畳込みニューラルネットワークを用いた音響特徴量変換とスペクトログラム高精細化による声質変換

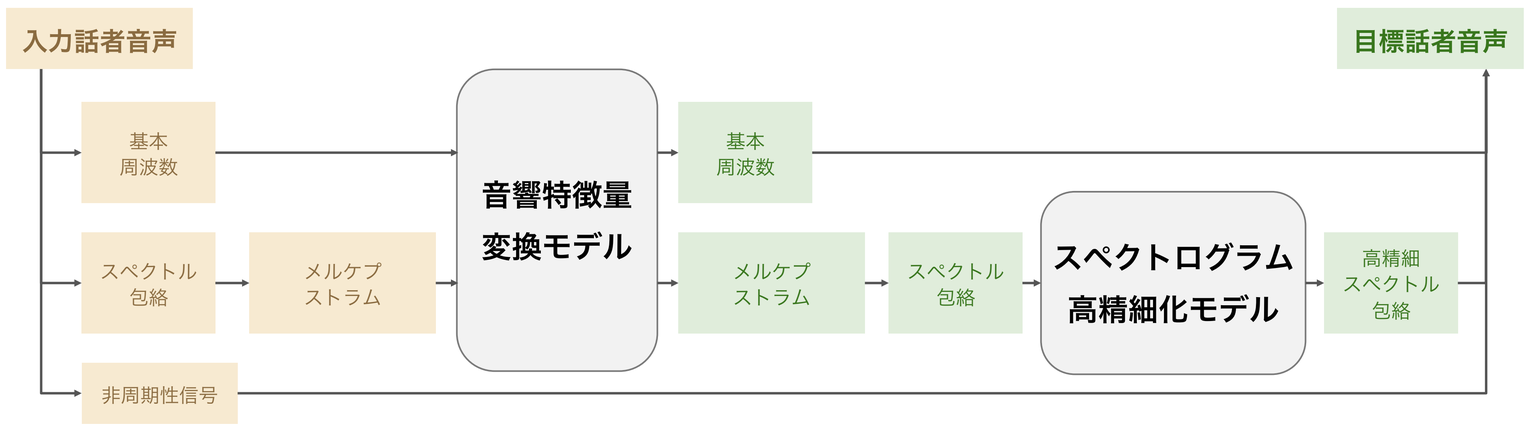
廣芝 和之 *1, 能勢 隆 *2, 宮本 颯 *2, 伊藤 彰則 *2, 小田桐 優理 *1

*1 株式会社ドワンゴ　*2 東北大学

本研究の GitHub リポジトリ

概要

変換と高精細化を異なるモデルとして学習することで、目標話者のみの音声データ数に応じて品質が向上する手法を提案

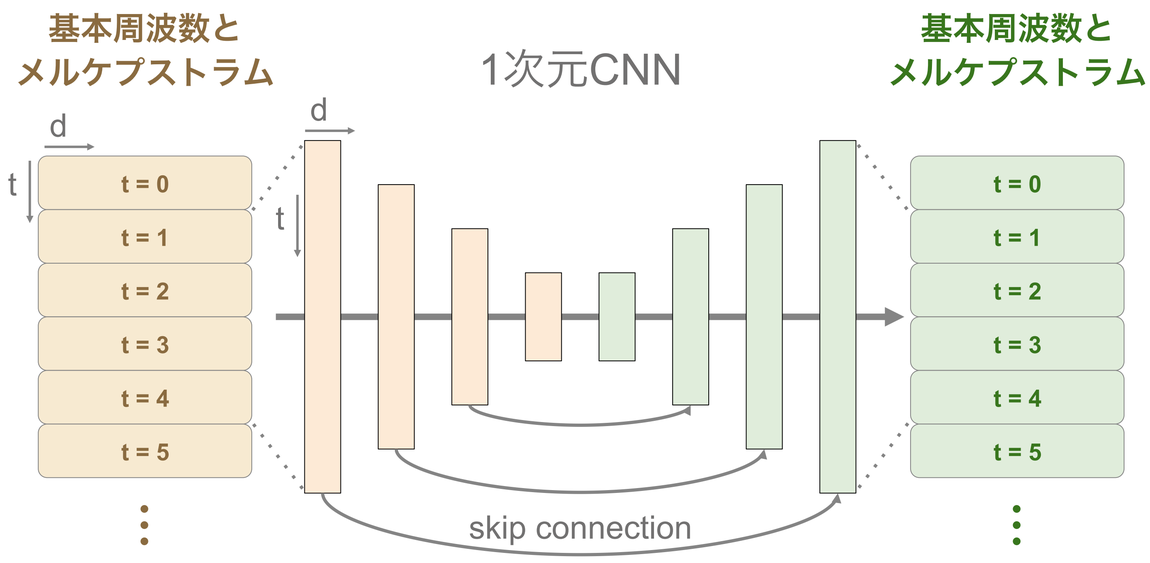


提案手法
変換と高精細化の２段階に分解．スペクトログラム高精細化は目標話者のデータのみで学習可能

変換モデルのみの学習
精度を上げるためには大量のパラレルデータが必要

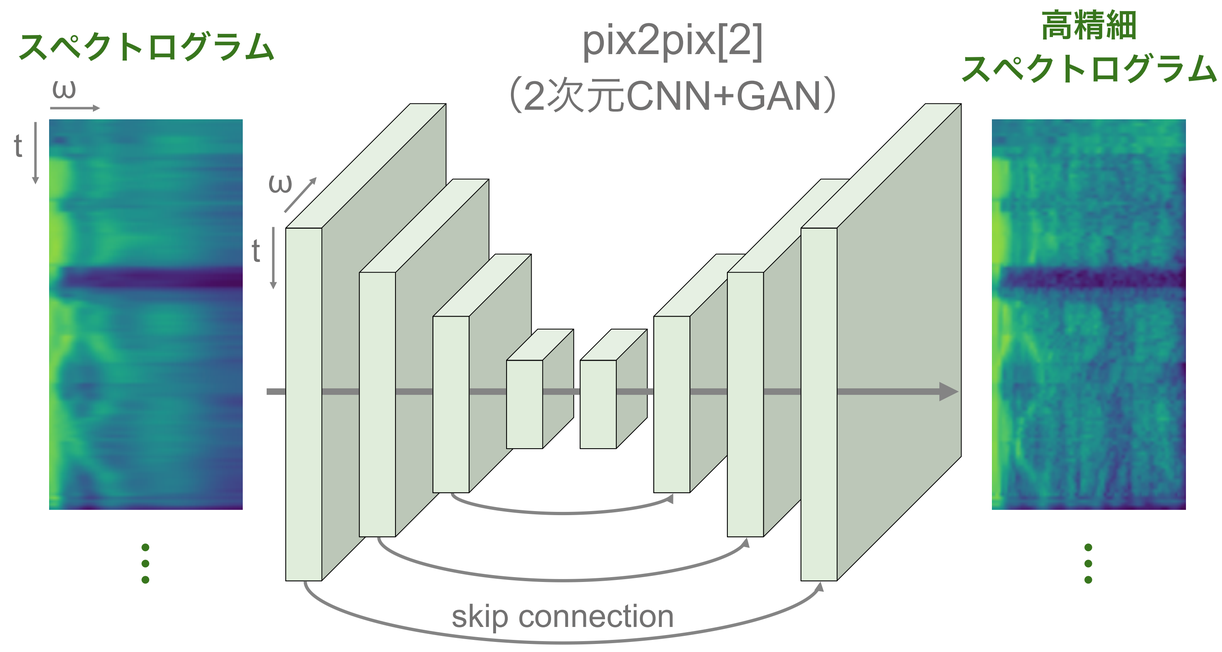
方法

音響特徴量変換モデル



- 1次元CNNを用いることで，フレームごとではなくシーケンスごとに変換する．
- 特徴量をチャンネルとみなし，時間方向に畳み込む．
- U-Net構造[1]（中間層同士にスキップ接続）にすることで，細部まで高精細な出力を生成する．

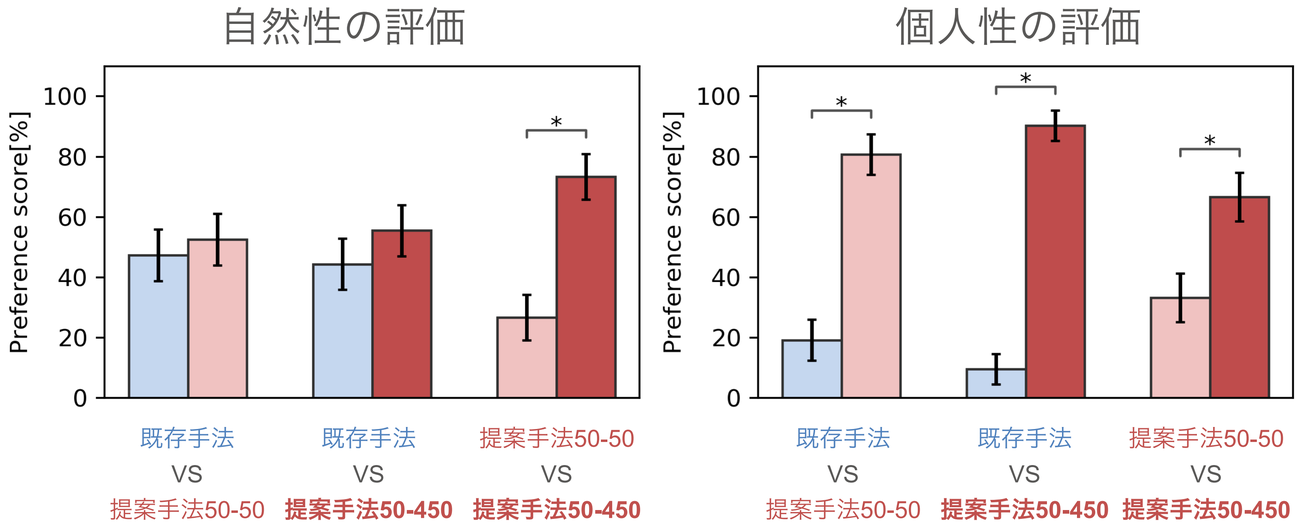
スペクトログラム高精細化モデル



- 多層2次元CNNと敵対的生成ネットワーク（GAN）を利用したpix2pixモデル[2]を用いることで，時間・周波数方向に大局的な特徴を捉えつつ変換結果を高精細化する．
- 学習には，目標話者の音声から得たスペクトログラムと，低次元のメルケプストラムを介して再構成したスペクトログラムのペアを用いる．

実験

評価実験により提案手法と既存手法を比較した．既存手法には差分スペクトルを用いたGMMベースの手法[3]を用いた．ATR日本語音声データベースセットBのMHT男声話者とFKN女声話者の音声データを用いて，MHT話者からFKN話者に声質変換した．音声データの発話内容は同一の音素バランス文503文である．本実験で用いる音響特徴量はWORLD分析[4]により得た．50，450文を学習に，残り53文を評価に用いた．スペクトル包絡は513次元，メルケプストラムは9次元とした．



- 提案手法の変換音声は，従来手法の変換結果に比べて，同程度の自然性と，高い個人性を持つことがわかった．
- 学習に用いる目標話者の音声データ数を増やすことで，より自然性・個人性の高い変換音声が得られた．

参考文献

[1] Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, MICCAI (2015).
[2] Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A.: Image-to-image translation with conditional adversarial networks, CVPR (2017).
[3] Kobayashi, K., Toda, T. and Nakamura, S.: F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential, IEEE SLT (2016).
[4] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE T INF (2016).