論 文 審 査 報 告 書

氏　　名	松永悟行（まつながのりゆき）
学位の種類	博士（工学）
学位記番号	博知第13号
学位授与日	令和 3年 3 月 20 日
論文題目	計算資源が限られた音声合成システムに用いる 深層学習モデルの学習法に関する研究
論文審査委員	（主査）富山県立大学 教 授 平 原 達 也 教 授 神 谷 和 秀 教 授 小 柳 健 一 准教授 モクタリ パーハム 北陸先端科学技術大学院大学 教 授 赤 木 正 人

内 容 の 要 旨

本論文は、深層学習モデル（DNN : Deep Neural Network）による音声合成システムにおいて、DNNに入力される言語特徴量が外れ値を取らないことを保証する正規化法、合成音声の品質を劣化させない音声特徴量の生成モデルをDNNに学習させる損失関数、および生成的敵対ネットワークを用いた先験知識が不要なDNNの学習法を提案するとともにその性能を評価し、計算資源が限られた計算環境においても、順伝搬型のDNNだけで構成した音声特徴量予測部を用いて頑健かつ高速に高品質の音声を合成できることを示したものである。

テキストから音声を合成するテキスト音声合成システムはさまざまな用途で利用されており、近年では、DNN を用いた統計モデル方式の音声合成システムの研究が盛んに行われている。音声合成システムは、入力テキストから言語特徴量を抽出する言語解析部、言語特徴量から音声特徴量を算出する音声特徴量予測部、音声特徴量から音声波形を生成する波形生成部の3つのサブシステムで構成される。各サブシステムの入出力関係をDNNを用いて個別に学習することにより、人智に基づいた規則や変換処理を用いずに、入力テキストから音声波形を合成できる。また、これらの DNN を連結した大規模なDNNを一貫学習すれば、入力テキストから音声波形を直接生成することもできる。

音声合成システムには、高品質な音声を合成することだけではなく、入力テキストに対する頑健な音声の合成、単語の読み方の指定、合成音声の韻律や声質の制御、遅延の少ない音声出力などが要求され

る。また、音声合成システムの実装には、演算装置の性能や記憶装置の容量などのハードウェアの制約もある。そのため、システムへの要求やハードウェアの制約が満たせない場合、高品質な音声を合成する音声合成システムであっても、それは製品やサービスに利用されない。例えば、大規模な DNN による音声合成システムは必ずしも多くの製品やサービスに適しているとはいえない。一方、低性能な計算機でも処理を削減すれば高速に音声を合成できるが、合成音声の品質は劣化する。 そこで、計算資源が限られた計算環境でも DNN 音声合成処理を利用して頑健かつ高速に高品質な音声を合成できる技術が求められている。

本論文の各章の内容は以下のとおりである。

第 1 章は、序論で、音声合成技術の歴史を概観するとともに、本論文で取り組む問題を明らかにし、目的および本論文の構成について述べている。

第 2 章は、DNN 音声合成システムの全体構成、音声特徴量予測部の詳細な処理、および本論文で利用した音声コーパスについて述べている。

第 3 章では、音声特徴量予測部の処理時間について検討し、計算資源が限られた計算環境では、音声特徴量予測部は FFNN（Feed-Forward Neural Network）と呼ばれる順伝搬型の DNN だけで構成する必要があることを示している。従来の音声特徴量予測部は、FFNN と後処理の音声特徴量の動的特徴量の尤度を最大化する音声特徴量の生成法（MLPG : Maximum Likelihood Parameter Generation）で構成される。もう一つの従来の音声特徴量予測部は、後処理の MLPG を必要としない再帰型の DNN（RNN : Recurrent Neural Network）だけで構成される。これら従来の音声特徴量予測部と FFNN だけで構成される音声特徴量予測部を用いて 5 秒分の音声特徴量を生成するのに要した時間を比較した結果、FFNN だけで構成される音声特徴量予測部の処理時間は、前世代の主演算装置でも百数十ミリ秒であり、従来の音声特徴量予測部の処理時間の 1/8 から 1/30 であった。

第 4 章では、DNN を用いた音声特徴量予測部が音声特徴量を頑健に予測するために、言語特徴量から音声特徴量を予測する DNN に外れ値を入力させない言語特徴量の正規化法について述べている。音声合成システムに入力される文章は、DNN の学習に用いる文章と文章構成が異なるため、言語特徴量に外れ値が含まれる可能性が高い。従来法では、学習データから算出した最小値と最大値に基づいて言語特徴量を正規化するが、正規化範囲を超える入力値は外れ値となる。この問題を解決するために、入力された一文章内の言語特徴量の任意の二つの属性値の比を用いる言語特徴量の正規化法を提案した。言語特徴量との関連が強い音声特徴量の一つである基本周波数を提案法と従来法を用いてモデル化し、合成した音声の品質を聴取実験で比較評価した結果、提案した正規化法による合成音声の評点は従来の正規化法による合成音声の評点よりも有意に高かった。提案した言語特徴量の正規化法は、音声特徴量予測部が入力文章に対して基本周波数を頑健に予測することを可能にした。

第 5 章では、FFNN だけで構成される音声特徴量予測部において、音声特徴量の時間構造に関する複数の誤差を考慮した損失関数による FFNN の学習法について述べている。音声特徴量をモデル化すると

きに用いる一般的な損失関数は、教師となる音声特徴量と予測された音声特徴量の時間フレームごとの誤差を計算する。提案する損失関数は、この誤差に加えて、音声特徴量の動的特徴量の誤差、系列全体の分散の誤差および共分散の誤差、いくつかの時間フレームにおける分散の誤差および共分散の誤差、次元間の関係を表した誤差を計算する。これらの誤差を逆伝播させることで、音声特徴量の構造を多角的に捉えたモデルが獲得できる。この MATS（Multiple Attributes of Temporal Sequence）損失関数を用いて学習した FFNN だけで構成された音声特徴量予測部と、一般的な損失関数を用いて学習した FFNN あるいは RNN および後処理で構成される音声特徴量予測部で基本周波数とメルケプストラムをモデル化し、合成音声の聴取実験による比較を行った。その結果、提案する MATS 損失関数を用いた学習法を用いて合成した音声の評点は、従来の学習法を用いて合成した音声の評点よりも有意に高かった。MATS 損失関数は音声特徴量に関する先験知識を用いて設計する必要があるが、FFNN だけで構成される音声特徴量予測部が合成音声の品質を劣化させない音声特徴量を予測することを可能にした。

第 6 章では、FFNN だけで構成される音声特徴量予測部における、音声特徴量に関する事前知識を用いずに音声特徴量の構造をモデル化する生成的敵対ネットワーク（GAN：Generative Adversarial Network）による FFNN の学習法について述べている。従来の GAN による学習法は、時間フレームごとの音声特徴量の生成誤差と識別誤差を用いるため、音声特徴量の系列全体の特徴を捉えることができない。提案する GAN による学習法は、時間フレームごとの音声特徴量の生成誤差と音声特徴量のグラム行列の識別誤差を用いることで、時間フレームごとの音声特徴量と音声特徴量の系列全体の特徴を考慮することができる。提案する GAN による学習法と、識別モデルに FFNN あるいは CNN（Convolutional Neural Network）を用いた従来の GAN による学習法でメルケプストラムをモデル化し、合成音声の聴取実験による比較を行った。その結果、提案する GAN による学習法を用いて合成した音声の評点は、従来の GAN による学習法を用いて合成した音声の評点よりも有意に高かった。提案する GAN による学習法は、FFNN だけで構成される音声特徴量予測部が事前知識を用いず自動的に音声特徴量の構造をモデル化することを可能にした。

第 7 章は結論で、本論文で明らかにした結果をまとめて述べている。

以上、本論文では、計算資源が限られた計算環境において DNN による音声合成システムを頑健かつ高速に動作させるためには音声特徴量予測部を FFNN だけで構成する必要があることを示すとともに、DNN に入力する言語特徴量が外れ値を取らないことを保証する正規化法、音声特徴量を高速に予測する DNN の損失関数、および生成的敵対ネットワークによる学習法を考案するとともに、それらの性能を評価した。その結果、提案した言語特徴量の正規化法は DNN を用いた音声特徴量予測部が音声特徴量を頑健に予測すること、提案した MATS 損失関数による FFNN の学習法と提案した GAN による FFNN の学習法は、計算資源が限られる計算環境においても FFNN が音声特徴量の時間構造を考慮したモデルを獲得することを示した。すなわち、計算資源が限られた計算環境においても、DNN 音声合成システムが頑健かつ高速に高品質な音声を合成できるようになった。

審 査 の 結 果 の 要 旨

本論文は、計算資源が限られる計算環境においても頑健かつ高速に動作する音声合成システムを実現することを目的として、深層学習モデル（DNN : Deep Neural Network）を用いて言語特徴量から音響特徴量を算出する音声特徴量予測処理に関する研究をまとめたもので、全 7 章で構成される。

文章を音声に変換するテキスト音声合成（TTS: Text-to-speech synthesizer）技術はさまざまなシステムで利用されている。コンピュータの性能向上に伴って統計モデル方式の TTS 技術は発展し、合成音声の品質も向上している。入力文章と出力音声の関係を大規模な DNN を用いて一括してモデル化する TTS システムも実現されている。しかし、このような一括方式の TTS システムは、モデル学習時に大規模な音声コーパスが必要であり、また、CPU 性能やメモリ容量が限られたコンピュータを用いて高速に音声を合成することやユーザーの要求に応じて合成音声の韻律や声質を変更することは容易ではない。そこで、本論文では、言語処理部－音声特徴量予測部－音声合成部で構成される TTS システムにおいて、言語特徴量と音響特徴量の関係を DNN を用いてモデル化する音声特徴量予測部を頑健かつ高速に動作させるために、音声特徴量予測処理の DNN の構成を検討し、DNN に外れ値が入力しない言語特徴量の正規化法を考案するとともに、音声特徴量予測部で用いる、新たな損失関数を用いた DNN の学習法と生成的敵対的ネットワークによる DNN の学習法を考案し、それらの性能を評価している。

第 1 章は序論で、音声合成技術の歴史と現状を概観して取り組む課題を明らかにし、本研究の目的について述べている。

第 2 章では、本研究で用いた TTS システムの構成、DNN を用いた音声特徴量予測処理の詳細、および音声コーパスについて述べている。

第 3 章では、DNN を用いた音声特徴量の予測処理時間を短縮する方法について述べている。まず、音声特徴量を予測する DNN および後処理にかかる演算時間を比較し、再帰型の DNN である RNN（Recurrent Neural Network）処理、順伝搬型の DNN である FFNN（Feed-Forward Neural Network）を用いた音声特徴量予測処理において合成音声の品質を確保するために必要な MLPG（Maximum Likelihood Parameter Generation）処理、RNN あるいは FFNN を用いた音声特徴量予測処理において合成音声の品質を確保するために必要なケプストラム強調処理、の演算時間が長いことを示している。そして、後処理を用いずに FFNN だけで音声特徴量予測部を構成することにより、言語特徴量から音声特徴量を少ない記憶容量で高速に算出できることを示している。

第 4 章では、 DNN が外れ値に対して脆弱である問題の解決法について述べている。すなわち、音声特徴量予測部において、一文の言語特徴量時系列ベクトル内の二つの属性値の比を取る正規化法を考案するとともに、文章構造と密接に関係する基本周波数の予測誤差と、聴取実験によって求めた合成音声の韻律の主観評価値を、提案法と従来法で比較している。その結果、提案した正規化法は、学習データが少ない場合でも言語特徴量の各属性と基本周波数の関連性を保ったまま外れ値を発生させずに言語特徴量を正規化でき、合成した音声の韻律の主観評価値は従来法の主観評価値と比べて有意に高いことを示している。この結果は、DNN を用いて基本周波数を頑健に予測できることを示している。

第 5 章では、従来の FFNN が時間フレームごとに音声特徴量をモデル化しているために、FFNN だ

けで構成した音声特徴量予測部を用いると合成音声の品質が低下するという問題の解法について述べている。すなわち、生成モデルを FFNN だけで学習させるための、音声特徴量時系列の複数の属性を先験知識を利用して組み合わせた MATS（Multiple Attributes of Temporal Sequence）損失関数を提案し、提案法と従来法で学習した FFNN で予測した基本周波数とメルケプストラムの予測誤差と、合成音声の韻律と音質の主観評価値を比較している。その結果、MATS 損失関数を用いて学習した FFNN は、後処理を伴う従来法と同等以上の知覚的に優れた音声特徴量を予測できることを示している。この結果は、合成音声の音質を低下させずに、第 3 章で述べた FFNN だけで構成される高速な音声特徴量予測部を実現できることを示している。

第 6 章では、生成的敵対ネットワーク（GAN：Generative Adversarial Network）を用いて、先験知識無しに、音声特徴量の時間構造と次元間の関係を考慮した生成モデルを FFNN に獲得させる学習法について述べている。すなわち、時間フレームごとの音声特徴量の生成誤差と音声特徴量のグラム行列の識別誤差を用いて音声特徴量時系列の相関関係を考慮した GAN による DNN の学習法を提案するとともに、その性能を従来法と比較評価している。その結果、提案した GAN による学習法が、音声特徴量の時間構造や次元間の関係を考慮した生成モデルを自動的に FFNN で学習することを可能にし、従来法と同等以上の知覚的に優れた音声を合成するメルケプストラムを予測できることを示している。この結果は、先験知識を用いずに、第 3 章で述べた FFNN だけで構成される高速な音声特徴量予測部を実現できることを示している。

第 7 章では、本論文の結論を述べている。

以上、本研究では、計算資源が限られたテキスト音声合成システムにおいて、DNN を用いて言語特徴量から音声特徴量を頑健かつ高速に予測する方法を提案し、提案法が有効であることを示している。本研究で提案された音声特徴量予測法を用いることにより、計算資源が十分にないシステムにおいても、DNN を用いた統計モデル方式のテキスト音声合成技術を利用できるようになることが期待される。本研究は、その研究手法および結果において独創性が認められ、結果の解釈も適切である。本研究の結果は工学的観点から有効であり、音声工学の発展に寄与すると評価される。本論文に関連する学術論文誌掲載論文は 2 編で、いずれも申請者が筆頭著者である。

令和 3 年 1 月 27 日に博士論文の審査および最終試験を実施した結果、申請者は当該分野に関する十分な全般的知識を有し、学術研究にふさわしい討論ができ、かつ独立して研究を遂行する能力を有すると判断され、本論文は博士（工学）の学位論文として合格であると認められた。