7N-07

Cycle GAN を用いた音声匿名加工に必要な WORLD 音響特徴の構成

近藤 伊佐直 †
和歌山大学システム工学研究科 †

西村 竜一 ‡
和歌山大学データ・インテリジェンス教育研究部門 ‡

1 はじめに

教育・研究の現場において，グループワークや問題解決型学習（PBL）の導入が進んでいる．その際，対話を収録したデータを共有するため，プライバシー保護や相互評価の公平性の担保を目的に，収録データの話者匿名性の確保が必要となっている．2017 年個人情報保護法改正により，匿名加工情報を本人の同意を得ることなく利活用できるようになったことで，対話音声を匿名加工処理する技術の需要は高まると考えられる．

前報 [1] では，変換前の話者を特定できないように声質変換する音声匿名化システムの開発を報告した．開発システムは，マイクロホンアレイでの複数話者音声の収録，CycleGAN-VC[2] を用いた声質変換を行う．評価実験では，変換後音声の品質及び聴取による話者特定の困難性を示した．

本稿では，追加の検討課題として，音声変換処理で入力とする音声分析合成システム WORLD[3](D4C edition[4]) の抽出特徴の構成を検討した。特に，抽出特徴のうち，非周期性指標（**Ap**）が，話者特定に寄与する影響を調査する．

2 音声匿名化システム（開発システム）

本研究で開発している音声匿名化システムの構成について述べる．本システムの収録部では，マイクロホンアレイを用いて対話の音声を収録する．声質変換部では，入力信号の音響特徴を抽出，声質に相当する特徴量を深層学習で非線形変換し，変換で得た特徴量から音声信号を再合成する．

Configuration of WORLD acoustic features for speaker anonymization based on Cycle GAN

† Isanao Kondo, Wakayama University, Graduate School of Systems Engineering
‡ Ryuichi Nisimura, Wakayama University, Data Intelligence Education Research Division

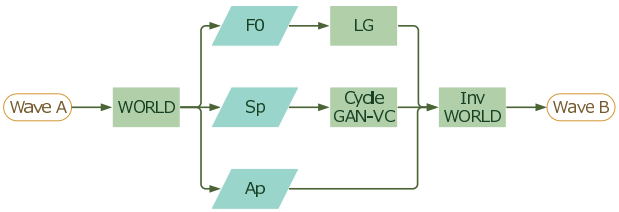

図 1　声質変換部の構成

開発システムの声質変換部の構成を図 1 に示す．**WaveA** が入力信号，**F0**，**Sp**，**Ap** が基本周波数，スペクトル包絡，非周期性指標の各特徴量ベクトル，**WaveB** が出力信号である．図中の各処理のうち，**WORLD**，**Inv WORLD** は，WORLD(D4C edition) による分析・再合成，**LG** が基本周波数変換，**CycleGAN-VC** がスペクトル包絡の変換を示す．

開発システムでは，声質変換のアルゴリズムに CycleGAN-VC を採用している．CycleGAN-VC は，2 話者間の声質変換をノンパラレルなデータセットで学習できる深層学習手法である．[2] では，音声分析合成システム WORLD(D4C edition) が推定した基本周波数（**F0**）とスペクトル包絡（**Sp**）の各特徴量を変換している．なお，非周期性指標（**Ap**）は変換しておらず、入力信号から抽出したベクトルをそのまま再合成に使用する．

3 特徴量構成の検討

本研究では，話者の特定に寄与する音響特徴量を実験で調査した．特に，前述のように，先行研究では変換処理をしていない非周期性指標（**Ap**）が話者特定に与える影響を確認することが目的である．

なお，前述の特徴量のうち，基本周波数（**F0**）を調査対象から除外している．基本周波数（**F0**）の値及び変動パターンや揺らぎといった動的な特徴は，一般に，個人に依存しており，個人性の特定に影響することが知られている．そのため，開発システムの声質変換部では，現時点で，基本周波数（**F0**）に変換処理を適用しており，あらためて調査する必要がないと考える．

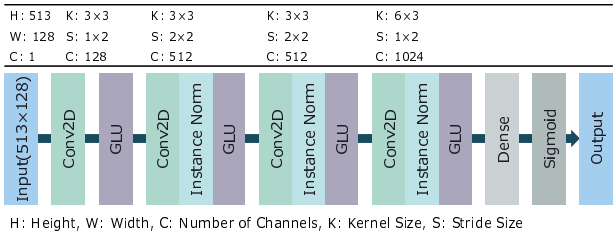

図 2　CNN 分類器の層構造

表 1　学習条件

項目	値
学習データセット	男女各 2 話者 × 10 文
検証データセット	男女各 2 話者 × 10 文
損失関数	L1 ノルム損失
最適化手法	Adam
特徴量次元数	513
フレーム長	128
フレームシフト	1
バッチサイズ	4096
学習ステップ	537,000 Iteration

3.1　実験内容

今回の実験では，機械学習による自動話者識別の正解率を比較することで，非周期性指標（**Ap**）とスペクトル包絡（**Sp**）の各特徴が，話者の特定に寄与する影響を考察する．

複数の条件で実験を行ったが、JVS corpus[5] の男性・女性各 2 話者（合計 4 話者）の音声データセットを用いたときの典型的な結果を示す．下記に実験の手順を示す。

手順 1.　入力音声から非周期性指標（**Ap**）とスペクトル包絡（**Sp**）を抽出

手順 2.　抽出した **Ap** 及び **Sp** の各ベクトルを次元毎に標準化

手順 3.　畳み込みニューラルネットワーク（CNN）を用いた 2 値分類器により話者を識別

上記の CNN 分類器は，CycleGAN-VC の Discriminator で用いられている 2 値分類器の入力層の次元数のみを 513 × 128 に変更したものを用いた．層構造を図 2，学習条件を表 1 に示す．

3.2　結果

自動識別の結果を表 2 に示す．SP がスペクトル包絡を入力に用いた場合，AP が非周期性指標を入力に用いた場合の正解率である．異なる男声間の 2 値識別，異なる女声間の 2 値識別，そして，男声と女声の 2 値識別の 3 通りすべての組み合わせにおいて，スペクトル包絡（**Sp**）を用いた方が，非周期性指標（**Ap**）の正解率を上回った．

表 2　自動識別の正解率（2 話者識別）

使用データ	SP(%)	AP(%)
男声，男声	**81.5**	67.0
男声，女声	**94.7**	81.2
女声，女声	**83.6**	67.9

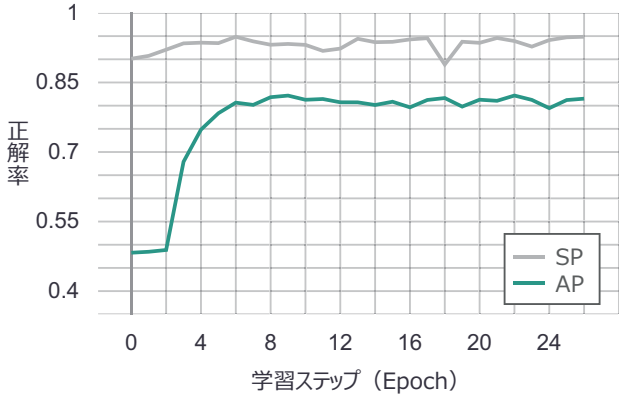

図 3　Epoch 毎の正解率の遷移

男声と女声を識別する条件において、学習の Epoch 毎の正解率の遷移を図 3 に示す．スペクトル包絡（**Sp**）では，1 つの Epoch で学習が大きく進み，正解率が 9 割を超えているのに対し，非周期性指標（**Ap**）では，学習の収束が遅く，かつ，正解率も 8 割程度となっている．

4　おわりに

本稿では，音声匿名加工において必要となる音響特徴の話者特定に対する寄与を調査した．男女すべての組み合わせにおいて，非周期性指標（**Ap**）を用いた場合の正解率は，スペクトル包絡（**Sp**）に比べて低い結果となった．この結果から，非周期性指標（**Ap**）からの話者特定は困難であることを確認した．開発システムにおいて，非周期性指標（**Ap**）を変換処理の特徴に追加導入する必要性は，これまでに引き続き，低いと考える。

謝辞 本研究は，JSPS 科研費 JP18K02862 の助成を受けて実施したものである．

参考文献

[1] 近藤, 西村, 第 19 回情報科学技術フォーラム (FIT 2020), vol. 2, pp. 17–20, 2020.

[2] Kaneko and Kameoka, Proc. EUSIPCO, pp. 2114–2118, 2018.

[3] Morise et al., IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[4] Morise, Speech Communication, vol. 84, pp. 57–65, 2016.

[5] Takamichi et al., arXiv preprint, 1908.06248, 2019.