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

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

区　　分　：甲

論 文 内 容 の 要 旨

調音・音声間変換は音声の生成および知覚の解明の一助となるのみならず，音声障害者のための代用音声，非母語言語や聴覚障害者の発話訓練など，種々の工学的応用も期待できる．調音・音声間変換には調音器官の運動パタンに関する情報である調音情報と音声を同時に記録したデータを基に，調音情報と音声の間の関係をモデリングするデータ駆動型のアプローチがある．このアプローチは物理モデルによる方法とは異なり，有効な周波数帯域が制限されず，通常発話の音声に対しても適用できる利点がある．また，近年では深層学習が様々な問題に対して有効であることが報告されつつある．調音・音声間変換に関しても例外ではないが，初期的な手法を導入するにとどまっているのが現状である．そこで，本研究では，大規模日本語調音・音声パラレルデータの収集，調音・音声間変換における音源情報の活用，深層学習の導入による調音・音声間変換の精度向上，実応用のための新たな調音情報の検討という4つのアプローチを通じて，調音・音声間変換のさらなる発展を目指す．

まず，データ駆動型調音・音声間変換を実現するための日本語調音・音声パラレルコーパスの収集を行った．既存の音素バランス文に加えて，一定の総モーラ数の中で，できるだけ多様なトライフォンが登場するように Minoux の改良貪欲法によって文選択を行った．得られた音素バランス文を既存のものと同時に用いることで，2回以上出現するトライフォンの種類数が 1.7 倍となり，より多様な音素文脈の収録が可能となった．この音素バランス文をもとに，磁気センサと口唇動画の収録を行った．磁気センサでは，日本語男性話者1名の約1時間のデータ，口唇動画では日本語話者男女1名ずつの各 4.8 時間のデータとなり，日本語として唯一であるだけではなく，英語の公開データベースと比較しても，話者あたりの発話時間が比較的長期となるデータを収集することができた．

続いて，磁気センサによる調音情報から音声を得る調音―音声順変換の検討を行った．ここでは，双方向再帰型ニューラルネットワークによって，声道のフィルタ特性だけではなく音源情報も推定する調音―音声変換を新たに提案した．先行研究と同一のコーパスを用いて比較を行ったところ，客観評価では提案法は多くの音声特徴量の推定誤差を改善できることがわかった．さらに，主観評価においては提案法の自然性に関する MOS 評点が 3.115 となり，先行研究と比べてより自然性の高い音声を得られる可能性が示唆された．

さらに，磁気センサによる調音情報を音声から得る音声―調音逆変換の検討を行った．本研究では新たに Residual Time-Delay Neural Network による音声―調音逆変換法を提案し，公開されている英語コーパスと本研究で収集した日本語データの計 10 名に関しての磁気センサデータを評価に用いた．各話者に関して話者依存の逆変換モデルをそれぞれ構築した上で先行研究との比較を行ったところ，すべての話者に関して提案法が最も調音軌道の推定精度が高いという結果が得られた．また，提案法はモデルの層数と学習データ量に対して頑健であることがわかった．さらに，提案法をもとに，音声―調音逆

変換に最適な特徴量を検討したところ，メル周波数ケプストラム係数が最も推定精度が良いこと，音源フィルタ分離は音声―調音逆変換に有効ではないこと，メルフィルタバンクと離散コサイン変換による特徴量の次元削減は限られたデータ量を扱う音声―調音逆変換には有効であることが示された．

最後に，口唇動画による調音情報から音声を得る調音―音声順変換の検討を行った．本研究では新たに畳み込み系列変換モデルを基とした調音―音声変換を提案し，本研究で収集した男女 2 名の日本語話者のデータ用いて，客観評価指標により評価を行った．まず，より良いメルスペクトログラム時系列を得るための改善法に関して検討を行ったところ，動的特徴を考慮した損失，Self-Attention の導入，Scheduled sampling による入力特徴量の劣化が有効であった一方，音源情報に関するマルチタスク学習は結果に悪影響を及ぼすことがわかった．改善した手法をすべて取り入れることで，合成音声の音素誤り率(PER)が 28.51%から 13.02%に改善された．また，ネットワーク構造に関して先行研究と提案法を比較した結果，特に PER に関して 4 から 10 ポイントの改善が見られ，提案したネットワーク構造がより了解性の高い音声を合成できることを示した．最後に，1 つのネットワークを複数話者のデータで学習する複数話者口唇動画―音声変換について予備的な検討を行ったところ，話者性や発話リズムなど大局的な特徴はよく表現できたが，音韻性の再現性という点に関しては PER が 9.22%から 15.68%に低下し，話者依存モデルよりも大きく劣化したため，さらなる検討が必要である結果となった．