End-to-End 系列変換型声質変換への高速ニューラル波形生成モデル導入の検討 *

◎山下陽生 [1,2]，岡本拓磨 [2]，高島遼一 [1]，大谷大和 [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

近年，声質変換 (voice conversion:VC) 技術が発展 [1, 2] しており，特に話速や韻律を制御可能な系列変換モデル [3] が注目されている．以前は音響モデルに Conformer-based Fastspeech (CFS2)[4] を使用し，Vocoder に Parallel WaveGAN[5] を用いて別々に学習させるパイプライン方式が一般的であったが，岡本らが提案した JETS-VC[6] は，End-to-End TTS モデルである JETS[7] の入力をソース話者のメルスペクトログラムに変更したモデルで，End-to-End で学習可能な系列変換 VC モデルとなっている．JETS-VC は従来のパイプライン方式よりも高品質な音声を生成可能であった．しかし，JETS-VC は HiFi-GAN[8] を使用しているため，低計算資源下ではリアルタイム合成が難しい．本稿では，HiFi-GAN を高速化するモデルである MS-iSTFT-HiFiGAN[9] や MS-FC-HiFiGAN[10] を導入し，高品質な音声を保ちつつ高速化を行った．また，JETS-VC の Reduction factor[11] に関しては，これまで日本語に限定した調査が行われていたが，本稿では英語音声における Reduction factor の変更による品質の比較を行った．さらに，JETS-VC の Transformer block をその発展モデルである E-Branchformer Block[12] に変更し，品質向上の可能性を検証した．

2 モデル詳解

2.1 VC モデル：JETS-VC

JETS-VC は End-to-End で学習可能な Seaquence-to-Sequence のテキスト音声合成モデルであり，JETS の入力をソース話者の音声に変更したモデルである．Fig. 1 に示すように，Modified variance adaptor を使用しており，ソース音声から分析される $\log f_0$ やエネルギーは使用せず，代わりに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 $\log f_0$ や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行われ，Alignment module は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 との最大の違いは，波形生成モデルまで一貫して学習する点である．これにより，1 つのモデルで音声の変換が可能となり，変換品質の向上を実現している [6].

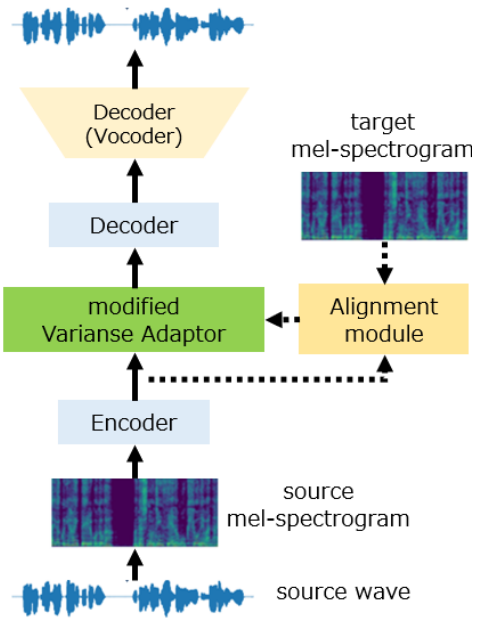

Fig. 1 JETS-VC model architecture.

2.1.1 Reduction factor[11]

VC モデルでは入力と出力の両方が音響特徴量であり時間分解能が高いため，attention 機構が上手く学習が出来ない．そこで，Reduction factor を用いて，音響特徴量内の隣接フレームを 1 フレームにスタックすることで時間軸を低減する．これによって atention 機構が上手く学習するようになる．

2.2 E-Branchformer Block

E-Branchformer Block は文献 [12] にて提案されたモデルであり，Branchformer Block[13] の発展モデルである．JETS-VC の Encoder/Decoder に用いられている Transformer Block の代わりに用いる．Fig. 2 に示すように，E-Branchformer Block はグローバルコンテキスト情報を抽出するグローバルブランチと，ローカルコンテキスト情報を抽出するローカルブラ

* Exploring the Introduction of High-Speed Neural Waveform Generation Models to End-to-End Sequence-to-Sequence Voice Conversion. by YAMASHITA, Haruki[1,2], OKAMOTO, Takuma[1], TAKASHIMA, Ryoichi[1], Yamato Ohtani[2], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], KAWAI, Hisash[1] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

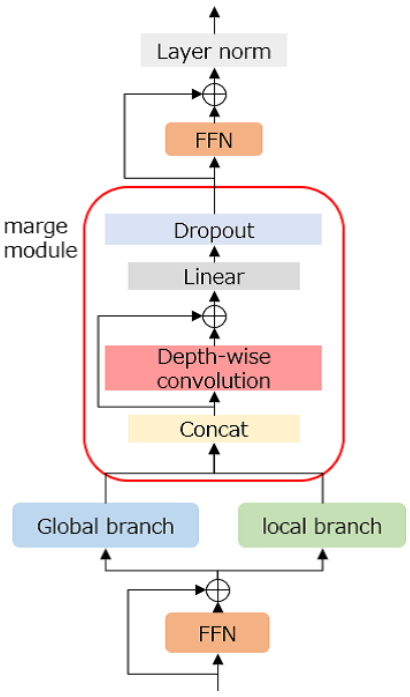

Fig. 2　E-Branchformer Block architecture.

ンチの二つのモジュールと，それらの情報をマージするモジュールから構成されている．マージモジュールには Depth-Wise Convolusion を用いることで，ローカルモジュールとグローバルモジュールの出力を逐次的かつ並列的にマージできるようになっており，音声認識タスクにて Branchformer よりも高い認識率を実現している．

2.3　ニューラル波形生成モデル

2.3.1　HiFi-GAN

HiFi-GAN の Generator は，入力のメルスペクトログラムを転置畳み込みと ResBlock を複数回用いてアップサンプリングし，それによって音声波形を生成する．また，HiFi-GAN は 2 つの優れた Discriminator を使用しており，これによって Generator を軽量化することが可能であり，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．

2.3.2　MS-iSTFT-HiFiGAN[9]

MS-iSTFT-HiFiGAN は，HiFi-GAN を品質を保ったまま高速化したモデルである Multi-stream HiFi-GAN[14] と iSTFTNet[15] を組み合わせ，Multi-stream 構造の間で iSTFT によるアップサンプリングを行うことで HiFi-GAN の合成品質を落とさずに約 4 倍の高速化を達成したモデルである．このモデルは VITS における高速化手法として提案されたモデルであるが，文献 [10] によって分析合成タスクでも HiFi-GAN と同等の合成品質となることが分かった．

2.3.3　MS-FC-HiFiGAN[10]

MS-FC-HiFiGAN は著者らによって提案されたモデルで，MS-iSTFT-HiFiGAN の iSTFT 部を単純な全結合層 (Fully Connected : FC) に変更することでトレーニング可能にし品質向上を目指したモデルである．文献 [10] によって分析合成タスク，TTS タスクの両方において MS-iSTFT-HiFiGAN を上回る音声品質となることが示されている．

3　実験

本実験ではまず各声質変換モデルと Reduction factor に関して客観評価実験を行う．その後，客観評価実験にて品質の良いモデルを選び，それらに対して主観評価実験を行う．

3.1　客観評価実験条件

データセット：英語音声データセットである CMU ARCTIC[16] から男性話者 (bdl) と女性話者 (slt) をそれぞれ 1 名ずつ選び，男性話者から女性話者への変換と，女性話者から男性話者への変換を行った．学習データは 1091 文とし，サンプリング周波数は 24 kHz とした．

モデル設定：Encoder/Decoder に Transformer を用いた変換モデルでは，Vocoder 高速化の影響を調べるために，Vocoder に HiFi-GAN とその高速モデルである MS-iSTFT-HiFiGAN，MS-FC-HiFiGAN の 3 種類を用いた．Vocoder の各モデルの ResBlock のカーネルサイズは HiFi-GAN が (8, 8, 2 ,2)，MS-iSTFT-HiFiGAN，MS-FC-HiFiGAN が (4, 4) とした．次に，Reduction factor の影響を調べるために，それぞれのモデルでは Encoder の Reduction factor (erf) を 2，3，4，8 の 4 種類とした．このとき，Decoder の Reduction factor (drf) は 1 である．erf，drf ともに 1 のモデルも学習したが，これらは変換が上手くできなかったため今回の実験には含めなかった．また，drf がどのように作用するかを調べるために Vocoder が HiFi-GAN のモデルについて (erf,drf)=(3,2)(4,2)(4,3) とした 3 つのモデルも学習し客観評価を行った．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．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 1000 epoch 学習した．

評価方法：合成品質の客観評価には，メルケプストラム歪み (MCD)，対数 f_0 の二乗平均誤差 ($\log f_0$ RMSE)，また変換による音声の崩れを評価するために Conformer ベースの音声認識モデルによる文字誤り率 (CER) 用いた．MCD と $\log f_0$ RMSE の計算に

はオープンソフトの ESPNet2-TTS[17] を利用した．CER は librispeech で学習した conformer ベースの音声認識モデルで測定した．推論速度の評価は RTF を使用し，AMD EPYC 7542 を 1PUC のみを使用して推論時の速度を測定した．客観評価には学習に用いていないデータのうち 20 文を使用した．

3.2 客観評価実験結果

JETS-VC 高速化については，RTF の結果から，MS-iSTFT-HiFiGAN，MS-FC-HiFiGAN を用いることで約 4 倍の高速化が達成されていることが分かる．

次に CER において erf=2 の場合において各手法ともに CER が低くなる一方，erf，drf をともに大きい値に設定すると CER が悪化する傾向がみられた．

また，高速 Vocoder を用いても変換後の CER は大きく悪化することはないこともわかった．

3.3 主観評価実験条件

客観評価実験の結果を Table 1 に示す．表から，主観評価には erf=8 のものと，(erf,drf)=(3,2)(4,3) の 3 つの条件を除いたモデルを用いた．合成品質の主観評価には分析合成実験と同様に平均オピニオン評点 (MOS) を用いた．また，話者類似性の評価には一対比較を行い，聴者が確かに同じ話者といえるかどうかを調べた．主観評価には各モデルから 5 文ずつを抜き出し，日本人聴者 7 名がヘッドホンを用いて行った．

3.4 主観評価実験結果

主観評価の結果を Table 2 に示す．E-Branchformer を用いたモデルは女性 → 男性変換，男性 → 女性変換によって erf の値は変わるが，どちらにおいても Transformer を用いるよりも高い品質になっていた．Transformer+HiFi-GAN においては erf=3 の時が自然性の評価値は最も高くになっており，女性 → 男性変換においては話者類似性も最も高くなった．MS-FC-HiFiGAN を用いた場合は erf=2 において自然性も話者類似性も共に最もよい評価値となった．MS-iSTFT-HiFiGAN では erf=2 だけでなく erf=4 においても高い品質となることが分かった．

以上の結果から，MS-iSTFT-HiFiGAN を除くモデルにおいては erf=2,3 が最も良い品質になり，CER の結果も踏まえると erf=2 が安定して良い品質になると考えられる．

4 おわりに

MS-iSTFT-HiFiGAN と MS-FC-HiFiGAN は JETS-VC (erf=2, 3) においても高品質な音質を維持したままの高速化が可能であることが分かった．また，E-Branchformer の導入によって変換品質が向上することが分かった．このとき，推論速度が Transformer よりも遅くなるが，MS-FC-HiFiGAN や MS-iSTFT-HiFiGAN を利用することで，この速度低下を抑えることができると考えられる．また，多くのモデルにおいて，CER，MOS ともに erf=2,3 が最も良くなっていた．しかし，女性 → 男性や男性 → 女性といった条件によって，モデルごとに Reduction factor の最適解が違い，条件に合わせた選択が必要であると考えられる．

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

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

[5] 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.

[6] T. Okamoto *et al.*, “E2E-S2S-VC: End-to-end sequence-to-sequence voice conversion,” in *Proc. Interspeech*, Aug. 2023.

[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.

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

[9] K. Masaya *et al.*, “Lightweight and High-Fidelity End-to-End Text-to-Speech with Multi-Band Generation and Inverse Short-

Table 1　Result of Objective Evaluation.

					female → male			male → female		
Encoder/Decoder block	Vocoder	erf	drf	RTF	CER	MCD[dB]	logf_o RMSE	CER	MCD[dB]	logf_o RMSE
Transformer	HFG	2	1	1.005	2.4	5.38 ± 0.37	0.17 ± 0.05	2.5	5.86 ± 0.39	0.20 ± 0.07
		3	1	1.006	1.9	5.36 ± 0.43	0.18 ± 0.06	3.9	5.93 ± 0.35	0.20 ± 0.04
		3	2	1.028	5.2	5.48 ± 0.42	0.17 ± 0.04	4.5	6.07 ± 0.38	0.18 ± 0.06
		4	1	1.001	2.4	5.43 ± 0.45	0.21 ± 0.08	4.7	5.86 ± 0.36	0.19 ± 0.06
		4	2	1.015	1.8	5.26 ± 0.32	0.19 ± 0.06	3.7	6.07 ± 0.36	0.20 ± 0.05
		4	3	1.017	13.6	5.78 ± 0.42	0.18 ± 0.06	15.1	6.48 ± 0.48	0.20 ± 0.07
		8	1	0.981	6.6	5.52 ± 0.40	0.20 ± 0.05	9.6	6.22 ± 0.47	0.20 ± 0.05
	MS-iSTFT	2	1	0.292	1.3	5.54 ± 0.34	0.17 ± 0.05	2.7	6.13 ± 0.38	0.21 ± 0.05
		3	1	0.290	2.0	5.34 ± 0.33	0.17 ± 0.05	5.1	5.92 ± 0.39	0.18 ± 0.06
		4	1	0.284	2.1	5.40 ± 0.39	0.20 ± 0.07	7.1	6.12 ± 0.36	0.18 ± 0.06
		8	1	0.276	6.5	5.54 ± 0.37	0.18 ± 0.05	17.7	6.65 ± 0.40	0.21 ± 0.05
	MS-FC	2	1	0.289	2.8	5.50 ± 0.40	0.17 ± 0.05	2.6	6.13 ± 0.35	0.18 ± 0.06
		3	1	0.287	2.8	5.37 ± 0.40	0.19 ± 0.06	5.4	6.36 ± 0.31	0.20 ± 0.06
		4	1	0.283	2.4	5.49 ± 0.37	0.19 ± 0.05	5.5	6.17 ± 0.36	0.22 ± 0.06
		8	1	0.282	7.8	5.58 ± 0.33	0.19 ± 0.05	22.1	6.60 ± 0.30	0.20 ± 0.06
E-Branchformer	HFG	2	1	1.070	1.8	5.30 ± 0.37	0.19 ± 0.04	3.4	5.76 ± 0.42	0.19 ± 0.06
		3	1	1.068	1.9	5.35 ± 0.40	0.18 ± 0.08	3.8	5.80 ± 0.42	0.20 ± 0.06
original	-	-	-	-	2.4	-	-	2.5	-	-

Table 2　Result of Subjective Evaluation.

				female → male		male → female	
Encoder/Decoder block	Vocoder	erf	drf	MOS	similarity[%]	MOS	similarity[%]
Transformer	HFG	2	1	3.83 ± 0.31	73.3	3.57 ± 0.37	56.7
		3	1	3.91 ± 0.27	**83.3**	3.77 ± 0.30	53.3
		4	1	3.80 ± 0.30	66.7	3.63 ± 0.30	70.0
		4	2	3.51 ± 0.37	63.3	3.26 ± 0.40	53.3
	MS-iSTFT	2	1	3.89 ± 0.36	73.3	3.43 ± 0.45	63.3
		3	1	3.11 ± 0.38	70.0	3.51 ± 0.43	50.0
		4	1	3.74 ± 0.39	46.7	4.03 ± 0.32	**80.0**
	MS-FC	2	1	3.74 ± 0.37	70.0	3.83 ± 0.32	**80.0**
		3	1	3.74 ± 0.33	63.3	3.71 ± 0.34	73.3
		4	1	3.11 ± 0.37	63.3	3.66 ± 0.36	63.3
E-Branchformer	HFG	2	1	**4.09 ± 0.29**	75.9	3.54 ± 0.44	50.0
		3	1	3.80 ± 0.38	73.3	**4.20 ± 0.30**	73.3
original	-	-	-	4.14 ± 0.39	-	4.66 ± 0.19	-

Time Fourier Transform, "in *Proc. ICASSP*, Jun 2023, pp. 1–5.

[10] 山下ら，“全結合層型アップサンプリングを導入した高速ニューラル波形生成モデル”，信学技報, vol. 123, no. 8, pp. 73–78, June 2023.

[11] W. Huang *et al.*, “ Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining, ” in *Proc. Interspeech,* Oct. 2020, pp.4676–4680.

[12] K. Kwangyoun, *et al.*, “ E-branchformer: Branchformer with enhanced merging for speech recognition,” in *Proc. SLT*, 2022. pp.84–91

[13] Y. Peng, *et al.*, “ Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding, ” in *Proc. ICML*, 2022.

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

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

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

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