時間系列を考慮した声質変換の検討

Examination of voice conversion considering time series

和田　楓也

Fuya Wada

広島市立大学院　言語音声メディア工学研究室

Language and speech research Laboratory, Hiroshima City University

概要

本研究では，既存の声質変換手法に「時間系列を考慮」し，性能向上を目的とする研究を行う．既存手法である CycleGAN-VC2[1]のネットワークに新たに RecycleGAN[2]のネットワークである Predictor(予測器)を導入することで，既存手法と比べ長い区間を学習データとして利用することができ，時間系列を考慮することで，生成された音声が向上しているか検討する．

1. はじめに

現在，娯楽または趣味として，実況動画やゲーム配信などの自分の声を世界に発信する機会が増えてきている．しかし，中には自分の声に自信がなく配信したくてもできない人もいる．声質変換を用いて好きな人の声，お気に入りの声に自由に変換できれば，有用なシステムになると考えられる．

人の声を別に人の声に変える技術のことを声質変換と言う．近年では，深層学習を利用した手法が多く研究されており，これまでの統計的変換手法からの進歩が著しい．

しかし，既存手法の声質変換には，時間系列が考慮されていないという問題がある．そこで本研究では，朗読調の男女の音声を用意し，既存手法である CycleGAN-VC2 の学習器に RecycleGAN の学習器である Predictor(予測器)を導入し，時間系列を考慮することで，生成された音声が向上しているか検討する．

2. 研究目的

本研究は，声質変換に必要な音響特徴量として，短時間フーリエ変換から計算した Mel-Spectrum を用いる．これは，人間の聴覚特性を考慮した特徴量である．

次にモデルについて述べる．本研究では，CycleGAN-VC2 に RecycleGAN のネットワークを追加した深層学習モデルを用いる．CycleGAN-VC2 は GAN(Generative Adversarial Networks，敵対生成ネットワーク)による声質変換手法である．そして，RecycleGAN とは，CycleGAN を動画に最適化した手法である．すなわち，時間情報と空間情報を組み合わせ，ターゲットドメインのスタイルを維持しながらクロスドメイン変換を実現する．空間情報のみに焦点を当てている CycleGAN と比較し，動画変換に最適化されている．

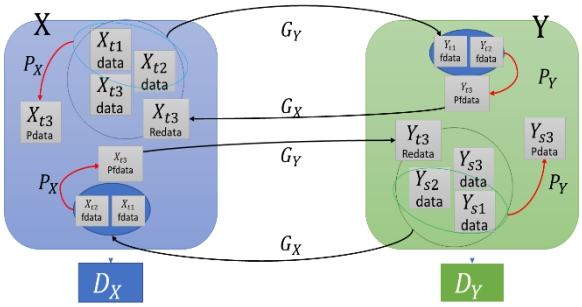

図 1：CycleGAN+Predictor のネットワーク構造

次に，RecycleGAN の構造を図 1 に示す．

G(Generator)と D(Discriminator)を 2 組用意し，Y から X の逆変換の学習も行う．具体的には，本物 x, y から，G によって偽物 y, x が生成される．さらに，その偽物から G によってまた偽物が生成される．これを繰り返し，本物か偽物か判別しにくい音声を生成できるよう学習する．

本研究の目的は，声質変換の対象として，男女の朗読調の音声を用意し，時間系列を考慮した CycleGAN-VC2 を用いて声質変換を行い，生成された音声が向上しているかどうかを検討することである．

3. 実験

実験として，CycleGAN-VC2 及び CycleGAN-VC2 に RecycleGAN のネットワークを追加し，時間系列を考慮した学習器で実験を行う．

学習データとして，The Voice Conversion Challenge 2018 の VCC2SM3(SM)，VCC2TF1(TF)を使用した．ここで，S, T, M, F はそれぞれソース，ターゲット，男性，女性を表す．各話者毎に，サンプリング周波数を 22.05kHz に設定した 64 個の発話を用い，epoch 数は 6,172 で実験を行った．そして，学習結果の評価には，学習時とは別の 35 発話を用いた．

客観的評価には，メル尺度上で目標音声と変換音声がどれだけ似ているかを表す尺度(Mel-Cepstral Distortion，MCD)[3]を使用する．値が低ければ似ている音声となり，性能が高いことを意味する．

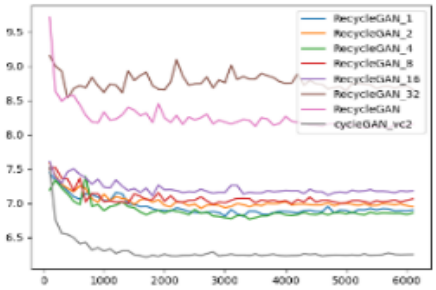

図 2：100epoch ごとの MCD の平均値の推移

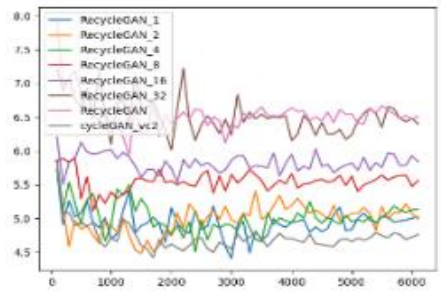

図 3：100epoch ごとの MCD の最小値の推移

3.1 実験 1：入力データの重なり具合の変更

学習時 Predictor で予測する際，入力するデータの重なり具合により，生成される音声が良くなることが予測される．重なり具合とは，連続する 3 つの入力データを重ね合わせ，固定数値でずらす(shift する)ことである．shift 幅が小さいほど，重なり具合は大きくなる．

以下の図は，重なり(shift 幅)1～32frame，なし，CycleGAN-VC2 の 100epoch ごとの MCD の平均値，最小値の推移のグラフである．

図 1 より，MCD の平均は既存手法が良かった．図 2 より，最小値で比べると同等かそれ以上の値が出ている．

3.2 実験 2：Predictor による予測されるデータの変更

Predictor にデータを入力する際，連続する 3 つのデータを用意する(前から順に 1，2，3 とする)．一つ目は 1, 2 から 3 を予測．二つ目は，1, 3 から 2 を予測．三つ目は，2, 3 から 1 を予測する．重なり具合は実験 1 より，結果が良かった shift 幅 1 を使用する．以下の図は，100epoch ごとの MCD の平均値の推移のグラフである．

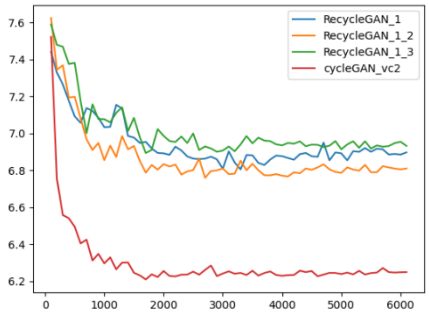

図 4：100epoch ごとの MCD の平均値の推移

図 3 より，1, 2 から 3 を予測する手法より，1, 3 から 2 を予測する手法が良いことが見て取れる．

3.3 考察

実験 1 より，MCD の平均値は既存手法より提案手法が劣っている．しかし，最小値は既存手法と同程度，もしくは優れる．平均値が高くなっている要因は，学習評価時に用意したデータ 35 個の内，20 個以上の音声が 3 秒以上の音声である．最小値の音声は 1 秒以下の音声(一言程度)であり，それよりも長い音声が多くある為，平均値が高くなっていると考えられる．

実験 2 より，通常の提案手法(1, 2 から 3 を予測する手法)と比較し，MCD の値が良くなったことにより，双方向からの補完が可能であることが分かる．

4. まとめと今後の課題

本研究では，時間系列を考慮した声質変換の検討を行った．MCD の平均値は，既存手法よりも低い値を取ってしまったが，最小値の音声は同等かそれ以上の音声もあったため，アプローチそのものは有効であると考える．

今後は，より有効な重み付けや，今回は学習データとして朗読調の英語の音声を使用したため，日本語や自然発話での音声で研究を行うことが課題である．また，Prallel-WaveGAN の先行研究より，様々な周波数帯域の Spectrogram の入力を行うこと．CycleGAN-VC2 のモデルを 2 つ用意し，音響的特徴量の変更・新たな loss の追加することで既存手法の性能向上を目指すことも課題である．

5. 参考文献

[1]T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, “CycleGAN-VC2:Improved CycleGAN-based Non-parallel Voice Conversion,” 2019, pp. 6820-6824 Apr.2019

[2]A. Bansal, S. Ma, D. Ramanan, Y. Sheikh “Recycle-GAN:unsupervised video retargeting,” 2018, In European Conference on Computer Vision (ECCV).

[3]L.Sun, K.Li, H.Wang, S.Kang, H. Meng, "Phonetic posteriorgrams for many-to-onevoice conversion without parallel data training," 2016, IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6.