G-10

# 周波数-時間スペクトラムの機械学習による音声データ特徴変換

## Feature Transformation for Frequency-Time Spectrum of Speech Data Using Machine Learning

○平澤慶樹[1]，往蔵隆成[1]，塚本新[2]
*Keiju Hirasawa[1]，Ryusei Ohkura[1]，Arata Tsukamoto[2]

**Abstract :** In this study, we attempted to improve speaker anonymity by transforming speech data using CycleGAN to reduce speaker-specific information while retaining linguistic information. As a measure to verify the retention of linguistic information in the converted speech, we proposed the intelligibility evaluation using ZER of MCEP. Furthermore, to verify the reduction of speaker-specific information, speaker classification using the k-means method was performed, and based on this index, speech conversion in this study was shown to be a method that can reduce the individuality of input speech.

### 1　はじめに

1990 年代頃まではテレビやラジオなどによる限られた人が情報発信を行っていた. しかしこの 20 年, 情報端末の急速な進化と爆発的な普及により, インターネットを介して一般的に情報を発信できるようになった. それに伴い, 発信される音声や画像データに含まれる多種多様な情報から発信者の意思によらず個人が特定されるようなプライバシーリスクが生じるようになったことで, 匿名性を確保しつつ情報発信を行う技術への関心が高まっている.

本研究は, 音声の情報発信におけるプライバシーリスクに対し, 言語情報 (メッセージ) を保存しつつ, 話者固有の情報を低減させる音声変換モデルの形成により, 発信者の匿名性を向上させることを目的とする. 本稿では, 音声データに対して周波数-時間スペクトルの特徴量を推定し, 深層学習により個人性を低減する特徴量の変換モデルの作成を行い, 評価した. そして, 変換された特徴量から生成された音声データに対して, 話者の非識別化を検証するために話者分類を行った.

### 2　主要技術

#### 2.1　CycleGAN

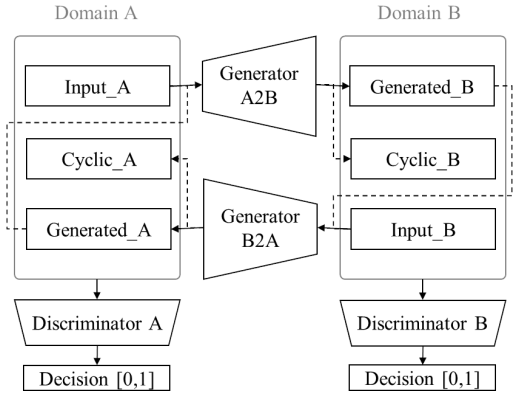


**Figure 1.** CycleGAN architecture

Figure 1. に本検討で用いる CycleGAN[1] のアーキテクチャを示す. Domain A の入力データ (Input_A) は Generator A2B により Domain B のデータ (Generated_B) に変換される. このデータは Generator B2A によって Domain A のデータ (Cyclic_A) に変換される. ドメイン間で一貫した変換を行うために, Cyclic_A は元のデータ (Input_A) に類似した分布である必要があるため, Discriminator A によってこれらのデータは偽物, 本物 [0,1] に分類される. Domain B の入力データ (Input_B) に対しても同様に変換, 分類が行われる.

### 3　実験方法

#### 3.1　音響特徴量の推定

メモリ消費の削減のため, 音声データは 24 kHz に変換した. 特徴量推定には WORLD[2](D4C edition[3]) を用いた. 音声データから WORLD の要件である基本周波数 F0, スペクトル sp, 非周期性指標 ap の推定を行った. F0 は対数化の後, 平均と標準偏差を評価した. sp は 24 次元でメルケプストラム (MCEP) に変換し, 標準化を行った.

#### 3.2　CycleGAN-VC2 を用いた MCEP の学習

CycleGAN-VC2[4] は CycleGAN を用いて音響特徴量の変換を行う先行研究で, 言語情報を保持しつつ, ペアとなる話者固有情報を相互変換できることが客観的に示されている. 本実験ではこのモデルアーキテクチャを用いて, 話者の非識別化を行うために MCEP の学習を行った.

#### 3.3　音声生成

変換元の MCEP を学習済みの Generator に入力して変換先と類似させた MCEP を生成し, sp に変換した. F0 を平均と標準偏差から変換先のピッチへと変換し, ap は変換せず使用し, WORLD により変換先と類似させた音声の生成を行った. 生成音声の言語情報の保持を検証するための指標として, MCEP(1 次) の零交差率 (ZCR) を用いた音声明瞭度の評価を提案した.

#### 3.4　話者分類

話者固有情報の非識別化を検証するために, 入力音声と生成音声の各データに対して F0, sp の平均値を算出し,

1:日大理工・院 (前)・情報　2:日大理工・教員・電子

k-means 法により話者数で分類し, その割合について評価を行った. 以降, 本稿における k-means 法はこれらの指標を用いた分類である.

### 3.5　データセット

Table 1 に実験 3.1-3.4 に用いたデータセットを示す.

**Table 1.** Dataset

| Item | Domain A | Domain B | Total time[s] |
|---|---|---|---|
| 1 | Male | Female | 600 |
| 2 | Male | Female(3 mixed) | 300 |

## 4　結果・考察

### 4.1　音声生成

Figure 2 は元音声, 生成音声に対してそれぞれ MCEP(1 次) の ZCR を算出した結果である. Fig.2 の凡例に示す X_fake は, 学習済み Generator により Tab.1, Item 1 の Domain X に変換した生成音声を表し, X_real は Domain X の元音声を表す. C, D は Tab.1, Item 2 の Domain B のうち, B_real を除く 2 者の元音声を表し, 1to3 は学習済み Generator により Domain B にした生成音声を表す. epoch 数はその時点における生成音声を表す. Fig.2 より平均値を比較すると, 生成音声は epoch 数の増加に従って MCEP(0 次) の ZCR が減少しており, 元音声は 0.64~0.66 付近でほぼ一定であることがわかった. また, 生成音声を聞いた結果, 2000 epoch では発話内容が不明瞭であったが, 20000 epoch では濁音や破裂音などの音圧が増加し, 発話内容の明瞭度が向上したことがわかった. このことから MCEP(1 次) の ZCR は, 音声の明瞭度と負の相関が存在すると考えられる. そして, いずれの epoch 数の生成音声についても, 元音声と同じ言語情報を一部または全部理解することができたことから, 本検討の生成音声における言語情報の保持は音声の明瞭度に依存することがわかった. Fig.2 の結果から, 5000 epoch 以上における生成音声は元音声に近い分布を取り, 言語情報を保持していると考える.

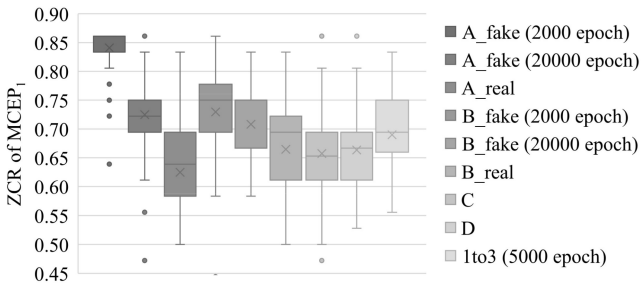


**Figure 2.** ZCR of MCEP (dim = 1)

### 4.2　話者照合

Table 2 は Tab.1, Item 1 のデータセットに用いた話者における元音声, 変換音声の他に, 音声変換に用いていない女性話者 C, D の 2 名を加えた 6 種類の音声について, k-means 法により話者数である K = 4 でクラスタリングを行った結果である. Tab.2 中の Other は変換音声 (fake) が A_real, B_real, C, D のいずれにも分類されなかった割合を表す. 変換音声が元音声と同じクラスタに分類された割合は A, B のいずれも 0 %であり, 本実験に用いたデータセットにおいて, 変換音声は元音声と異なる話者情報を持つと言える. ただし, 変換先音声の話者に分類されるため, 変換先音声も個人性を低減したものであることが望ましい.

そこで, Domain B を複数話者にし, 変換先音声の個人性低減を試みた. Table 3 は 1to3 と変換に用いた 4 名の話者である A, B, C, D に対して k-means 法により K = 4 でクラスタリングを行った結果である. 1to3 が話者 A に分類された割合は 0 %であることから, 1to3 の音声は話者 A とは異なる話者情報を持つと言える. また話者 B, C, D と同じクラスタに分類された割合の合計より, Other へ分類された割合が高いことから, 1to3 は話者 B, C, D の個人性を低減した音声と考えられる.

**Table 2.** Classification of dataset item 1 (K=4)

| | A_real | B_real | C | D | Other |
|---|---|---|---|---|---|
| A_fake | 26 | **0** | 0 | 0 | 74 |
| B_fake | **0** | 41 | 49 | 5 | 5 |

**Table 3.** Classification of dataset item 2 (K=4)

| | A | B | C | D | Other |
|---|---|---|---|---|---|
| 1to3 | **0** | 22 | 10 | 2 | 66 |

## 5　まとめ

本研究は, 音声の情報発信における話者の匿名性を向上させる目的に対し, CycleGAN を用いて言語情報を保持しつつ, 話者固有情報を低減する音声変換を行うことで達成を試みた. 変換音声の言語情報の保持を検証するための指標として, MCEP(1 次) の ZER による明瞭度評価を提案した. さらに, 話者固有情報の低減を検証するために k-means 法による話者分類を行い, 本指標に基づき, 本研究の音声変換は入力音声の個人性を低減可能である一方法として示した.

### 参考文献

[1] Jun-Yan Zhu, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[2] Morise Masanori, et al. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, pp. 1877–1884, 2016.

[3] M. Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, Vol. 84, pp. 57–65, 2016.

[4] Takuhiro Kaneko, et al. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.