AutoVCを用いたZeroshotリアルタイム声質変換

金沢工業大学
4EP3-77 1095117 鈴木大志

1. 序論

背景

近年、深層学習を用いたゼロショット声質変換という、学習データに用いたか否かを問わず、どのような声質間でも変換が可能な深層学習モデルが開発されている。

問題点

現在のゼロショット声質変換の技術では変換に多くの時間と計算資源を用いるために、配信などのリアルタイム性のある用途に用いることが困難である。また、リアルタイムが可能な場合でも、高性能なGPUなどが必要で、一般的な利用は困難である。

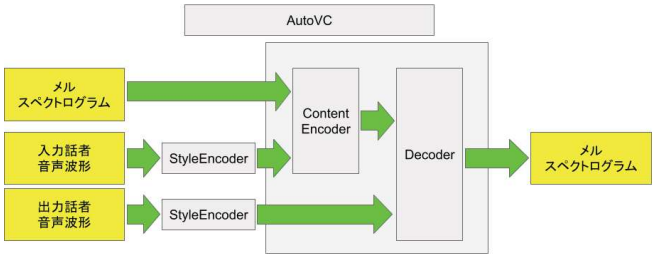
解決法

そこで本研究では、上記のような問題点が発生するのは位相推定の処理が困難であるというのが原因であると考え、位相推定を用いない特徴量を利用した音声変換を行いリアルタイム化をしたいと考える。
※ここでいうリアルタイムとは、入力音声が入力音声の秒数以下で変換が可能であることを指す。

2. 既存手法について

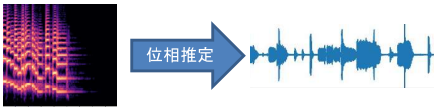
AutoVCについて

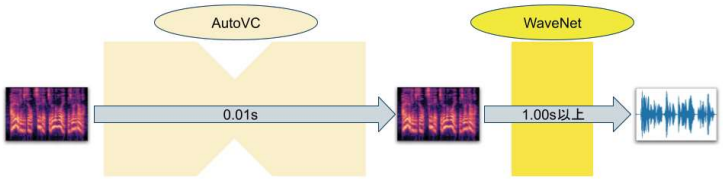
AutoVC[1]は、音声変換に特化したオートエンコーダーを用いた声質変換モデルで、初めてゼロショット音声変換を実現したモデルだ。入力には入力話者の声のメルスペクトログラム、入力話者の声質を表す埋込ベクトル、出力話者の声質を表す埋込ベクトルを用いる。すると出力として、目的の話者の声のメルスペクトログラムが出力される。



メルスペクトログラムから音声への変換

メルスペクトログラムとは、入力音声をSTFT(短時間フーリエ変換)してできる、スペクトログラムの振幅のみを、メルフィルタバンクという行列とかけることで得られる、人間の耳の特性を考慮した特徴量である。
この特徴量は位相情報を含まないため、位相推定によって復元をする必要がある。その方法として、Griffin-Lim法[2]があるが、リアルタイムが困難である上に、低品質な音声になってしまう。そこでAutoVC[1]ではWaveNet[3]を用いることで高品質な変換を実現している。





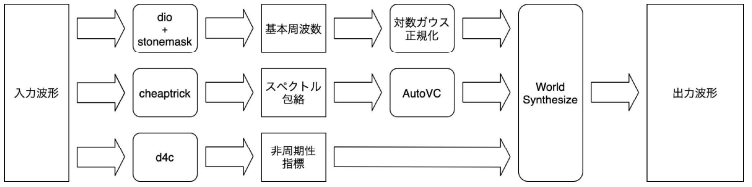
初のZeroShotリアルタイム声質変換であるConVoice[4]ではWaveNet[3]の改良版であるWaveGlow[5]を用いている。

既存手法の問題点

既存手法の多くでリアルタイム化が達成できない理由として、位相推定が困難で多くの計算資源や時間がかかってしまうことが理由であると考える。ConVoice[4]の場合リアルタイムは達成できているものの、高性能なGPUを利用していることが前提となっている。

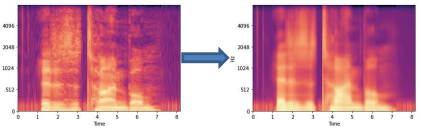
3. 提案手法について

既存手法では、位相推定の計算が困難であることが原因で、リアルタイム化が難しいことがわかった。そこでWorld[6]で提案された手法(音声波形を基本周波数、スペクトル包絡、非周期性指標に分解し、それらを部分的に変換し、再合成することで声質変換ができるという仕組み)を応用し、基本周波数は対数ガウス正規化を用いたピッチ変換を用いて変換、非周期性指標はそのまま、スペクトル包絡のみをAutoVCによって変換し、それを再合成することで位相推定を用いず、既存手法よりも高速な変換が可能になると考えた。



4. 今の状況

上記の手法を、元論文と同じ教師データを用いて同じエポック数学習させた。元のAutoVC[1]では、3秒の音声の変換に9分23.3秒程度かかる。提案手法を用いた場合1.5秒程度で変換が完了した。これは実に370倍以上の高速化が実現したと考えられる。
しかし、私の主観では変換後の音声の品質がやや下がっていると感じた。
右の図は実際に男声と女声を変換した際のスペクトル包絡だ。
実際の音声は下記のページに記載した。



https://suzukidaishi.github.io/pd3-tyuukan/

5. 検証方法

学習済みの変換手法に対して、以下の方法で検証して精度を比較する。

- 主観評価
  - ABXテストによる変換後の声質の類似度検証
  - MOS評価実験による音声品質検証
- 客観評価
  - 手法ごとのSN比を測定する。
  - 出力音声と入力音声のSpeakerEmbedding間のCos類似度を算出する。

6. 今後の展望

今後の展望として、以下の目標を達成したいと考える。

1. AutoVC[1]と同程度かそれ以上の品質の声質変換
2. より高速化をするための工夫(量子化,枝刈りなど)
3. 実際に使いやすいようなUI/UXの検討

参考文献

[1] AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss, https://arxiv.org/abs/1905.05879 (参照 2021/9/6)

[2] Signal estimation from modified short-time Fourier transform, https://ieeexplore.ieee.org/document/1164317 (参照 2021/9/6)

[3] WaveNet: A Generative Model for Raw Audio, https://arxiv.org/abs/1609.03499 (参照 2021/9/6)

[4] ConVoice: Real-Time Zero-Shot Voice Style Transfer with Convolutional Network, https://arxiv.org/abs/2005.07815 (参照 2021/9/6)

[5] WaveGlow: A Flow-based Generative Network for Speech Synthesis, https://arxiv.org/abs/1811.00002 (参照 2021/9/6)

[6] WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, https://www.jstage.jst.go.jp/article/transinf/E99.D/7/E99.D_2015EDP7457/_article (参照 2021/9/6)