End-to-End 系列変換型声質変換の高速化およびノンネイティブ話者変換の検討 *

◎山下陽生 [1,2]，岡本拓磨 [2]，高島遼一 [1]，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

近年では話者の声を別の話者に変換する声質変換 (Voice conversion: VC) の技術が発展し，様々な分野で応用されている [1, 2]。CycleGAN ベースのモデル [3] のようなフレームバイフレームの方式と比べて，Sequence-to-sequence(S2S) 型の系列変換型モデルは話速や韻律も制御可能であり有用なモデルである [4]。

従来方式は，ソース話者の音響特徴量をターゲット話者の音響特徴量へと変換する音響モデル Conformer-based FastSpeech2(CFS2)[5] と，変換された音響特徴量を音声波形へと変換する波形生成モデル Parallel WaveGAN(PWG)[6] の 2 つを別々に学習する必要があるパイプライン方式であるのに対して，岡本らによって提案された End-to-end(E2E) モデルでは，E2E テキスト音声合成モデルである JETS[7] を VC 化にしたものであり，従来のパイプラインモデルモデルよりも高品質な音声をより高速で推論が可能であった [8]。

この E2E-S2S-VC モデルでは，1CPU で高品質な音声が合成可能なニューラル波形ボコーダ [9] である HiFi-GAN[10] を用いていたが，HiFi-GAN の高速モデル [11, 12, 13] がいくつか提案されている。そこで本論文では，それらを HiFi-GAN の代わりに使用することで VC 分野においても合成品質を損なわない高速化が有効であるかの検討を行う。また，この E2E-VC モデルではネイティブ (L1) 話者からネイティブ話者への変換を検討しているが，ノンネイティブ (L2) 話者からネイティブ話者の発音へ変換が可能になれば国際的な場においてのコミュニケーションがより容易になっていくと考えられるため，重要な課題と言える。そこで本論文では，さらに，E2E-S2S-VC を用いた L2 話者から L1 話者への VC の検討も行う。

2 VC モデル

2.1 CFS2+PWG

CFS2+PWG は，Fig. 1(a) に示すように CFS2 の入力をテキストからソース話者音声に変えることでソース話者の音響特徴量からターゲット話者の音響特徴量へ変換し，得られた音響特徴量を CFS2 とは別途に学習した PWG に入力することで変換された音声を生成するモデルである。ソース話者の音声からメルスペクトログラム以外に $\log f_0$，エネルギーを分析し Variance adaptor に入力する。トレーニング時のアライメントは教師とした Voice Transformer network の予測に従い，推論時は Variance addaptor が行う [4]。

2.2 JETS-VC

JETS-VC は E2E-S2S 型テキスト音声合成モデルである JETS の入力をソース話者音声へ変更したモデルである。Variance adaptor の代わりに用いられている Modified variance adaptor では，ソース音声から分析される $\log f_0$ やエネルギーを利用せず，ソース話者のメルスペクトログラムのみからターゲット音声の $\log f_0$，エネルギーを予測する。アライメントはトレーニング時のみ Aligment module がモノトニックアライメントによってターゲットのメルスペクトログラムから予測を行う。CFS2+PWG との最も大きい違いとして，波形生成モデルまで一貫学習することがあげられる。これによって 1 つのモデルで音声の変換が可能になり，変換品質の向上を実現している [8]。

本論文では，波形生成モデルにおいては HiFi-GAN のほかに，その高速モデルである Multi-stream HiFi-GAN[11]，iSTFTNet[12]，FC-HiFi-GAN[13] をそれぞれ用いることで，JETS-VC においても合成品質を損なわず高速化を実現できるかを検討する。

---

* Acceleration of end-to-end sequence-to-sequence voice conversion and investigation of non-native speaker voice vonversion. by YAMASHITA, Haruki[1,2], OKAMOTO, Takuma[1], TAKASHIMA, Ryoichi[1], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], KAWAI, Hisash[1] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

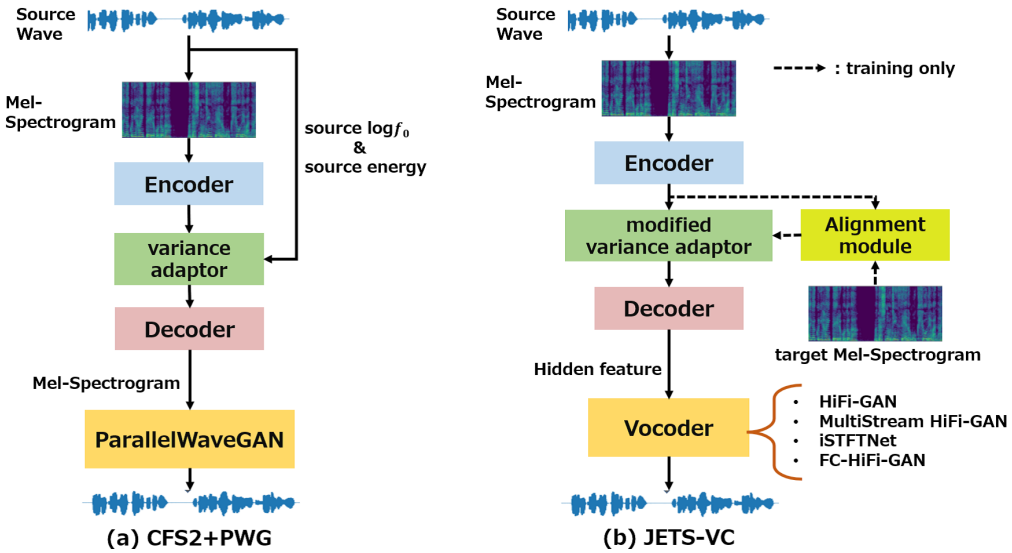

Fig. 1　(a) conventional pipeline VC model and (b) end-to-end sequence-to-sequence VC model.

3　ニューラル波形生成モデル

3.1　HiFi-GAN

HiFi-GAN の Generator は入力メルスペクトログラムを転置畳み込み+ResBlock によるアップサンプリングを複数回実行することによって音声波形を生成する。HiFi-GAN は 2 つの優れた Discriminator を用いており，それよって HiFi-GAN は Generator を軽くすることができ，1CPU などの低計算資源でも高品質な音声波形を生成できる。本稿では，転置畳み込み+ResBlock によるアップサンプリングを (8, 8, 2, 2)，カーネルサイズを (16, 16, 4, 4)，隠れチャンネルを 512 とする HiFi-GAN V1 を用いた。また，以下に示す HiFi-GAN の高速モデルにおいても Discriminator は HiFi-GAN と同じものを使用した。

3.2　MultiStream HiFi-GAN[11]

HiFi-GAN の最後の 4 倍アップサンプリング層を，4 帯域のサブバンド波形に分けるフィルタとそれらを再合成するサブバンド合成フィルタに置き換える処理を MultiBand 化と呼び，それを学習可能な畳み込み層にすることで制約を緩め学習が進むようにした処理を Multi-stream 化という。Multi-stream HiFi-GAN は，HiFi-GAN に Multi-stream 化を適用したモデルであり，HiFi-GAN の合成品質を保ったままの高速化に成功した [11]。本検討でも，合成フィルタのカーネルサイズは 63 とした。

3.3　iSTFTNet[12]

iSTFTNet は HiFi-GAN の高速モデルであり，HiFi-GAN の最後の 4 倍アップサンプリングを逆短時間フーリエ変換 (iSTFT) によるアップサンプリングに変更することで合成品質を保ったまま合成速度の向上に成功したモデルである [12]。64 倍アップサンプリング後の 1 次元畳み込み層から短時間フーリエ変換 (STFT) の出力である位相成分と振幅成分を出力することで iSTFT を行えるようにした。ここで，iSTFT は fft サイズは 16，window サイズは 16，hop サイズは 4 とした。

3.4　FC-HiFi-GAN[13]

FC-HiFi-GAN は，iSTFTNet の iSTFT 部をフーリエ基底の固定フィルタとみなし，それを学習可能な線形結合層 (Fully Connected: FC) に置き換えたモデルであり，著者らによって提案されている。これは Multi-stream HiFi-GAN において，信号処理フィルタを学習可能にすることで学習が上手く進むようになり精度が向上したという知見を iSTFTNet に適応したものである [13]。

4　実験

4.1　実験条件

L2 話者から L1 話者への話者変換において，CFS2+PWG と JETS-VC における合成品質を比較し，E2E-S2S-VC が適切に機能するかを確認する。JETS-VC は波形生成部を HiFi-GAN，Multi-stream HiFi-GAN，iSTFTNet，FC-HiFi-

GAN の 4 モデルを比較し，HiFi-GAN の高速化が Jets-VC にも適応するかを調べる。

データセットには L1 話者として CMU-ARCTIC[15] から男性話者 (bdl)，女性話者 (slt) を 1 名ずつ (24 kHz，各 1131 文) 用いた。L2 話者には，L2-ARCTIC[16] の 24 話者のうち，音声認識モデルを用いて最も CER が良かった女性話者 (NJS)1 名 (24 kHz, 1131 文) を選び，そのうちトレーニングには 1091 文を用いた。音声認識モデルは Confomer 型モデル [5] を LibriSpeech[14] で学習したものを用いた。VC モデルの学習には Pytorch ベースのオープンソースである ESPnet2-TTS[17] を利用した。推論速度と合成品質の客観評価には学習に用いていないデータのうち 20 文を使用し，主観評価にはそのうち 10 文を用いた。音響特徴量は 8 kHz まで帯域制限したメルスペクトログラムとした。各モデルは 1000epoch 学習し，ESPnet2-TTS を用いて実装や推論を行った。

合成品質の主観評価には平均オピニオン評点 (MOS) を用いた。MOS の測定には各モデルから学習に用いていない 10 文と，L1 話者音声各 10 文を合わせた合計 220 文を 7 名の日本人話者にヘッドフォン聴取を行い求めた。

合成品質の客観評価には，メルスペクトログラム歪み (MCD)，f_0 の対数二乗平均誤差 ($\log f_0$ RMSE)，conformer ベースの音声認識モデルによる文字誤り率 (CER) を用いた。MCD，$\log f_0$RMSE の計算には ESPNet2-TTS[17] を利用した。推論速度の評価は Real Time Factor(RTF) を使用し，Intel Xeon Gold6152 CPU 2.1GHz を 1 コアでの推論時の速度を測定した。

4.2 実験結果

Fig. 2 から，すべての条件で JETS-VC(HiFi-GAN) が最も評価値が高くなっており，合成品質が良いことが分かる。また，L1→L1 の変換，L2→L1 の変換の両方において，CFS2+PWG が MultiStream HiFi-GAN と同程度の評価値であり、さらに iSTFTNet や FC-HiFi-GAN よりも高い結果となった。しかし，Table 1 の CER の結果から，L2→L1 への変換において，E2E-S2S-VC モデルを用いることによってより文字誤り率が大幅に改善されることが分かる。これらのことから，CFS2+PWG は高速 HiFi-GAN を導入した JETS-VC に比べてノイズが乗っていないなど音質こそ悪くないものの，実際の人間の音声とは全く違った音声が出力されており，E2E-S2S-VC モデルを用いることで人間の発音により近い音声を出力できることが分かる。また，Table 1 から，JETS-VC においても HiFi-GAN から他の高速モデルに変えることで推論速度が向上していることが分かる。しかし，音質は下がっているため，高品質化が今後の課題となる。

5 おわりに

L2 話者と L1 話者での声質変換において，E2E-S2S-VC により合成品質の向上に大きく貢献していることが分かった。また，波形生成部に高速化モデルを適用することにより高速化が可能となるが，音質とのトレードオフの関係があり，今後の課題である。

参考文献

[1] S. H. Mohammadi and A.Kain, “An overview of voice conversion systems,” *Speech Commun.*, vol.88, pp.65–82, Apr. 2017.

[2] B. Sisman *et al.*, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE/ACM Trans.Audio, Speech,Lang.Process.*, vol.29, pp.132–157, 2021.

[3] T. Kaneko *et al.*, “CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks”, in *Proc. EUSIPCO*, 2018, pp.2114–2118.

[4] T. Hayashi *et al.*, “Non-autoregressive sequence-to-sequence voice conversion,” in *Proc. ICASSP*, June 2021, pp.7068–7072.

[5] P. Guo *et al.*, “Recent developments on ESPnet toolkit boosted by Conformer,” in *Proc. ICASSP*, June 2021, pp. 5874–5878.

[6] 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.

[7] D. Lim *et al.*, “JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech,” in *Proc. Interspeech*, Sept. 2022, pp.21–25.

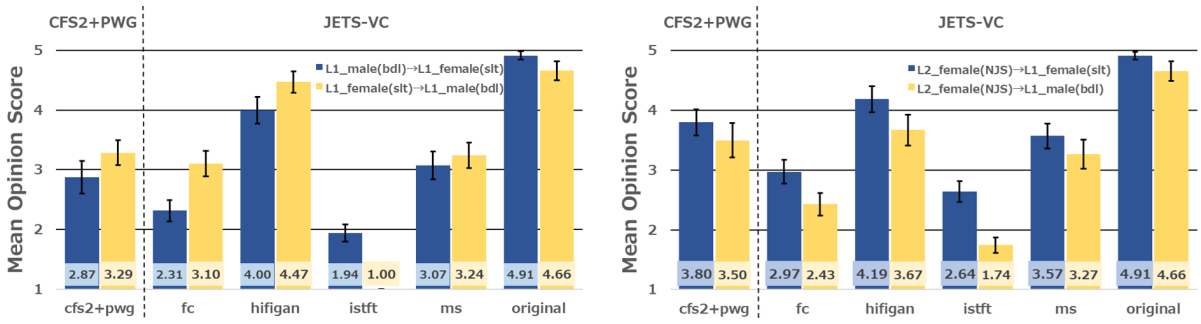

Fig. 2　Mean Opinion Score

Table 1　Objective Evaluation

	L1 female (slt) → L1 male (bdl)				L1 male (bdl) →L1 female (slt)		
	RTF	MCD [dB]	logf_o RMSE	CER [%]	MCD [dB]	logf_o RMSE	CER [%]
original	N/A	N/A	N/A	0.5	N/A	N/A	1.2
CFS2+PWG	3.44	5.76	0.20	3.5	5.23	0.20	3.0
JETS-VC(HiFi-GAN)	0.78	5.80	**0.19**	3.9	5.26	0.18	2.3
JETS-VC(MS HiFi-GAN)	0.50	**5.75**	0.21	4.2	**5.04**	**0.17**	**0.7**
JETS-VC(iSTFTNet)	0.50	9.16	0.23	4.6	5.47	0.18	1.6
JETS-VC(FC-HiFi-GAN)	**0.48**	5.87	0.20	**2.8**	5.56	0.18	1.6

	L2 female (NJS) → L1 male (bdl)				L2 female (NJS) → L1 female (slt)		
	RTF	MCD [dB]	logf_o RMSE	CER [%]	MCD [dB]	logf_o RMSE	CER [%]
original	N/A	N/A	N/A	0.5	N/A	N/A	1.2
CFS2+PWG	3.44	7.47	0.22	44.2	6.54	0.20	46.8
JETS-VC(HiFi-GAN)	0.78	6.77	**0.21**	**20.9**	6.04	0.20	**16.9**
JETS-VC(MS HiFi-GAN)	0.50	**6.56**	0.22	21.2	**5.90**	**0.17**	21.8
JETS-VC(iSTFTNet)	0.50	6.93	0.23	25.0	6.61	0.23	25.5
JETS-VC(FC-HiFi-GAN)	**0.48**	6.92	0.26	24.4	6.39	0.18	21.8

[8] 岡本ら, “E2E-S2S-VC：End-to-end 系列変換型声質変換”, 音講論, Mar. 2023.

[9] 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, June 2022.

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

[11] T. Okamoto *et al.*, “Multi-stream HiFi-GAN with data-driven waveform decomposition,” in *Proc. ASRU*, Dec. 2021, pp. 610–617.

[12] T. Kaneko *et al.*, “iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform,” in *Proc. ICASSP*, May 2022,

[13] 山下ら, “FC-HiFi-GAN：全結合層型アップサンプリングを導入した高速 HiFi-GAN”, 音講論, Sept. 2022, pp. 1133–1136.

[14] V. Panayotov *et al.*, “Librispeech: an ASR corpus based on public domain audio books,” in *Proc.ICASSP*, 2015, pp.5206–5210.

[15] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” in *SSW5*, 2004, pp.223–224.

[16] G.Zhao *et al.*, “L2-ARCTIC: A Non-native English Speech Corpus,” in *Proc. Interspeech*, 2018, pp.2783–2787.

[17] T.Hayashi *et al.*, “ESPnet2-TTS: Extending the edge of TTS research,” *arXiv:2110.07840*, 2021.