

ホウノ ユキヤ

氏名 法野 行哉

学位の種類 博士（工学）

学位記番号 博第1246号

学位授与の日付 2022年3月31日

学位授与の条件 学位規則第4条第1項該当 課程博士

学位論文題目 ACOUSTIC AND WAVEFORM MODELING FOR SINGING VOICE SYNTHESIS
BASED ON DEEP NEURAL NETWORKS
(深層ニューラルネットワークに基づく歌声合成のための音響・波形モデリング)

論文審査委員

主査	教授	徳田 恵一
	教授	本谷 秀堅
	准教授	酒向 慎司
	准教授	南角 吉彦
	准教授	橋本 佳
	教授	戸田 智基

(名古屋大学)

論文内容の要旨

日々の生活の中で「歌」はとても重要な存在である。我々人類は、遙か昔から、お互いに感情表現する手段の一つとして歌を用いている。これまでに、民謡から歌謡曲まで数え切れない程多くの曲が生み出され、様々な場面で歌われている。近年では、計算機やそれに伴う情報技術の発展により、コンピュータによって任意の歌声を作り出す歌声合成技術の開発が進められており、歌声合成システムによって合成された歌声を用いた楽曲の制作や、楽曲制作の補助としての歌声合成システムの利用等、様々な場面で応用されている。歌声合成は、人間らしい歌を合成するのにとどまらず、非常に早いテンポの歌唱や超高音での歌唱等といった、人間による歌唱が困難な曲での歌声の合成も可能であるため、楽曲制作時の表現や創造の可能性を広げることのできる技術としても注目されている。このような背景の中で、任意の楽譜から、より自然な歌声を合成できる歌声合成技術が求められていると言える。

歌声合成における代表的な枠組みとして、統計的パラメトリック歌声合成が挙げられる。この枠組みでは、楽譜を解析することで得られる楽譜特徴量から音声波形から抽出される音響特徴量という二つの中間表現を導入し、歌声合成を、楽譜特徴量から音響特徴量の予測と、音響特徴量から音声波形の生成という二つの問題の組み合わせとして捉える。楽譜特徴量と音響特徴量の対応関係は、これまで、隠れマルコフモデル (hidden Markov model; HMM) によって統計的にモデル化されてきたが、近年では深層ニューラルネット

ワーク (deep neural network; DNN) を用いる試みが広く行われている。歌声合成では、音符の音高や音符長、楽曲のテンポなど、楽譜によって与えられる情報に忠実に従った波形を生成する必要がある。その一方で、歌には、音高を周期的に揺らすビブラートのように、楽譜に明記することが困難な歌唱表現が存在することも特徴的である。従って、これらの歌唱表現を含む歌声の音響的特徴とその時間構造をどのようにモデル化するかは、歌声合成における重要な課題である。一方、波形生成の方法としては、これまで、信号処理に基づくボコーダが広く用いられてきた。近年では、それに代わる手法として、DNNに基づく波形生成モデルが数多く提案されており、統計的パラメトリック音声合成分野において広く用いられつつある。しかし、歌声合成では、与えられるピッチに忠実に従う歌声波形の生成が必要である等の理由により、それらの波形生成モデルを歌声合成システムにそのまま転用することが適切とは言えず、歌声波形の生成に最適な波形生成モデルが求められる。

これらの背景を踏まえ、本論文では、まず、歌声の歌唱表現を考慮しつつ音響特微量と時間構造を明示的にモデル化する DNN 歌声合成システムの枠組みを提案する。提案システムでは、歌声波形から推定した発声タイミングのずれと音素継続長を DNN により独立してモデル化する。また、歌声の対数基本周波数からビブラート成分をビブラートパラメータとして分離し、明示的にモデル化する。DNN 歌声合成は、統計的手法であることから学習データの傾向を再現しようとする。そのため、歌声データに調子外れな歌声データが含まれていた場合、調子が外れた歌声が合成されてしまう可能性があり、主観的な品質低下に繋がってしまう恐れがある。本論文では、音響モデルの学習過程で調子外れの原因である音高の逸脱を自動的に吸収、補正することで、この問題に対処する。

次に、自然性の向上を目指し、周期・非周期成分の分離を考慮した DNN 波形生成モデルを検討する。本論文では、音声波形が周期成分と非周期成分の和で表現されるとして、明示的な周期信号及び非周期信号を入力とする並列型や直列型のモデル構造を持った DNN 波形生成モデルを検討し、有効性を確認する。

最後に、歌声の音響特微量と時間構造を効率よく同時モデル化することを目指し、擬似音素境界を用いた sequence-to-sequence (seq2seq) 歌声合成を提案する。前述した DNN 歌声合成システムでは、歌声の音響特微量と時間構造を三つの独立した DNN を用いてモデル化していたが、全体最適化の観点では必ずしも適切であるとは限らない。近年、Attention 機構に基づく seq2seq モデルによって、楽譜特微量から音響特微量を单一の DNN を用いてモデル化する手法が提案されている。しかし、歌声の時間構造は同一歌詞であってもテンポやリズムに応じて大きく変化するため、時間解像度が大きく異なる楽譜特微量と音響特微量の変換を单一のモデルで直接表現することは、依然として容易ではなく、学習の困難性や自然性の低下を招く。提案法では、歌唱時の発声タイミングのずれの傾向を考慮しながら音符の音符長からヒューリスティックに擬似音素境界を求め、擬似音素境界から本来の境界への時間構造の変換を seq2seq モデルが担う。これにより、seq2seq 歌声合成における時間構造のモデル化の困難性の克服を目指す。

以上のように、本論文では、歌声合成システムのための、DNN を用いた音響特微量及び歌声波形のより適切なモデル化手法を提案し、評価実験によりその有効性を示す。

論文審査結果の要旨

我々人類は、遙か昔から、お互いに感情表現する手段の一つとして歌を用いており、多くの曲が生み出され、様々な場面で歌わされてきた。近年では、計算機やそれに伴う情報技術の発展により、コンピュータによって任意の歌声を作り出す歌声合成技術の開発が進められており、歌声合成システムによって合成された歌声を用いた楽曲の制作や、楽曲制作の補助としての歌声合成システムの利用等、様々な場面である。歌声合成は、人間らしい歌を合成するのにとどまらず、非常に早いテンポの歌唱や超高音での歌唱等といった、人間による歌唱が困難な曲での歌声の合成も可能であるため、楽曲制作時の表現や創造の可能性を広げることのできる技術としても注目されている。このような背景の中で、任意の楽譜から、より自然な歌声を合成できる歌声合成技術に関する研究を行うことは大変価値のあることと考えられる。

本論文では、歌声合成における代表的な枠組みの一つである統計的パラメトリック歌声合成に立脚した研究を行っている。この枠組みでは、楽譜を解析することで得られる楽譜特微量と、音声波形から抽出される音響特微量という二つの中間表現を導入することで、歌声合成を、楽譜特微量から音響特微量の予測（問題①）と、音響特微量から音声波形の生成（問題②）という二つの問題の組み合わせとして捉えることができる。歌声合成では、音符の音高や音符長、楽曲のテンポなど、楽譜によって与えられる情報に忠実に従った波形を生成する必要があるが、その一方で、歌には、音高を周期的に揺らすビブラートのように、楽譜に明記することが困難な歌唱表現が存在することも特徴的である。従って、これらの歌唱表現を含む歌声を統計的パラメトリック歌声合成の枠組みにおいて、如何にモデル化するかは、歌声合成を考える上で、重要な課題であると言える。

本論文では、まず、楽譜特微量と音響特微量の間の関係性のモデル化（問題①）に関して、深層ニューラルネットワーク（DNN）を用い、歌声の音響的特徴に加え、歌声の発声タイミングの時間的なずれやビブラートといった歌唱表現も明示的にモデル化する枠組みを提案している。提案法では、DNN を用いた柔軟なモデリングが可能となることから、学習データの傾向を精度良く再現でき、任意の楽譜から自然な歌声の合成を実現することができる。その一方で、再現性の高さ故に、楽譜の音符の音高から大きく逸脱した調子外れな歌声データが学習データに含まれる場合、合成される歌声も、調子外れに陥ってしまう可能性が考えられる。音高の正確性は、主観的な品質に大きな影響を与えると予想されるため、それらの逸脱を自動的に補正する枠組みについても重要な課題と考えられるが、本論文では、音響モデルの学習過程で調子外れの原因である音高の逸脱を自動的に吸収、補正することで、この問題に対処している。

音響特微量から歌声波形の生成方法（問題②）としては、これまで、信号処理に基づくボコーダが広く用いられてきた。近年では、主にテキスト音声合成分野において、深層波形生成モデルが広く利用されつつある。歌声合成では、テキスト音声合成とは異なり、与えられた音高やプレス（息の音）の高い再現性等が求められることから、本論文では、並列・直列構造を持つ、歌声合成に最適な深層波形生成モデルを提案しており、問題②における課題を解決している。

最後に、歌声の音響特微量と時間構造を効率よく同時モデル化することを目指し、擬似音素境界を用いた sequence-to-sequence (seq2seq) 歌声合成（問題③）を提案している。前述した DNN 歌声合成では、歌声の音響特微量と時間構造を複数の DNN を用いてモデル化していたが、全体最適化の観点では必ずしも適切であるとは限らない。近年、Attention 機構に基づく seq2seq モデルによって、楽譜特微量から音響特微量を単一の DNN を用いモデル化する試みがなされているが、歌声の時間構造は同一歌詞であってもテンポやリズムに応じて大きく変化するため、seq2seq モデルによるモデル化は依然として容易ではなく、学習の困難性や自然性の低下を招くといった課題があった。それに対して、提案法では、歌唱時の発声タイミングのずれの傾向を考慮しながら音符の音符長からヒューリスティックに擬似音素境界を求め、擬似音素境界から本来の境界への時間構造の変換を seq2seq モデルが担う手法により、seq2seq モデルを用いたモデル化の困難性を克服している。

以上より、本論文は、深層ニューラルネットワークに基づいた歌声合成システム構成法について網羅的な検討を行うことができており、博士論文として十分な内容をもっていることを確認した。