深層学習を用いたデータ駆動型調音・音声間変換に関する研究

田口, 史朗

（様式６－２）

氏　　名	田口 史朗
論　文　名	深層学習を用いたデータ駆動型調音・音声間変換に関する研究
論文調査委員	主　査　　九州大学　　教授　　鏑木 時彦 副　査　　九州大学　　准教授　吉永 幸靖 副　査　　長崎大学　　教授　　松永 昭一

論 文 審 査 の 結 果 の 要 旨

深い階層構造を有するディープニューラルネットワークは，音声情報処理においても有効であり，音声合成，音声認識，言語処理など幅広い分野で性能の改善が実現され，有効性が確かめられている。本研究は，音声の特徴を表現するパラメータとして，音声生成のプロセスに由来し生理学的な意味を持つ調音運動に着目し，この特徴パラメータと音声信号の間の相互変換を，ディープニューラルネットワークによって構築することを目的としたものである。このような調音・音声間変換は，人の音声コミュニケーションの理解に役立ち，代用発声法などの福祉的応用が可能であり，幅広い発展性が見込まれる。

しかしながら，このような調音・音声間変換の実現は容易ではなく，いくつもの問題をクリアしなければならない。まず１点目に，舌，唇，下顎，軟口蓋などの調音器官の運動パタンと，音声のスペクトル包絡特性の間の関係は，高度に非線形である。特に，音声から発話時の調音運動を推定する逆推定では，解の非一意性に由来した困難さが同時に生じる。2 点目に，音声を生成するには一般に声道フィルタのスペクトル特性とそのフィルタを駆動する音源信号とが必要となるが，調音運動とこの音源情報との間には，直接的な因果関係がほとんど存在しないことである。これらの問題に対処するには，深い階層構造を有するニューラルネットワークによって非線形な写像関係を表現し，さらに入力された調音運動の中長期的な時系列パタンから音声を得る必要がある。従って，多量のパラメールを内包するニューラルネットワークの学習のため，音声と調音運動を同時に記録した大規模なパラレルデータが必要とされる。3 点目の問題は，このような大規模なデータセットをいかに精度よく構築するかである。

本論文では，これらの問題を解決するための方法が明確かつ詳細に説明されており，高精度な調音運動観測手法である３次元磁気センサや口唇動画を用いた調音・音声パラレルデータの収集，最新の知見を取り入れたディープニューラルネットワークによる調音運動からの音声の生成，および，音声からの調音運動の推定法が提案され，これらの提案法の有効性が客観評価ならびに主観評価実験を通して検証されている。本研究により得られた主要な成果は，以下の通りである。(1)大規模な調音・音声パラレルデータを効果的に収集するため，音素バランスに考慮した文章セットを作成し，磁気センサでは男性話者１名の約１時間のデータ，口唇動画では男女各１名についてそれぞれ約５時間のデータセットを収集することができた。(2)双方向再帰型ニューラルネットワークを用いて，磁気センサデータから音声を合成する変換モデルを構築し，混合ガウスモデルなどの従来法よりも推定精度が優れていることを示した。(3)音声から磁気センサデータを推定する逆変換に関して，残差構造を有する時間遅延ニューラルネットワークを新たに提案し，その有効性を 10

名分のデータを用いて示した。(4)口唇の動きだけから音声を合成するため，畳み込み系列変換モデルを基とした合成モデルを提案し，従来法よりも優れた性能を有することを示した。

本研究の内容は，サイレント音声インターフェースとして近年注目を集めているものであり，日本語音声を対象として，調音・音声間変換を高精度に実現可能であることを示したものである。本研究が特に優れた研究であると判断される理由としては，規模の大きさにおいて随一と言える調音・音声パラレルデータを構築したこと，従来は大多数が試行的なレベルに留まっていた変換モデルを，ディープニューラルネットワークに関する最新の知見を取り入れて構築したこと，さらに，提案した変換モデルの性能評価を，従来法との比較のもとにきわめて精緻にかつ横断的に実施したことが挙げられる。そもそも，音源情報との関係が希薄な調音運動のみから音声を生成できること自体がユニークであり，なおかつ，口唇動画で与えられる情報は唇の動きだけであり，最も重要な調音器官である舌の運動が含まれていない。それにも関わらず，舌に調音点を持つ種々の母音，子音を含めて了解可能な音声が得られることは，代用発声技術への応用可能性を示唆するものであり，福祉音響技術としてのサイレント音声インターフェースの展開や，音声情報処理技術の学術的発展において大きな貢献が認められる。

以上より，学位審査を厳正に実施した結果，本論文は博士（芸術工学）の学位に値するものとして認めた。