# 構音障害者のための高明瞭度音声合成における HiFi-GAN を用いた品質改善 *

△松原圭亮 [1,2]，◎高島遼一 [1]，岡本拓磨 [2]， 滝口哲也 [1]，戸田智基 [3,2]，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症状である。本研究で対象とするアテトーゼ型脳性麻痺は，発声器官に加えて手足の運動障害も引き起こすため，手話のような代替手段が取れず，音声コミュニケーションに頼らざるを得ない患者も多い。構音障害者の音声コミュニケーションを支援する技術として，テキスト音声合成 (Text-to-Speech; TTS) システムが期待されている。しかし，一般的な TTS システムはユーザとは別人の音声が生成されるため，より本人らしいコミュニケーションを実現するため，本人らしく（話者性の維持）かつ聞き取りやすい（高明瞭度）音声を合成する技術が求められている。

構音障害者本人の話者性を維持しつつ高明瞭度な音声を生成する研究として，大きく分けて以下三つのアプローチが存在する。一つ目は，構音障害を引き起こす前のユーザ発話を収録し，TTS モデルを学習するアプローチである [1]。このアプローチは筋萎縮性側索硬化症のような進行性の病気や喉頭摘出など，後天的な構音障害に対して効果的であるが，アテトーゼ型脳性麻痺のような先天性の構音障害を持つ患者に対して適用することはできない。

二つ目は，声質変換 (Voice conversion; VC) を用いて，障害者のユーザ発話から健常者らしい音声へと変換するアプローチである [2, 3]。ただし，単純に障害者から健常者への声質変換を行うと話者性が失われてしまうため，文献 [2] では子音のみ，文献 [3] では韻律のみを変換することで話者性を維持している。

三つ目は，二つ目のアプローチとは逆に，健常者の音声から明瞭度を維持しつつ障害者のユーザらしい音声へ声質変換するアプローチである [4, 5, 6]。一般に声質変換は入力音声の音韻情報を維持し，話者情報などその他の情報を変換することを目的として開発された技術である。一方，二つ目のアプローチは一部の音韻情報が欠落した障害者音声から，欠落した音韻情報を復元する，つまり音韻情報の変換を目的としているため，声質変換の枠組みでこれを行うことは困難であると考えられる。逆に，音韻情報が揃っている健常者音声から障害者への声質変換であれば，声質変換技術の性質上，音韻情報が維持されるため，明瞭度を損なわずに話者性のみを変換可能であると考えられる。変換元音声として，文献 [4, 5] では TTS によって生成された健常者音声を使用し，文献 [6] では，障害者音声から声質変換して得られた健常者音声を使用している。どちらも生成される音声は高明瞭度な障害者音声であるが，前者は TTS，後者は VC をアプリケーションとして想定した手法となっている。

本研究では前述した三つ目のアプローチ，特に我々の先行研究 [5] に焦点を当て，生成音声のさらなる品質改善を目的とする。先行研究では，声質変換によって得られた障害者ユーザの音響特徴量に対して，LPCNet ボコーダ [7] を用いて音声波形へと変換している。ここで LPCNet について，明瞭度の低い障害者ユーザの収録音声で学習した特定話者モデルを用いた場合，変換によって得られる音響特徴量は明瞭度が向上したものであるため，学習データとのミスマッチにより生成波形の品質劣化が生じる。そのため先行研究では，複数の健常者音声と構音障害者音声を混ぜて学習した不特定話者 LPCNet ボコーダを使用することでミスマッチを低減していたが，それでも品質には課題があった。

近年，不特定話者でも高品質に合成可能なニューラルボコーダが複数提案されており [8]，中でも HiFi-GAN [9] は CPU でもリアルタイム動作可能なボコーダとして注目されている。また我々はさらに幅広い音声を表現可能な HiFi-GAN の改良モデル [10, 11] を検討してきた。本研究では，HiFi-GAN ベースのニューラルボコーダを LPCNet に置き換えることで，生成音声の品質向上を検討する。4 名のアテトーゼ型脳性麻痺者を対象に高明瞭度音声合成の実験を行い，生成音声の自然性，話者性および明瞭性の観点で評価を行う。

## 2 TTS，VC，ボコーダの統合による構音障害者の高明瞭度音声合成システム

先行研究 [5] および本研究で検討する，構音障害者の高明瞭度音声合成システムを Fig. 1 に示す。本シ

* Quality improvement using HiFi-GAN on high-intelligibility speech synthesis for dysarthric people. by MATSUBARA, Keisuke[1,2], TAKASHIMA, Ryoichi[1], OKAMOTO, Takuma[2], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2] and KAWAI, Hisashi[2] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

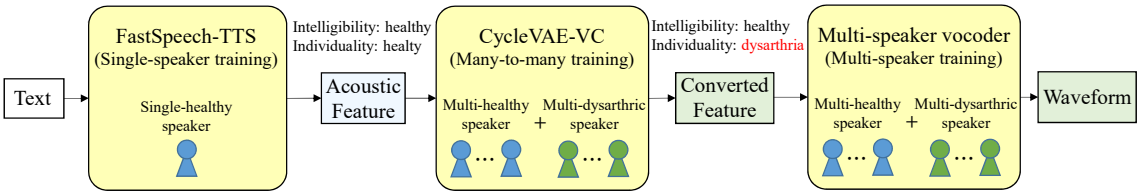


Fig. 1　Overview of our high-intelligibility speech synthesis system based on TTS, VC, and vocoder.

ステムは大きく TTS，VC，ボコーダの 3 モジュールで構成される。

TTS では，入力テキストをもとに，健常者の音響特徴量を生成する。TTS は 1 話者で学習された特定話者モデルを使用する。先行研究では TTS モデルとして Transformer-TTS [12] が使用されている。本研究では，Transformer-TTS の合成速度の遅さと不安定な継続長推定を改善したモデルである FastSpeech [13] を使用する。

TTS が出力した健常者の音響特徴量は VC によって構音障害者の音響特徴量へ変換される。VC には CycleVAE [14] を用いる。CycleVAE はエンコーダ・デコーダにより構成される多対多の VC モデルである。エンコーダは，入力音響特徴量から話者情報を取り除いた潜在特徴量へと変換する。デコーダは，外部入力される話者コードを受けて，潜在特徴量から対応する話者性を持った音響特徴量に変換する。

ボコーダは，VC が出力した音響特徴量から音声波形へと変換する。本研究では，以下の三種類のボコーダについて比較評価を行う。

**LPCNet**:　先行研究では LPCNet を使用している。これは，LPCNet が比較的少量の学習データでも高品質なモデルを学習可能なことが報告されている [15] ため，大量の学習データを収録困難な構音障害者音声のボコーダとして適していると考えられたためである。

**HiFi-GAN**:　HiFi-GAN は敵対的生成ネットワークをベースとするニューラルボコーダである。生成器では転置畳み込みを用いて入力音響特徴量をアップサンプリングしながら音声波形に変換する。識別器は，合成音声を複数のサンプリング周波数において真偽を識別する Multi-scale discriminator と，合成音声を様々な間隔でリサンプリングすることで複数の受容野から真偽を識別する Multi-period discriminator で構成される。また，複数話者音声で学習された HiFi-GAN は未知の話者であっても高品質に波形生成できることが報告されている [9]。

**Harmonic-Net** [11]:　Harmonic-Net は HiFi-GAN をベースに，さらに生成可能な音声表現の幅を広げることを目的として改良したモデルである。具体的には，生成したい音声の $F_0$ に対する高調波を励起信号として，畳み込みにより段階的にダウンサンプリングし，各ダウンサンプリング信号を対応する HiFi-GAN 生成器のアップサンプリング（転置畳み込み）層へ入力することで，合成音声の基本周波数を制御可能としている。手法の詳細については，本論文集に掲載の文献 [11] を参照されたい。本研究において基本周波数の制御は目的ではないが，1 章で述べた通り，推論時は障害者の高明瞭度な音響特徴量が入力されるため，学習時の音響特徴量とのミスマッチが生じる。そのため，HiFi-GAN よりも表現の幅が広い本モデルを使用することで，ミスマッチの影響がさらに軽減されることを期待している。

以上で述べた 3 種類のボコーダはいずれも複数の健常者音声および複数の障害者音声を用いて学習する。これは前述の音響特徴量のミスマッチの影響を軽減することが目的である。また，TTS や VC によって生成された音響特徴量は実際の音声から抽出される音響特徴量と比べて歪みが含まれる。そこで，歪みによるミスマッチにも対応するため，各ボコーダは個別に学習した後，CycleVAE が出力した音響特徴量を用いてファインチューニングを行っている。

## 3　評価実験

健常者音声として，JSUT コーパスおよび JVS コーパス [16] に収録される日本人話者音声を使用し，障害者音声として，4 名のアテトーゼ型脳性麻痺患者が ATR 日本語データベース [17] のテキスト 503 文を読み上げた音声を収録して使用した。サンプリング周波数はいずれも 24 kHz である。LPCNet を用いる場合は先行研究 [5] と同様に 30 次元バークケプストラム，ピッチ周期，およびピッチ相関を音響特徴量として使用し，HiFi-GAN および Harmonics-Net を用いる場合は，WORLD [18] によって抽出された 50 次元のメルケプストラム係数，3 次元非周期性指標と対数連続 $F_0$ を音響特徴量として使用した。

FastSpeech-TTS はフルコンテキストラベルを入力とし，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 文を用いて

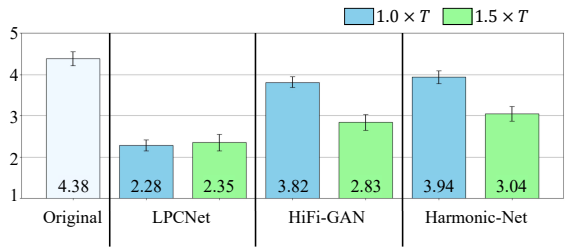


Fig. 2　MOS scores on naturalness evaluation

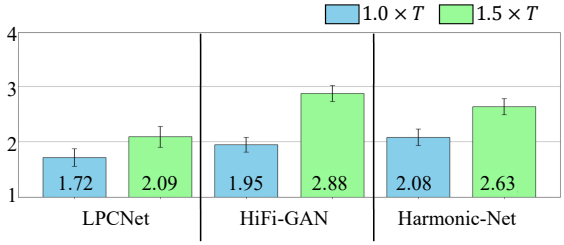


Fig. 3　MOS scores on similarity evaluation

学習した[1]。評価時に用いるテキストは ATR データベースより 20 文を使用した。音声合成を実施する際，FastSpeech が推定した音素継続長をそのまま用いて合成した場合（$1.0 \times T$）と，継続長を 1.5 倍に引き延ばした上で合成した場合（$1.5 \times T$）の 2 パターンで実験を行った。継続長を引き延ばした理由は，本来ゆっくり話す構音障害者の音声を健常者と同じ話速で合成すると，聴取実験者にとって違和感が生じ，話者性の評価が低くなったためである。

CycleVAE-VC の学習には，健常者音声として TTS からの出力音響特徴量 100 文，JSUT コーパスより 100 文，JVS コーパスより 7 名 ×100 文の計 900 文を使用し，障害者音声として 4 名 ×100 文の計 400 文を使用した。LPCNet 用の音響特徴量の変換は先行研究 [5] と同様に行い，HiFi-GAN および Harmonic-Net 用の音響特徴量の変換は，メルケプストラムのみを CycleVAE で変換し，基本周波数は話者ごとの平均と分散を用いて線形変換した。非周期性指標は変換を行わない。

全てのボコーダは，まず JVS コーパスより 96 名の計 12,477 文の音声を用いて事前学習を行った。次に，CycleVAE が出力した再構成特徴量を用いてファインチューニングを行った。具体的には，元話者から同じ話者への変換を行って得られた特徴量を 1,300 文，元話者から異なる話者に一度変換した上で元話者に逆変換して得られた特徴量を 11,700 文使用した。

評価音声は，各手法につき，構音障害者 4 名 ×5 文の計 20 文である。自然性および話者性の評価は聴取実験により行った。被験者は 10 名の日本人話者である。明瞭性については，音声認識誤り率による客観指標を用いて評価した。

### 3.1　自然性ついての評価実験

Fig. 2 に聴取実験による平均オピニオン評点 (MOS) を示す。HiFi-GAN および Harmonic-Net は LPCNet に比べた高いスコアとなった。これは，LPCNet と比べて HiFi-GAN ベースのボコーダは未知の音響特徴に対して頑健であったためと考えられる。同様に，Harmonic-Net が HiFi-GAN より高いスコアを示していることから，より幅広い音声を表現可能な Harmonic-Net が，未知の音響特徴に対してさらに頑健であると考えられる。

### 3.2　話者性についての評価実験

話者性評価の実験では，被験者は最初に障害者の元音声を参照音声として聞き，その後評価音声を聞いて，(4: 参照音声の話者に非常に近い，3: どちらかといえば近い，2: どちらかといえば違う，1: 全く違う) の 4 段階で評価した。

Fig. 3 に結果を示す。全てのボコーダにおいて，音素継続長を 1.5 倍に伸ばした方が高い評価値が得られた。Fig. 2 の HiFi-GAN および Harmonic-Net の結果でも示される通り，話速を遅くすることで，自然な音声からはかけ離れ，自然性評価のスコアは低下する。これは，自然性の低い音声ほど話者性が高い音声と評価されている，つまり被験者にとって自然性と話者性を切り離して評価することが困難であることを示唆している。1.5 倍の条件で比較したとき，Harmonic-Net よりも HiFi-GAN の方が高いスコアを示しているが，これも Harmonic-Net より HiFi-GAN の方が自然性が低いことによる裏返しとも解釈できる。

### 3.3　明瞭性についての評価実験

音声認識実験を実施し，認識誤り率により明瞭性を客観評価した。音声認識モデルには，ESPNet [19] にて公開されている事前学習済みの Transformer モデルを使用した。Fig. 4 に各条件における文字誤り率 (Character error rate; CER) を示す。TTS が出力した健常者の合成音声 (7.7%) と比べて変換音声はいずれも CER が悪化しているものの，障害者のオリジナル音声 (90.4%) より大幅に CER が低減されていることが確認された。また，Harmonic-Net が最も低い CER を示したことから，最も明瞭性の高い音声が生成されていることが示唆された。

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

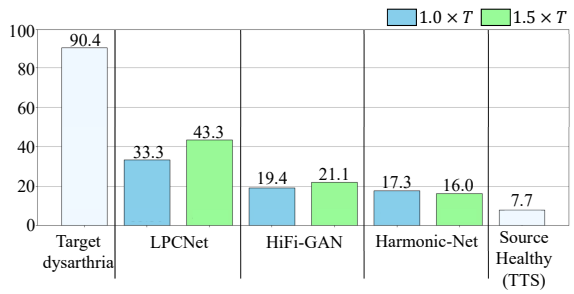


Fig. 4　Character error rates [%] for each condition.

## 4　おわりに

本稿では，TTS，VC，ボコーダを統合した構音障害者の高明瞭度音声合成システムにおいて，HiFi-GAN およびその改良である Harmonic-Net をボコーダとして使用することで，先行研究で使用されていた LPCNet と比較して自然性，話者性，明瞭性のいずれも改善することを確認した。しかし話者性の評価において，自然性と話者性を切り離して評価することが困難であることが示唆されたことから，純粋に話者性のみを評価する方法について，今後検討する。

## 参考文献


[1] 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.

[2] R. Aihara *et al.*, “Individuality-preserving Voice Conversion for Articulation Disorders Using Phoneme-categorized Exemplars,” *ACM Trans. on Accessible Computing*, vol. 6, issue 4, no. 13, pp. 1–17, Jun. 2015.

[3] D. Wang *et al.*, “Learning Explicit Prosody Models and Deep Speaker Embeddings for Atypical Voice Conversion,” in *Proc. Interspeech*, pp. 4813–4817, Aug. 2021.

[4] R. Nanzaka and T. Takiguchi, “Hybrid Text-to-Speech for Articulation Disorders with a Small Amount of Non-Parallel Data,” in *Proc. APSIPA*, pp. 1761–1765, Nov. 2018.

[5] K. Matsubara *et al.*, “High-intelligibility Speech Synthesis for Dysarthric Speakers with LPCNet-based TTS and CycleVAE-based VC,” in *Proc. ICASSP*, pp. 7058–7062, Jun. 2021.

[6] W.-C. Huang *et al.*, “A Preliminary Study of a Two-stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion,” in *Proc. Interspeech*, pp. 1329–1333, Aug. 2021.

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

[8] 岡本, “ニューラルネットワークに基づく音声波形生成モデル,” 音響誌, vol. 78, no. 6, pp. 328–337, Jun. 2022.

[9] J. Kong *et al.*, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, pp. 17022–17033, Dec. 2020.

[10] 松原ら, “Period-HiFi-GAN: 基本周波数を制御可能な高速ニューラルボコーダ,” 音講論, pp. 901–904, Mar. 2022.

[11] 松原ら, “Harmonic-Net+: 高調波入力と Layerwise-Quasi-Periodic 畳み込みを用いた基本周波数制御可能な高速ニューラルボコーダ,” 音講論, Sept. 2022.

[12] N. Li *et al.*, “Neural Speech Synthesis with Transformer Network,” in *Proc. AAAI*, pp. 6706–6713, Jan. 2019.

[13] Y. Ren *et al.*, “FastSpeech: Fast, Robust and Controllable Text to Speech,” in *Proc. NeuIPS*, pp. 3165–3174, Dec. 2019.

[14] P. L. Tobing *et al.* “Non-parallel Voice Conversion with Cyclic Variational Autoencoder,” in *Proc. Interspeech*, pp. 674–678, Sep. 2019.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] T. Hayashi *et al.*, “ESPnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-end Text-to-speech Toolkit,” in *Proc. ICASSP*, pp. 7654–7658, May 2020.