5R-02

声優データを用いた不特定話者感情音声変換 *

北村　健太郎 †,　　伊藤 克亘 †,

1　序論

現在の音声変換技術は明確な変換ターゲットのある対データでは，高品質な変換はできるが対のないデータでは変換することができず，コーパスに含まれない発話に対しての変換性能は低下する．日本語話者の感情音声のデータは通常の朗読のコーパスデータより数が限られており，未知の発話の感情変換は難しい．このことを解決するために本研究では話し方を学習，転用することにより，変換ターゲットがなくても変換することを期待する．また，システムの特徴として，1）変換話者のターゲット感情データのいらない，不特定話者の感情音声変換が可能 2）変換音声の統計的特徴量を，合成時に用いることで話者性を残した変換が可能が挙がる．

2　GAN を用いた音声変換

GAN の強みは変分オートエンコーダー (VAE) よりもより高品質な音声を生成することができる点である．VAE では，KL 誤差を導入し潜在変数の分布が正規分布になるように制約を掛けた．VAE では，この制約により，高品質な音声を生成することができる．その結果入力にある，細かな違いは変換の過程で無視されて，平均的な生成物が出力され，もやがかかったような音声が生成される．GAN ではこのような制約がないので，よりはっきりとした出力を得ることができる．

2.1　CycleGAN

CycleGAN はペアとなるデータがない画像間翻訳モデルを学習する技術である．このモデルは，ソースデータとターゲットデータの画像データを用いて教師なし方式で学習されるが，学習データの画像は 1 対 1 で関連付けられている必要がない．

2.2　StarGAN

StarGAN[3] は CycleGAN と同じく，ペアとなるデータがない画像間翻訳モデルを学習する技術であるが，異なる点として，単一のモデルを用いて複数ドメインの画像間変換を行うことができる技術である．1 つのネットワーク内でドメインの異なる複数のデータセットの同時学習が可能となり，既存のモデルと比較して翻訳画像の品質に優れ，入力画像を任意のターゲットドメインに柔軟に翻訳することができるという新しい機能を備えている．

3　GAN を用いた感情音声の生成実験

3.1　CycleGAN を用いた感情音声の生成実験

特徴量変換には CycleGAN を用いる．入力には平常時の音声のメルスペクトログラムと，f0 を与える．出力は感情が付与された音声のスペクトルグラムと f0 である．今回は平常音声と感情音声の f0 と，音色 (メルケプストラム) のスタイルを学習する．音色と f0 を別々に学習することにより，話者性のない感情表現と話者性のある音色を分けて考慮できる．f0 とメルケプストラムの算出と音声合成には WORLD を用いる．学習した変換モデルを用いて，平常時の特徴量を喜びや怒りの特徴量へ変換する．変換した特徴量と非周期指標をもとに WORLD で音声合成する．

感情音声が豊富な声優統計コーパスを用いた感情音声の生成実験を行った．女性声優 A の音声を学習データとした．女性声優 A の平常時の音声を喜びの音声に変換した．

3.2　StarGAN を用いた感情音声の生成実験

メルケプストラム係数と，話者性を表すためのワンホットベクトルを入力とした StarGAN での音声変換実験を行った．生成に使った音声ソースデータは，変換後の話者性を確認するために別話者の JSUT コーパス [7] の BASIC5000 を用いた．またターゲットデータは感情付与効果を確認するために声優統計コーパスの女性声優 A の音声を用いた．

3.3　コーパス

声優統計コーパス [6] はプロの女性声優 3 名が 平常，怒り，喜びの 3 パターンの感情で読み上げた音声である．総長約 2 時間の JVS[10] 音素バランス文を読み上げたものである．

JSUT[7] は，日本語のテキストと朗読形式の音声で構成されている．音声データは 48kHz でサンプリングされ，無響室にて録音されている．このコーパスには，様々な読み上げによるデータからなる 10 時間分の音声が含まれている．今回は BASIC5000 を学習に用いる．

4　予備実験

4.1　CycleGAN を用いた音声変換

変換後の音声の f0 軌跡を次に示す (図 1).

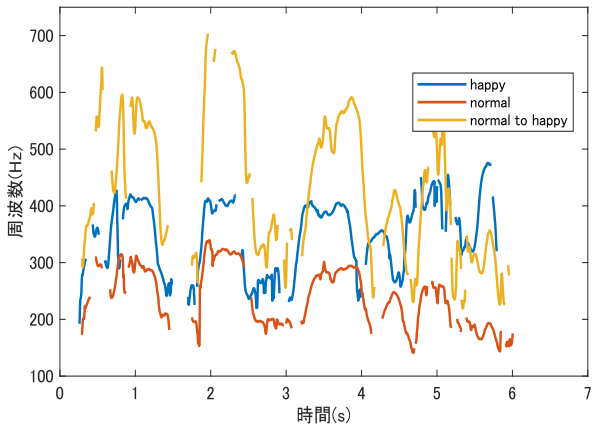

図 1. f0 の軌跡 (赤：平常時　青：喜び　黄：平常 → 喜び)

*　：English Title *Speaker independent emotional voice conversion using voice actor data (Hosei Univ.) et al.*

†法政大学 情報科学部

図 1 では，平常時から喜びに変換された音声は f0 の平均が上がっているのが確認できる．しかし，喜びの音声の文末での特徴的なアクセント変化は反映されておらず，声の高さだけ変わった印象を受けた．別の人の音声を女性声優 A の音声で学習したモデルを使い音声変換した．女性声優 B の音声 f0 のレンジが広くなり，声のトーンが高くなったことが確認できた．しかし，話者性が損失しており，女性声優 B の声と認識できなかった．

実装した音声統計コーパスを用いた音声変換では，個人性の損失や，細かなアクセント表現の欠落などの問題があった．これは，CycleGAN が学習において non-Parallel データを扱うので，生成音声で直接モデルを最適化できないことが原因であると考えられる．この問題を解決するために，学習時にソース・ターゲット話者を特定する情報を与える．

4.2 StarGAN を用いた音声変換

StarGAN での変換音声のスペクトログラムを載せる 2.

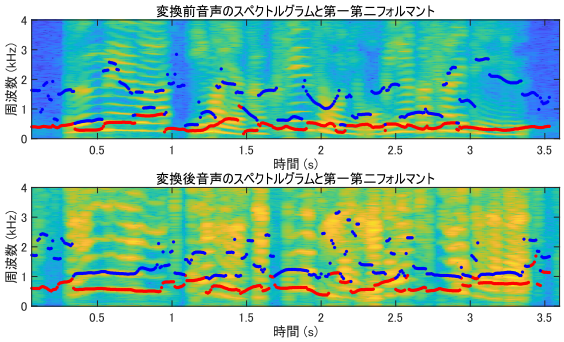

図 2. JSUT,BASIC5000 の音声を女性声優 A のモデルを使い，喜びへと感情変換したものの，スペクトログラムとフォルマントの位置． 下：生成音声，上：元音声

図 2 では倍音構造が，表現されていることが確認できる．しかし，フォルマントの生成ができておらず，言語を識別できない．

同じモデルを用いた VCTK コーパスの男性話者 p262 と女性話者 p272 を使った音声変換は，言語内容を理解できる変換ができたが，声優統計コーパスを用いた変換では，言語が聞き取ることができなかった．VCTK コーパスと同じ，読み上げ文章で，日本人話者の lNAS コーパス [9] を用いて音声変換実験を行った．生成音声のスペクトルグラムと第一第二フォルマントを (図 3) に載せる．図 3 より，倍音構造は男性の音声であるが，日本語発声のフォルマントが生成されておらず，音声言語として認識できない．これは，生成音声の第一フォルマントが元音声の基本周波数に影響されていることが原因だと考えられる．学習するために，フィルタをかけて特徴量を強調する．

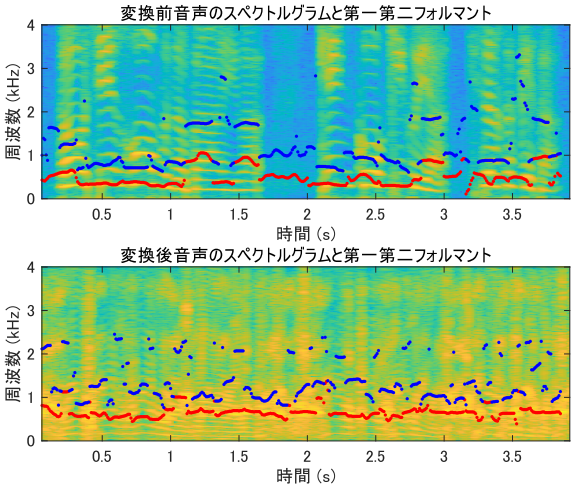

図 3. JNAS, 女性話者の音声を男性話者へと音声変換したものの，スペクトログラムとフォルマントの位置．下：生成音声，上：元音声

4.3 結論

CycleGAN と StarGAN を用いて，変換音声を生成することができた．生成音声を改善するために，1）データ量を増やす．2）loss 関数の制約を緩和する．3）フレームシフトの幅を狭め細かい変化に対応する．などの工夫をし，より明瞭的な音声を生成したい．

参考文献

[1] 小池和仁，et al. ”感情音声の合成”, (1998-SLP-024)

[2] K. Hirokazu , et al. ”StarGAN-VC Non-parallel many-to-many voice conversion with star generative adversarial networks”

[3] Y. Choi, et al. ”StarGAN: Unified generative adversarial networks for multidomain image-to-image translation” (CVPR), 2018, pp. 8789-8797

[4] P. Isola, et al. “Image-to-image translation with conditional adversarial networks, ”

[5] J. Zhu, et al. ”Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”, ICCV 2017 paper

[6] y benjo and MagnesiumRibbon, “ Voice-actress corpus, ” http://voice-statistics.github.io/.

[7] S. Ryosuke, et al. ”JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint, 1711.00354, 2017.

[8] C. Veaux, et al. ”The CSTR VCTK Corpus”

[9] The Acoustical Society of Japan. ”ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS)”. Speech Resources Consortium, National Institute of Informatics.

[10] T. Shinnosuke, et al, ”JVS corpus: free Japanese multi-speaker voice corpus,” arXiv preprint, 1908.06248, Aug. 2019.