早稲田大学大学院情報生産システム研究科

博士論文審査結果報告書

論　文　題　目

Study on Symbolic Singing Annotation and Monophonic Music Generation Using Deep Neural Networks

申　請　者

Xiao FU

情報生産システム工学専攻
ニューロコンピューティング研究

2024 年　6 月

近年、人工知能を適用した音楽分野、特に歌唱アノテーションと音楽生成の分野で大きな進歩が見られる。これにより、作曲と演奏に対する理解が深まっただけでなく、音楽 AI（Artificial Intelligence）分野における創造的表現やイノベーションの新たな道が開かれた。歌唱アノテーションとは、歌唱に関連する注釈を加えることである。例えば、歌唱音声に対応する各音符に「pitch, duration, lyrics, onset, offset」という注釈を付加することで、研究者やミュージシャンが歌唱パフォーマンスを理解できるだけではなく、音楽生成にも利用できる。従来の方法では、アノテーターとして人間の専門家に作業しており、コストと時間がかかる。最近では、機械学習技術を利用して歌唱音声から直接アノテーションを行うアプローチが注目されている。しかし、教師データが少ないなどの課題も残されている。一方、音楽生成は、歌唱アノテーションの一つのアプリケーションと考えられ、計算機的手法を用いて自動的に楽曲を作成することを目指している。これまでのアプローチの一つは、シーケンス予測モデルを利用して音符単位で作曲するものである。このアプローチでは、シーケンス内の局所的な一貫性を確立し、個々の音符間の滑らかさを確保することができる。しかし、音楽フレーズ全体の包括的な一貫性に対処するには不十分である。その結果、生成された音楽は断片的で、全体的な構造や一貫性に欠けることが多い。そこで本論文では、教師データが少ないという困難を克服し、深層学習、さらに深層転移学習を用いて、Coarse-to-fine 2 段階によるシンボリック歌唱アノテーションを行っている。そして、音楽フレーズの全体的な一貫性を向上させるモンテカルロ木探索（Monte-Carlo Tree Search: MCTS）に基づくモノフォニック音楽生成フレームワークを提案している。

以下に、本論文の構成と各章の概略について述べ、評価を与える。

第 1 章では、歌唱アノテーションと音楽生成の背景および関連研究を紹介し、本研究の動機・目的を明らかにしている。

第 2 章では、深層畳み込みニューラルネットワークに基づく、楽譜と歌詞ファイル情報を利用した Coarse-to-fine 2 段階シンボリック歌唱アノテーションの提案について述べている。深層ニューラルネットワークでは、end-to-end モデルが主流であるが、歌唱アノテーションにおいては利用可能な教師データが限られており、複雑な end-to-end モデルを構築するには不十分である。本論文では、より単純なモデルで実装可能な Coarse-to-fine 2 段階からなる歌唱アノテーション法を提案している。まず、楽譜と歌詞のラベリング法を適用して、楽譜と歌詞ファイル情報から Coarse アノテーションを生成する。次に、歌唱音声と Coarse アノテーションから歌唱ラベルを自動校正することで、正確なアノテーションを推定する。具体的に、歌唱ラベル自動校正フレームワークは、データ前処理、時間シフト推定、ラベル補正という 3 つの相互に関連するモジュールから構成される。データ前処理モジュールは、モデルの入力の前処理として、歌唱音声と Coarse アノテーションをメル・スペクトログラムに変換する。時間シフト推定モジュールは、設計された完全畳み込みニューラルネットワークを用いて、Coarse アノテーションと正確なアノテーションの間の時間的

オフセットを予測する。ラベル補正モジュールは、Coarseアノテーションをさらに精緻化し、アノテーション精度を向上させるために、2つのキャリブレーション手法を組み込んでいる。さらに、提案法の性能を数値例で評価するために、手動でHSD（X. Fu et al., 2022）という歌唱アノテーションデータセットを独自に開発している。提案法をHSDデータセットに適用し、従来の手動キャリブレーション手法と比較した実験結果では、平均ラベリングL1誤差が0.388から0.348に減少し、アノテーション精度が向上した。さらに、計算時間も37%短縮された。これらのことから、提案法の有効性が明らかとなっている。

第3章では、容易に入手可能な楽器音を利用した、転移学習による改良型歌唱アノテーション法の提案について述べている。第2章で述べたCoarse-to-fine 2段階シンボリック歌唱アノテーション法では、訓練において教師データを増強することにより、歌唱音声とCoarseアノテーションから正確なアノテーションを推定する精度を向上させることができる。本論文では、楽器音を活用した改良型歌唱アノテーション法を提案している。提案法は楽器ラベルキャリブレーション（Musical Instrument Label Calibration: MILC）という事前学習と自動歌唱ラベルキャリブレーション（Automatic Singing Label Calibration: ASLC）というファインチューニングの2つのフェーズから構成される。MILC事前学習フェーズでは、まず、教師データセットの出力ラベルである正確なアノテーションから、楽譜と歌詞ラベリング法に基づくシミュレーションでCoarseアノテーションを生成する。次に、ピアノ等の多様な楽器で楽器音を生成する。そして、これらの楽器音による教師データセットで深層畳み込みニューラルネットワークモデルの事前学習を行う。ASLCファインチューニングフェーズでは、歌唱音声による教師データセットで深層畳み込みニューラルネットワークモデルをファインチューニングする。このような事前学習・ファインチューニングという転移学習を行うことで、教師データが不十分という困難を克服し、高性能な自動歌唱アノテーションを実現している。性能評価の数値例では、提案法を前述のHSDデータセットに適用して、転移学習のない方法と比較した結果、予測精度が0.845から0.860に向上し、ラベリングL1誤差が0.258から0.248に減少した。このようにより良い精度が達成できたことから、提案法の有効性が明らかとなっている。

第4章では、シーケンス予測モデルとMCTSアルゴリズムを組み合わせた、モノフォニック音楽生成フレームワークの提案について述べている。シーケンス予測モデルを用いた従来の音楽生成手法では、全体的な構造と一貫性に欠ける音楽が生成されることが多い。本論文では、MCTSの強力なシミュレーションと評価機能を活用し、生成される音楽フレーズの全体的な一貫性を高めるモノフォニック音楽生成フレームワークを提案している。提案した音楽生成フレームワークは、音符候補生成と音符候補評価という2つのモジュールから構成される。音符候補生成モジュールでは、Self-Attention Mechanismに基づくトランスフォーマーを用いたシーケンス予測モデルを構築し、複数の音符候補を

生成する。音符候補評価モジュールでは、MCTS アルゴリズムが各音符候補に対して後続の音符をシミュレートし、その音符候補を含んだ音楽フレーズとして評価を行い、最終的に音符候補を選択する。提案された音楽生成フレームワークでは、MCTS に合理的な評価関数を提供するため、音楽フレーズの一貫性を評価できるバリューネットワークをミスマッチ法で学習して構築している。さらに、MCTS 検索プロセスを平滑化するために、改良型 PUCT (Polynomial Upper Confidence Trees) アルゴリズムを提案している。性能評価のための数値例では、提案の音楽生成フレームワークと従来の音符シーケンス予測モデルを 10 のテスト曲に適用して比較し、予測精度においては 8 のテスト曲で、提案法が従来法より優れていた。例えば、テスト曲 song-3 に対して、従来法の top-1 予測精度は 43.92%であるのに対して、提案法では 47.04%である。また、ユーザーによる調査実験において、提案アプローチは、創造性・一貫性・音楽性・全体的な印象の点で、従来法の MidiNet (L.C. Yang et al., 2017)や SSMGAN (H. Jhamtani et al., 2019)等より優れている。これらのことから、提案法の有効性が明らかとなっている。

第 5 章では、結論として、本研究の成果について総括し、今後の研究課題を論じている。

以上を要約すると、本論文では、1)畳み込みニューラルネットワークモデルに基づく Coarse-to-fine 2 段階からなるシンボリック歌唱アノテーション法を提案している。2)容易に入手可能な楽器音を活用した転移学習に基づく、改良型高性能自動歌唱アノテーションを実現している。3)音楽フレーズの全体的な一貫性を向上させるモンテカルロ木探索に基づく、モノフォニック音楽生成フレームワークを提案している。また、提案法をユーザー調査やベンチマークデータセットによって評価することでその有効性を明らかにしている。その他、「pitch, duration, lyrics, onset, offset」というラベルを含む歌唱アノテーションデータセットを独自に開発し、公開したことも研究者コミュニティへの貢献として認められる。これらの成果は深層学習、ニューロコンピューティング分野の発展に寄与するところ大である。よって、本論文は、博士（工学）の学位論文として価値あるものと認める。

2024 年 5 月 13 日

審査員

主査	早稲田大学	教授	博士(情報工学)(九州工業大学)	古月 敬之
副査	早稲田大学	教授	工学博士(早稲田大学)	吉江 修
	早稲田大学	教授	博士(工学)(早稲田大学)	藤村 茂