Seqtrument：DAW 上での使用を想定した 自然な楽器音の逐次的生成手段の検討

小平卓実[†1]　西本一志[†1]

概要：コンピュータでアコースティック楽器を用いた楽曲を制作する際，現実の楽器を再現したソフトウェア音源が用いられることがある．しかし既存の音源を用いた場合，各種のパラメタをかなり緻密に調整しても，実際に人間が演奏した音に比べて不自然な音色になることが多い．これに対して近年 AI を使った解決が試みられ，かなり自然な音色の生成が実現されてきている．しかし，既存研究の多くは楽曲全体や一区切りの旋律など，まとまった範囲で処理しているため，楽曲の先頭から逐次的に楽譜を読み込む一般的な音楽制作ソフト（DAW）上で直接動かすことが難しい．そこで本研究では，AI による自然な音色を，既存のソフトウェア音源と同様に DAW 上で直接生成できるように，楽譜情報から逐次的に楽器音を生成する機械学習モデル Seqtrument を提案する．本稿では，システムの構成とこれを用いた現在進行中の予備的な実験結果について報告する．システム全体の有効性の検証結果については，インタラクション 2026 にて報告する予定である．

1. はじめに

近年，コンピュータの普及や安価なソフトウェア音源の登場により，誰でも手軽に作曲を行えるようになった．ソフトウェア音源が楽器の音色を生成する方法として，楽器の 1 音 1 音を全て録音したサンプラー方式や，楽器の物理的特性を再現する物理モデリング方式といったものが存在する．しかし，これらの手法で生成された楽器音をそのまま単純に使用すると，実際に演奏された音に比べて不自然な音色になることが多い．自然な音色に近づけるためには，楽曲の進行に合わせて音量や音高の変化といった様々なパラメタを調整する必要がある．しかしながらこのパラメタ調整は非常に複雑かつ微妙であり，手間がかかる困難な作業となるため，誰でも手軽に実行できるというわけにはいかない．

そこでこの課題を解決するために，AI の活用が試みられている．人間の歌声を対象としたものとして，法野らが発表した Sinsy [1]では DNN をベースとすることで，従来の HMM ベースのもの[2]よりも自然な歌声が合成できることを示した．また，Liu らが発表した DiffSinger [3]では拡散生成モデルによってメルスペクトログラムを生成することで，敵対的生成を用いた手法[4]より高い品質が得られることを示した．楽器を対象としたものでは，Hao らが発表した ViolinDiff [5]等があり，拡散ベースのモデルにピッチベンド情報を明示的に組み込むことで，他の拡散ベースの手法[6]より生成品質が上回った．Deep Performer [7]では Transformer ベースのモデルを用い，ピアノ等の多声音を含む演奏音を高品質に生成できることを示した．

これらの既存手法は，楽曲全体や一区切りの旋律など，まとまった範囲を対象としてバッチ的に処理するものがほとんどである．一方，一般的に用いられている音楽制作ソフト（DAW: Digital Audio Workstation）では，楽曲の先頭から逐次的に楽譜を読み込んで処理を行う．それゆえ，既存の音色合成手法を直接 DAW 上で動かすことができない．そのため，音楽制作者は DAW と DAW 以外の複数のソフトウェアとの間を行き来しながら音楽制作を行う必要があり，作業が煩雑になる．そこで，本研究では逐次的に波形を生成する機械学習モデルを構築することで，AI による自然な楽器音を従来のソフトウェア音源と同様に DAW 上で直接生成させる手段を構築することにより，作曲者のメロディ探索を支援することを目指している．

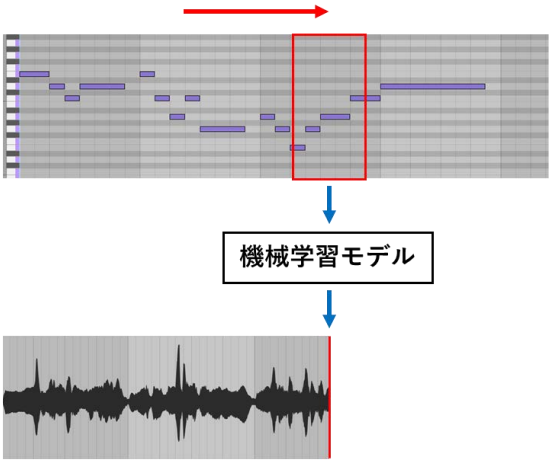

図 1　DAW の再生に合わせて上の楽譜情報から下の波形を逐次的に生成するシステムの動作イメージ

2. Seqtrument

2.1 システムの概要

図 1 は最終的なシステムの動作イメージである．DAW の再生に合わせて，機械学習モデルに現在演奏中の箇所の前後数秒の楽譜情報を入力として与え続けることで，逐次的に波形の生成を行う．

2.2 F0 を条件とした VAE によるスペクトルの圧縮

図 2 に VAE（Variational Autoencoder）の構成を示す．VAE

†1 北陸先端科学技術大学院大学 先端科学技術研究科
Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology

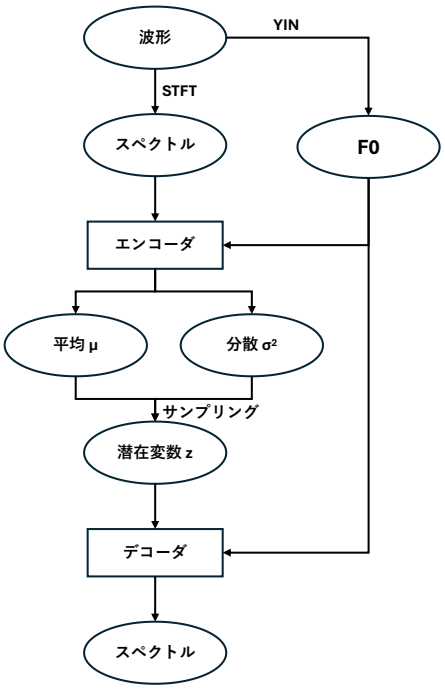

図 2　VAE の構成

表 1　STFT パラメタ

サンプリング周波数	16kHz
FFT サイズ	2048
フレーム間隔	512
窓関数	hann
周波数軸の次元数	1025

[8]は，通常のオートエンコーダとは異なり，入力データを潜在空間へ確率的に圧縮することで，新しいデータを生成することができる．本研究ではまず，表 1 のパラメタを用いた短時間フーリエ変換によって波形データから変換した時間フレームごとの 1025 次元のスペクトルを圧縮するために VAE を用いる．音高情報をモデルに明示的に与えるため，Yin アルゴリズム[9]によって，波形データから基本周波数（F0）を事前に推定した．これを図 2 に示すようにエンコーダとデコーダ両方の入力に組み込むことで，条件付き VAE を構成した．

エンコーダは, 各時間フレームにおけるスペクトルと F0 を入力とし，潜在変数 z の平均 μ と分散 σ^2 を出力する．デコーダは，平均 μ と分散 σ^2 から式（1）によってサンプリングした潜在変数 z と F0 を入力とし，元のスペクトルを出力する．

$$z = \mu + \sigma \cdot \epsilon, \quad \varepsilon \sim N(0,1) \tag{1}$$

損失関数は，式（2）に示すように，再構成誤差と，潜在変数分布と事前分布との間の KL ダイバージェンス（KL 項）の和として定義した．

$$loss = (x - \hat{x})^2 + \beta \frac{1}{2}(\mu^2 + \sigma^2 - \log \sigma^2 - 1) \tag{2}$$

ここで，x は教師データ，$\hat{x}$はデコーダの出力，μ，σ^2 はそれぞれエンコーダが出力する潜在変数における平均，分散，β は KL 項の重みを表す．

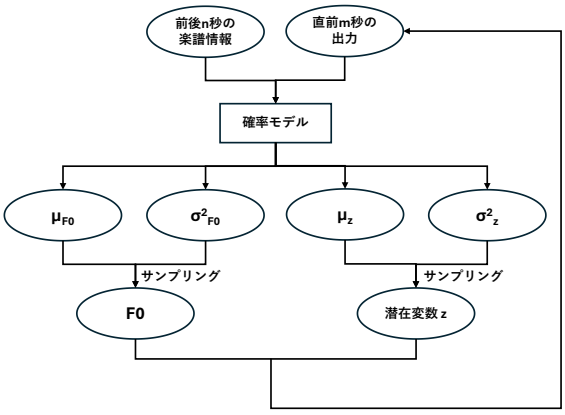

図 3　確率モデルに対する入出力の構成

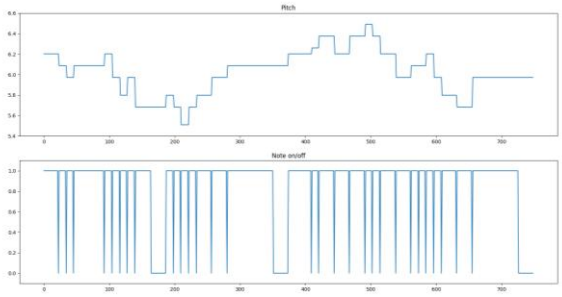

図 4　確率モデルに入力される楽譜情報の形式
上がピッチ，下がノート on/off

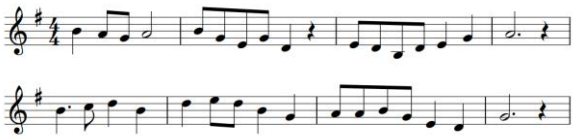

図 5　童謡「七つの子」の楽譜の一部

学習は 1000 エポック実行し，バッチサイズは 64 とした．学習を安定させるため，KL アニーリングを導入し，KL 項の重み β を最初の 200 エポックかけて 0 から 0.01 まで線形に増加させ，潜在変数が過度に事前分布へ収束することを防ぐため Free-bits を導入し, KL 項が 0.5 を下回らないよう制約を設けた．潜在変数の次元は大きくするほど再構成品質は向上するが，潜在変数側も F0 に関する情報を持つようになり，条件 F0 との分離性が低下する．これらのバランスから，潜在変数の次元数を 8 次元に設定した．

2.3　楽譜情報に基づく F0，潜在変数の確率的予測

図 3 に確率モデルに対する入出力の構成を示す．生成箇所の前後 n 秒の楽譜情報と直前 m 秒の出力をモデルの入力とし，楽譜情報（楽譜上の音高データ）と実際の演奏データから推定した F0 の差分，及び 2.2 のエンコーダによって圧縮した潜在変数 z の平均を出力する．楽譜情報は MIDI から各時間フレームにおける以下の形式に変換する．

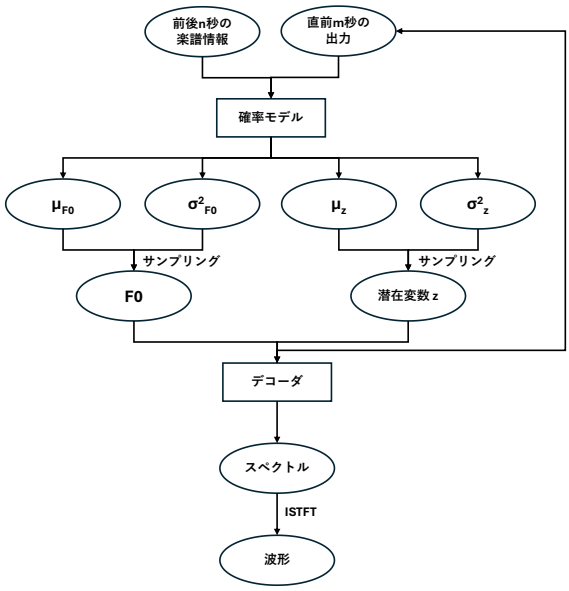

図 6　システムの全体構成

- ピッチ：楽譜上の音の高さ（Hz）を対数化した実数値
- ノート on/off：ノートの on/off に対応したバイナリ値

例として図 4 は図 5 の楽譜を変換したものである．

損失関数には，式 (3) に示す Gaussian Negative log likelihood Loss を用いることで，生成箇所における F0，z の確率分布を学習する．

$$loss = \frac{1}{2}\left(\log \sigma^2 + \frac{(y-\mu)^2}{\sigma^2}\right) \quad (3)$$

ここで，y は教師データ，μ，σ^2 はそれぞれモデルが出力するガウス分布における平均，分散を表す．

学習は 200 エポック実行し，バッチサイズは 64 とした．VAE 同様，最適化アルゴリズムには adam を用い，学習率は 0.001 に設定した．

2.4　システムの全体構成

システムの全体構成を図 6 に示す．推論時は 2.2 のデコーダと 2.3 の確率モデルを図 6 のように繋げて使用する．確率モデルによって，前後の楽譜情報，および直前の出力から F0，潜在変数 z の平均 μ，分散 σ^2 を出力し，式 (4) によってサンプリングする．

$$F0 \sim N(\mu_{F0}, {\sigma^2}_{F0})$$
$$z \sim N(\mu_z, {\sigma^2}_z) \quad (4)$$

サンプリングで得られた，F0，潜在変数 z をデコーダに入力し，出力されたスペクトルから逆短時間フーリエ変換 (ISTFT) によって波形へ変換する．

3.　予備実験

本研究では，最終的に逐次的に楽器音を生成することを目標としているが，現段階のモデルでは，VAE が学習する対象を振幅スペクトルのみに限定しており，位相情報は明示的に含まれていない．そのため，生成された振幅スペクトルから波形を復元する際に Griffin-Lim アルゴリズム[10]を用いて位相の推定を行っている．しかし，Griffin-Lim アルゴリズムでは，スペクトル全体に対して反復的な処理が必要となるため，現段階では全てが逐次的ではなく，スペクトルから波形に復元する際に一括的な処理が含まれている．本節で示す実験は，逐次処理に対する直接的な評価ではなく，今後のモデル改良の方向性を確かめるための予備的実験として実施した．

表 2　実験で提示した音源

Ground Truth	ピッチ，タイミング修正を施した実際のバイオリン演奏データ
VAE (WIP)	VAEによってGround Truthを再構成した音源
Seqtrument (WIP)	本システム(途中経過)によって生成
Sampling	サンプリング音源によって生成
Physical Modeling	物理モデリング音源によって生成

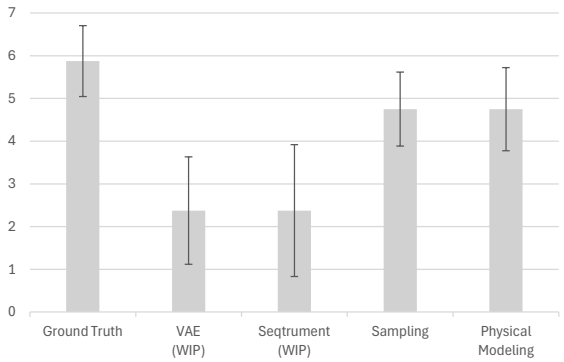

図 7　自然さ

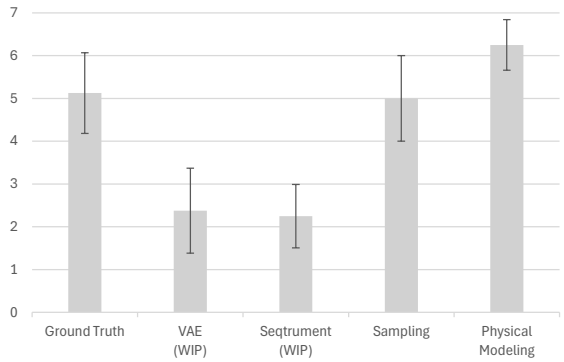

図 8　音質

3.1　データセット

本学の学生による「七つの子」や「蛍の光」などの童謡 10 曲分のバイオリン演奏を各 3 セット（約 10 分×3 セット）録音し，ピッチ補正ソフトによって意図しないピッチやタイミングのずれに対して修正を行った．DAW のトランスポーズ機能を用いて，元の音源から半音上げたデータを追加し，9 割を学習データに用い，残りの 1 割をテストデータとした．16bit/44.1kHz からモノラル 16bit/16kHz にダウンサンプリングして使用した．

3.2　評価方法

被験者に対して表 2 の音源を聴き比べてもらい，以下の 2 つの項目について 7 段階の評価アンケートを実施した．

- 自然さ：人間が演奏しているように感じますか？
- 音質：ノイズやひずみがなくきれいに聴こえますか？

3.3 結果・考察

図 7，図 8 はそれぞれ自然さ，音質について 8 人の回答結果をまとめたものである．現段階では，本システムによって生成された音は，自然さ，音質の両面において既存のソフトウェア音源より低い評価となった．一方で，VAE により再構成した音源と，システム全体を通して生成した音源の評価に有意差が認められなかったことから，VAE がボトルネックになっていると考えられる. 被験者からは, 両音源に対して「音が途切れているように感じる」といった意見が多く得られた. 前述の通り，現段階での VAE には位相情報を与えておらず，各フレームは明示的に時系列間の情報を持たない. そのため，復元する際に時間軸方向の再構成精度が十分に確保されなかったことで，各評価の低下につながった可能性がある．そこで，今後は VAE の精度改善に重点を置き，逐次的に位相を推定するアルゴリズムの実装，または位相情報を含めたスペクトルの圧縮を試みる予定である．

4. まとめ

本研究では DAW 上での使用を想定した楽器音生成モデル Seqtrument を提案した．今後はさらにモデルの改良を試みた上で，逐次的に生成した楽器音の音色の自然さや生成品質について，既存の手法と比較実験を行う予定である．

謝辞 実験にご協力いただいた協力者の皆さんに厚くお礼申し上げます. 本研究は JSPS 科研費 JP24K02976 の助成を受けたものです.

参考文献

[1] Yukiya Hono, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda：Sinsy: A Deep Neural Network-Based Singing Voice Synthesis System, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.29, pp.2803-2815, 2021.

[2] Keijiro Saino, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, and Keiichi Tokuda : An HMM-based singing voice synthesis system, Ninth International Conference on Spoken Language Processing, pp.1141-1144, 2006

[3] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao : DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism, Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[4] Jie Wu, and Jian Luan：Adversarially Trained Multi-Singer Sequence-To-Sequence Singing Synthesizer, INTERSPEECH 2020, pp.1296-1300, 2020

[5] Daewoong Kim, Hao-Wen Dong, and Dasaem Jeong : ViolinDiff: Enhancing Expressive Violin Synthesis with Pitch Bend Conditioning, ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing （ICASSP）, 2025.

[6] Ben Maman, Johannes Zeitler, Meinard Müller, and Amit H. Bermano : Performance Conditioning for Diffusion-Based Multi-Instrument Music Synthesis, ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

[7] Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley : Deep Performer: Score-to-Audio Music Performance Synthesis, ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing （ICASSP）, 2022.

[8] Diederik P Kingma, and Max Welling : Auto-Encoding Variational Bayes, International Conference on Learning Representations (ICLR), 2014.

[9] Alain de Cheveigné, and Hideki Kawahara : YIN, a fundamental frequency estimator for speech and music, The Journal of the Acoustical Society of America, Vol.111, pp.1917-1930, 2002.

[10] D. Griffin, and Jae Lim：Signal estimation from modified short-time Fourier transform, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol.32, Issue2, pp.236-243, 1984