提案方式が Deep Performer を上回り，合成楽器音の品質が向上したことが示唆された．一方で，Ground Truth や市販の音源の品質には及ばないという結果となった．

今後の課題として，Inversion Model の推定精度の向上が挙げられる．最終的な楽器音信号が出力される部分の品質を向上させることが合成楽器音の高品質化に直結するため，今後は 5 章で述べた歌声合成用のニューラルボコーダモデルの調査をおこなっていく予定である．

参考文献

[1] 鮫島俊哉, “楽器音の合成における基盤技術 ——物理モデルによる楽器音の合成—— (小特集—話す・歌う・奏でる音の合成技術—),” 日本音響学会誌, Vol. 75, No.7, pp. 412–418, 2019.

[2] S. Ji et al., “A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions,” ACM Computing Surveys, Volume. 56, No. 7, pp. 1–39, 2021.

[3] A. Oord et al., “WaveNet: A Generative Model for Raw Audio,” SSW, 2016.

[4] J. Engel et al., “Neural audio synthesis of musical notes with WaveNet autoencoders,” ICML, Vol. 70 pp. 1068–1077, 2017.

[5] Y. Ren et al., “FastSpeech: Fast, Robust and Controllable Text to Speech,” NeurIPS, vol. 32, pp. 3171–3180, 2019.

[6] H. Dong et al., “Deep Performer: Score-to-Audio Music Performance Synthesis,” ICASSP, pp. 951–955, 2022.

[7] A. Défossez et al., “SING: Symbol-to-instrument neural generator,” NeurIPS, Vol. 31, 2018.

[8] P. Esling et al., “Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces,” ISMIR, pp. 175–181, 2018.

[9] N. Jonason et al., “The Control-Synthesis Approach for making Expressive and Controllable Neural Music Synthesizers,” Conference on AI Music Creativity, 2020.

[10] J. Engel et al., “DDSP: Differentiable digital signal processing,” ICLR, 2020.

[11] N. Jonason et al., “DDSP-based Neural Waveform Synthesis of Polyphonic Guitar Performance from String-wise MIDI Input,” arXiv preprint, arXiv:2309.07658, 2023.

[12] J. Koguchi and M. Morise, “Neural electric bass guitar synthesis framework enabling attack-sustain-representation-based technique control,” EURASIP Journal on Audio, Speech, and Music Processing, 2024.

[13] A. Vaswani et al., “Attention is all you need,” NeurIPS, Vol. 31, pp. 5998–6008, 2017.

[14] J. Kong et al., “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” Advances in Neural Information Processing Systems(NeurIPS), Vol. 34, pp. 17022–17033, 2020.

[15] C. Raffel, “Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching,” PhD Thesis, Columbia University, 2016.

[16] R. Yoneyama et al.,, “Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder,” ICASSP, 2023.