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 *

☆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]，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

音響特徴量から音声波形を生成する信号処理に基づくソースフィルタ型ボコーダ (e.g. WORLD [1]) と比較して，近年ではニューラルネットを用いた方式 (ニューラルボコーダ) の合成品質の向上が目覚ましい [2]。当初提案された自己回帰モデル (e.g. WaveNet ボコーダ [3]，WaveRNN [4]) は合成速度が遅いという問題に対して，合成品質を保ちつつ 1CPU でも音声合成可能な推論速度を持つパラレル生成モデルとして HiFi-GAN [5] が提案され，音素系列から直接音声波形を 1 つのニューラルネットで高品質に生成可能な End-to-end テキスト音声合成モデルにも採用されている [6, 7]。

HiFi-GAN は音声波形を生成する Generator および優れた 2 つの Discriminator により構成される。HiFi-GAN の Generator では転置畳み込み+ResBlock によるアップサンプリングを計 4 回行うことでメルスペクトログラムから音声波形を生成するが，さらなる高速化として提案された Multi-Stream HiFi-GAN [8] や iSTFTNet [9] では最後の 2 層の転置畳み込み+ResBlock によるアップサンプリングをゼロ挿入型アップサンプリング+畳み込みや iSTFT のような高速処理に置き換えることで，合成品質を保ちつつ高速化に成功している。これらは HiFi-GAN の Generator においては，転置畳み込み+ResBlock によるアップサンプリングを 4 回も行わずとも推論のための十分な学習ができることを示し，別の処理に置き換えるた場合においても高速化が可能であることを示している。

これらの結果を踏まえ，本稿では Multi-Stream HiFi-GAN および iSTFTNet において提案されたアイデアを合わせ，新たに HiFi-GAN における最後 2 層の転置畳み込み+ResBlock によるアップサンプリングの代わりに単純な全結合層を導入した FC-HiFi-GAN を提案する。FC-HiFi-GAN と Multi-Stream HiFi-GAN，iSTFTNet について，合成品質，合成速度について比較実験を行う。

2 HiFi-GAN と応用モデル

2.1 HiFi-GAN

HiFi-GAN の Generator はメルスペクトログラムを入力として受け取り，転置畳み込み+ResBlock によるアップサンプリングを複数回行うことで音声波形を生成する。HiFi-GAN では優れた 2 種類の Discriminator を導入することにより Generator サイズを大きくせず，音声波形の生成時には Generator のみを使用するため，高品質かつ高速な音声合成を実現している。HiFi-GAN は，Fig. 1(a) に示すように，転置畳み込み+ResBlock によるアップサンプリングは (8, 8, 2, 2) とし，カーネルサイズは (16, 16, 4, 4) である。以下の改良モデルでは，隠れチャンネルは 512 とする HiFi-GAN V1 を基本にしたうえで，最後の 2 層の転置畳み込み+ResBlock を別の高速処理に置き換えることにより高速化を実現している。

2.2 Multi-Stream HiFi-GAN

HiFi-GAN と同様のアップサンプリング型ニューラル波形生成モデルとして Mel-GAN が提案されているが，音質を保ちつつ合成速度をさらに高速化する方式として Multi-Band MelGAN [10] が提案された。Multi-Band MelGAN では，最後の 4 倍アップサンプリングの代わりに 4 帯域にサブバンド化した波形を出力し，信号処理に基づくサブバンド合成フィルタを用いてフルバンド波形を得る。しかし，この方式を HiFi-GAN にそのまま適用した場合，サブバンドの制約が強すぎるため生成波形に対する Discriminator 損失が下がらず，学習が進まないという問題がある。

それに対して，Fig. 1(b) に示す Multi-Stream HiFi-GAN [8] は，信号処理に基づくサブバンド合成フィルタがゼロ挿入型アップサンプリング+固定フィルタによる畳み込み演算であることに着目し，固定値のサブバンド合成フィルタを学習可

*FC-HiFi-GAN: High-speed HiFi-GAN with fully connected layer-based upsampling. by YAMASHITA, Haruki[1,2], OKAMOTO, Takuma[1], TAKASHIMA, Ryoichi[1], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[1], KAWAI, Hisash[1] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

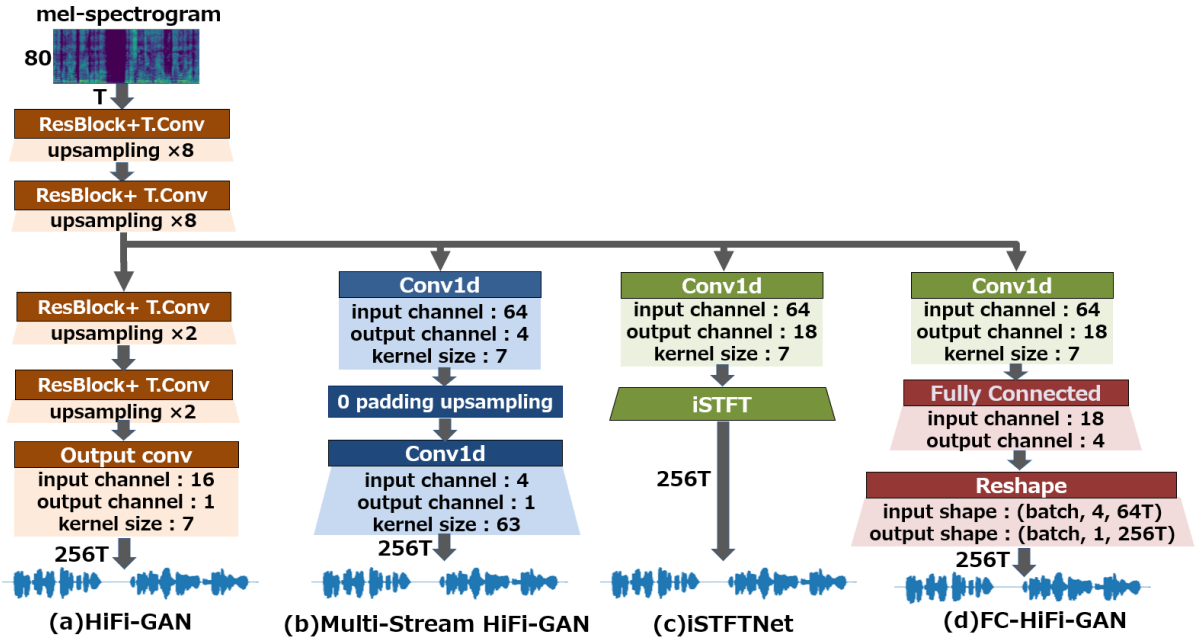

Fig. 1　model outline

能な畳み込みに置き換えたモデルである。これにより，高速化を実現しつつ，HiFi-GAN と同等の合成品質を実現できる。Multi-Band MelGAN のサブバンド合成フィルタ長を合わせるため，カーネルサイズは 63 としている。

2.3　iSTFTNet

iSTFTNet [9] は Fig. 1(c) に示すように，HiFi-GAN における最後の 4 倍アップサンプリングを逆短時間フーリエ変換 (iSTFT) に置き換えたものであり，Multi-Stream HiFi-GAN と同様，合成品質を保ちつつ，合成速度を改善させることに成功している。ここで，Fig. 1(c) の iSTFT は fft サイズは 16，window サイズは 16，hop サイズは 4 である。入力されたメルスペクトログラムは出力する音声波形の周波数特性を含んでいることに着目し，64 倍アップサンプリング後の Conv1d 層にて音声波形の短時間フーリエ変換 (STFT) の振幅成分と位相成分を出力するよう設計することで，iSTFT を通すことにより音声波形を出力する。文献 [9] では様々な条件が検討されたが，Fig. 1(c) に示す構造が HiFi-GAN と同等の最も高音質を実現している。

2.4　提案法：FC-HiFi-GAN

提案手法である FC-HiFi-GAN は，Fig. 1(d) のように，HiFi-GAN における最後の 4 倍のアップサンプリングを全結合層に置き換えたものである。提案モデルは Multi-Stream HiFi-GAN および iSTFTNet を基に考えられている。iSTFTNet では HiFi-GAN における最終 2 層の転置畳み込み+ResBlock を iSTFT に置き換えていたが，iSTFT は重みをフーリエ基底とした全結合層+オーバーラップ加算したものである。全結合層は学習可能な線形変換であるため，iSTFT の代わりに学習可能な全結合層を導入したモデルとして FC-HiFi-GAN を考案する。また，Multi-Stream HiFi-GAN において重みが固定のサブバンド合成フィルターの代わりに学習可能な畳み込みを用いることで高精度化に成功したように，iSTFT を学習可能にすることで精度の向上が期待できる。iSTFTNet と同様に，Conv1d の出力は振幅成分と位相成分に相当する 18 次元とし，学習可能な全結合層により 4 次元の信号を出力し，最後に並び替えにより 1 次元とすることにより 4 倍のアップサンプリングを実現している。

なお，iSTFT の演算と同じように全結合層の出力をオーバーラップ加算させる手段も考えられるが，単純な全結合層だけであっても十分な合成品質と推論速度を保っていることが確認されたため，本稿ではこのモデルを用いて実験を行った。

3　実験

3.1　実験条件

Multi-Stream HiFi-GAN，iSTFTNet，および提案法である FC-HiFi-GAN の合成品質，推論速度の比較を行う。データセットには 130 文を読み

Table 1　RTF，WARP-Q，MCD

	HiFi-GAN	MS-HiFi-GAN	iSTFTNet	FC-HiFi-GAN
RTF	0.74	0.45	0.35	0.35
WARP-Q	0.88	0.87	0.90	0.84
MCD	4.93	4.94	5.03	4.93

上げた100人の話者の音声が入っているJVSコーパス [11](サンプリング周波数24kHz) を用いた。そのうちjvs011からjvs100までの90名のパラレル100文およびノンパラレル30文の合計11,697文を学習に用いた。推論速度の測定と客観評価実験においては学習に使用していないjvs001からjvs010の10名のノンパラレル30文の合計300文を用いた。これにより，学習に含まれていない話者および発話の未知話者合成の評価となる。音響特徴量は8 kHzまでに帯域制限した80次元のメルスペクトログラムとした。それぞれのモデル学習におけるパラメータ更新250万回行った。モデルの実装及び学習，推論にはPyTorchベースのオープンソース[1]を用いた。

モデルの客観評価は，HiFi-GAN，Multi-Stream HiFi-GAN，iSTFTNet，FC-HiFi-GANの4つのモデルについて，WARP-Q [12] とメルケプストラム歪み (MCD) を用いた品質評価と推論速度の測定を行った。WARP-Qはニューラル波形生成モデルの評価手法の1つであり，音質評価における平均オピニオン評点 (MOS) [13] との関係が確認されており，WARP-Q値が小さければ小さいほどMOS値が良くなる傾向にある [12]。推論速度の測定にはIntel Xeon Gold 6152 CPU 2.1GHzを1コア用いて生成を行い，その時の実時間係数 (RTF:Real Time Factor) を計測した。モデルの主観評価には，11名の正常な聴力を有する日本人成人を実験参加者とする聴取実験によるMOS評価を行った。学習に用いていない話者のうち，jvs001，jvs003，jvs004，jvs008(男女2名ずつ) からそれぞれ10文ずつ抜き出した合計40文について，Multi-Stream SHiFi-GAN，iSTFTNet，FC-HiFi-GANから推論された音声と，原音声を合わせた160文を用いた。実験参加者は静かな環境でヘッドホンにより音声を聴取した。

[1]https://github.com/kan-bayashi/ParallelWaveGAN

3.2　実験結果

3.2.1　客観評価実験

RTF，WARP-Q，MCDの値をTable 1に示す。RTFの結果から，HiFi-GANが最も合成速度が遅く，Multi-Stream HiFi-GAN，iSTFTNet，FC-HiFi-GANのそのモデルも高速化されていることが確認できた。高速化されたモデルの中では，Multi-Stream HiFi-GANが最も合成速度が遅く，FC-HiFi-GANとiSTFTNetが同等の合成速度であることが分かった。FC-HiFi-GANとiSTFTNetはどちらも処理が線形変換となっているため同じ速度であり，すべてのモデルの中で最も速くなっていると考えられる。Multi-Stream HiFi-GANでは合成フィルターのカーネルサイズを63と大きくしているためMulti-Stream HiFi-GANがiSTFTNet，FC-HiFi-GANよりも遅くなっていると考えられる。

次に，合成品質についての結果を確認する。WARP-Qの結果を見ると，iSTFTNetの値が最も高くなり，高速化されたモデルの中で唯一HiFi-GANよりも悪い結果を示している。FC-HiFi-GANは4つの中で最も良い結果を示した。MCDの結果を見ると，WARP-Qと同様にiSTFTNetが最も悪い結果となり，HiFi-GAN，Multi-Stream HiFi-GAN，FC-HiFi-GANの3つのモデルは差がないという結果となった。

FC-HiFi-GANとiSTFTNetの比較から，iSTFTを学習可能な全結合層に変更するだけで，合成速度を維持したままより合成品質を上げることが可能であると考えられる。これは，iSTFTが単純な線形変換の重ね合わせであるのに対して，全結合層は学習可能な線形変換であり，複数話者に適応するように学習した結果，iSTFTよりも柔軟に未知話者予測が可能となったためであると考えられる。

3.2.2　主観評価実験

Fig. 2にMOS評価の結果を示す。結果より，WARP-Qでの結果と違い，Multi-Stream HiFi-GANが最も合成品質が良く，FC-HiFi-GANが

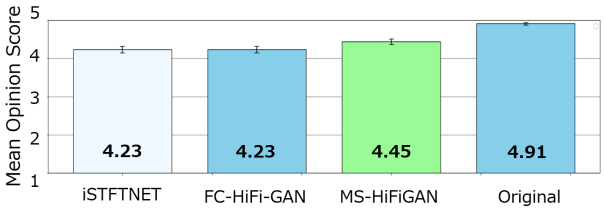

Fig. 2 Mean Opinion Score(MOS), Confidence level of the error base was 95%.

と iSTFTNet の合成品質では FC-HiFi-GAN は同じ値で，有意な差はないという結果となった。

客観評価実験の結果と合わせて，Multi-Stream HiFi-GAN が最も合成品質が良いものの合成速度は遅く，FC-HiFi-GAN と iSTFTNet は合成速度，合成品質ともに差がないが，Multi-Stream HiFi-GAN よりも合成速度が速いという結果になった。

合成品質は Multi-Stream HiFi-GAN が最も良かったが，本来の HiFi-GAN の高速化という面を考えると，FC-HiFi-GAN が合成速度と合成品質のバランスにおいて最も良いことが分かった。

FC-HiFi-GAN と iSTFTNet においては合成速度，合成品質ともに同じであったが，単純な信号処理である iSTFT よりも全結合層の方が改良の余地があるため，FC-HiFi-GAN の方が今後の発展に期待できる。

また，本稿で用いた Multi-Stream HiFi-GAN はカーネルサイズが 63 となっており，このサイズが大きいために合成速度が遅くなっている。そのため，カーネルサイズを減らし，iSTFTNet，FC-HiFi-GAN と同程度の合成速度に合わせたのちに合成品質を比較する必要もあると考え，これは今後の検証課題である。

4 おわりに

iSTFTNet の iSTFT によるアップサンプリングを学習可能な全結合層に置き換えた FC-HiFi-GAN を提案した。HiFi-GAN，Multi-Stream HiFi-GAN および iSTFTNet との比較実験を行い，提案法の有効性を確認した。

参考文献

[1] 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, July 2016.

[2] 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.

[3] A. Tamamori *et al.*, “Speaker-dependent WaveNet vocoder,” in *Proc. Interspeech*, Aug. 2017, pp. 1118–1122.

[4] N. Kalchbrenner *et al.*, “Efficient neural Audio Synthesis,” in *Proc. ICML*, June 2018, pp. 2410–2419.

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

[6] J. Kim *et al.*, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. ICML*, July 2021, pp. 5530–5540.

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

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

[9] T. Kaneko *et al.*, “iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform,” in *Proc. ICASSP*, May 2022, pp. 6207–6211.

[10] G. Yang *et al.*, “Multi-band MelGAN: Faster waveform generation for highquality text-to-speech,” in *Proc. SLT*, Jan. 2021, pp. 492–498.

[11] S. Takamichi *et al.*, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[12] W. A. Jassim *et al.*, “WARP-Q: Quality prediction for generative neural speech codecs,” in *Proc. ICASSP*, June 2021, pp. 401–405.

[13] ITU-T Recommendation P. 800, *Methods for subjective determination of transmission quality*, 1996.