Vol.2018-MUS-118 No.23
Vol.2018-SLP-120 No.23
2018/2/21

CycleGANを用いたクロスリンガル声質変換

房 福明[1]　Jaime Lorenzo-Trueba[1]　山岸 順一[1,2]　越前 功[1]

概要：CycleGAN は敵対的学習法を利用したニューラルネットワークにおいて，出力情報から入力情報を再現できるように一部の情報を保つ Cycle consistency 基準を加えたモデル構造であり，対にならないデータからでも異なるドメインへの変換関係を自動的に学習できるという特徴が知られている．この CycleGAN をノンパラレル声質変換に応用した場合，パラレル声質変換と同等もしくはそれ以上の話者性の変換が実現できることがわかっている．異なる言語対の音声データから声質変換モデルを学習する “クロスリンガル声質変換” は，このノンパラレル声質変換の特別なケースである．そこで，本研究では CycleGAN をクロスリンガル声質変換に利用し，その有効性を示す．

1. はじめに

声質変換は，ある話者の音声信号を変換し，あたかも別の話者の声に聞こえるように話者性を変換する技術である．クロスリンガル声質変換は，この声質変換のモデルを異なる言語対の音声データから学習することを意味する．この技術を利用することで，自動音声翻訳システムの出力音声をパーソナライズすることや，俳優本人の声での吹き替え映画を作成できると期待される．また，自身の音声を聞きながら外国語の発音練習する第二言語学習などへの応用も期待できる．

クロスリンガル声質変換を実現するため，これまで話者間の類似音響特徴のマッピング [1], [2], [3]，話者性の分離と置換 [4]，話者適応 [5] などが試みられてきた．通常のパラレル声質変換は DTW によりアライメントをとることでマッピングを定義できるが，この様なクロスリンガル声質変換ではマッピングを定義するフレーム対やセグメント対を異言語間で探し出し，定義する必要があった．しかし音素セットなどが大きく異なる言語間においてはその様なマッピングを適切に定義できないこともあるため，変換モデルに悪い影響を与えるという欠点があった．また，適応に用いる初期モデルを作成するためにパラレルデータベースやバイリンガルデータベースを必要とする手法もあるが，これらはシステムを柔軟に構築できないという欠点があった．

これらの問題点を解決するため，著者らは CycleGAN[6] を利用したノンパラレル声質変換の手法を提案し，高い話者性の変換を実現した [7]．CycleGAN は敵対的学習法を利用したニューラルネットワークにおいて，出力情報から入力情報を再現できるように一部の情報を保つ Cycle consistency 基準を加えたモデル構造であり，対にならないデータからでも異なるドメインへの変換関係を自動的に学習できるという特徴が知られている．クロスリンガル声質変換では，変換元の話者と変換先のターゲット話者は異る言語を話すため，ノンパラレル声質変換の特別なケースと見なすことができる．そこで，本研究は CycleGAN を利用したノンパラレル声質変換をクロスリンガル声質変換に利用し，その有効性を示す．

[1] 国立情報学研究所
〒 101-8430，東京都千代田区一ツ橋 2-1-2
[2] エジンバラ大学

2. CycleGAN に基づいた声質変換

図 1 に CycleGAN の構造を示す．CycleGAN は識別器 $D_{\mathbf{x}}$ と $D_{\mathbf{y}}$ 及び生成器 G と F を持つ．$\mathbf{x}$ と $\mathbf{y}$ は入力データであり，$\hat{\mathbf{y}} = G(\mathbf{x})$ と $\hat{\mathbf{x}} = F(\mathbf{y})$ は生成器により変換した出力データである．図に示しているように CycleGAN は 2 通りの変換方向がある．フォワードの $\mathbf{x} \to \hat{\mathbf{y}} \to \hat{\mathbf{x}}$ 及びバックワードの $\mathbf{y} \to \hat{\mathbf{x}} \to \hat{\mathbf{y}}$ である．これらの変換により，入力データは生成器を通して出力を再現しつつ，一部の入力情報を保つことが期待できる．言い換えれば，$\mathbf{x}$ と $\mathbf{y}$ をそれぞれソース話者とターゲット話者の音響特徴量として定義した場合，フォワード変換後の音響特徴量 $\hat{\mathbf{y}}$ は入力音響特徴量 $\mathbf{x}$ と同じ言語内容になっていることを期待しており，バックワード変換後の音響特徴量 $\hat{\mathbf{x}}$ と $\mathbf{y}$ も同じ言語内容であることを期待している．また，敵対的な学習により $\hat{\mathbf{y}}$ の分布は $\mathbf{y}$ の分布に近くなり，$\hat{\mathbf{x}}$ の分布は $\mathbf{x}$ の分布に近くなる．声質変換において，$\hat{\mathbf{y}}$ で表す話者性はターゲット話者に類似し，$\hat{\mathbf{x}}$ で表す話者性はソース話者に類似する．

Vol.2018-MUS-118 No.23
Vol.2018-SLP-120 No.23
2018/2/21

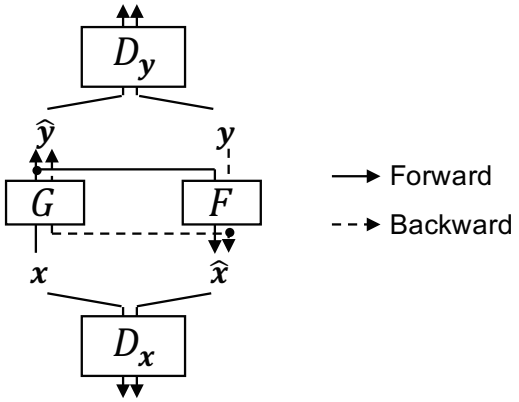

図 1 CycleGAN の構造．$D_{\mathbf{x}}$ と $D_{\mathbf{y}}$ は識別器であり，G と F は生成器を表す．$\mathbf{x}$ と $\mathbf{y}$ はそれぞれソース話者とターゲット話者の音響特徴量であり，$\hat{\mathbf{x}}$ と $\hat{\mathbf{y}}$ は変換したソース話者とターゲット話者の音響特徴量である．

その結果，対にならないノンパラレルデータを用いても声質変換モデルを学習できるようになる．そして，フォワードとバックワード変換により，ソース話者からターゲット話者への変換 G とターゲット話者からソース話者への変換 F を同時に実現できる．

クロスリンガル声質変換に応用する際に，フォワードとバックワード変換はソース話者とターゲット話者の異なる言語内での話者性変換を表す．つまりソース話者の言語を A とターゲット話者の言語を B とした場合，フォワードとバックワード変換はそれぞれ言語 A および言語 B 内での変換であり，CycleGAN を用いて異なる言語対の変換モデル同時に学習できると考えられる．本研究は [7] と同様に，音響特徴量の F_0，非周期成分とメルケプストラムを抽出し，各特徴量成分を独立的に変換を行い，変換した特徴量を再び合成する．F_0 については，ソース話者の $\log F_0$ をターゲット話者と同じ平均及び同じ分散となるように線形変換を行う．非周期成分については，変換を行わず直接ターゲット話者の特徴として使用する．メルケプストラムはさらに高次元と低次元成分を分けて変換を行う．高次元成分はスペクトル微細構造に対応するため，言語情報と話者情報をあまり含有していないことから，この成分を変換せず直接ターゲット話者の特徴量として使用する．低次元成分のメルケプストラムは，スペクトル包絡に対応し，言語情報と話者情報を明らかに反映するため，この成分は CycleGAN を用いて学習と変換を行う．

3. 実験

3.1 実験条件

本実験では，英語男性話者バラク・オバマ（第 44 代アメリカ合衆国大統領）をターゲット話者とした．ソース話者は日本語とアメリカ英語のバイリンガル女性話者であり，この話者による日本語音声データ，英語音声データ，日本語と英語を混ぜたデータの 3 種類をソースデータとした．これらの 3 種類のデータから変換モデルを学習し，変換した音声の品質，及び，ターゲット話者との類似性について 5 段階の MOS（mean opinion score）により評価した．

上述のバイリンガル話者の学習データは，アラジン日英・日中バイリンガル独話データベースに含まれている EJF101 話者であり，日本語発話と英語発話をそれぞれ 671 文章選択して構築した．ターゲット話者であるバラク・オバマの学習データはインタビューなど様々な音源から収集された．オバマの音声データにはノイズや反響が存在するため，まず SEGAN [8] に基づいた音声強調手法 [9] を用い，雑音および反響音の除去を行った．次に SN 比を計算し 50dB 以上の音声信号（2359 文章）をモデル学習に使用した．テストデータは，学習データに含まれていない 20 文章であり，バラク・オバマの就任演説を EJF101 話者スタジオで読み上げたものである．

全ての音声データは 16KHz にダウンサンプリングし，WORLD[10] と SPTK[11] を用いて音響特徴量を抽出した．メルケプストラムは 60 次元であり，最初の 35 次元を低次元成分，残りの 25 次元を高次元成分とした．動的特徴としてメルケプストラムの 1 回と 2 回微分も利用した．従って，CycleGAN を学習するための入力データは 105 次元のベクトル列である．CycleGAN は Tensorflow[12] より実装した．生成器と識別器はそれぞれ 6 層の全結合ニューラルネットワークを用いた．隠れ層のニューロン数はそれぞれ 128,256,256,128 に設定し，隠れ層の活性化函数は sigmoid 関数である．生成器の各隠れ層は Batch Normalization [13] を行った．生成器と識別器を更新するため，学習率はそれぞれ 0.001 と 0.0001 に設定し，バッチサイズはそれぞれ 128 と 4096 フレームである．エポックサイズは 50 に設定した．音声を合成する前に MLPG（maximum likelihood parameter generation）[14] とポストフィルタ [15] を用いて音響特徴量をスムージング・強調した．

3.2 実験結果

変換音声は被験者 9 人により評価された．結果を図 2 に示す．この図より，ソース話者は日本語話者であっても英語話者であっても，学習したモデルの性能にはあまり大きな差を見られなかったため，CycleGAN はクロスリンガル声質変換に有効であると考えられる．また，日本語と英語を混ぜて学習したモデルは一番高い性能を得た．これは学習データ量を増加したためと考えられる．

興味深いことに言語ペアにかかわらず変換性能はあまり差がなかったが，話者類似性のスコアは全体的に低かった．これは 2 つの原因が考えられる．1 つ目はターゲット話者オバマの音声データに雑音や反響音がいまだ存在し，モデルを精密的に学習することが難しいことである．2 つ目はイントネーションが言語ごとに異なるため，単純な線形変

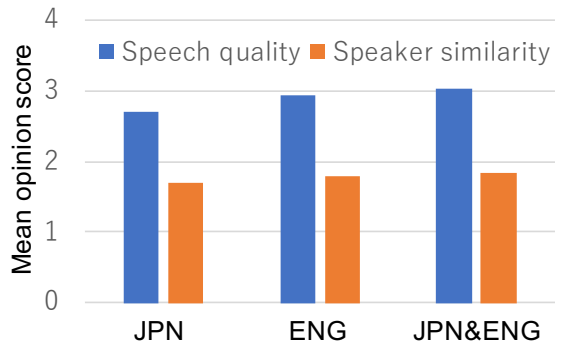

図 2　変換音声の品質及び話者類似性に関する被験者の評価結果．「JPN」，「ENG」と「JPN&ENG」はそれぞれ日本語ソース話者，英語ソース話者，日本語と英語を混ぜたソース話者により学習したモデルを表す．

換を行うだけでは F_0 を正しく変換できないことである．

4. まとめと課題

本研究では CycleGAN をクロスリンガル声質変換へ応用することを試みた．日本語話者と英語話者のデータを用いて変換の有効性を検証した．今後の課題は話者類似性の改善や様々な言語を用いた検証などが挙げられる．

謝辞

本研究の一部は MEXT 科研費 15H01686，16H06302 と 17H04687 の助成を受けたものである．

参考文献

[1] Masanobu Abe, Kiyohiro Shikano, and Hisao Kuwabara, “Cross-language voice conversion,” in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 345–348.

[2] Daniel Erro and Asunción Moreno, “Frame alignment method for cross-lingual voice conversion,” in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[3] David Sündermann, Harald Höge, Antonio Bonafonte, Hermann Ney, and Julia Hirschberg, “Text-independent cross-language voice conversion.,” in *INTERSPEECH*, 2006.

[4] Victor Popa, Jani Nurminen, and Moncef Gabbouj, “A study of bilinear models in voice conversion,” *Journal of Signal and Information Processing*, vol. 2, no. 02, pp. 125, 2011.

[5] Malorie Charlier, Yamato Ohtani, Tomoki Toda, Alexis Moinet, and Thierry Dutoit, “Cross-language voice conversion based on eigenvoices,” in *INTERSPEECH*, 2009.

[6] J. Zhu, T. Park, P. Isola, and A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[7] 房 福明, 山岸 順一, and 越前 功, “Cyclegan を用いた高品質なノンパラレル声質変換,” in 第 *119* 回音声言語情報処理研究会, 2017.

[8] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, “Segan: Speech enhancement generative adversarial network,” *arXiv preprint arXiv:1703.09452*, 2017.

[9] Xin Wang Jaime Lorenzo-Trueba and Junichi Yamagishi, “Stealing your vocal identity from the internet: cloning obama’s voice from found data using gan and wavenet,” in 第 *120* 回音声言語情報処理研究会, 2018.

[10] M. Morise, F. Yokomori, and K. Ozawa, “World: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[11] SPTK Working Group et al., “Speech signal processing toolkit (SPTK),” *http://sp-tk.sourceforge.net*, 2009.

[12] “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, Software available from tensorflow.org.

[13] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, 2015, pp. 448–456.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1315–1318.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis,” *Systems and Computers in Japan*, vol. 36, no. 12, pp. 43–50, 2005.