氏　　名	井上　勝喜
授与した学位	博　士
専攻分野の名称	統合科学
学位授与番号	博甲第　　６６６５　　号
学位授与の日付	２０２２年　3月　２５日
学位授与の要件	ヘルスシステム統合科学研究科　　ヘルスシステム統合科学専攻 （学位規則第４条第１項該当）
学位論文の題目	深層学習に基づく感情音声合成のための少量データを用いた学習方式の研究
論文審査委員	教授　横平　徳美　　教授　五福　明夫　　教授　髙橋　規一　　教授　阿部　匡伸

学位論文内容の要旨

深層学習に基づく音声合成は，医療分野やエンターテインメントの分野などで幅広く利用されている。音声合成システムの恩恵をさらに享受するためには，感情表現の再現が重要となる。従来の音声合成では数十時間の学習用データを必要とするが，一般的な話者にとって長時間に及ぶ感情音声データの収録は困難である。このため，少量の感情音声データから深層学習モデルを構築することは重要な課題である。

本研究では，合成目標となる話者の音声データが最大でも数十分しかない場合を考える。数十分の音声では，必要とされる量の 20 分の 1 以下であり，深層学習モデルの構築は不可能である。この問題を解決するために，データ量の不足を異なる話者の音声で補完させる学習方式を用いる。これにより，合成目標となる話者の少量データを用いて深層学習に基づく感情音声合成を実現させることを目的とする。構築した音声合成の深層学習モデルにより，少量データを用いた感情音声合成システムの実現が可能であることを示した。

はじめに，Deep Neural Network (DNN) 音声合成における感情外挿方式を検討した。ここでは，平静音声しか収録していない話者を目標話者とする。外挿とは，ある既知の数値データを基にして，そのデータの範囲の外側で予想される数値を求めることである。目標話者の話者性を持つ合成音声に対して，他の話者が発声した少量の感情音声から抽出した感情表現を付与することで実現する。主観評価実験から，他の話者の感情表現を付与することにより目標話者の話者性をもつ感情音声を生成可能であることが示された。

次に，sequence-to-sequence (seq2seq) 音声合成における半教師あり二段階感情適応方式を検討した。ここでは，少量の感情音声を発声した話者を目標話者とする。半教師あり学習とは，学習データの一部を自身または別のモデルにより生成しつつ学習する枠組みである。また，二段階感情適応とは，大量データで構築した事前学習モデルに対し，話者性と感情表現を順次適応する方式である。まず，大量データを収録した話者の音声で事前にモデルを学習させる。その後，自動音声認識から得られるテキストを用いて目標話者の感情音声に二段階でモデルを fine-tuning することで実現する。半教師あり話者適応に関する主観評価実験から，自動音声認識から得たテキストを用いたとしても，人手によるテキストを用いた場合と同等の話者性を再現できることを示した。また，二段階感情適応に関する主観評価から，目標話者の話者性への適応を経由することで，事前学習モデルから直接適応する場合に比べ，感情音声の再現性が向上することが示された。

論文審査結果の要旨

本論文は，テキストからの音声合成(Text-to-Speech: TTS)において感情音声が合成できる TTS 方式を提案している。TTS は文章(テキスト)を入力して音声を出力するシステムであり，ニュース記事やメールの読み上げ等に利用されている。しかしながら，様々な声質での音声合成，場面に応じた発話スタイルの使い分け，感情の表出，などは実現できているとは言い難い。これらの機能は，人間とシステムが音声でやり取りする音声対話では必須であり，スマートフォン，スマートスピーカー，人間型ロボットなどの普及とともに要望が高まってきている。本論文は，このような背景から TTS の応用範囲を拡大することを目的としており，社会の要請を適切に捉えている。

提案方式のポイントは，目的とする話者の少量の音声データを用いて感情音声合成を実現する点にある。近年の TTS の品質の改善は著しく，そのブレイクスルーの鍵の１つは深層学習の適用にある。しかしながら，一般に深層学習は大量なデータを用いて学習するため，大量データの収集がシステム開発のボトルネックになるケースが多い。提案方式は，この点を解決するために２つの方式を提案している。

第 1 の方式は，ある話者の感情音声データから感情表現モデルを構築し，その感情表現モデルを別の話者に適用する。これは，声優やナレータのように感情表出に秀でた話者の感情表現を一般人に適用できるとともに，一般人の感情音声を収録する必要はない。評価実験の結果から，提案方式の有効性が示された。

第 2 の方式は，大量な音声データとそれに対応するテキストデータを用いて音声認識モデルと音声合成モデルを構築し，このモデルを利用して少量の音声データしかない話者の感情音声合成を実現する。少量の音声データに対応するテキストデータは不要である。評価実験の結果，目的とする話者のデータが深層学習によって TTS を構築できないほど少量であっても，目的とする話者の話者性と感情表現が実現できることが示された。

以上のように本論文は，TTS が求められている課題を的確に捉えていると共に，最も有望な深層学習の枠組みを利用しながら，ボトルネックとなるデータ量の問題を解決する方式を提案し，その有効性を明らかにしている。

学位審査会においては，上記の博士論文の内容が要約されて発表されており，質疑応答も適切に行うことができた。

以上のことを考慮し，本学位審査会は，井上勝喜氏が，博士の学位が授与されるのに相応しいと判断する。