論 文 の 内 容 の 要 旨

論文題目　　話者空間基底を用いた特徴量分解とそれに基づく
パラレルデータフリーDNN型声質変換

氏　　名　　橋本　哲弥

声質変換はある話者の音声を，その言語情報を損なうことなく，発話者や発話様式などの声質のみを変換する技術であり，テキスト音声合成など多くのアプリケーションに応用されている．声質変換は，変換元の話者（入力話者）の音声と変換先の話者（出力話者）の音声に対して特徴量空間上でのマッピングを構築するタスクと考えることができ，これまでにも，ガウス混合モデル (GMM)，ディープニューラルネットワーク (DNN)を用いた手法や，非負値行列分解を用いた事例ベースアプローチなどが研究されている．

従来の統計的変換モデルを学習する際には，一般的に入出力話者による同一文の読み上げ音声データ（パラレルデータ）が必要となる．しかし，パラレルデータを使用した変換モデルの問題点として，学習データの発声内容が大きく制限される点や，学習を行った特定の話者間にしか変換モデルを適用できないという点が挙げられる．データ効率や実用面を考えると，発声内容に制限の無い，どのような音声データでも学習に使用できるパラレルデータフリー声質変換システムが望ましい．

任意の入出力話者に対して少量の学習データで適応可能な声質変換は多く研究されており，GMMにおいてはパラメータ適応による多対多声質変換が提案されている．しかしDNNを用いた多対多声質変換は，モデル中の各パラメータが表す情報が明示的で無いために，GMMと比べて数が少ない．一方でDNNを用いた声質変換はGMMを用いた声質変換よりも精度が高いという研究結果も存在している．

そこで，本研究では高精度かつ柔軟な声質変換を実現するために，DNN声質変換に対して多対多変換を可能とする手法を目的としている．

本稿では，初めにDNNによる柔軟かつ高精度な変換の足がかりとして，複数のサブネットワークを用いたDNN多対一声質変換を提案した．この手法では，DNNによる声質変換では浅い層では特徴量抽出などの入出力話者に依存しない処理が集中しており，深い層では入出力話者に依存した処理が行われているという仮定を用いている．この仮定を基に，DNNを前半と後半で区切り，後半のネットワークを変換先の話者によって分岐させた上で学習を行った．これにより，前半の層には話者に非依存な処理が，後半の沿うには話者に依存した処理が明示的に集中し，前半の層の汎化性能が向上することが期待される．実験の結果，pre-trainingに複数話者を用いることによる変換精度の向上と，提

案手法による学習データ中の話者と未知話者の両方の入力に対する変換精度の向上を確認し，提案手法が入力話者に対して柔軟な変換が可能であることを示した．

一方で，ここで提案したサブネットワークを用いた変換手法は，入力に関しては適応を行うことなく未知話者を用いた変換が可能であるものの，出力に関しては未知話者に関して適応を行うことができないという問題があった．

そこで，より柔軟かつ高精度な多対多声質変換を目的として，固有声GMM（EVGMM）とDNNを組み合わせた話者空間基底を用いた多対多声質変換手法を提案した． 初めにEVGMMのパラメータを用いて，全事前収録話者の特徴量を「平均話者」と「話者基底成分」に相当する特徴量へと変換する．この「平均話者」と「話者基底成分」に相当する特徴量と，元々の特徴量の組み合わせを擬似的なパラレルコーパスの様に使用することで，DNNによって元々の特徴量から「平均話者」と「話者基底成分」への分解を行うような変換を学習する．最終的な出力特徴量はこの分解された特徴量と，出力話者固有の重みの積によって表現される．任意話者変換を扱う場合の未知出力話者の重みに関しても，教師無し適応によって求めることができる．

他の多対多声質変換手法との比較実験から，客観評価においては最も高い性能が得られ，手法の有用性が示された．自然性と個人性を評価する主観評価においては一部他の多対多変換手法を僅かに下回る結果が見られたものの，それらはほぼ同程度の結果であり，学習データ量が多い場合の自然性においては最も良い結果が得られた．

加えて，提案手法の拡張として，パラレルデータフリーな共分散推定手法を導入することで全学習過程においてパラレルデータフリーな多対多声質変換手法を提案した．この手法に関して，教師あり学習を行ったGMMとの比較を行った結果，客観評価においては学習データ量が少ない場合でも提案手法がGMMを僅かに下回っていると言う結果が得られた．その一方で，主観評価においては，学習データ量が少ない場合には自然性と個人性の両尺度において提案手法がGMMを上回っていた．