2018年度 システム情報科学実習 Project6 最終発表　FUN Project Learning_Project 6 Final presentation

AIするディープラーニング AI Love Deep Learning

メンバー：濱口和希*　白鳥孝幸　山田大貴　齋藤匠　三浦幸泰　坂下夏槻　和田孝喜　川村一世　松元健　*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

Member：Kazuki Hamaguchi* / Takayuki Shiratori / Daiki Yamada / Takumi Saitou / Kouta Miura / Katsuki Sakashita / Kouki Wada / Issei Kawamura / Takeru Matsumoto * : Leader

Supervisor：Takashi Takenouchi / Yuichi Katori / Kengo Terasawa / Yasuhiro Katagiri / Atsuko Tominaga

概要 Overview

ディープラーニングとは What's Deep Learning

- **人口の脳神経回路網(Artificial Neural Network：ANN)を元に作られた機械学習手法**
  Machine learning method based on Artificial Neural Network
  - **より層を深く(ディープに)したANNを用いる**
    Deep Learning uses ANN with deeper layer
  - **より高精度な学習が可能**
    Deep Learning can learn more accurately
- **従来の手法では不可能だった課題を解決できる**
  Deep Learning can solve problems could not be done with previous technology
  - **強化学習** Reinforcement Learning
  - **自然言語処理** Natural Language Processing
  - **画像認識** Image Recognition
  - **生成モデル** Generative Model

目標 Purpose

- **ディープラーニングを使用し応用技術の開拓に挑戦する**
  We challenge to the development of applied technology by deep learning

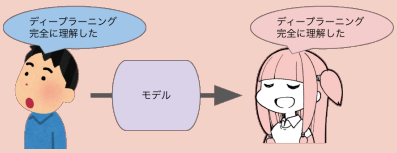
Group A：声質変換
Voice Conversion

Group B：姿勢推定
Pose Estimation

声質変換 Voice Conversion

背景と目的 Background and Purpose

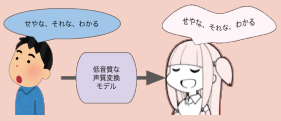
- **個人の動画配信、バーチャルYouTuberの活動において、声質変換の需要が高まっている**
  In personal video streaming and virtual YouTuber activities, the demand of voice conversion is increasing
- **ディープラーニングによって、自身の音声を、VOICEROIDである琴葉茜[1]の音声に変換する**
  We'll try to convert particular voice to voice of Kotonoha Akane by using DeepLearning



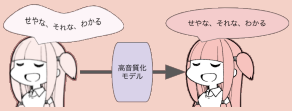
[1]「VOICEROID2 琴葉茜・葵」https://www.ah-soft.com/voiceroid/kotonoha/ (参照2018-11-30).

各モデル Each models

- **1段階目のモデル** First model
  - **自身の音声を、琴葉茜の音声(低音質)に変換**
    It converts particular voice to Akane
  - **学習にはパラレルデータ*が必要**
    It is needed parallel datas to learn



- **2段階目のモデル** Second model
  - **低音質な琴葉茜の音声を、高音質な琴葉茜の音声に変換**
    It converts low-quality Akane voice to high-quality Akane Voice
  - **学習には大量の琴葉茜の音声が必要**
    It is needed a lot of voice datas of Akane to learn

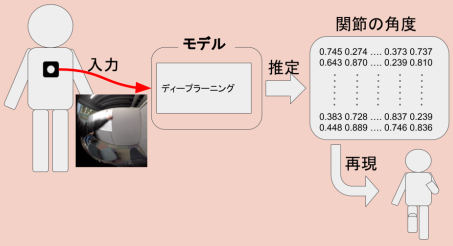


*パラレルデータ：入力話者と出力話者について同時に同じ内容を発話した音声データ

姿勢推定 Pose Estimation

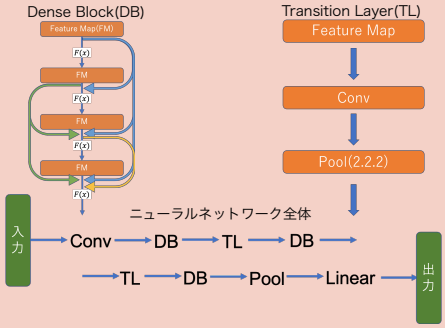
背景と目的 Background and Purpose

- **現状、モーションキャプチャーには様々な機材や広いスペースを必要とする**
  Now, motion capture needs various equipments and large space
- **第三者視点のカメラも大掛かりな機材も必要ないモーションキャプチャの手法を提案**
  We propose a motion capture technique without neither large equipments nor third person camera.



実験と結果 Experiments and Results

- **学習には72層のDenseNetを使用した**
  We used DenseNet-72



- **十分な精度を持つモデルを作ることはできなかった**
  We could not construct good estimator

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

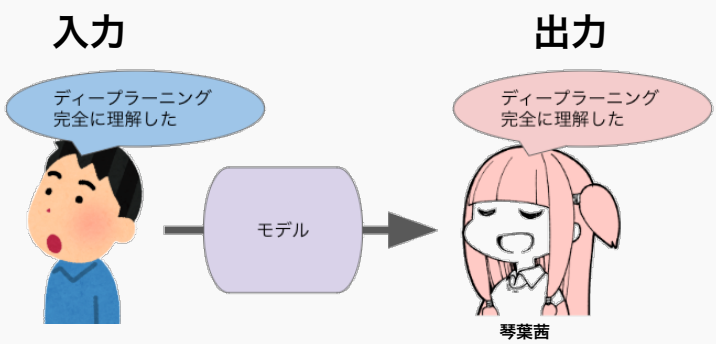
背景

・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

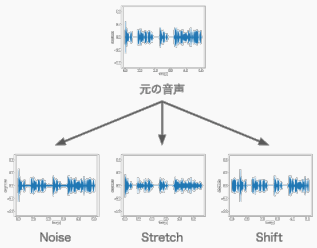
[2]廣芝 和之, 能勢 隆, 宮本 颯, 伊藤 彰則, 小田桐 優理: 畳込みニューラルネットワークを用いた音響特徴量変換とスペクトログラム高精細化による声質変換, 音楽シンポジウム, 2018.

[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).

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

[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

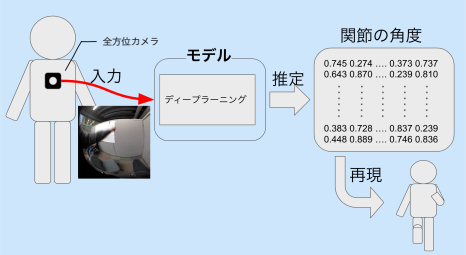
・変換した音声はスライド発表中にデモを行うので、
ぜひ発表をご覧ください

AIするディープラーニング Group B
全方位カメラによるモーションキャプチャ

メンバー：三浦幸泰* 坂下夏槻 和田孝喜 川村一世 松元健 *：グループリーダー
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

目的

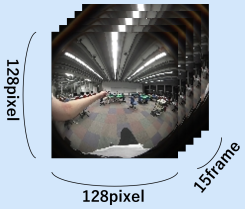
第三者視点のカメラも大掛かりな機材も必要ないモーションキャプチャの手法を提案



実験

モデルの入出力

- 入力: 全方位カメラからの映像
- 出力: 関節の角度



データセットの作成方法

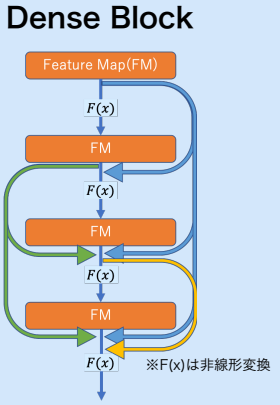
- 被験者はカメラを胸に装着
- 被験者の前方にKinectを設置
- 全方位カメラの装着者がKinectの前で足踏み
- 全方位カメラからは映像、Kinectからは関節の角度を同じタイミングで取得
- 今回は6000個のデータセットを作成

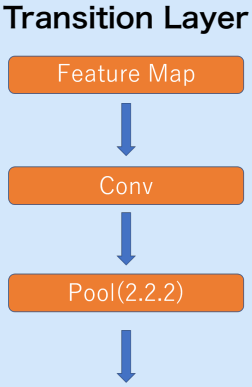
実装したモデル

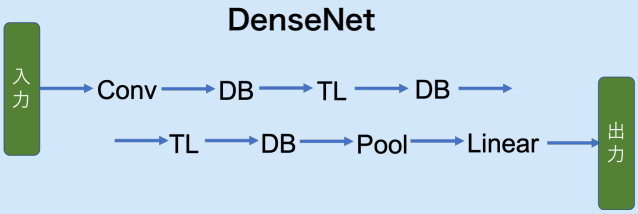
- CNNでは表現力が足りないため、層を増やせるResNet[1]を実装した
- ResNetでも表現力が足りなかったため、DenseNet[2]を実装した

モデルの説明

- 72層DenseNet
- Dense Block(DB)とTransition Layer(TL)を交互に組み合わせてより深い層を実現







[1]https://arxiv.org/pdf/1512.03385.pdf
[2] http://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.pdf

結果と考察

結果

- 使用したモデルではいい精度が出なかった
- 下図のような姿勢から細動する状態になった

結果に対する考察

- 今回のタスクに対して十分な表現力のあるモデルを構築できなかったから
- 学習データの中に手足の動き以外の余計な情報が多く含まれていたから
- タスク達成に必要な情報が全方位カメラからのみでは不足していた