CycleGANを用いたパラレルデータフリー声質変換*

◎金子卓弘, 亀岡弘和 (NTT)

1 はじめに

言語情報を保ちながら非言語またはパラ言語情報を変換する技術を声質変換 (Voice Conversion: VC) と呼ぶ。声質変換はソース音声からターゲット音声への変換関数を推定する回帰問題として定式化することができ，従来手法の多くは，時間的にアライメントのとれたパラレルデータを用いてこの変換関数を学習していた。しかし，このようなデータを収集するのは容易ではなく，また，タスクによっては収集が難しいこともあった。さらに，パラレルデータを集められたとしても，時間的なアライメントを自動的に完璧に行うことは容易ではなく，手作業によるアライメントの補正やデータの選別などが行われていた。

以上の背景を動機として，本研究ではパラレルデータフリーな声質変換問題を扱う。特に，本研究では，高品質で汎用性のある (トランスクリプトや参照データなどの外部データ，ASR などの外部モジュールを用いない) パラレルデータフリー声質変換手法 (**CycleGAN-VC**) を提案する。本手法で用いる CycleGAN [2] では，順変換と逆変換のマッピング関数を Adversarial Loss [3] と Cycle-Consistency Loss [4] とともに同時学習することにより，ペアデータがない状況下で，擬似的なペアデータを見つけ出しながら学習することを可能にする。また，Adversarial Loss には統計的平均化の影響を避ける効果があることが知られており [5, 6, 7, 8]，これにより，統計的手法によって得られた音声の品質劣化の主要な原因の一つである過剰な平滑化の問題を緩和することができる。さらに，CycleGAN を Gated CNN [9] を用いて構成し，Identity-Mapping Loss [10] とともに学習することで，言語情報を保持しながら，時系列的・階層的構造を捉えることを可能にする。なお，本稿は 2017 年 11 月に arXiv に投稿した原稿 [1] の日本語版である。本研究と同時期 (2017 年 12 月の SLP 研究会) に，扇ら [11] も CycleGAN を用いた声質変換手法を提案しているが，当手法はフレームごとの変換のため時系列構造については考慮しておらず，また，言語情報の保持については将来課題として残されていた。

提案手法の評価は，Voice Conversion Challenge 2016 (VCC 2016) データセット [12] を用いて行った。客観評価では，CycleGAN-VC によって得られた特徴量系列ではターゲットに近い系列内変動 (Global Vairiance: GV) [13] と変調スペクトル (Modulation Spectra: MS) [14] が得られていることを示す。主観評価では，パラレルデータを用いて学習した GMM 声質変換 [13] に匹敵する，または上回ることを示す。

2 CycleGANを用いたパラレルデータフリー声質変換

本研究の目的は，ソースデータ $x \in X$ からターゲットデータ $y \in Y$ への変換関数をパラレルデータを要することなく学習することである。本研究では，この問題を CycleGAN [2] をベースにして解く。本章では，第 2.1 節で CycleGAN の概要を説明し，第 2.2 節で提案手法 (CycleGAN-VC) について説明する。

2.1 CycleGAN

CycleGAN では，変換関数 $G_{X\to Y}$ と逆変換関数 $G_{Y\to X}$ を Adversarial Loss [3] と Cycle-Consistency Loss [4] の二つの目的関数を用いて学習する。

Adversarial Loss: Adversarial Loss は以下の式で表される。

$$\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = \mathbb{E}_{y\sim P_{\mathrm{Data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x\sim P_{\mathrm{Data}}(x)}[\log(1 - D_Y(G_{X\to Y}(x)))] \quad (1)$$

生成器 $G_{X\to Y}$ は，この目的関数を最小化することによって識別器 D_Y がターゲットデータ y と区別できないようなデータを生成できるようにする。一方，D_Y は，この目的関数を最大化することによって，$G_{X\to Y}$ が生成したデータと y とを見分けられるようにする。このように敵対的な条件下で最適化することによって，変換データの分布 $P_{G_{X\to Y}}(x)$ をターゲットデータの分布 $P_{\mathrm{Data}}(y)$ に近づけることができる。なお，CycleGAN では逆変換 $G_{Y\to X}$ についても同様に Adversarial Loss，$\mathcal{L}_{adv}(G_{Y\to X}, D_X)$ を考える。

Cycle-Consistency Loss: Adversarial Loss は変換データがターゲットデータの分布に従うかどうかという制約しか与えないため，変換前と変換後でコンテキストが保持されているかどうかは保証しない。この問題に対処するため，CycleGAN では以下の式で表される Cycle-Consistency Loss を導入する。

$$\begin{aligned}&\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X})\\ &= \mathbb{E}_{x\sim P_{\mathrm{Data}}(x)}[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1]\\ &+ \mathbb{E}_{y\sim P_{\mathrm{Data}}(y)}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1]\end{aligned} \quad (2)$$

上式により，変換して逆変換すると元に戻るという構造的な制約を与えることができる。これにより，コンテキストを保持しながら変換することが可能になる。

2.2 CycleGAN-VC

CycleGAN を音声データに適用するために，本研究では Gated CNN [9] を用いて構成し，Identity-Mapping Loss [10] とともに学習する手法 (CycleGAN-VC) を提案する。

Gated CNN: 音声データの重要な性質の一つに，有声・無声，音素・形態素といった時系列的・階層的構造があることが挙げられる。本研究では，Gated CNN を用いてこの構造をモデル化する。Gated CNN は系列データに対して並列化が可能なため高効率であるのに加え，言語モデル [9] や音声モデル [5] として高い性能を持つことが示されている。Gated CNN では，Gated Linear Unit (GLU) が活性化関数として用いられ，$l+1$ 層の出力 $\boldsymbol{H}_{l+1}$ を l 層の出力 $\boldsymbol{H}_l$ とモデルパラメータ $\boldsymbol{W}_l$，$\boldsymbol{V}_l$，$\boldsymbol{b}_l$，$\boldsymbol{c}_l$ を用いて計算する。

$$\boldsymbol{H}_{l+1} = (\boldsymbol{H}_l * \boldsymbol{W}_l + \boldsymbol{b}_l) \otimes \sigma(\boldsymbol{H}_l * \boldsymbol{V}_l + \boldsymbol{c}_l) \quad (3)$$

$\otimes$ は要素積で，σ はシグモイド関数である。ゲート構造を用いることで情報を選択的に伝播することができ，時系列的・階層的構造のモデル化が可能である。

Identity-Mapping Loss: Cycle-Consistency Loss は変換時に構造的な制約を与えるが，言語情報を保持するかどうかということに対しては十分な制約を与えない。しかし，声質変換においては言語情報の保持は重要である。そこで，この問題を解決するため，本研究では，以下の式で表される Identity-Mapping Loss を用いる。

$$\begin{aligned}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) &= \mathbb{E}_{y\sim P_{\mathrm{Data}}(y)}[\|G_{X\to Y}(y) - y\|_1]\\ &+ \mathbb{E}_{x\sim P_{\mathrm{Data}}(x)}[\|G_{Y\to X}(x) - x\|_1]\end{aligned} \quad (4)$$

---

*Parallel-data-free voice conversion using CycleGAN. by KANEKO, Takuhiro and KAMEOKA, Hirokazu (NTT). 本稿は [1] の日本語版である。

3　実験的評価

3.1　実験設定

提案手法の有効性を検証するために，VCC 2016 データセット [12] を用いて実験を行った。VCC 2016 データセットは米国英語話者 10 人 (男性 5 人，女性 5 人) のパラレルな 216 発話 (約 13 分) から構成されており，これらのうち 162 文は学習データ，54 文は評価データとして提供されている。音声データは 16 kHz にダウンサンプリングし，WORLD [15] を用いて 5 ms ごとに 24 次のメルケプストラム (MCEP)，対数基本周波数 ($\log F_0$)，非周期性指標 (AP) の抽出を行った。これらの音響特徴量のうち，メルケプストラムは提案手法によって変換し，対数基本周波数は平均と分散をもとに線形変換し，非同期性指標はソース音声のものをそのまま用いた。

ネットワーク構造としては，先行研究 [5] を参考に生成器には 1D CNN を用い，識別器には 2D CNN を用いた。具体的には，生成器は 2 層の Downsampling 層，6 層の Residual 層，2 層の Upsampling 層を含み，特に，Upsampling 層では Pixel Shuffler [16] を用いて Upsampling を行った。識別器には 4 層の Convolution 層，1 層の Fully Connected 層を含む構造を用いた。また，生成器，識別器ともに入力層と出力層以外は，Instance Normalization [17] を用いて正規化を行った。

学習データとしては，パラレルデータフリーな条件下で評価を行うため，学習データ 162 文のうち，前半 81 文をソース音声，後半 81 文をターゲット音声として学習に用いた。前処理としては，学習データの統計情報を用いてメルケプストラムを各次元ごとに正規化した。GAN の目的関数としては学習の安定化のため LSGAN [18] を用いた。最適化には Adam [19] を用い，バッチサイズは 1 とした。バッチは，音声データからランダムに抽出したセグメント (128 フレーム) を用いて構成した。以下では，スペースの関係でソース音声に SF1，ターゲット音声に TF1 を用いた場合の変換結果について述べる。他の変換例についてはデモページ [1] で公開しているので参照されたい。

3.2　客観評価

本実験では，提案手法はメルケプストラムの変換に適用されているため，変換メルケプストラムの品質について評価を行った。評価指標としては，主観評価とも相関が高いと言われている GV [13] と MS [14] を用いた。また，比較手法としてはパラレルデータありの声質変換としてよく使われている GMM ベースの手法 (GMM-VC) [13] を用いた。なお，GMM-VC は学習時にパラレルデータが必要であるため，学習データ 162 文全てを用いて学習を行った。これは，CycleGAN-VC と比べて有利な条件 (データ量 2 倍，パラレルデータあり) であることに留意されたい。Fig. 1(a) にメルケプストラム次数ごとの GV の比較，Fig. 1(b) に変調周波数ごとの MS の比較を示す。これらの結果より，提案手法 (CycleGAN-VC) によって得られた特徴量系列ではターゲット音声に近い GV と MS が得られていることが分かる。

3.3　主観評価

主観評価は，VCC 2016 に従い，自然性については MOS 評価を行い，話者性についてはソース・ターゲット音声との類似度について評価を行った。自然性の評価にはランダムに選択した 20 文を用い，話者性の評価にはランダムに選択した 10 ペア文を用いた。被験者としては英語教育を十分に受けた 9 人が参加した。なお，サンプル音声はデモページ [1] で公開しているので参照されたい。また，ベースラインとしては VCC 2016 のベースライン音声 (GMM ベースの手法 [13]) [2] を用いた。Fig. 2(a) に自然性の MOS 評価の結果，Fig. 2(b) にソース・ターゲット音声との類似度評価の結果を示す。結果より，提案手法はベースラインを上回るか匹敵するかであると言える。なおベースラインは有利な条件下 (データ量 2 倍，パラレルデータあり) で学習していることに留意されたい。

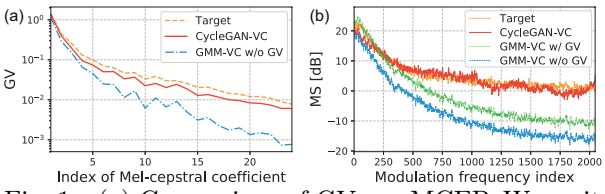

Fig. 1　(a) Comparison of GV per MCEP. We omit GMM-VC w/ GV in this figure because it directly estimates GV. (b) Comparison of MS per mudulation frequency (11th Mel-cepstrum).

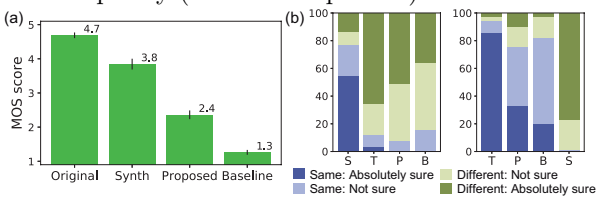

Fig. 2　(a) MOS for naturalness with 95% confidence intervals. (b) Similarity to source speaker and to target speaker (S: Source, T: Target, P: Proposed, and B: Baseline.)

4　おわりに

本研究では，高品質で汎用性のあるパラレルデータフリー声質変換の実現のため，CycleGAN-VC と呼ぶ CycleGAN を Gated CNN を用いて構成し，Identity-Mapping Loss とともに学習する方法を提案した。客観評価では，GV と MS についてターゲット音声に近い変換音声が得られていることを示し，主観評価では，パラレルデータを用いて学習した GMM 声質変換に匹敵する，または上回ることを示した。提案手法は汎用的な手法であり，将来研究としては，スペクトルや波形などの他の特徴量への応用，音声強調や発音変換など他のタスクへの応用を検討している。

謝辞　本研究は JSPS 科研費 17H01763 の助成を受けて実施された。

参考文献

[1] T. Kaneko, H. Kameoka, *arXiv preprint arXiv:1711.11293*. 2017.
[2] J.-Y. Zhu et al., *Proc. ICCV*, pp. 2223–2232, 2017.
[3] I. Goodfellow et al., *Proc. NPIS*, pp. 2672–2680, 2014.
[4] T. Zhou et al., *Proc. CVPR*, pp. 117–126, 2016.
[5] T. Kaneko et al., *Proc. INTERSPEECH*, pp. 1283–1287, 2017.
[6] T. Kaneko et al., *Proc. ICASSP*, pp. 4910–4914, 2017.
[7] T. Kaneko et al., *Proc. INTERSPEECH*, pp. 3389–3393, 2017.
[8] Y. Saito et al., *Proc. ICASSP*, pp. 4900–4904, 2017.
[9] Y. N. Dauphin et al., *Proc. ICML*, pp. 933–941, 2017.
[10] Y. Taigman et al., *ICLR*, 2017.
[11] 房ら, **情報処理学会研究報告**, pp. 1–6, 2017.
[12] T. Toda et al., *Proc. INTERSPEECH*, pp. 1632–1636, 2016.
[13] T. Toda et al., *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.
[14] S. Takamichi et al., *Proc. ICASSP*, pp. 290–294, 2014.
[15] M. Morise et al., *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.
[16] W. Shi et al., *Proc. CVPR*, pp. 1874–1883, 2016.
[17] D. Ulyanov et al., *arXiv preprint arXiv:1607.08022*. 2016.
[18] X. Mao et al., *Proc. ICCV*, pp. 2794–2802, 2016.
[19] D. Kingma, J. Ba, *Proc. ICLR*, 2015.

[1] CycleGAN-VC のサンプル音声: `http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc`

[2] VCC 2016 のベースライン音声: `http://dx.doi.org/10.7488/ds/1575`