RAVE と DDSP を用いた二段式楽器音色変換法
Two-stage Instrument Timbre Transfer Method Using RAVE and DDSP

胡棣 (HU DI)*

法政大学 情報科研究科 情報科学専攻
di.hu.8x@stu.hosei.ac.jp

Abstract

Recently, the Real-time Audio Variational autoEncoder (RAVE) method was developed for high-quality audio waveform synthesis. The RAVE method is based on a Variational AutoEncoder (VAE) and employs a two-stage training strategy. However, the RAVE model still has limitations in timbre transformation, especially when converting between instruments with significantly different timbres. Issues such as pitch instability, inaccurate timbre reproduction, and severe degradation in sound quality can arise. To enhance timbre transfer performance, we propose a two-stage timbre transfer method using RAVE and Differentiable Digital Signal Processing(DDSP), which involves applying two timbre transfer models to perform a dual transformation on the original input audio. To evaluate the proposed method, we trained the model and tested its performance using audio generated from MIDI and SoundFont2 sound sources. The results demonstrate that the proposed method improves the timbre transfer compared to the single-stage RAVE model.

1 はじめに

近年、人間の声における音色変換の分野で大きな進展が見られた。しかし、音色の違いが大きい場合、変換結果は満足のいくものではなく、特に異なる楽器間の音色変換において顕著である。楽器音色変換の分野には、依然として解決すべき多くの課題が存在する。

楽器音色変換の目的は、ある楽器が生み出す音楽の音色を別の楽器の音色に変換することである。この際、メロディやダイナミクスなどの他の重要な音楽的特徴を可能な限り保持する必要がある。入力楽器の音色をターゲット楽器の音色に変換する際には、変換された音色がターゲット楽器の特徴を正確に反映しつつ、音楽の品質やその他の内容を維持することが重要である。

複数の入力音色をターゲット音色に変換しながら、その他の情報を保持するためには、ターゲット音色のみを学習し、モデルを構築する手法が求められる。本論文では、単一楽器の音声を学習し、音色変換モデルを構築するための主な方法として、オートエンコーダ（Autoencoder）を用いた教師なし学習アプローチを採用する。

音色の差異が大きい楽器においては、変換された音声に多くのノイズが含まれる場合がある。この問題を解決するために、本論文では 2 段階の楽器音色変換法を提案する。まず、入力音声のメロディなどの音楽的内容を保持する能力に優れた DDSP を用いて、入力楽器の音色をターゲット楽器の音色に近い中間楽器音色に変換する。次に、RAVE を用いて、中間音色を最終的なターゲット音色に変換する。

直接的な楽器音色変換と比較して、本手法は音色特性が大きく異なる楽器間でも効果的に利用できる汎用性を備えている。また、変換された音色の忠実度をより正確に保持し、その他の重要な音楽情報を維持することが可能である。

本論文の構成は以下の通りである。第二章では関連研究を概説する。第三章では提案手法を提示する。第四章では実験の設定と結果を詳述する。第五章では本論文の結論と今後の研究への提言を述べる。

2 関連研究

2.1 音色変換

近年、ディープラーニングモデルが徐々に音色変換タスクに応用されるようになっている。特定の楽器ペア間の音色変換においては、教師あり学習法、特に生成的対向ネットワーク（Generative Adversarial Networks, GAN）が一般的に使用されている。例えば、Héctor Martel は Pix2Pix アーキテクチャを利用し、対向ネットワークを通じて楽器音スペクトラム間の変換関係を学習し、楽器音色変換を実現した。しかし、この手法は一対一の楽器音色変換に限定されており、多対多の音色変換のニーズを満たすことができない。また、この方法でモデルを訓練するためには、異なる音色を持つ複数の楽器が同一の音楽内容を演奏したデータセットが必要であるが、これは一部の音色では実現が困難である。このため、ターゲット音色のオーディオデータセットのみを必要とする教師なし学習法に注目が集まっている。

* 指導教員：伊藤克亘 教授

教師なし学習において、変分オートエンコーダー（Variational Autoencoder, VAE）[9] は、その優れた性能から広く注目を集めており、特に音色変換タスクにおいて重要である。Tatar らは、音色潜在合成に関する研究 [12] において、VAE の潜在空間の特性が高品質な音色合成と変換を実現する方法を探求し、音声構造を損なうことなく音色変換が可能であることを示した。生成モデルとしての VAE は、複雑な音声データ分布を学習し、新しいサンプルを生成する能力を持ち、音色変換において重要な役割を果たす。その主な利点は、連続的な潜在空間を学習する能力にあり、異なる音色間を滑らかに遷移させることが可能である。この連続性により、音色変換がより自然になり、元の音声のメロディやリズムなどの重要な特徴を保持しながら、音色を成功裏に変更できる。

さらに、VAE の教師なし学習能力は、大量のアノテーションされていない音声データから価値ある音色特徴を抽出することを可能にし、アノテーションデータが不足している場合や取得コストが高い場合に特に重要である。Bonnici らは、VAE と Cycle-Consistent GAN（CycleGAN）を組み合わせたモデルを提案し、音色変換の安定性と忠実性を向上させた [3]。この手法では、CycleGAN を利用して音声特徴の一貫性を維持しつつ、VAE の潜在空間の特性を活用して柔軟な音色変換を実現している。

また、ベクトル量子化変分オートエンコーダー（Vector-Quantized Variational Autoencoder, VQ-VAE）[6] のような技術を取り入れることで、VAE は音色変換の品質と安定性をさらに向上させることができる。VAE のエンコードおよびデコードプロセスにおける優れた再構成能力により、音色が変換される間も元の音声の他の音楽的特徴を最大限に保持することが可能である。その生成能力、連続的潜在空間、教師なし学習能力、および拡張性により、VAE は音色変換タスクにおける理想的なツールとなり、高品質な音色変換を実現するための堅固な基盤を築いている。

2.2 RAVE モデル

VAE を拡張したもう一つの手法として、リアルタイム音声変分オートコーダー（Real-Time Audio Variational Autocoder, RAVE）[4] が挙げられる。これは高品質な音声合成を目的に開発されたものである。既存研究の限界に対応し、より良い合成品質を得るために、RAVE は二段階の学習プロセスを導入している。すなわち、表現学習フェーズと敵対的ファインチューニングフェーズである。

表現学習フェーズにおいては、RAVE は VAE モジュール内のエンコーダーとデコーダーネットワークを学習する。その後、敵対的ファインチューニングフェーズにおいてデコーダーをファインチューニングする。RAVE トレーニングの模式図は図 1 に示されている。

2.3 DDSP モデル

DDSP（Differentiable Digital Signal Processing）[5] は、従来のデジタル信号処理（DSP）の手法を微分可能な形式で再構築したフレームワークであり、深層学習と DSP の利点を統合することを目的としている。DDSP は、音楽音声信号の特徴的な物理的構造を明示的にモデル化し、機械学習モデルがこれらの構造を利用できるようにすることで、音声生成や変換タスクの品質と効率を大幅に向上させる。

従来の音色変換手法の多くは、スペクトル領域での信号表現に依存しており、信号の時間的・周波数的特徴を捉える際に限界がある。これに対して、DDSP は信号生成の過程を微分可能な形でモデル化することで、物理的特徴や信号構造を直接利用する音色変換を実現する。このアプローチは、信号の基本周波数（F0）や振幅包絡といった高次特徴量を利用するため、音色変換の精度が向上し、よりリアルで自然な変換が可能になる。

DDSP を用いた音色変換の代表的な手法は、楽器音声における音色モデリングである。楽器音声は通常、ハーモニクス構造やノイズ成分を含む複雑な信号で構成される。DDSP は、ハーモニクス合成モデルや残差ノイズモデルを組み合わせることで、これらの特徴を明示的に表現し、ターゲット音色への変換を実現する。

さらに、DDSP は、事前学習モデルを利用することで、教師なし学習環境でも有効である。これにより、注釈付きデータが不足している場合でも、ターゲット音色の学習と変換を可能にする。また、DDSP は微分可能な特性を持つため、深層学習フレームワークと統合が容易であり、音色変換タスクにおけるエンドツーエンド学習をサポートする。

このように、DDSP は伝統的な DSP の物理的洞察と深層学習の柔軟性を組み合わせることで、音色変換タスクに新たな可能性をもたらしている。その高い再現性と効率性から、DDSP は今後、音楽生成や音声処理分野でさらに広範に応用されることが期待される。

2.4 まとめ

楽器音色変換の分野では大きな進展が見られる一方で、いまだに未解決の課題が残っている。例えば、音色変換の過程で旋律やダイナミクスといった他の音楽的特徴をどのようにより良く保持するか、また変換された音色の品質をどのようにさらに向上させるかといった問題である。これらの課題は、今後の研究の方向性を示すものである。

総括すると、既存の手法は音声および楽器音色変換においてかなりの進展を遂げたものの、大きな音色差を扱う能力や音楽コンテンツの品質を維持する点で制約がある。本論文では、DDSP と RAVE を組み合わせた新しい 2 段階アプローチを提案し、これらの制約を克服し、より正確で高品質な音色変換を実現する方法を示す。

3 提案手法

本論文では、二段階の楽器音色変換手法を提案する。まず、目標楽器の音色を設定し、最終的に入力楽器の音色を目標楽器の音色に変換することを目的とする。

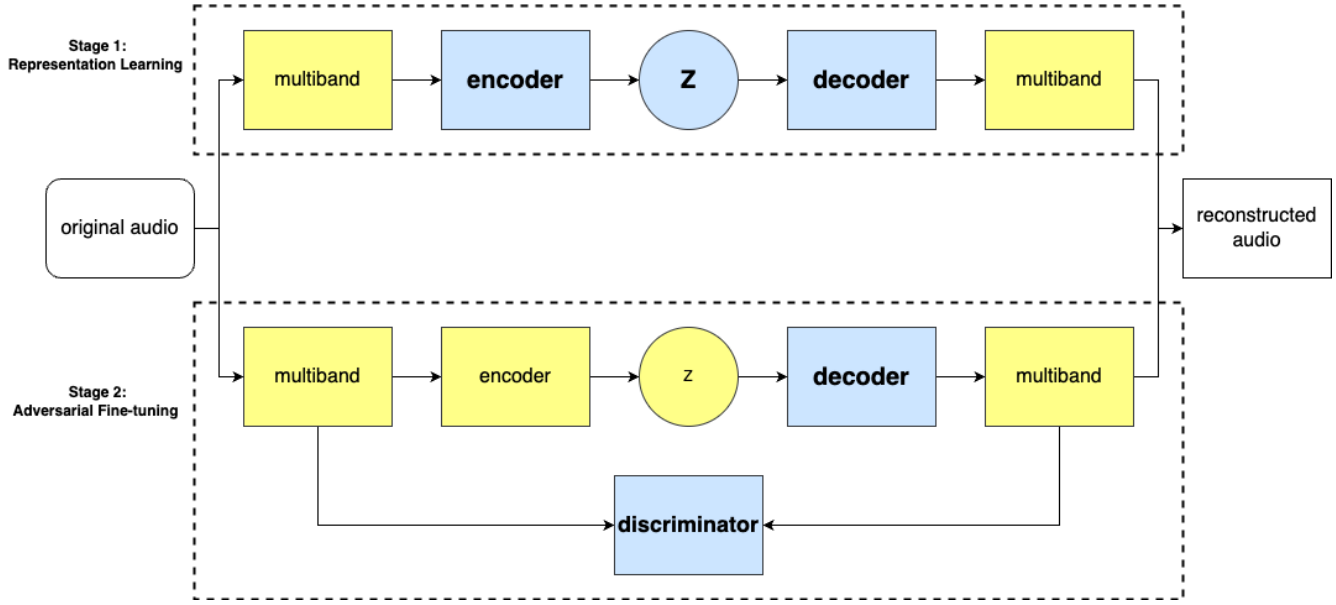


図 1: RAVE モデルのネットワーク構造。青色のブロックは各トレーニングステップで最適化され、黄色のブロックは固定またはフリーズされる。構造の詳細については、[4] を参照する

3.1 手法の概要

提案する二段階の音色変換プロセスを図 2 に示す。

第一段階：中間音色の選定と中間音色モデルの学習次に、目標音色との差が小さい中間楽器の音色を選定する。そして、中間音色の音声データを用いて、DDSP モデルで中間音色モデルを学習する。このモデルの役割は、入力楽器の音色を目標音色に近い安定した中間音色に変換することである。

第二段階：目標音色変換モデルの学習まず、目標音色の音声データを用いて、RAVE モデルで目標音色変換モデルを学習する。このモデルの役割は、任意の入力音色を目標音色に変換することである。

音色変換の際、元の音色を中間音色モデルに入力して中間音色を得た後、その中間音色を目標音色モデルに入力して最終的な目標音色を得る。

3.2 中間音色の選択

入力音声が汎用楽器の場合、入力音色と中間音色の差が大きい場合において、DDSP 音色変換モデルは入力楽器の倍音列を目標楽器の倍音列に効果的に変換できる。同時に、入力音声と同じ音高および音量を保持することで、高品質な中間音色の出力を保証する。一方、RAVE 音色変換モデルは、VAE の特性上、入力音色と目標音色の差異の影響を受けやすい。そのため、目標音色にできるだけ近い中間音色を RAVE 音色変換モデルの入力として選択する必要がある。これにより、音色差異が RAVE モデルの音色変換に与える影響を最小限に抑えることができる。具体的な中間音色の選択方法は図 3 に示されている。

1 つの目標音色に対して RAVE 変換モデルを用意する。目標音色と中間音色の音声をこのモデルに入力し、それぞれの潜在変数を取得する。その後、各潜在変数の重みを計算し、それぞれの重みを潜在変数に掛ける。中間音色の潜在変数が目標音色の潜在変数に最も近いものを比較・選定する。

3.3 変分オートエンコーダ（VAE）の詳細

変分オートエンコーダ（VAE）を音色変換に利用する原理は、以下の主要なステップと概念に基づく。

エンコーディングとデコーディングのプロセス：VAE は、エンコーダとデコーダという 2 つの主要な構成要素を持つ。音色変換のタスクにおいて、エンコーダは入力された音声信号（例えば特定の楽器の音色）を圧縮し、潜在空間の低次元表現（潜在ベクトル）に変換する。この潜在ベクトルは、入力音声の主要な特徴を表す。次に、デコーダはこの潜在ベクトルから新しい音声信号を生成し、目標とする音色に近い音声を出力することを目指す。

潜在空間での分布の学習：従来のオートエンコーダとは異なり、VAE は入力データを特定の潜在ベクトルに直接マッピングするのではなく、確率分布（通常はガウス分布）のパラメータ（平均と分散）にマッピングする。この分布からランダムにサンプリングすることで、VAE は複数の潜在ベクトルを生成し、これをデコーダに入力して新しい音声を生成する。この確率分布の学習により、VAE は潜在空間内での複数の音色変換パスを探索できるようになる。

再構成損失と KL ダイバージェンス：VAE の訓練目標は、再構成損失と KL ダイバージェンス（Kullback-Leibler Divergence）の 2 つの損失関数を最小化すること [9]。再構成損失は、元の音声と生成された音声の差を測定し、デコーダが入力に似た音声を生成できるようにする。KL ダイバージェンスは、潜在空間の分布と標準正規分布との違いを測定し、VAE

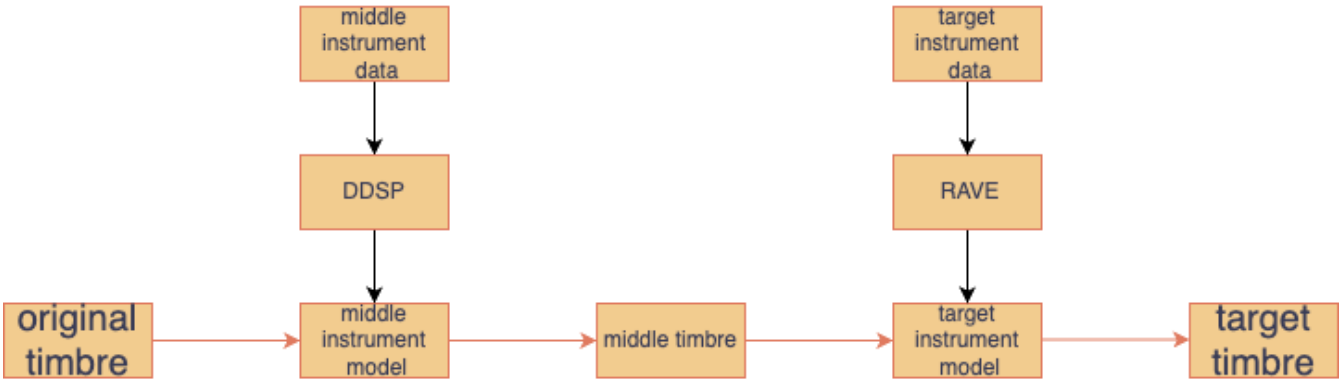


図 2: 最初の入力音声は、2 つの音色変換モデルを順番に通過し、最終的な音声を得る

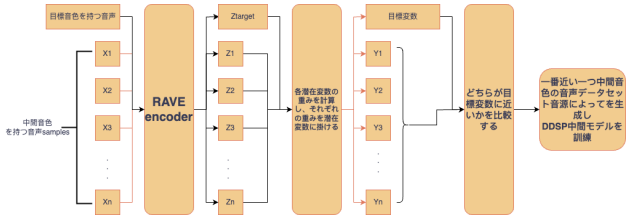


図 3: 中間音色の選択

が潜在空間内で適切に分布する潜在表現を生成できるようにする。

音色変換の実装：音色変換タスクでは、VAE は訓練を通じて潜在空間内で異なる音色の特徴をキャプチャする。変換時には、入力音声を潜在変数にエンコードし、この潜在変数の分布を調整することで新しい音色を生成する。デコーダは調整された潜在変数を目標音色の音声信号に変換する。このプロセスにおいて、VAE はその生成モデルの能力を活用し、ある音色から別の音色への自然な変換を実現する。

VAE を用いた音色変換の詳細なプロセスを図 4 に示す。

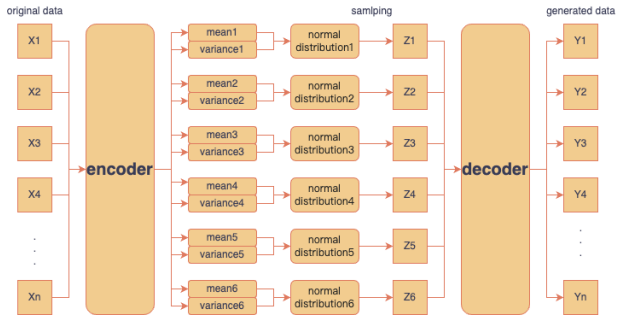


図 4: VAE は、入力音声サンプルに対して複数の潜在変数を持つ正規分布を構築し（この図では潜在変数の数は 6）、そこからサンプリングして新しい音声サンプルを再構成する

利点と制限：VAE の利点は、多様な音色変化を生成できる点にあり、特に複雑な音色空間におけるさまざまな変換経路を探索するのに適している。しかし、潜在空間におけるランダム性のため、生成される音声がぼやけたり歪んだりすることがあり、音色変換の過程でノイズが入りやすいという欠点がある。

3.4 ワッサースタインオートエンコーダ（WAE）

RAVE における 2 段階の学習プロセスは、VAE だけでなくワッサースタインオートエンコーダ（WAE）[13] や他のオートエンコーダにも適用できる。

音色変換タスクにおいて、変分オートエンコーダ（VAE）とワッサースタインオートエンコーダ（WAE）は、それぞれ異なる強みを持ち、適した応用シナリオが異なる。VAE は、データを潜在空間の分布パラメータ（例えば平均と分散）にエンコードし、この分布からサンプリングしてデータを生成する方法を取る。真のデータ分布を近似するために、証拠下界（ELBO）[9] を最大化することを目指す。潜在空間でのサンプリングが確率的であるため、VAE は多様な音色を生成する点に優れるが、特に複雑な音声信号を扱う場合には、生成される音声がぼやけることがあり、音質が低下する場合がある。

一方、WAE は潜在空間におけるワッサースタイン距離を最小化することにより、エンコーダが生成する分布が目標分布により近づき、より鮮明で高品質な音声サンプルを生成する。音色変換において、WAE は生成される音色の正確性と一貫性を重視し、生成される音声のノイズや歪みを効果的に低減するため、高忠実度の音色変換が求められるタスクに適している。

VAE と WAE は、それぞれ異なる応用に適している。VAE は潜在空間の広範な探索や多様な音色の生成が必要なシナリオに理想的であり、特に実験的な研究や詳細よりも創造性が重要な音声生成タスクでよく使用される。一方、WAE は高忠実度な音声生成や音色変換の品質が重要なタスク（音楽制作や高品質な音声合成など）に適している。

さらに、VAE の生成プロセスにはランダム性が高いため、生成される音声にはより多くのノイズが含まれる可能性があり、音色変換の際に音質が望ましくない形で低下する場合がある。WAE は生成プロセス中に距離測定を最適化することで、通常はノイズが少ない音声を生成し、音色変換の全体的な音質とユーザー体験を向上させる。

まとめると、VAE と WAE は音色変換においてそれぞれ異なる強みと弱みを持つ。VAE は多様な音色の探索に適しており、WAE は高忠実度の音色変換に優れる。どちらのモデルを選ぶかは、音色変換タスクの具体的な要件や生成される音声の品質に対する期待に依存する。

3.5 二段階音色変換が一段階変換より優れている理由

音色変換の課題において、二段階で変換された音声は、一段階で処理された音声よりも、しばしば高い音質と目標音色への近さを実現する。この優位性は、音色変換の複雑性をより効果的に管理し、ノイズや歪みを軽減し、全体的な音質を向上させる二段階アプローチの特性によるものである。以下に、その改善の理由を詳述する。

3.5.1 複雑なタスクの分解

特に音色差が大きい楽器間の音色変換は非常に複雑な課題である。二段階変換では、最初の段階で入力音色を中間音色に変換する。この中間音色は、目標音色に近づきながらも、元の音色の特徴をいくらか保持している。その後、第二段階でこの中間音色をさらに目標音色へと精密に変換する。このように、変換を段階的に分割することで、モデルは音色変化の複雑性を効果的に処理し、重要な詳細を失ったり過剰なノイズを導入したりすることを避けられる。

3.5.2 ノイズと歪みの軽減

一段階変換では、全ての音色変化を一度に達成しようとするため、ランダム性が大きくなり、ノイズや歪みが発生しやすい。一方、二段階変換では、第一段階で音色変化の大部分を処理しつつ、高い音質を維持する。その後、第二段階で第一段階で発生したノイズや歪みを軽減し、最終出力の品質をさらに向上させる。このプロセスにより、音色の変化がより滑らかになり、一貫性があり自然な音声が得られる。

3.5.3 潜在空間の効果的な利用

第一段階では、モデルが入力音色を潜在空間の表現にマッピングする。この表現は目標音色に近づく一方で、まだ入力の特徴をいくらか保持している場合がある。第二段階では、この潜在表現をさらに精密化し、目標音色の特性をより正確に捉えることで、高品質な音声を生成する。この段階的な精密化により、単一段階での大きな変化を試みた場合に発生するモード崩壊（出力が過度に似通うか歪む現象）を回避できる。

3.5.4 生成プロセスの段階的最適化

第一段階は大まかな音色変換を行い、主に広範な変化を捉えることに集中する。一方、第二段階では細部の違いを捉え、音質を最適化する。この段階的な最適化プロセスにより、最終的に生成される音声が目標音色の詳細な特性により近づく。また、二段階アプローチはモデルの汎化能力を向上させ、新しい音色変換タスクにも適応しやすくなる。

3.5.5 複数モデルの強みの活用

二段階変換では、異なる種類のモデルを組み合わせることも可能である。例えば、第一段階では大まかな変換のために DDSP を使用し、第二段階では微調整のために RAVE を使用する。このような組み合わせにより、各モデルの強みを活かした、より堅牢で効果的な音色変換を実現する。

4 実験

4.1 トレーニングデータの生成

トレーニングデータ品質を確保するため、音声ファイルは仮想楽器と MAESTRO データセット [7] の MIDI ファイルを使用して合成する。

データセットのオーディオは MIDI ファイルから合成されており、異なる楽器から一貫した信頼性のあるデータを得ている。これにより、2 つのオーディオの間で唯一の変化が音色であることを保証していますが、これにも制約があります。

通常、MIDI ファイルから音楽ファイルを合成するためには、音源とシンセサイザーソフトが必要である。

ここでは fluidsynth シンセサイザーと SoundFont のソースファイルを用いて音楽を合成している。

MIDI と音源を通してギターからピアノを変換するためのトレーニングや実験用の音声を生成するフローチャート例を図 5 に示す。

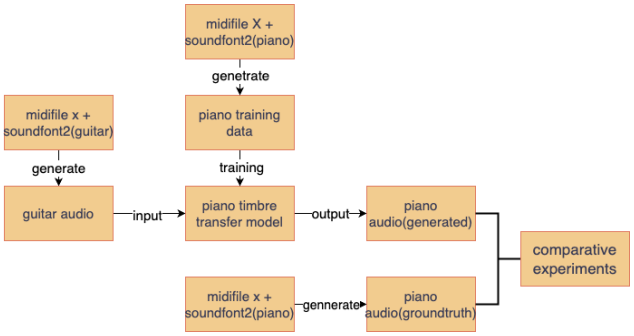


図 5: MIDI と音源でトレーニングや実験用の音声を生成する

4.2 訓練設定

この実験では、DDSP でいくつかの中間楽器音色モデルと RAVE でピアノ音色の変換モデルをトレーニングした。

DDSP でのトレーニングは、約 30 分の楽器音声を使用して、DDSP ツール [10] を用いて音色変換モデルを 30000 steps で訓練した。

ピアノ音色の変換モデルを個別に訓練するため、約 23 時間分のピアノ音声を使用して、RAVE ツール [1] を用いて VAE ベースおよび WAE ベースの変換モデルを訓練した。他の入力音色をピアノ音色に変換することを目的とする。ピアノは音域が広く、豊かな倍音を持つため、音色変換タスクの代表的な選択肢となる。

VAE ベースおよび WAE ベースの両方の変換モデルにおいて、表現学習段階では 100 万回、敵対的ファインチューニング段階では 200 万回の反復を行う。バッチサイズは 8 に設定し、入力データは 3 秒間のセグメントに切り取って処理する。

4.3 評価

本研究における音色変換手法の評価は、客観的評価と主観的評価の 2 つの観点から行う。客観的評価では、変換後の音声と目標音色との類似度を数値的に測定するために、Fréchet Audio Distance（FAD）や Jaccard Distance（JD）などの指標を用いる。一方、主観的評価では、リスナーによる聴覚テストを実施し、変換後の音声の目標音色との一致度を評価して、多角的な観点から分析を行う。

4.3.1 Fréchet Audio Distance (FAD)

Fréchet Audio Distance (FAD)[8] は、音声データを評価するための有用な指標であり、実データと生成データの類似性を測定する。この指標は、元のデータセットに含まれる実際の音声と生成された音声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、元のデータと生成された音声データの分布の差を測定する。

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、音声の知覚品質と特徴の一致を評価する。この距離が小さいほど、生成された音声データが元のデータに近く、高品質な音声生成を示す。したがって、FAD は生成音声の品質向上の指標となる。計算式は式 (1) に示す。

$$FAD = |\mu_r - \mu_g|^2 + \mathrm{tr}\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right) \quad (1)$$

ここで、(μ_r, Σ_r) および (μ_g, Σ_g) は、それぞれ実データと生成データに対応する埋め込みの平均と共分散を示す。記号 tr はトレース演算を表す。

FAD 値が小さいほど、生成されたサンプルはより現実的であることを示す。

4.3.2 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輪郭の一致を測定する。この手法は [11] に基づき、Essentia ライブラリ [2] 内で実装され、2 つのピッチ集合 A と B の不一致を評価するために Jaccard 距離を使用する。

特に、Jaccard 距離は、2 つの集合の交差と和の比率を計算することで測定する。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く、高いコンテンツ保持能力を示す。高いピッチ輪郭の一致は、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し、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ピッチ変化やメロディーの一致を詳細に分析し、生成モデルの改善点や課題を特定することが可能となる。これにより、音声生成モデルのコンテンツ保持能力を正確に評価することができる。計算式は式 (2) に示す。

$$JD(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Jaccard 距離が小さいほど不一致が少なくなり、生成されたピッチの類似性が高いことを示す。

4.3.3 主観的評価

主観的な評価には平均意見スコア（MOS）を使用する。MOS スコアは、50 人の匿名リスナーによる聴取テストに基づいて計算する。このスコアリングシステムは 1（低い）から 5（高い）の範囲で評価し、以下の 3 つの次元を含む。

スタイル変換の成功度 (ST): 変換後の音色が目標とする音色にどれだけ近いかを評価する。

コンテンツ保持度 (CP): 変換後の音楽コンテンツが元のバージョンとどれだけ一致しているかを評価する。

音質 (SQ): オーディオ全体の音質を評価する。

4.4 実験結果

実験は 2 つの部分に分けて実施する。まず、異なる楽器の単音オーディオをトレーニングされたピアノ音色変換モデルに入力し、異なる入力音色が変換に与える影響を比較する。次に、単一段階の変換と二段階の変換による出力オーディオの品質を比較する。

4.5 単音の変換

トレーニングに使用したピアノ音源と 5 つの異なる楽器群の音源を用いて、基準となるピアノ音と異なる音色の 5 つのオーディオサンプルを合成する。これらを 1 秒間の C4 音高の単音の MIDI ファイルと組み合わせる。これらのサンプルをピアノ RAVE 変換モデルに個別に入力する。出力オーディオのスペクトルを図 6 に示す。

例えば電子ギターのような楽器をピアノ音色変換モデルに入力すると、周波数帯域が消失しやすくなる。

バイオリンのような楽器を入力した場合、単音内の微妙な音高変化が複数音として解釈されることがあり、複数のオーディオセグメントが生成されやすくなる。

これらの実験を通じて、ピアノ音色変換モデルでは、入力楽器の音色自体の特性やそれがピアノ音色と異なる程度が、音色変換の結果に影響を与えることが分かる。

4.6 二段式楽器音色変換効果の評価

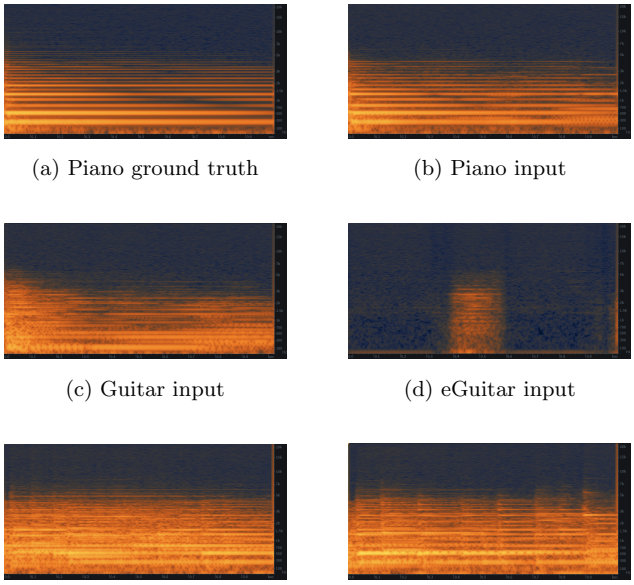
4.7 客観的評価

5 つの異なる楽器で 2 段階の音色変換を実施し、それぞれの変換後のフレシェ音声距離（FAD）スコアを計算して、音声品質の忠実度を評価した。FAD スコアが低いほど、変換された音声が元の音声に近く、音質の忠実度が高いことを示す。テスト対象の楽器には、バイオリン、アコースティックギター、エレクトリックギター、サックス、ピアノ、クラリネットが含まれる。

実験では、最初に音声を RAVE で音色変換を行い、次に音声を二段式音色変換を行った。両方の変換後に FAD スコアを計算し、2 つの変換間の忠実度性能を比較した。

実験結果を表 2 に示す。1 は RAVE の単回変換、2 は二段階変換を表す。

結果の分析から、二段階変換を施した場合、特に FAD スコアの改善が顕著であり、音色の忠実度が向上していることが確



(a) Piano ground truth (b) Piano input

(c) Guitar input (d) eGuitar input

(e) Sax input (f) Violin input

図 6: 異なる音色の単音を入力として、ピアノ音色変換モデルに通した後の単音スペクトログラムの比較

表 1: 直接および二段式音色変換後の FAD、JD スコア

楽器	FAD1	FAD2	JD1	JD2
Violin	9.78	2.32	0.69	0.20
Acoustic Guitar	7.05	4.26	0.82	0.84
Electric Keyboard	16.36	2.80	0.92	0.93
Saxophone	10.71	2.88	0.84	0.19
Clarinet	5.28	2.99	0.36	0.21

認された。また、JD スコアにおいても、特に三つの楽器で改善が見られた。これらの改善は、二段階変換が音高保持の問題を一定程度解決し、音色変換の内容保持においてより精度が高くなったことを示唆している。

4.8 主観的評価

被験者は 38 名の 20 代〜30 代の学生と社会人とし、できるだけ明確的な評価を得るようにした。収集したデータは統計的手法（分散分析 ANOVA）を用いて解析し、RAVE 生成音声と二段式生成音声の間で有意な差があるかを検討した。結果として、二段式生成音声の方が音色の類似性（ST）、音質（SQ）、コンテンツ保持度 (CP) の評価が高い傾向を示し、RAVE 生成音声よりも基準音声に近い音色を再現できていることが確認された。

実験結果を表 3 に示す。1 は RAVE の単回変換、2 は二段階変換を表す。

表 3 のデータから、二段階生成音声は各主観評価指標において、直接 RAVE を用いて生成された音声よりも顕著に優れていることが確認できる。

5 結論と今後の課題

本稿では、二段階の音色変換手法の有効性を示し、単一段階変換と比較して、変換された音声の知覚品質と精度が向上することを確認した。中間的な変換を通じて音色を徐々に洗練することにより、目標とする音色のより忠実な再現が可能になることを実験結果が示している。

中間音色の選択が最終出力品質に大きな影響を与えるため、変換プロセスにおいて適切な中間音色を選択することの重要性が強調される。また、二段階変換手法はノイズの低減や音声品質の向上に効果的であり、目標音色により近い出力を実現することができた。

今後の課題としては、変換プロセスの有効性をさらに高めるため、より最適な中間音を探索することに注力する予定である。中間音の選択が最終出力品質に与える影響を考慮しつつ、最適化を進める。また、変換プロセス中の潜在空間の動態や変動を深く分析することを計画している。例えば、入力音色から目標音色へモデルが変換する過程を詳細に理解することで、変換プロセスのさらなる最適化が可能となり、より堅牢で多用途な音色変換モデルの実現につながると考えられる。

さらに、楽器音色の固有特性が音色変換プロセスに与える影響を検討する予定である。入力音色のエンベロープを目標音色に一致させる調整を行うことで、より正確で効果的な音色変換を目指す。この手法により、音声の時間的なダイナミクスが目標楽器の特性により適合することで、変換の品質が一層向上すると期待される。

参考文献

[1] acids ircam. Rave: Realtime audio variational autoencoder, 2023. Accessed on Mar. 5, 2024.

[2] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra, et al. Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR), 2013.

[3] R. S. Bonnici, M. Benning, and C. Saitis. Timbre transfer with variational auto encoding and cycle-consistent adversarial networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

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

表 2: 直接および二段式音色変換後の主観評価スコア（転置）

	Violin	**Acoustic Guitar**	**Electric Keyboard**	**Saxophone**	**Clarinet**
SQ1	1.58 ± 0.83	1.71 ± 0.84	1.45 ± 0.98	1.76 ± 0.91	1.58 ± 0.92
SQ2	3.18 ± 0.80	2.89 ± 0.95	2.92 ± 0.91	3.16 ± 1.03	2.92 ± 1.00
ST1	1.68 ± 0.90	1.71 ± 0.90	1.21 ± 0.47	2.08 ± 1.00	1.71 ± 0.98
ST2	3.55 ± 1.01	2.84 ± 0.92	3.08 ± 0.88	3.58 ± 1.03	3.55 ± 1.01
CP1	1.53 ± 0.69	1.79 ± 0.74	1.18 ± 0.51	1.50 ± 0.60	1.58 ± 0.83
CP2	3.42 ± 1.13	2.74 ± 0.89	3.18 ± 0.87	3.53 ± 1.01	3.47 ± 1.06

[5] J. Engel, L. Hantrakul, C. Gu, and A. Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

[6] C. Gârbacea, A. van den Oord, Y. Li, F. S. Lim, A. Luebs, O. Vinyals, and T. C. Walters. Low bit-rate speech coding with vq-vae and a wavenet decoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 735–739. IEEE, 2019.

[7] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck. Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv preprint arXiv:1810.12247*, 2018.

[8] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

[9] D. P. Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[10] Magenta. Ddsp: Differentiable digital signal processing, 2023. Accessed on Mar. 5, 2024.

[11] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE transactions on audio, speech, and language processing*, 20(6):1759–1770, 2012.

[12] K. Tatar, D. Bisig, and P. Pasquier. Latent timbre synthesis: Audio-based variational auto-encoders for music composition and sound design applications. *Neural Computing and Applications*, 33:67–84, 2021.

[13] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.