2018年度 システム情報科学実習 Project6 最終発表

AIするディープラーニング Group A

声質変換

メンバー：濱口和希*　白鳥孝幸　山田大貴　齋藤匠　*リーダー
担当教員：竹之内高志　香取勇一　寺沢憲吾　片桐恭弘　冨永敦子

概要

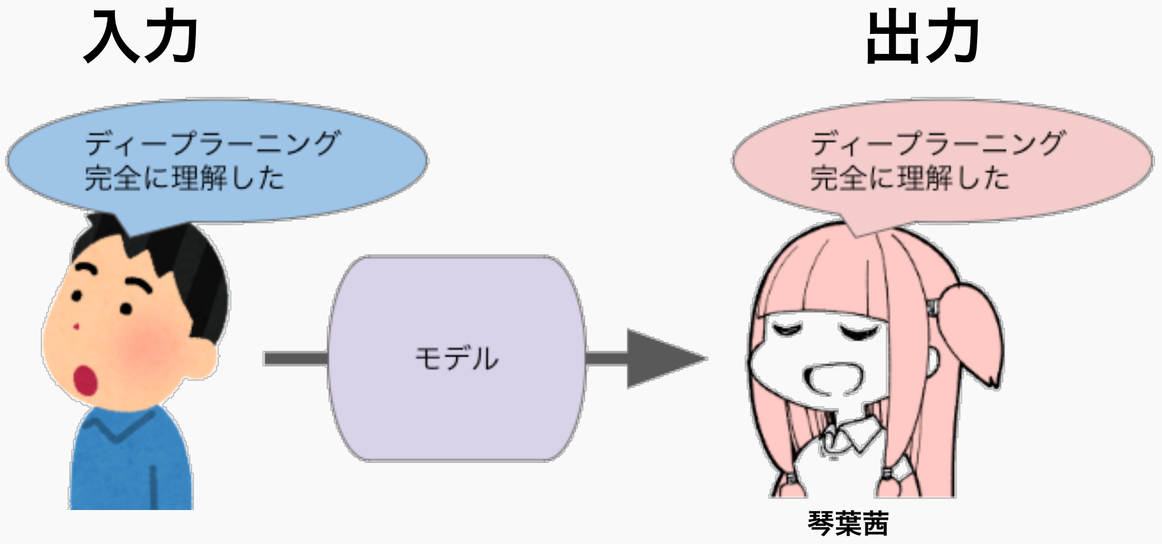
背景

・個人の動画配信、バーチャルYouTuberの活動において、
声質変換の需要が高まっている

目標

・ディープラーニングを用いて、自身の音声を、
VOICEROIDである琴葉茜[1]の音声に変換する

[1]「VOICEROID2 琴葉茜・葵」https://www.ah-soft.com/voiceroid/kotonoha/ (参照2018-11-30).



モデルと工夫

1つ目の工夫点

・2つのモデルを利用[2]
- 画像分野において効果的である手法を、声質変換に適用
- 変換と高精細化の2段階に分けて、高音質な画像を生成する手法[3][4]

各モデルの動作

・１段階目のモデル
- 自身の音声を、低音質な琴葉茜の音声に変換
- 学習にはパラレルデータ*が必要

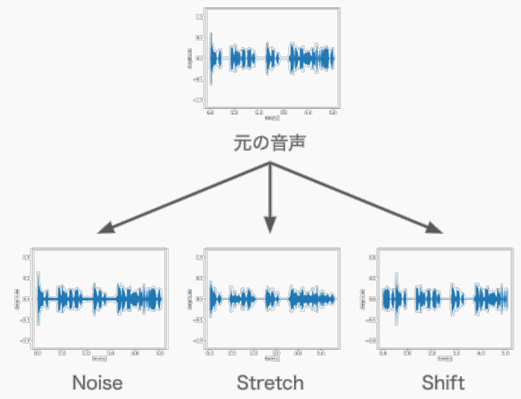
・2段階目のモデル
- 低音質な琴葉茜の音声を、高音質な琴葉茜の音声に変換
- 学習には大量の琴葉茜の音声が必要

*パラレルデータ：入力話者と出力話者について同時に同じ内容を発話した音声データ

・パラレルデータの作成は難しく時間がかかる
- 琴葉茜と同じ文章を録音するのに時間がかかる
  - 500個作成するのに約１５時間ほど要する
  - 読み上げの際、話す速度、話すタイミング、話し方を真似る必要がある

2つ目の工夫点

・１段階目の学習を行う際に、データかさ増しに取り組んだ
- Audio data augmentation[5]を利用
  - 元の音声を加工し、３種類の音声データを作成
    - Noise：一定の大きさの雑音を加える
    - Stretch：再生速度を変更
    - Shift：再生の開始位置をずらす



- １段階目のモデルにおいて、作成した音声データを様々な組み合わせで学習し、評価を行った

１段階目のモデルに対する評価

元の音声のみで学習したモデルとの比較

学習に利用したデータの組み合わせ	主観的評価	日本語として聞き取れるか
元の音声, noise	さ行がかすれて聞こえた	○
元の音声, shift	言葉が崩れ聞き取れなかった	×
元の音声, stretch	rawとほとんど変わらなかった	○
元の音声, noise, shift	言葉が崩れ聞き取れなかった	×
元の音声, noise, stretch	母音が連続する場所が聞き取れなかった	○
元の音声, shift, stretch	言葉が崩れ聞き取れなかった	×
元の音声, noise, shift, stretch	言葉が崩れ聞き取れなかった	×

- 実際に元の音声とStretchで学習したものは、
元の音声のみで学習したモデルと比べ、違和感が少なかった

[2]廣芝 和之, 能勢 隆, 宮本 颯, 伊藤 彰則, 小田桐 優理: 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, 音楽シンポジウム, 2018.

[3]Furusawa, C., Hiroshiba, K., Ogaki, K. and Odagiri, Y.: Comicolorization: semi-automatic manga colorization, SIGGRAPH Asia Technical Briefs, p. 12 (2017).

[4]Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X. and Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, ICCV, pp. 5907–5915 (2017).

[5](2017)「Audio data augmentation」,https://www.kaggle.com/CVxTz/audio-data-augmentation (参照2018-11-30).

まとめ

結果

・琴葉茜のような音声に変換することができた

実際の音声

・変換した音声はスライド発表中にデモを行うので、
ぜひ発表をご覧ください