The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

GANを用いた声質変換における周波数帯域ごとの分析

Frequency Analysis in Voice Conversion Using Generative Adversarial Networks

和田　楓也[*1]　黒澤　義明[*1]　目良　和也[*1]　竹澤　寿幸[*1]
Fuya WADA　Yoshiaki KUROSAWA　Kazuya MERA　Toshiyuki TAKEZAWA

[*1] 広島市立大学大学院　情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

In recent years, deep learning has enabled high-quality speech synthesis and voice quality conversion. Traditional methods use a GAN (Generative Adversarial Network) to perform voice conversion. However, the generated speech sounds a little muffled compared to actual speech. There are also some shortcomings regarding the generated 2D features. Therefore, in this study, the generated spectrogram is divided into several frequency bands, and the Mel-Cepstrum Distortion (MCD) of each frequency band to investigate and analyze which frequency bands are well generated. Analysis showed that the low frequency of the generated Spectrograms were well generated, but the mid/high frequency were not well generated. In addition, we found that although the linguistic information was reproduced, the reproduction of speaker characteristics was insufficient.

1. はじめに

現在，Mask-CycleGAN[Kaneko 21]を用いた声質変換手法が研究されており，高品質な合成音声が生成可能となっている．しかし，生成された音声は本物の音声と比べ，ややこもったような音声となっており，生成された2次元特徴量に関しても再現が不十分な箇所はある．

そこで本研究では，朗読調の男女の音声を用意し，その音声データセットを用いて男女音声間の声質変換を行う．そして，生成される2次元特徴量を数個の周波数帯域ごとに分割し，それぞれの周波数帯域ごとの MCD（Mel-Cepstrum Distortion）を計算し，どの周波数帯域がうまく生成されているか検討・分析する．この分析結果を利用することで，深層学習による声質変換・音声合成の音声品質向上を目指す．

2. 関連研究

本章ではGANによる声質変換の関連の研究について紹介する．また，音声特徴量を音声に変換する手法について述べる．

2.1 CycleGAN-VC

CycleGAN-VC[Kaneko 17]は，GANによる画像変換手法であるCycleGAN[Zhu 17]を声質変換手法に適応させた手法である．この手法では，2つの画像の対応する各ピクセルの関係を学習するのではなく，2つの異なる画像データセットのドメイン間の関係を学習して画像変換を実現する．CycleGANの基本構造を図1に示す．

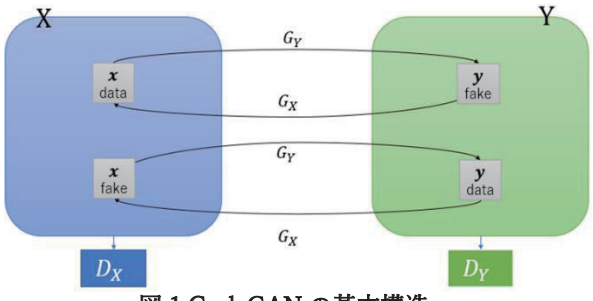

図1 CycleGANの基本構造

（参考文献[Kaneko 17]より改良し引用）

この図は，変換と逆変換の循環(cycle)構造となっている．GeneratorとDiscriminator（以下GとDとする）をそれぞれ2組用意し，ドメインY→ドメインXの逆変換の学習を行う．図1の例では，Gによってdata xからfake yが，data yからfake xがそれぞれ生成される．fakeとは，Generatorから生成された偽物のデータのことである．さらにそのfakeからGによってまたfakeが生成される．この繰り返しの変換が循環構造になっている．そして，双方向の変換を保証するドメイン間の関係を学習する

音声の例で表現すると，男性の声を入力とし，Gによって偽の女性の声が生成される．さらに，その偽の女性の声を入力とし，Gによって偽の男性の声が生成される．これを繰り返して，それぞれのDを騙すような音声を生成するように学習する．

2.2 Mask-CycleGAN-VC

Mask-CycleGAN-VC[Kaneko 21]はCycleGAN-VC2[Kaneko 19]を発展させた手法である．構造を図2に示す．Generatorを学習させる際に，最初に入力させる画像にマスクを用いて意図的に欠損を作り，欠損を復元しつつ変換するようにモデルを学習させる．このようにすることでモデルに時系列の情報を学習させ，精度の向上を目指した研究である．その結果，自然さと話者の類似性の主観的評価によりCycleGAN-VC2と比較し同様のモデルで優れていることが分かった．

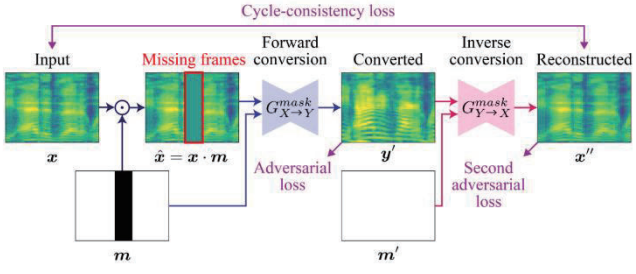

図2 Mask-CycleGAN-VCによる声質変換手法[Kaneko 21]

連絡先：黒澤義明，広島市立大学院情報科学研究科，広島県広島市安佐南区大塚東3-4-1，kurosawa@hiroshima-cu.ac.jp

本研究では，Mask-CycleGAN-VCを使用し，既存手法の性能向上のために，生成された 2 次元特徴量の分析・実験を行う．

2.3 Vocoder

Vocoder とは，音声データから音響特徴量の取得と特徴量から元の音声に変換を行うシステムのことである．

Mask-CycleGAN-VC では，学習データとして音声を Mel-Spectrogram という 2 次元特徴量に変換としている．波形を Mel-Spectrogram に変換するには短時間フーリエ変換を行う．一方，Mel-Spectrogram をフーリエ逆変換行う際，位相などの情報が欠けているため高品質な音声を得ることはできない．そこで Vocoder を用いることで，Mel-Spectrogram から高品質な音声を復元することができる．

2.4 MelGAN

MelGAN[Kumar 19]は音声特徴量から音声波形を生成するモデルである．音声特徴量から音声波形を生成するモデルは Vocoder とよばれ，MelGAN も Vocoder の一種である．

DNN を利用した Vocoder として WaveNet[Oord 16]が有名であり，非常に高品質な音声生成が可能である．しかし，自己回帰で推論する構造のため波形生成に非常に時間がかかることが問題となっている．

一方，MelGAN は自己回帰ではない並列計算が可能な構造で高速かつ高品質を目指したモデルである．通常音声特徴量から波形は一意に求まらない設定が多く，通常の回帰モデルでは綺麗な音声の生成が難しいとされているが，MelGAN は GAN を利用して音響特徴量の条件付きの分布からサンプリングする生成モデルとして学習することで実データに近い予測を可能にしている．

Mask-CycleGAN-VC では学習された 2 次元特徴量を MelGAN を用いて音声へと変換している．

3. 提案手法

本研究では，GAN を用いた声質変換において，生成された 2 次元特徴量に対し，低中高の 3 つの周波帯域ごとに分割した 2 次元特徴量を使用し分析を行う．そして，分析された結果を用い，深層学習を行う際にどの周波数帯域の loss を追加すればよいかの検討を行う．

4. 実験・評価

本実験では，音声を 2 次元特徴量に変換し，得られた 2 次元特徴量を入力データとして実験を行った．4.1 節では，実験条件，データ数，組み合わせについて述べ，4.2 節では評価手法について述べる．

4.1 実験条件

本実験では，[Kaneko 21]の手法を用いて声質変換を行う．実験に使用するコードは[1]をベースにして作成した．入力データとして後述する 2 次元特徴量を用いる．また，実験の epoch 数は従来研究と同条件にするために，6,172epoch とするし，以下に実験条件を述べる．

[1] https://github.com/GANtastic3/MaskCycleGAN-VC

[2] https://datashare.ed.ac.uk/handle/10283/3061

表 1 データ数

	SM2, SM3, TF1, TF2
train	256(各 64)
test	140(各 35)

表 2 STFT パラメータ

パラメータ名	パラメータ数
window size[sample]	512
hope length[sample]	80

(1) 音声コーパス

本研究では，音声として単数話者による発話を用いて実験を行った．使用した音声コーパスは，The Voice Conversion Challenge 2018[Trueba 18](以降 VCC2018)の VCC2SM2(SM2)，VCC2SM3(SM3)，VCC2TF1(TF1)，VCC2TF2(TF2)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，女性を表している．

(2) データ数

学習・評価に用いたデータ数を表 1 に示す．

(3) Mel-Spectrogram

Short-time Fourier Transform(STFT)によって表現され，音声波形より振幅のみを取り出し，横軸を時間，縦軸を周波数として表現された 2 次元音声特徴量を振幅 Spectrogram という．STFT は対話音声において，言語情報や話者性などを表現しており，声紋分析で使われている音声特徴量である．その Spectrogram の周波数軸がメル尺度上で表されているものが Mel- Spectrogram である．本実験で使用する STFT のパラメータを表 2 に示す．

(4) 組み合わせ

CycleGAN の枠組みでは，声質変換を行う際，入力音声として 2 人セットの話者データが必要である．本実験において先行研究である Mask-CycleGAN を用いた声質変換と比較を行うために，本研究では VCC2018[2]より SM3 と TF1，SM2 と TF2 の 2 つの組み合わせによる実験を行う．

4.2 評価方法

(1) 音声評価指標

本実験では，音声を 2 次元特徴量に変換し，それを入力データとして実験を行っている．本実験では音声評価指標である MCD(Mel-cepstral distortion)[Sun 16]を用いて評価を行った．式を 1.1 に示す．

$$MCD[dB] = \frac{10}{ln10}\sum_{t}^{T}\sqrt{2\sum_{d=1}^{F}\left(mc_d^{(y)}(t) - \widehat{mc}_d^{(y)}(t)\right)^2} \quad (1.1)$$

ここで，T は時間伸縮後の総フレーム数，F は周波数を表す．$mc_d^{(y)}(t)$ は目標音声の t フレーム目の Mel-Cepstram の d 次元目，$\widehat{mc}_d^{(y)}(t)$ は変換音声の t フレーム目の M-Cep の d 次元目を表わす．MCD は Mel-Cepstram 間における距離に基づく尺度であり，MCD の値が小さければ小さい程，変換音声のスペクトルが目標音声のスペ

クトルへと近づく. MCD は主に声質変換の評価指標として使用されている.

(2) 周波数帯域の分析

本実験では, 生成された Mel- Spectrogram を低中高の 3 つに分割し, それぞれの周波数帯域の分析を行う. 各周波数帯域の最大値, 最小値, 平均値, MCD の値を計算し, どの周波数帯域が生成できているか・できていないかを分析する.

5. 実験・結果

Mask-CycleGAN を用いた声質変換で生成された Spectrogram に対し, 低中高の 3 つの周波数帯域ごとに分割したデータの分析を行う. 結果を表 3~6 に示す. また, 各数値の reral と fake の平均の差を表 7 に示す. 分析に使用する Mel-Spectrogram は 100epoch ごとの MCD の平均値が最小であった 6,100epoch(390,400iteration)で生成された Mel-Spectrogram を分析に使用する.

5.1 各声質変換の分析・考察

表 7 より, どれも低周波数帯域の Mel-Spectrogram の MCD は平均で 1.239 と低く良い結果が出ている. また, 平均値や最大値, 最小値の real と fake の差に関しても 0.003, 0.092, 0.086 とほぼ差が見られず, うまく再現できていることが確認できる. 高周波帯域は M→F, F→M どちらの変換も MCD は 6.276 と高く, ほかの周波数帯域と比べ, real と fake の最大値の差が 0.650 と大きな差がみられる. real と比較し最大値が低いことより, 音の高さの再現が不十分であることが分かる. しかし, 最小値はどの変換も-9.210 と同じ値となっており, 差も 0 のことから, 暗くなる部分はうまく再現できていることが見て取れる. 中周波数帯域は, MCD の平均も 2.587 とよくうまく生成されている. しかし, 最小値の差の平均がほかの周波数帯域と比べ 0.302 と高く, 低くしなければならないところが再現されていないことが考えられる.

5.2 生成された Mel-Spectrogram

次に実際に生成された Mel-Spectrogram を図 3, 4 に示す. どちらも左が本物の音声の Mel-Spectrogram, 右が生成された偽物の音声の Mel-Spectrogram である. どちらも波形の特徴を捉え, うまく生成できているように見える. しかし, 全体的に Mel-Spectrogram の明るさが real と比較し, 再現できていないことが分かる. 特に 2,024Hz~8,192Hz の区間は, ほかの周波数帯域と比べ不十分な再現となっている. 実際に生成された音声を聞くと, 本物の音声に比べ, ややこもったような音声になっている原因と考えられる. また, Mask-CycleGAN では学習時に意図的に欠損部を作り, 双方向からの欠損の修復をすることで時系列が考慮され, 性能が向上した. しかし, 欠損部分の修復の際, 本来は明るく修復しなければいけないのに対し, 双方向が暗くなってしまっていることにより, real と比べ全体的に生成された Mel-Spectrogram が暗くなったのではないかと考える.

表 3 SM3→TF1 の分析(6,100epoch)

	SM3→TF1					
freq	low		mid		high	
M-Spec	real	fake	real	fake	real	fake
MCD	1.062		2.608		6.365	
ave	0.002	0.003	-0.010	-0.023	-0.510	-0.533
max	0.398	0.350	0.942	0.614	5.228	4.594
min	-0.402	-0.394	-0.995	-1.115	-9.210	-9.210

表 4 SM2→TF2 の分析(6,100epoch)

	SM2→TF2					
freq	low		mid		high	
M-Spec	real	fake	real	fake	real	fake
MCD	1.167		2.457		5.637	
ave	0.003	0.002	0.004	-0.015	-0.579	-0.647
max	0.400	0.405	1.003	0.882	4.756	4.292
min	-0.344	-0.386	-1.001	-0.878	-9.210	-9.210

表 5 TF1→SM3 の分析(6,100epoch)

	TF1→SM3					
freq	low		mid		high	
M-Spec	real	fake	real	fake	real	fake
MCD	1.008		2.501		7.098	
ave	-0.005	0.003	0.004	-0.005	-0.483	-0.666
max	0.503	0.336	0.925	0.503	5.064	4.096
min	-0.519	-0.274	-1.285	-0.519	-9.210	-9.210

表 6 TF2→SM2 の分析(6,100epoch)

	TF2→SM2					
freq	low		mid		high	
M-Spec	real	fake	real	fake	real	fake
MCD	1.720		2.784		6.006	
ave	0.002	-0.001	0.001	0.003	-0.611	-0.619
max	0.687	0.539	1.117	0.933	5.078	4.546
min	-0.543	-0.488	-0.917	-0.718	-9.210	-9.210

表 7 各数値の real と fake の差の平均値

freq	low	mid	high
MCD	1.239	2.587	6.276
ave_diff	0.003	0.011	0.070
max_diff	0.092	0.264	0.650
min_diff	0.086	0.302	0

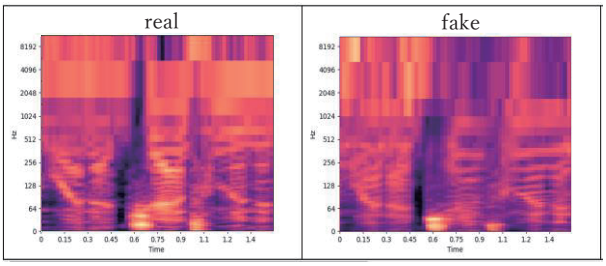

図 3 SM3→TF1 で生成された Mel-Spectrogram

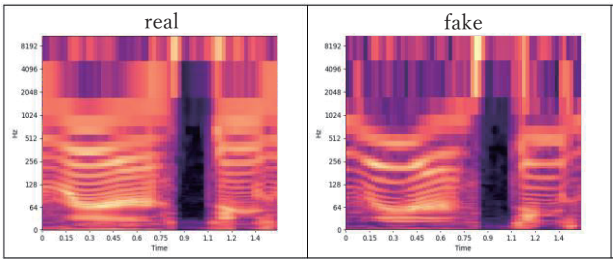

図 4 SM2→TF2 で生成された Mel-Spectrogram

6. まとめと今後の課題

本研究では，GAN を用いた声質変換において，生成された 2 次元特徴量に対し，低中高の 3 つの周波帯域ごとに分割した 2 次元特徴量を使用し分析を行った．

実験の結果より，低周波数帯域の生成は MCD の値も低くうまくできていたが中・高周波帯域はまだ MCD の値が高く，生成された Mel-Spectrogram も不十分であった．また，生成された音声は言語情報の学習はうまく生成できているが，話者性の学習がまだ不十分であることが分かった．この分析結果を利用し，今後，高周波帯域を学習させるような新たな loss を加えることによって声質変換の性能向上を図る必要がある．生成された Mel-Spectrogram も全体的に暗く生成されているため，波形を学習するだけでなく，全体の輝度を学習するような loss を追加したいと考えている．

参考文献

[Kaneko 17] Kaneko, T. and Kameoka, H.: Parallel-data-free voice conversion using Cycle-consistent adversarial networks, European Signal Processing Conference (EUSIPCO) (2017)

[Kaneko 19] Kaneko, T., Kameoka, H., Tanaka, K. and Hojo, N.: CycleGAN-VC2: Improved CycleGAN based non-parallel voice conversion, International Conference on Acoustics, Speech, & Signal Processing (ICASSP) (2019)

[Kaneko 21] Kaneko, T., Kameoka, H., Tanaka, K. and Hojo, N.: MaskCycleGAN-VC: Learning non-parallel voice conversion with filling in frames, International Conference on Acoustics, Speech & Signal Processing(ICASSP) (2021)

[Kumar 19] Kumar, K., Kumar, R., Boissiere, de T., Gestin, L., Teoh, W.Z., Sotelo, J., Brebisson, de A., Bengio, Y and Courville, A.: MelGAN: Generative Adversarial Networks for conditional waveform synthesis, Electrical Engineering and Systems Science. Audio and Speech Processing(eess.AS) (2019)

[Oord 16] Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A generative model for raw audio, computer science system design(cs.SD) (2016)

[Sun 16] Sun, L., Li, K., Wang, H., Kang, H. and Meng, H.: Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. (2016)

[Trueba 18a] Trueba, J. L., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T. and Ling, Z.: The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods, Electrical Engineering and Systems Science. Audio and Speech Processing(eess.AS) (2018)

[Zhu 17] Zhu, J. Y., Park, T., Isola, P. and Efros, A. A.: Unpaired Image-to-Image translation using cycle-consistent adversarial networks, International Conference on Computer Vision (ICCV) (2017)