CycleGAN を用いた合成音声から自然音声への波形変換*

◎田中 宏，金子 卓弘，北条 伸克，亀岡 弘和（NTT)

1 はじめに

多くのテキスト音声合成システムや音声変換システムでは，ボコーダ方式により音声特徴量から音声が合成される．ボコーダ方式による音声合成は，コンパクト（低次元）な特徴量表現が可能であり，特に限られた数の学習サンプルしか得られない状況において大変有益である．一方で，ボコーダ特有の機械的な音質や生成・変換された音声特徴量の過剰な平滑化などにより，合成音声の音質劣化が問題であった．

この問題に対して，音声特徴量空間上でより自然な音声特徴量へ補正する方法が提案されている．例えば，テキスト音声合成や音声変換において加工された音声特徴量の変調スペクトル（Modulation Spectrum: MS）を自然な音声の MS へ補正する手法 [1] や，加工・変換した音声特徴量に対して，Generative Adversarial Networks （GAN）を用いて自然性を向上させる成分を足しこむことで自然な音声の音声特徴量へと補正する手法 [2] が提案されている．上述の手法は，一定量の音質改善を達成しているもののコンパクト空間での補正であることは変わりなく，また最終的な音声合成部はボコーダを通るため，音質改善の潜在的な限界が存在する．一方で，GAN を用いて音声波形に対する直接的な補正を行う手法（SEGAN）[3] も提案されている．音声波形を入力として直接補正を行うため，音声特徴量空間上での補正と比較するとより大きな品質改善が見込まれる．しかしながら，適用場面が限られており，学習データ中の入力波形と理想とする目標波形の間で理想的なアライメントが取られている場合において有効である．例えば，理想環境で収録された音声に対して，計算機上で雑音を重畳し雑音環境下音声を生成したのち雑音除去を行う場合は，入力音声である雑音環境下音声と目標音声である理想環境で収録された音声のアライメントは完璧であるため，音質改善が可能である．しかしながら，テキスト音声合成や音声変換において生成された合成音声から自然な音声への補正は，上述のアライメント問題により [3] の単純適用では品質改善が難しい．

本稿では，cycle-consistent adversarial networks を用いて合成音声から自然音声へ音声波形を直接補正する手法を提案する．実験的評価結果より，ボコーダ特有の機械的な音質が改善され，音質に関する自然さの大幅な改善を達成していることを示す．

2 SEGAN [3]

ソース音声（例えば，計算機上で雑音を重畳した雑音環境下音声）を $\boldsymbol{x}$，ターゲット音声（例えば，理想環境で収録された自然音声）を $\boldsymbol{y}$ とする．SEGAN [3] では，目的関数 $\mathcal{L}_{\mathrm{SEGAN}}$ が次式で与えられる．

$$\mathcal{L}_{\mathrm{SEGAN}} = \mathcal{L}_{\mathrm{adv}}(G_{\boldsymbol{x}\to\boldsymbol{y}}, D_{\boldsymbol{y}}) + \lambda_{\mathrm{add}}\mathcal{L}_{\mathrm{add}}, \quad (1)$$

ここで，

$$\begin{aligned}\mathcal{L}_{\mathrm{adv}}&(G_{\boldsymbol{x}\to\boldsymbol{y}}, D_{\boldsymbol{y}}) \\ &= \mathbb{E}_{y\sim P_{\mathrm{Data}(y)}}[\log D_{\boldsymbol{y}}(\boldsymbol{y})] \\ &\quad + \mathbb{E}_{x\sim P_{\mathrm{Data}(x)}}[\log(1 - D_{\boldsymbol{y}}(G_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x}))], \quad (2)\end{aligned}$$

$$\mathcal{L}_{\mathrm{add}} = ||G_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}||_1, \quad (3)$$

であり，$G_{\boldsymbol{x}\to\boldsymbol{y}}$ はソース音声 $\boldsymbol{x}$ からターゲット音声 $\boldsymbol{y}$ へと変換する変換関数であり，$D_{\boldsymbol{y}}$ は変換された音声 $G_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x})$ とターゲット音声 $\boldsymbol{y}$ を見分ける識別関数である．目的関数が $\mathcal{L}_{\mathrm{adv}}(G_{\boldsymbol{x}\to\boldsymbol{y}}, D_{\boldsymbol{y}})$ のみである通常の GAN では音韻性を維持する保証がないため，$\mathcal{L}_{\mathrm{add}}$ を導入することで学習が安定する．λ_{add} は追加項を制御する重みパラメータである．変換関数 $G_{\boldsymbol{x}\to\boldsymbol{y}}$ と識別関数 $D_{\boldsymbol{y}}$ はニューラルネットで記述されており，$G_{\boldsymbol{x}\to\boldsymbol{y}}$ は $D_{\boldsymbol{y}}$ を騙せるように，$D_{\boldsymbol{y}}$ は $G_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x})$ を識別できるように競合しながら学習される．

3 CycleGAN を用いた合成音声から自然音声への波形変換

3.1 コンセプト

ソース音声とターゲット音声のアライメントが完璧である場合，SEGAN [3] は大幅な音質改善が可能である．一方で，ソース音声とターゲット音声がパラレルなデータでない場合（テキスト音声合成や音声変換において生成された合成音声から自然な音声への補正など）は，音声の音韻性が崩れ，単純適用が難しいことを予備実験で確認した．音声の音韻性が崩れる要因の一つに，学習時において，アライメント問題によりソース音声とターゲット音声の音韻が違うにも関わらず，式 (3) のような目的関数を最小化しようとすることがある．

CycleGAN [4] は，cycle-consistent adversarial networks を導入することで，2 つの異なるドメイン間の変換を学習する GAN である．Cycle-consistent adversarial networks は，その名の通り，ターゲットドメインへ変換されたのち，元のドメインへ再度変換された際の再構成誤差を含む目的関数を用いて学習される．すなわち，ソース音声と，ターゲットドメインへ変換されたのち元のドメインへ再度変換された音声とのアライメントは完璧であり，音韻の不一致が存在しない．それゆえ，音韻性を崩さずに，音質の自然さを改善できる可能がある．

*Synthetic-to-Natural Speech Transformation Based on CycleGAN. by TANAKA, Kou, KANEKO, Takuhiro, HOJO, Nobukatsu, KAMEOKA, Hirokazu (NTT)

3.2 CycleGAN を用いた波形変換

CycleGAN [4] では，目的関数を次式で与えられる．

$$\mathcal{L} = \mathcal{L}_{\text{adv}}(G_{\boldsymbol{x}\to\boldsymbol{y}}, D_{\boldsymbol{y}}) + \mathcal{L}_{\text{adv}}(G_{\boldsymbol{y}\to\boldsymbol{x}}, D_{\boldsymbol{x}}) + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}. \quad (4)$$

ここで，

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G_{\boldsymbol{y}\to\boldsymbol{x}}, D_{\boldsymbol{x}}) &= \mathbb{E}_{x\sim P_{\text{Data}(x)}}[\log D_{\boldsymbol{x}}(\boldsymbol{x})] \\ &+ \mathbb{E}_{y\sim P_{\text{Data}(y)}}[\log(1 - D_{\boldsymbol{x}}(G_{\boldsymbol{y}\to\boldsymbol{x}}(\boldsymbol{y}))], \quad (5)\\ \mathcal{L}_{\text{cyc}} &= \mathbb{E}_{x\sim P_{\text{Data}(x)}}[||G_{\boldsymbol{y}\to\boldsymbol{x}}(G_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x})) - \boldsymbol{x}||_1] \\ &+ \mathbb{E}_{y\sim P_{\text{Data}(y)}}[||G_{\boldsymbol{x}\to\boldsymbol{y}}(G_{\boldsymbol{y}\to\boldsymbol{x}}(\boldsymbol{y})) - \boldsymbol{y}||_1], \quad (6)\end{aligned}$$

であり，$\mathcal{L}_{\text{adv}}(G_{\boldsymbol{y}\to\boldsymbol{x}}, D_{\boldsymbol{x}})$ は 2 節で述べた目的関数 $\mathcal{L}_{\text{adv}}(G_{\boldsymbol{x}\to\boldsymbol{y}}, D_{\boldsymbol{y}})$ のソース音声とターゲット音声を入れ替えた場合と同等であり，$\mathcal{L}_{\text{cyc}}$ は再構築誤差に関する制約項，λ_{cyc} は再構成誤差を制御する重みパラメータである．$\mathcal{L}_{\text{cyc}}$ を最小化するためには，変換した音声 $G_{\boldsymbol{y}\to\boldsymbol{x}}(\boldsymbol{y})$ および $G_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x})$ がそれぞれ元の音声 $\boldsymbol{x}$ および $\boldsymbol{y}$ を再構築できるだけの情報を保持している必要がある．したがって，学習が成功した際の $G_{\boldsymbol{y}\to\boldsymbol{x}}$ および $G_{\boldsymbol{x}\to\boldsymbol{y}}$ は，$\boldsymbol{x}$ および $\boldsymbol{y}$ に共通する構造を保ったまま変換する関数となる．例えば，合成音声と自然音声に共通する音韻情報は保ったまま，合成音声と自然音声とを聞き分ける要因になる成分（ボコーダ方式による合成音声が持つボコーダ特有の機械的な音質など）が変換されることになる．

4 実験的評価

4.1 実験条件

女性話者の音声データセット 437 文中，提案手法を学習するために 407 文（約 1 時間）を，評価データとして 30 文（4 分）を用いて，音質の自然さに関する主観評価実験を行った．音声のサンプリングレートは 22.05 kHz である．適用先の手法として， [2] 中における DNN テキスト音声合成手法 **SYN** を用いた．提案手法のモデル構造として，$G_{\boldsymbol{y}\to\boldsymbol{x}}$ および $G_{\boldsymbol{x}\to\boldsymbol{y}}$ は 3 層の CNN，6 層の residual block（1 residual block につき 2 層の CNN），3 層の CNN を積み重ねたネットワークを，$D_{\boldsymbol{x}}$ および $D_{\boldsymbol{y}}$ は 4 層の CNN を用いた．

4.2 音質の自然さに関する聴取実験

表 1 に，自然音声（**Natural**），提案手法適用前音声（**Baseline**），提案手法適用後音声（**Proposed**）に対する音質の自然さに関する 5 段階オピニオンスコア（1: 合成音声らしい ~5: 自然音声らしい）の評価結果を示す．提案手法を用いることで，音質の自然さが大幅に改善されていることがわかる．さらに，「ボコーダ特有の機械的な音質が十分に改善されている」と被験者からのフィードバックを得ている．しかしながら，自然音声と比較すると未だギャップがあること

Table 1 音質の自然さに関する 5 段階オピニオンスコアによる音声の聴取実験結果．（比較手法 3 手法 × 評価データ 30 文 × 被験者 10 名）

手法	平均値	95 % 信頼区間
Natural	4.733	± 0.093
Baseline	2.322	± 0.180
Proposed	3.904	± 0.145

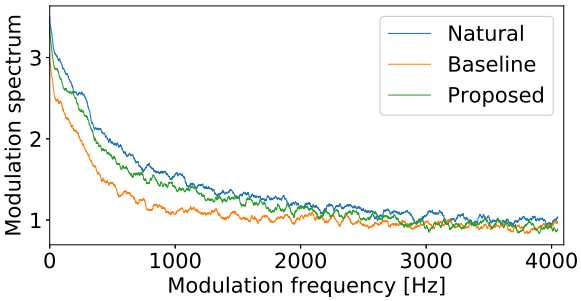

Fig. 1 横軸を変調周波数とした第 14 次メルケプストラム系列の変調スペクトル．

がわかる．このギャップの原因として，「提案手法適用後音声は，かすれた音声に聞こえる部分があり，この部分において自然音声と識別可能である」とのフィードバックも得ている．また，提案法により得られた波形から分析された第 14 次メルケプストラム系列の変調スペクトルを図 1 に示す．提案手法のような波形に対する直接的な補正においても，メルケプストラム系列の変調スペクトルが回復していることがわかる．

5 おわりに

本稿では，cycle-consistent adversarial networks を用いて合成音声から自然音声へ音声波形を直接補正する手法を提案した．実験的評価結果より，音質の自然さが大幅に改善されていることを示した．今後は，提案手法適用により生じる「かすれた音声に聞こえる部分」の改善，および，ボコーダ方式の音声合成システムの研究に取り組む予定である．

謝辞 本研究の一部は，JSPS 科研費 17H01763 の助成を受け実施したものである．

参考文献

[1] Shinnosuke Takamichi *et al.*, in *Proc. ICASSP*, 2014, pp. 290–294.

[2] Takuhiro Kaneko *et al.*, in *Proc. ICASSP*, 2017, pp. 4910–4914.

[3] Santiago Pascual *et al.*, *arXiv preprint arXiv:1703.09452*, 2017.

[4] Jun-Yan Zhu *et al.*, *arXiv preprint arXiv:1703.10593*, 2017.