Emotional voice conversion via facial expression recognition

17 顔表情認識による音声変換技術

顔の表情でリアルタイムに声の表情をつくる

どんな研究	対話や講演において、心理的な緊張状態や能力的な限界などにより思うような話し方で話せない場面があります。本研究では、入力音声の話し方の雰囲気（表情）を、顔の表情や体の動作により制御することを目的としたクロスモーダル音声表情変換の問題に初めて取り組みました。
どこが凄い	話し方の表情は声質・抑揚・リズムによって決まります。従来技術の多くは声質のみの変換を行いますが、われわれの音声変換技術は、声質とともに抑揚やリズムの変換も可能にします。この技術と顔表情識別技術を組み合わせ、顔画像を用いて声の表情を変換する技術を実現しました。
めざす未来	人と人とのコミュニケーションには、物理的・能力的・心理的な状態に起因する様々な形の制約が存在します。本研究では、このような制約を取り除き、あらゆる人が自由に快適にコミュニケーションを行える環境を実現することをめざしています。

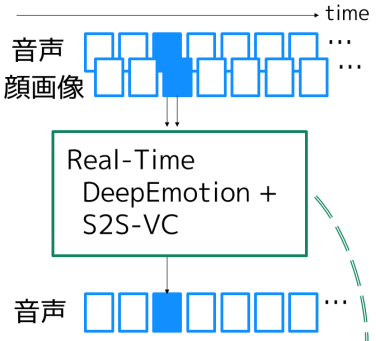
音声変換によるコミュニケーション機能拡張

音声変換技術を通じて多様なシーンにおいて人が自由にコミュニケーションできる手段を創出

顔画像による音声表情のリアルタイム制御

顔表情認識技術と音声変換技術を組み合わせたクロスモーダル音声変換

（例：顔画像の表情の時間的な変化に追従するようリアルタイムに音声の表情を変換）



要素技術

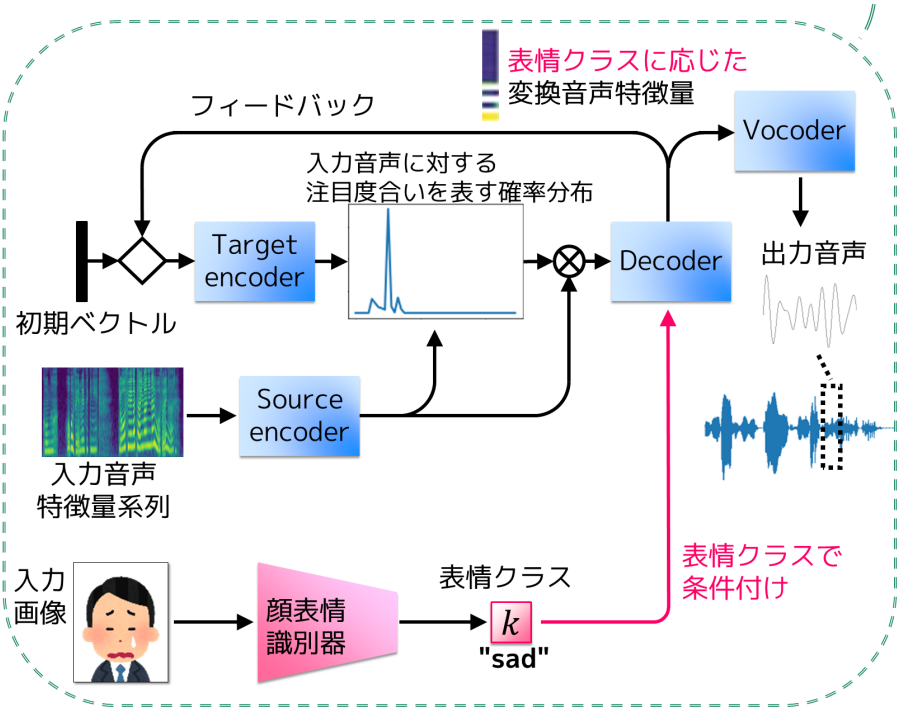
系列変換モデルを用いた音声変換

■系列変換(S2S)モデル
- 系列から系列への変換則を学習する深層学習の枠組（機械翻訳や音声認識などで有効性が知られる）
- エンコーダ／デコーダ構造と注意機構で長期依存関係を捉えた系列変換を扱える
- 変換元と変換先の系列の要素間の対応づけ規則を学習できる
- 通常は大規模な学習データが必要

■S2Sモデルを用いた音声変換
- 話し方の雰囲気や感情表現は、特に抑揚やリズムに色濃く現れる
- S2Sモデルにより声質だけでなく抑揚、リズム、話速の変換を扱える
- 少ない学習データでも学習可能な方式を考案

顔表情認識

■全層畳み込みネットワークを用いたクラス識別
- 顔表情識別結果を音声変換器に入力



関連文献

[1] H. Kameoka, K. Tanaka, T. Kaneko, N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1849-1863, June 2020.

[2] K. Tanaka, H. Kameoka, T. Kaneko, N. Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019)*, pp. 6805-6809, May 2019.

[3] M. Shervin, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network." *Sensors 21.9:3046*, 2021.

連絡先

田中 宏 （Kou Tanaka） メディア情報研究部 メディア認識研究グループ
Email: cs-openhouse-ml@hco.ntt.co.jp