CycleVAE 型声質変換を用いた構音障害者のための高明瞭度音声合成 *

◎ 松原圭亮[1,2]，岡本拓磨[2]，高島遼一[1]，滝口哲也[1]，戸田智基[3,2]，志賀芳則[2]，河井恒[2]
[1] 神戸大学，[2] 情報通信研究機構，[3] 名古屋大学

1 はじめに

構音障害とは，発声器官の障害や運動機能障害などによって正しい発声が困難となる障害である。構音障害者はその発声の聞き取りづらさから他者とのコミュニケーションが難しく，それによって引き起こされる社会活動での障壁を取り除くための支援技術の需要がある。本研究ではアテトーゼ型脳性麻痺による構音障害者を対象としたコミュニケーション支援のためのニューラル音声合成手法を提案する．

構音障害者へ向けたコミュニケーション支援の具体的なアプローチとしては，声質変換 (Voice conversion: VC) を用いて障害者音声の明瞭性を向上させる手法と，テキスト音声合成 (Text-to-Speech: TTS) を用いて障害者が入力したテキストを読み上げる手法に大別される。前者は近年の深層学習を用いた声質変換技術の発展に伴って様々な手法が提案されているが [1, 2]，障害者の言語的特徴が失われた音声からそれらを復元するタスクは一般的な声質変換よりも難しく，品質および明瞭性には限界がある。後者の手法は構音障害者向けに特別な処理を行う必要はないが，構音障害者からは自分らしい声でコミュニケーションをとりたいという需要が多く，彼らの話者性を持ちつつ明瞭性の高い音声が合成可能な TTS システムへの需要がある。このようなアプローチの代表例としては，山岸ら [3] の Voice banking project 等がある。

本研究では，一般的な声質変換タスクで高い精度を達成している既存の声質変換モデルを活用して構音障害者のための TTS システムを構築することを考える。前述のとおり，近年の深層学習の発展に伴って声質変換でも様々な手法が提案されている。特に敵対的生成ネットワーク (Generative Adversarial Network: GAN) や変分自己符号化器 (Variational Autoencoder: VAE) 等の深層生成モデルを用いた手法は，学習にパラレル発話を必要とせず比較的少量のデータでも学習が可能になっている。しかし，一般に音声の局所的な特徴 (言語特徴等) を維持しながら大局的な特徴 (話者性等) を変換すれば良いのに対し，構音障害者から健常者への声質変換で明瞭性を向上させるタスクは言語特徴などの変換も必要となるため，より問題は複雑化する。そこで本研究では，健常者から構音障害者への声質変換を行う。これは健常者音声に構音障害者の話者性を付与する課題であり，構音障害者音声の明瞭性改善よりも比較的簡単な課題であると考えられる。本研究では，Transformer 型音響モデル [4] による健常者 TTS，CycleVAE (CycVAE) [5] による健常者から構音障害者への声質変換，LPCNet ボコーダ (LPCN) [6] による音声生成を合わせた音声合成手法を提案する。CycleVAE は Voice Conversion Challenge 2020 (VCC2020) においてクロスリンガル声質変換タスクにおける有効性が確認されており [7]，健常者と構音障害者間の声質変換もクロスリンガル声質変換と同様の課題と捉えることができることから本研究の課題でも有効であることが期待される。

実験では，健常者の日本人女性話者 1 名から構音障害者の日本人男性話者 1 名への声質変換を行い，音声の自然性，話者性および明瞭性の観点における評価実験を行う。

2 提案手法および関連研究

Fig. 1 に提案手法の概略図を示す。提案手法は Transformer 型音響モデルを用いた健常者 TTS，CycleVAE を用いた健常者から構音障害者への声質変換，LPCNet を用いた音声合成の 3 つから構成される。

2.1 Transformer 型音響モデルを用いた TTS

TTS には，知覚品質において自然音声と同等の品質を達成している Transformer 型音響モデルを用いた [4]。本研究ではこの TTS モデルを日本人女性健常話者 1 名の音声で学習する。よってこの TTS モデルは明瞭性の高い健常者の音響特徴量が出力される。

2.2 CycleVAE を用いたノンパラレル VC

健常者から構音障害者への声質変換には CycleVAE (CycVAE) による声質変換を用いる。VAE に基づく声質変換は，入力される音響特徴量を言語的特徴のみをもつ潜在特徴量に変換するエンコーダと，話者コードを受け潜在特徴量を対応する話者性を持った音響特徴量に変換するデコーダから構成される。CycleVAE では変換された音響特徴量をさらにエンコーダへ入力し，入力音声の音響特徴量へ戻した際の再構成誤差も考慮して学習をすることで変換の安定化

* High-Intelligibility speech synthesis for dysarthric speakers with CycleVAE-based voice conversion by MATSUBARA, Keisuke[1,2], OKAMOTO, Takuma[2], TAKASHIMA, Ryoichi[1], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], SHIGA, Yoshinori[2] and KAWAI, Hisashi[2] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

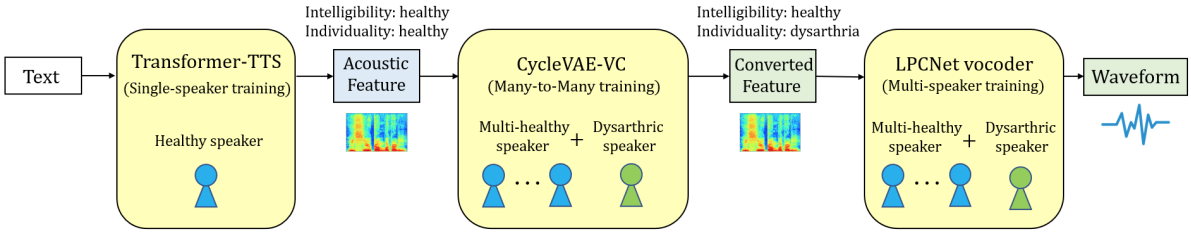

Fig. 1　The flow of the proposal method and training dataset on each component.

を図っており，VAE のみの場合よりも精度の高い声質変換が可能となっている。

CycleVAE による声質変換では，元話者と目的話者の言語が異なっているクロスリンガル声質変換でも精度の高い声質変換が可能であることが報告されている [7]。本研究ではこのクロスリンガル声質変換における有効性に着目する。構音障害者の発話は健常者の発話と比較して言語情報が欠落しているが，これにより健常者と構音障害者間の声質変換はクロスリンガル声質変換と同様に，話者間の言語が異なる声質変換の問題と考えることができる。しかし障害者から健常者への声質変換を行うと，障害者の言語的特徴を保持したまま健常者の話者性が付与されることになり，明瞭性の低い音声になってしまうと考えられる。そこで本研究では健常者から障害者への変換を行うことで，明瞭性は健常者音声のものを維持しつつ障害者の話者性を持った音声に変換が可能であると考える。本研究では目的の日本人構音障害者１名と複数の日本人健常話者の音声で学習を行う。

2.3　LPCNet ボコーダ

変換特徴量から音声波形の合成には LPCNet ボコーダ (LPCN) を用いる。LPCNet は WaveRNN に基づくニューラルボコーダで，線形予測分析 (Linear Prediction Coding: LPC) で推定した音声信号の残差信号をニューラルネットで推定するモデルである。また１時間程度の少量データでの学習が可能であることも確認されており [8]，発話動作の負担が大きいことから大量のデータを集めることが難しい構音障害者音声の合成も問題なく行えると考えられる。

本研究では LPCNet ボコーダの学習に際し，目的話者の構音障害者音声のみでの学習と複数の健常者音声も用いた学習の２条件で実験を行う。前者は構音障害者の話者性を保った音声のみが合成されることが期待されるが，健常者音声のような明瞭性の高い音声を合成しようとする際に品質が劣化する可能性がある。後者は合成音声の品質は改善されると考えられるが，目的話者の話者性が損なわれる可能性がある。

また [7] では，変換特徴量の劣化に対する頑健性を向上させるために，CycleVAE から出力される再構成された音響特徴量をニューラルボコーダの学習に用いることを提案している。本研究では，健常者音声の再構成音響特徴量を LPCNet の学習に用いることで合成音声の品質の向上を図る。

3　実験

3.1　実験条件

提案手法の有効性を確認するため，健常者から構音障害者への声質変換を行う。健常者音声には JSUT コーパス [9] より日本人女性話者 1 名の音声を用いた。また CycleVAE の学習には JVS コーパス [9] より健常者の日本人男性話者 4 名を追加話者として用いた。構音障害者音声は，アテトーゼ型脳性麻痺による構音障害を持つ日本人男性話者 1 名による，ATR 日本語データベース [10] のテキスト 503 文のうち 430 文を読み上げたものを用いた。すべての音声はサンプリング周波数 24 kHz として用いた。

テキスト音声合成では，フルコンテキストラベルを入力とする Transformer 型の音響モデルを用いた。入力特徴量は 38 次元の音素と HTS 形式のフルコンテキストラベルの A と F から 9 次元のアクセントラベルを抽出した 47 次元のベクトルとした。出力特徴量は LPCNet の公式実装 [6] を改良し，30 次元バークケプストラムおよびピッチ周期，ピッチ相関の 32 次元の特徴量を抽出して用いた。学習データは JSUT コーパスの内，手修正済みフルコンテキストラベルが提供されている 4,800 文を用いた[1]。モデルは ESPNet-TTS [11] での実装をベースに，フルコンテキストラベルを入力できるよう修正したものを用いた。

声質変換では，VCC2020 における CycleVAE 型声質変換モデルの実装 [7] を，使用する特徴量を修正して用いた。学習データには，元話者の音声を JSUT コーパスより 100 文と TTS による読み上げ音声 100 文，目的話者の構音障害者音声を 100 文と，学習を補助するための追加話者として JVS コーパスより健常男性話者 4 名の音声を各 100 文づつ用いた。

[1] https://github.com/sarulab-speech/jsut-label

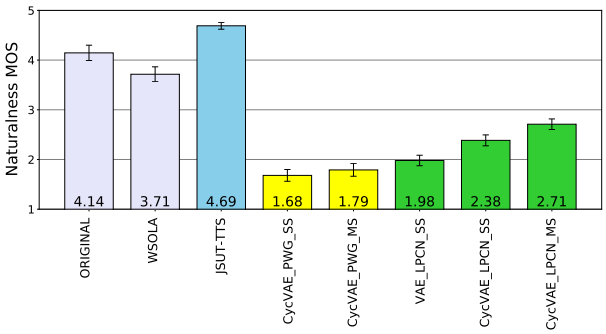

Fig. 2 Results of Naturalness MOS test with 10 listening subjects. Confidence level of the error bars was 95 %.

ニューラルボコーダには，LPCNet の公式実装をサンプリング周波数 24 kHz 用に改良して用いた [8]。学習には，目的話者の構音障害者音声のみでの学習 (Single Speaker: SS) と，複数の健常者音声および CycleVAE による再構成特徴量も用いた学習 (Multi Speaker: MS) の 2 つのデータ構成で実験を行った。前者は 370 文を学習データに用いた。後者はそれに加えて JVS コーパスより 100 名の健常者の計 3,000 文の音声と，CycleVAE によって得られた 5 名の健常者音声の音響特徴量と正解音声のペアを用いた。また構音障害者音声の再構成特徴量を学習に使うことも可能だが，こちらは事前実験によって音声品質の劣化が確認されたため，本研究では用いていない。

比較条件には，VCC2020 での CycleVAE の公式実装と通常の VAE による声質変換モデルを用いた。前者は特徴量に WORLD [12] を用いて抽出した 50 次元メルケプストラム，対数基本周波数，無声/有声ベクトルを使用し，ボコーダに Parallel WaveGAN (PWG) [13] を使用した。

評価実験には，合成音声の自然性および話者性の主観評価実験と音声認識モデルによる音素誤り率 (Phoneme Error Rate: PER) の計測を行った。また実験条件として，構音障害者音声の原音に加えその音声を WSOLA [14] によって発話長が健常者音声に一致させたもの用いた。これは構音障害者の発話が一般に間延びしており，それによる評価の変動を抑えるための処理である。

3.2 自然性の実験結果

音声の知覚品質を評価するために，聴取実験による平均オピニオン評点テストを行った。実験参加者は 10 人の成人日本語母語話者で，テストセット 20 文に対して 8 条件の計 160 文をヘッドホン聴取により評価した。Fig. 2 に聴取実験の結果を示す。原音や WSOLA 音声と比較して，提案手法は十分な品質には届かなかった。合成音声の条件を比較すると，LPCNet を用いた条件はすべて Parallel WaveGAN を用いた条件より優位な結果となった。これは LPCNet は自己回帰モデルで音声サンプルの生成に過去のサンプルを参照できるため，非自己回帰モデルの Parallel WaveGAN よりも変換特徴量の劣化に対して頑健であったと考えられる。LPCNet を用いた条件では，CycleVAE および MS 学習を用いた条件が最も高いスコアとなった。これによりサイクル損失の使用と，ニューラルボコーダの学習に健常者音声および再構成特徴量を用いることの優位性が確認できた。

Table 1 Results of individuality ABX test with 10 listening subjects.

Method	WSOLA	JSUT-TTS
CycVAE_PWG_MS	0.71	0.29
VAE_LPCN_SS	0.39	0.61
CycVAE_LPCN_MS	0.53	0.47

3.3 話者性の実験結果

合成音声に目的話者の構音障害者の話者性が含まれているかどうかを評価するために，話者性における ABX テストを行った。実験参加者は自然性の実験と同じく 10 名の成人日本語母語話者で，2 つの参照音声 (WSOLA，目的話者。 JSUT-TTS，元話者) を聞いた後に合成音声を聞き，参照音声のどちらに話者が近かったかを 2 段階で評価した。テストセット 20 文に対して 3 条件の計 60 文を評価した。

Table 1 に実験結果を示す。提案手法は目的話者，元話者間で有意差がない (p 値=3.6×10^{-1}) という結果となった。これは変換後の音声は明瞭性が健常者のものに近いため，明瞭性に起因する話者性が失われてしまうことが原因と考えられる。また CycVAE_PWG_MS は目的話者側に (p 値=6.5×10^{-9})，VAE_LPCN_SS は元話者側 (p 値=1.5×10^{-2}) に有意差ありという結果となった。前者は音声の自然性の劣化が大きいために明瞭性が構音障害者に近づいたことが原因と考えられる。後者は声質変換時にサイクル損失を考慮してないため，話者性が十分に変換されなかったと考えられる。これらの実験結果は，本研究の目的である明瞭性が向上した構音障害者の音声という明確な正解データのないものの品質評価が，通常の手法では難しい課題であることを示している。これらの正確な評価方法の策定は今後の課題とする。

3.4 明瞭性の実験結果

音声の明瞭性を評価するために，健常者音声で学習させた音声認識モデルによる音素誤り率 (PER) を計測した。Fig. 3 に実験結果を示す．構音障害者音声

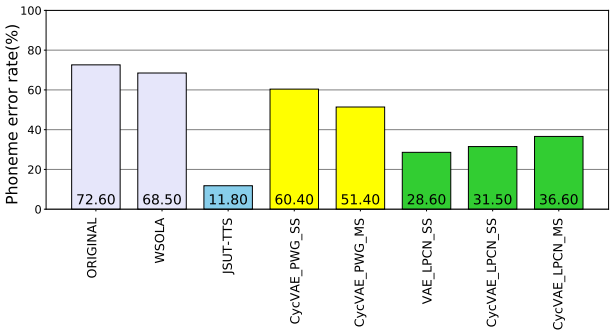

Fig. 3　PERs [%] for a variety of methods

の PER は 72.8 %と非常に高く，その他の実験条件はすべて PER が改善されるという結果となった。健常者 TTS 音声 (11.8 %) に次いで良い結果となったのは VAE_LPCN_SS で 28.6 %となった。これは 3.3 の結果より話者性の変換が十分に行われていないことから，その分明瞭性の損失も少なかったからだと考えられる。提案手法はそれに次いで良い結果となったが，MS が 36.6 %と SS の 31.5 % に対して悪い結果となった。この原因としては，MS の音声にはパルス状のノイズが含まれている場合が確認されており，それにより音声認識精度が低下したこと等が考えられる。詳しい原因の調査およびその改善は今後の課題とする。

4　おわりに

本研究では構音障害者のコミュニケーション支援のための，本人の話者性を持ちながら明瞭性の高い音声を合成可能な TTS システムの構築を目的とし，Transformer-TTS, CycleVAE, LPCNet を合わせたニューラル音声合成システムを提案した。実験結果より，健常者音声の明瞭性をある程度維持しつつ構音障害者の話者性を持った音声が合成可能であると分かったが，その精度にはまだ課題が残った。今後は音声品質の改善，他の声質変換モデルとの比較や他の構音障害者音声での実験等を検討する。

参考文献

[1] L.-W. Chen *et al.*, “Generative adversarial networks for unpaired voice transformation on impaired speech,” in *Proc. Interspeech*, Sept. 2019, pp. 719–723.

[2] 今井ら, “多数話者健常音声を用いた CycleGAN に基づく構音障害者音声の明瞭性改善”, 音 講論, Sept. 2020, pp. 931–934.

[3] J. Yamagishi *et al.*, “Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction” *Acoust. Sci. Tech.*, vol. 33, no. 1, pp. 1–5, Jan. 2012.

[4] N. Li *et al.*, “Neural speech synthesis with transformer network,” in *Proc. AAAI*, 2019, pp. 6706–6713.

[5] P. L. Tobing *et al.*, “Non-parallel voice conversion with cyclic variational autoencoder,” in *Proc. Interspeech*, Sept. 2019, pp. 674–678.

[6] J. Valin, J Skoglund, “LPCNet: Improving Neural Speech Synthesis Through Linear Prediction,” in *Proc. ICASSP*, May 2019, pp. 5891–5895.

[7] P. L. Tobing *et al.*, “Baseline system of Voice Conversion Challenge 2020 with cyclic variational autoencoder and Parallel WaveGAN” in *Proc. Joint Workshop for the Blizzard Challenge and voice Conversion Challenge 2020*, Oct. 2020, pp. 155–159.

[8] K. Matsubara *et al.*, “Investigation of training data size for real-time neural vocoders on CPUs,” in *Acoust. Sci. Tech.*, vol. 42, pp. 65–68, 2020.

[9] S. Takamichi *et al.*, “JSUT and JVS: free Japanese voice corpora for accelerating speech synthesis research,” in *Acoust. Sci. Tech*, vol. 41, pp. 761–768, 2020.

[10] A. Kurematsu, *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Commnun.*, col. 9, no. 4, pp. 357–363, Aug. 1990.

[11] T. Hayashi, *et al.*, “ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” in *Proc. ICASSP*, May 2020, pp. 7654–7658.

[12] M. Morise *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE trans, Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[13] R. Yamamoto *et al.*, “Parallel WaveGAN:A fast Waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in *Proc. ICASSP*, May 2020, pp. 6199–6203.

[14] W. Verhelst, *et al.*, “An overlap-add technique based on waveform similarity (WSOLA) for high quality timescale modification of speech,” in *Proc. ICASSP*, Apr. 1993, pp. 554–557.