ISS-SP-217

Auto Encoder による声質変換のための特徴量抽出および変換

蓮沼　勇太†　　長尾　智晴††
† 横浜国立大学 理工学部　　†† 横浜国立大学 大学院環境情報研究院

1. はじめに

従来の声質変換では，音声からメルケプストラムや包絡線といった特徴量を計算し変換を行うが，話者の発話内容によらない声質を取り出すことはできていない．そこで本研究では特徴量抽出手法として近年注目される Auto Encoder(AE)による声質変換のための特徴量抽出および変換を提案する．さらに，言葉によらない声質を表す特徴量を抽出することで声質の変換をわかりやすいものにする．

2. 提案手法

提案手法は AE による声質の特徴量抽出部分と変換部分に分かれる．固有部分と共通部分を話者ごとの AE で学習することで固有の声質と共通の言葉の特徴を獲得する．

2-1. 特徴量抽出

図 1 に特徴量抽出の概要を示す.複数話者のケプストラムを同時に学習する AE を構成する．ノード数を少なくしていくことで入力の情報を圧縮したものが中間層に現れる．最後の層を学習する際，話者ごとに別々の AE を学習し，その後ノードを追加して学習する．このとき追加前の結合荷重は更新せず，追加した結合荷重の更新は各 AE で共有する．このように学習することで各 AE には固有部分と共通部分ができ，それぞれ音声の固有の声質と共通の言葉の特徴を獲得する．

2-2. 特徴量変換

2-1 で得られた AE に対し，固有部分を入れ替えることによって声質の変換を行う．

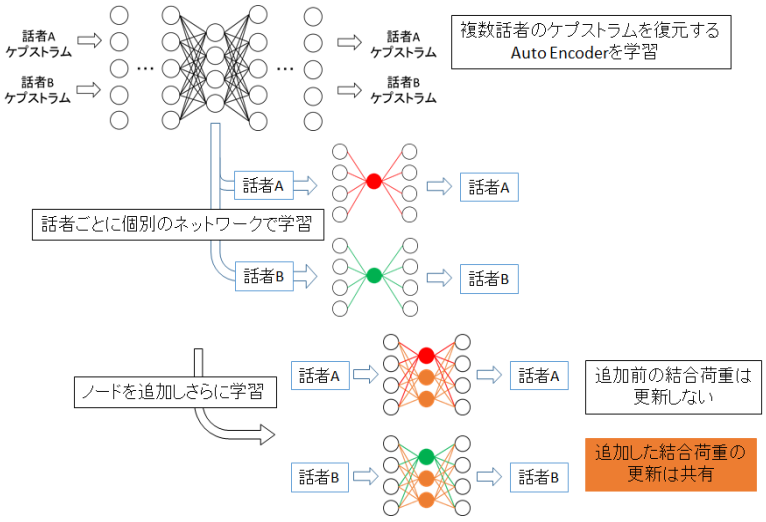

図 1. 特徴量抽出の概要

3. 実験条件

提案手法による声質変換の実験を行った．日本語話者の男性，女性各 2 名の単音(/a/，/i/，/o/)の音声データ(16kHz, 16bit)から FFT ポイント 1024，シフト 5ms のケプストラムを計算し，それぞれ 100 フレームを学習データとした．0 次を除いたケプストラム 512 次元を AE の入力とした．最後の層を学習する際，中間層のノード 1 つで学習し，その後ノードを 10 個追加して再学習した．音声復元の際には，ケプストラムの 0 次と位相情報は変換前のものをそのまま用いた．

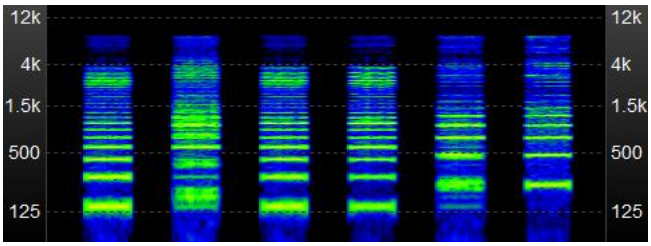

図 2. (左から順に)固有部分のみで男性を復元，共有部分のみで男性を復元，全部で男性を復元，男性原音声，男性から女性へ変換，女性原音声のスペクトログラム(すべて/a/の音声)

4. 実験結果

図 2 に実験で得られた音声のスペクトログラムの例を示す．固有部分のみではどの音声でも言語がはっきりとしない音声になり，言語によらない話者の特徴量が得られた．共通部分のみでは言語情報はあるものの話者が混ざったような機械的な音声が得られた．全てのノードを使うことで元の音声がよく復元できている．固定部分のみを交換したところ，もう一方の話者の音声に近づいているものの不自然さは残っていた．これは追加部分にも声質の特徴が含まれているためであると考えられる．

5. まとめ

AE による話者ごと音声の特徴量抽出および変換する手法を提案した．AE によって言葉によらない話者の音声の特徴が得られた．今後は，音声から声質と言語の分離をよりよく行うため，学習人物の増加，AE の学習法の検討を行っていきたい．

参考文献

[1] Seyed Hamidreza Mohammadi and Alexander Kain, “Voice conversion using deep neural networks with speaker-independent pre-training”, In Spoken Language Technology Workshop (SLT) IEEE,　pp. 19-23, 2014