CE-008

RAVE を用いた稀少楽器音色を利用するための二段式楽器音色変換法

Two-stage Instrument Timbre Transfer Method for Utilizing Rare Instrument Timbres Using RAVE

胡棣 † 伊藤克亘 ‡
HU DI† Katsunobu Ito‡

1 はじめに

楽器音色変換の目標は、ある楽器によって生成された音楽作品の音を、別の楽器による演奏のように変換することを目指すことである。このプロセスでは、旋律やダイナミクスなど、音楽に関連する他の重要な特性をできる限り維持することが求められる。つまり、音色が変わっても、元の楽曲のニュアンスや表現が損なわれないようにすることが重要である。

したがって、任意の入力音色を目標音色に変換し、他の情報を保持するためには、目標音色だけを学習してモデルを構築する方法が必要である。

しかし、変換対象となる音色が希少な楽器の場合、その楽器の演奏録音が限られているため、訓練データが不足することがある。このような場合、十分な訓練データが得られないため、訓練された音色変換モデルの効果が不十分である可能性が高くなる。この課題に対処するためには、より少ないデータでも高い性能を発揮できる手法が求められる。

本論文では、二段階の楽器音色変換手法に焦点を当てている。この手法の主な利点は、トレーニングデータが不足している状況でも、目標音色への変換を二段階に分けることで、より効果的な変換が可能となる点である。具体的には、まず初めに中間過渡音色への変換を行い、その後に最終的な目標音色への変換を行うことで、データ不足の問題を緩和し、変換の精度を向上させる。

さらに、本研究では、音色変換の効果と楽器自体の特性、そして楽器音色の差異との関係にも焦点を当てている。具体的には、音色変換の効果を最大化するために、中間過渡音色の選択方法についても研究を行っている。これにより、より自然で効果的な音色変換を実現するための新しいアプローチを提案している。

† 法政大学大学院 情報科学研究科, Hosei University Graduate School of Computer and Information
‡ 法政大学 情報科学部, Hosei University Faculty of Computer and Information

2 関連研究

文献にはいくつかの種類の音色変換技術が提案されている。[6] では CycleGAN が適用され、Constant-Q 変換を使用して対応のない転送を行い、音声は WaveNet モデルを通じて復元される。

[2] では既存の音楽の音色変換モデルと比較して、異なる楽器間で多対多の音色転送を実現することができる。提案された方法は、2 つの事前学習されたエンコーダと 1 つの非監督学習されたデコーダから成るオートエンコーダフレームワークに基づいている。

RAVE[1] と DDSP[4] といった技術は、目標とする音色の音声データのみを入力として訓練することができる。これにより、平行なデータセットを必要とせず、目標音色のデータのみを使用してモデルを訓練することが可能となる。これは、データ収集の手間を大幅に軽減し、モデルの訓練プロセスを効率化する上で非常に有益である。

3 提案手法

まず、目標となる音色が設定され、最終的な目標は他の楽器の音色をこの目標音色に変換することである。

第一段階：目標音色の変換モデルのトレーニング:

最初に、目標音色の音声データを使用して、RAVE（Realtime Audio Variational autoEncoder）で目標音色の変換モデルをトレーニングする。このモデルの役割は、任意の他の楽器音色の入力を目標音色に変換することである。しかし、目標音色の訓練データが不足している場合、出力の音質が安定しない可能性がある。特に、変換後の音色が一貫性を欠くことや、音質が低下するリスクがある。

第二段階：中間過渡音色の選定とモデルのトレーニング:

次に、目標音色と高品質なトレーニングデータを持つ他の音色との類似性を比較する。このプロセスでは、質の高いトレーニングデータを持ついくつかの音色を選択し、これらを中間過渡音色と定義する。中間過渡音色は、目標音色と似た特徴を持ちながら、

より安定した変換を可能にするものである。

中間過渡音色の音声データを使用して、RAVE で中間過渡音色の変換モデルをトレーニングする。このモデルの役割は、任意の他の楽器音色の入力を安定した中間過渡音色に変換することである。これにより、最初の段階で見られた音質の不安定さを軽減し、より一貫性のある変換結果が得られる。この過程は、図 1 に示されている。

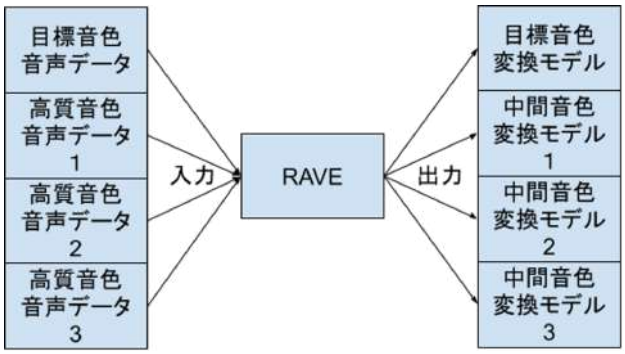


図 1: 目標音色と中間音色モデルのトレーニング

最終段階：二段階変換プロセス:

最後に、他の楽器の演奏音を中間過渡音色の変換モデルに入力し、中間過渡音色の音声を出力する。この出力された中間過渡音色の音声を、次に目標音色の変換モデルに入力することで、最終的な目標音色の音声を生成する。この二段階変換プロセスにより、最初の変換段階で安定性を確保し、最終段階で目標音色に近づけることができる。

生成された出力音の中から、最も目標音色に近いものを選択する。この手法により、音色の質と変換の精度が向上し、より自然で目標に近い音色変換が可能となる。提案手法の全体的な流れは図 2 に示されている。

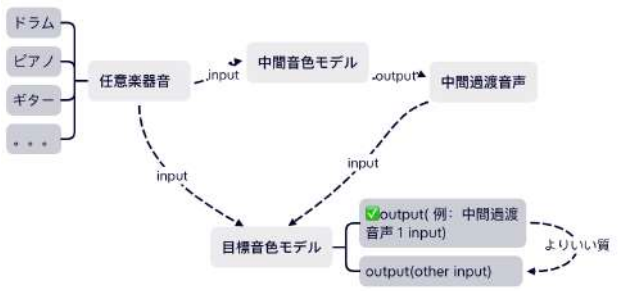


図 2: 提案手法の流れ

3.1 RAVE での音色変換モデルをトレーニング

RAVE は、高速で高品質なニューラル・オーディオ合成のための変分オートエンコーダである。論文 [1] では、RAVE のトレーニングプロセスは 2 つのステージ、すなわち表現学習と対抗的微調整に分かれている。この手法では、平行なデータセットではなく、目標音色の音声を直接入力して音色変換モデルをトレーニングすることができる。トレーニングが完了した後、モデルの出力は torch script ファイルとして保存することができる。

3.2 中間過渡音色と変換元音色、目標音色の関係

現在のアプローチは、複数の中間過渡音色変換モデルをトレーニングし、最終的に最も変換効果の高いものを選択するものである。しかし、音色の類似性と変換効果の関係はまだ完全には明らかになっていない。そのため、複数の中間音色変換モデルをトレーニングする際には、このトピックも同時に検討する必要がある。

[5] によると、FFT 音響記述子と機械学習手法を用いることで、音色の類似性と変換効果の関係を調査することが可能である。このアプローチの考え方は、目標音色とソース音色データに中間音色データを追加し、それらをクラスター分析によって評価するというものである。具体的には、追加された中間音色が目標音色および変換元音色とどれほど類似しているかを評価し、その類似性が変換効果にどのように影響するかを明らかにする。

この方法では、まず目標音色とソース音色データを準備し、それに対して中間音色データを追加する。次に、クラスター分析を使用して、これらの追加された音色と目標音色および変換元音色の間の類似性を評価する。クラスター分析により、音色の類似性が高いグループと低いグループに分類される。この類似性の評価に基づいて、どの中間音色変換モデルが最も効果的であるかを判断することができる。

このアプローチを通じて、音色の類似性と変換効果の関係を詳細に理解することができ、音色変換モデルの性能向上に寄与することが期待される。特に、FFT 音響記述子と機械学習手法を組み合わせることで、音色の微細な特徴を捉え、より精度の高い変換モデルの開発が可能となる。

4 評価手法と実験

音色変換後の音声の品質評価には、客観的および主観的な 2 つの側面から評価する必要がある。

ご予定の指標は以下の通りである。

4.1 客観的な評価

4.1.1 Fréchet Audio Distance (FAD)

Fréchet Audio Distance (FAD)[7] は、音声データの評価において有用な指標であり、真のデータと生成されたデータの間の類似性を測るために使用される。この指標は、元のデータセットに含まれる実際のオーディオと、生成されたオーディオとの間で計算される。FAD は、主に知覚的な類似性を評価するために設計されており、元の音声データと生成された音声データの分布の差異を測定する。

具体的には、FAD は、2 つのデータセット間の

Fréchet 距離を計算することで、音声の知覚的な品質や特徴の一致度を評価する。この距離が小さいほど、生成された音声データが元の音声データに近いことを示し、高い品質の音声生成が達成されていることを意味する。したがって、FAD は音声生成モデルの性能評価において重要な役割を果たし、生成されたオーディオの品質向上に向けた指標として利用される。

4.1.2 Jaccard Distance (JD)

Jaccard Distance (JD) は、音声生成モデルのコンテンツ保持能力を評価するための指標であり、特に生成された音声トラックのピッチコンターの一致度を測るために使用される。この手法は [3] に基づいており、ピッチの 2 つのセット A と B の不一致を Jaccard 距離を用いて評価する。

具体的には、Jaccard 距離は、2 つのセット間の交差部分と結合部分の比率を計算することで測定される。この値が小さいほど、生成された音声のピッチが元の音声のピッチに近いことを示し、コンテンツの保持能力が高いことを意味する。ピッチコンターの一致度が高いということは、生成された音声が元の音楽的内容や意図を忠実に再現していることを示すため、音声生成モデルの重要な評価基準となる。

JD を使用することで、生成された音声の具体的な音高の変化やメロディの一致度を詳細に分析することが可能となり、生成モデルの改善点や課題を明確にすることができる。このようにして、音声生成モデルのコンテンツ保持能力を精緻に評価することができる。

4.2 主観的な評価

音色変換の能力を評価するために、参加者による聴取テストが行われた。テストの条件および各パートの例の順序はランダム化され、参加者には公正な評価を促すために、特定の順序に依存しない形式が採用された。

具体的な評価基準としては、音色変換の能力を三つの主要な側面から評価することとした。まず第一に、音色還原の程度である。これは、変換後の音色がどれほど参照トラックに似ているかを評価するもので、参加者には原音と変換後の音を聴き比べてもらった。第二に、音質の好悪である。音質は、音色変換の技術が音のクリアさや豊かさにどのように影響を与えるかを測るものであり、参加者にはその音の品質についても評価してもらった。最後に、内容保持能力である。これは、音色変換が元の音楽的内容や意図をどれだけ忠実に保っているかを示すもので、参加者には音楽的な内容がどれほど元の意図通りであるかを確認してもらった。

これらの評価は、5 つの要素から成る Likert スケールを用いて行われ、参加者は各条件について参照トラックとの類似性を基準に評価を行った。このスケールを用いることで、評価が定量的に行われ、分析が容易となった。また、ランダム化された条件と例の順序により、評価のバイアスを最小限に抑えることができた。このテストの結果により、音色変換技術の現状の評価と、今後の改良点を明確にすることができる。

結果は全テストコースでの平均値。

4.3 現段階の実験: 二段式変換有効性検証

直接的に稀少な楽器音色変換モデルを使用して変換されたオーディオと、二段階式変換法を使用して変換されたオーディオを比較し、2 段階の変換法が音色変換の効果を向上させるかどうかを検証する。

現在の段階では、実験は以下の通りである。図 3、図 4、図 5 はそれぞれ変換元オーディオのスペクトログラム、音色変換モデルを直接使用した後のオーディオのスペクトログラム、および二段式音色変換法で変換した後のオーディオのスペクトログラムに対応している。

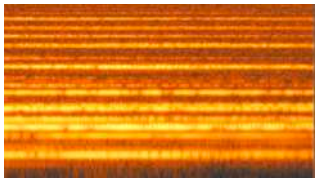


図 3: 変換元オーディオのスペクトログラム

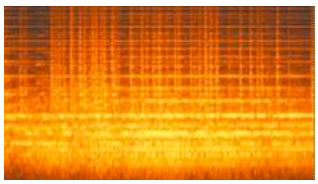


図 4: 変換後のスペクトログラム

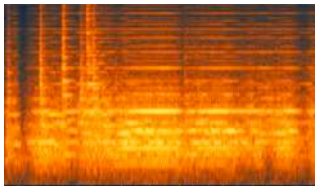


図 5: 変換後のスペクトログラム（二段式音色変換法による）

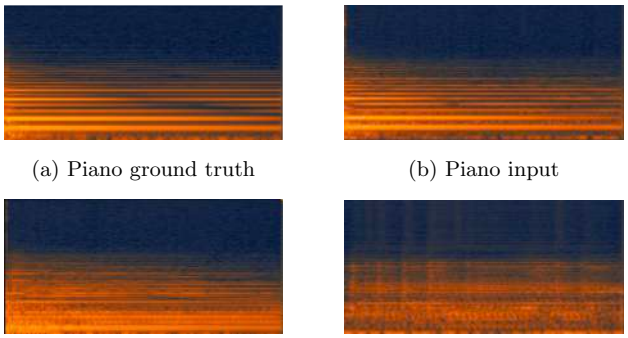
実験結果から見ると、直接変換と比較して、二段階の音色変換を使用すると、一部の音声のノイズが減少し、音声が安定する傾向があるが、同時に一部のスペクトルの欠落も引き起こすことがあった。

4.4 現段階の実験: 単音の音色変換

少ない訓練データ量で構築された音色変換モデルの場合、入力音色と変換目標音色の差異が変換結果に影響を及ぼす。音色の差異が変換効果にどのよう

な関係を持つかを研究するために、単音色変換実験を行った。

現在、2 時間ほどのピアノ音声を用いてピアノ音色変換モデルを訓練し、アコースティックギター、エレクトリックギター、ピアノの単音をこのモデルに入力して、出力結果を観察した。結果は図 6 に示す通りである。この実験では、異なる楽器の音色をピアノ音色に変換する過程を通じて、モデルの性能を評価した。各楽器音の入力に対する変換結果を詳細に分析することで、音色の違いが変換効果に及ぼす影響を明らかにしようとした。



(a) Piano ground truth　(b) Piano input

(c) Guitar input　(d) eGuitar input

図 6: Comparison of Spectrograms

同じ音高と音量であっても、楽器の音色の違いが楽器音色変換の効果に明らかな影響を与えることがわかりる。もし、目標とする音色に近い中間楽器モデルが存在し、さらに入力楽器の音色を高品質な中間楽器音色に変換できる十分な性能を持っている場合、楽器音色変換の効果は大幅に向上する。これにより、異なる楽器間での音色変換がよりスムーズかつ正確に行われることが期待される。

5 おわり

現在、二段階の音色変換法の一部の有効性を証明した。一部のスペクトルの欠落問題や、異なる楽器音色が変換結果に与える影響を解決するために、最適な中間音色を選択する方法に焦点を当てて研究を進めている。今後、WAE（Wasserstein オートエンコーダー）を用いた複数の音色変換モデルをトレーニングし、実験をさらに充実させる予定である。

5.1 WAE

WAE（Wasserstein オートエンコーダー）[8] は、潜在空間の分布をより正確に再現することができる技術である。Wasserstein 距離の最小化により、潜在変数がより理想的な分布に近づくため、生成されるデータの多様性や質が向上する。これに対して、従来の VAE（変分オートエンコーダー）は、通常、事前分布（通常はガウス分布）と潜在分布の間の KL ダイバージェンスを最小化するが、これにより生成されるデータの質や多様性が制限されることがある。

5.2 実験：楽器音色差異との関係

同じ音高で異なる音色の楽器のシングルノートを稀少楽器音色変換モデルに入力し、出力を元の音色と比較した。この比較を通じて、変換効果と楽器間の音色の違いの関係を分析した。さらに、音高を変えて実験を繰り返し、音高が変換結果に与える影響も検討した。

5.3 実験：音色クラスタリング法による中間音色の選択の有効性検証

音色クラスタリング法で選択された中間音色が、二段階の音色変換法を使用した実験で最も優れた変換効果を示すかどうかを検証した。この検証により、音色クラスタリング法によって選択された中間音色の有効性を評価し、最適な中間音色の選択方法を確立することを目指している。

参考文献

[1] Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021.

[2] Yu-Chen Chang, Wen-Cheng Chen, and Min-Chun Hu. Semi-supervised many-to-many music timbre transfer. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 442–446, 2021.

[3] Ondřej Cífka, Alexey Ozerov, Umut Şimşekli, and Gael Richard. Self-supervised vq-vae for one-shot music style transfer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 96–100. IEEE, 2021.

[4] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. In *International Conference on Learning Representations*, 2020.

[5] Yubiry Gonzalez and Ronaldo C Prati. Similarity of musical timbres using fft-acoustic descriptor analysis and machine learning. *Eng*, 4(1):555–568, 2023.

[6] Sicong Huang, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, and Roger B Grosse. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620*, 2018.

[7] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, pages 2350–2354, 2019.

[8] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.