5Q-04

敵対的生成ネットワークを用いた非パラレル声質変換

蓮沼 勇太 †　　長尾 智晴 ‡

† 横浜国立大学大学院環境情報学府　　‡ 横浜国立大学大学院環境情報研究院

1　はじめに

二者の異なる文章を発した音声データを学習データとする非パラレル声質変換において, 従来ではメルケプストラムまたはスペクトル包絡といった単一の音響特徴量のみを変換していた. メルケプストラムは次元が小さく変換の学習が容易であるが情報が圧縮されているため変換音声の自然性が損なわれる. 一方, スペクトル包絡はメルケプストラムよりも情報が多く, 変換音声の自然性は高められるが, 学習が容易でなく類似性を高めることが難しい. そこで本研究ではこれらを同時に変換し統合することで両者の長所を活かした, より品質の高い声質変換手法を提案する.

2　提案手法

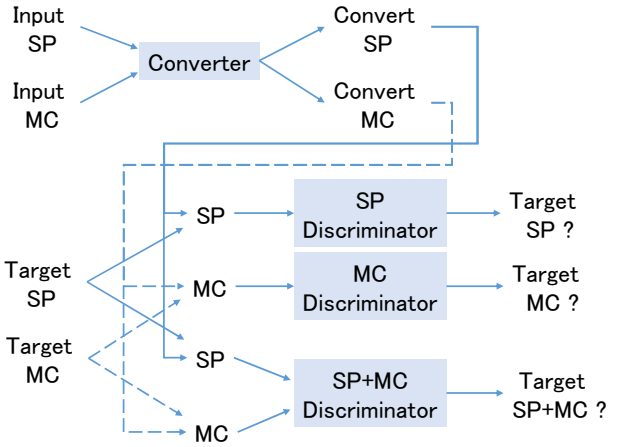
提案手法は敵対的学習によって特徴量を変換するネットワークを最適化する. 以下に概要を示す.

2.1　敵対的学習によるネットワークの最適化

図 1 に提案手法におけるネットワークの最適化の概要を示す. 提案手法ではメルケプストラム・スペクトル包絡の二種の音響特徴量を同時に変換する. これらを同時に変換する Converter, それぞれの特徴量が目標話者のものかを判定する Discriminator, 二種の音響特徴量の組み合わせを判定する Discriminator, 計 4 つのネットワークを敵対的学習によって最適化する. 3 つの Discriminator は入力された特徴量を正確に判定するように学習し, Converter は 3 つの Discriminator を騙すような特徴量を生成するように学習する.

2.2　スペクトル包絡へのノイズ付与

最適化の際, スペクトル包絡から極大値・極小値を検出し, 極大値を 1, 極小値を 0 として線形補間する. これを pk とする. その後以下の式 (1) によってスペクトル包絡にノイズを付与する. ここで sp はスペクトル包絡を表す. 図 2 にノイズ付加の結果を示す. これによりノイズは極大値付近は小さく, 極小値付近は大きくなる.



(SP: Spectrum envelope, MC:Mel-cepstrum)

図 1: 提案手法によるネットワーク最適化

こうすることでフォルマント周波数を重視した最適化が行われる.

$$sp_{noise} = sp + \mathcal{N}(0.0, 1.0) \otimes (1.0 - pk) \tag{1}$$

2.3　特徴量の統合

最適化した Converter によって入力話者の二種の音響特徴量を変換しその後統合する. 以下の式 (2) によって統合し新たなスペクトル包絡を得る. ここで, sp_{mc} はメルケプストラムから得られるスペクトル包絡を表す. 図 3 に統合結果を示す.　図 3 にあるようにメルケプストラムから得られるスペクトル包絡は低周波が細かく, 高周波では粗くなっている. 一方スペクトル包絡は高周波で細かくなっている. よって, 低周波ではメルケプストラム, 高周波ではスペクトル包絡の影響が大きくなるように両者のスペクトル包絡を統合し, 新たなスペクトル包絡を生成する. これにより両者の情報を補間しあうようなスペクトル包絡が生成される.

$$sp_{result} = sp \otimes (1.0 - c_{mel}) + sp_{mc} \otimes c_{mel} \tag{2}$$

$$c_{mel}(Hz) = \frac{d}{dHz} mel(Hz) \cdot 0.5$$

$$mel(Hz) = 1127.010480 \cdot \ln \frac{Hz}{700} + 1.0$$

Non-parallel voice conversion using generative adversarial networks
†Yuta hasunuma　‡Tomoharu Nagao
†Graduate School of Environment and Information Sciences, Yokohama National University
‡Faculty of Environment and Information Sciences, Yokohama National University

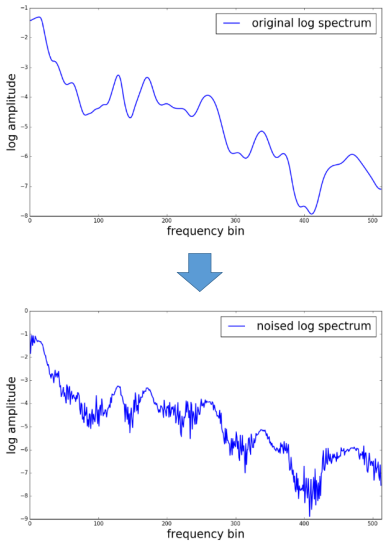


図 2: スペクトル包絡へのノイズ付加

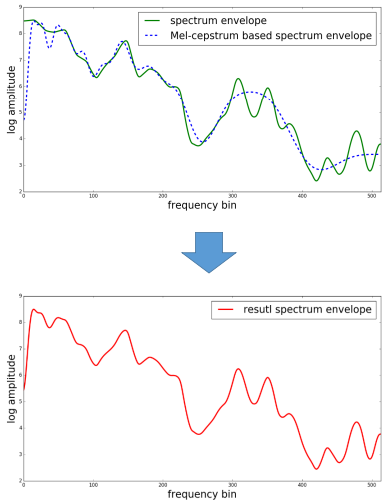


図 3: メルケプストラムとスペクトル包絡の統合

3 実験条件

Voice Conversion Challenge 2016[1] に収録されている英文を発した音声データ (サンプリング周波数:16kHz, 量子化ビット数:16bit) を用いて実験を行った. 入力話者・目標話者それぞれ 150 文を学習データ, 54 文をテストデータとした. 音響特徴量として 25 次元のメルケプストラム・513 次元のスペクトル包絡の連続 5 フレーム分を用いた. 4 つのネットワークは Convolutional Neural Network によって構成した. 基本周波数は従来と同様に logarithm Gaussian normalized transformation によって変換した.

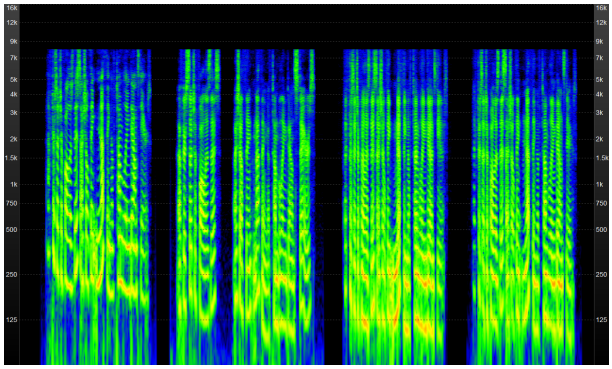


図 4: 実験結果 (左から入力音声, 目標音声, 従来手法による変換音声, 提案手法による変換音声のスペクトログラム)

4 実験結果

図 4 に入力音声・目標音声・従来手法 [2] による変換音声・提案手法による変換音声のスペクトログラムを示す. 入力話者として女性 (SF1), 目標話者として男性 (TM3) の音声を使用した. スペクトログラムにあるように従来手法は目標音声と比べ低周波成分が余分に強くなっているが, 提案手法では抑えられている. 音声を再生したところ従来手法よりも類似性は同程度だが, 自然性が向上し, 変換精度の向上がみられた.

5 まとめ

敵対的生成ネットワークを用いて二種の音響特徴量を変換し, 統合する声質変換手法を提案した. また, 英語話者の音声データを用いて声質変換の実験を行った. 今後はさらなる精度向上のため, 特徴量の統合方法やネットワークの構造の検討を行っていく.

参考文献

[1] Tomoki Toda *et al.*, “The Voice Conversion Challenge 2016 ”, INTERSPEECH 2016

[2] Chin-Cheng Hsu *et al.*, “ Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks ”, INTERSPEECH 2017