4S-08

SHAREVOX -多彩な表現が可能なテキスト音声合成ソフトウェアの開発とモデルアーキテクチャの改善-

芦田 裕飛 [*1]　寺内 健人 [1]　本部 勇真 [1]　柳井 啓司 [1]

[1] 電気通信大学

1 はじめに

近年, 深層学習の発展により, テキスト音声合成 (TTS) は大きく進展している. この例として, 2021 年 8 月に TTS を行うだけでなく, それを独自 UI を通して自分好みに調整することが可能である VOICEVOX[*1] というソフトウェアが個人より公開された. 自分好みに調整可能なのは, アクセント, イントネーション (F0), 音素長などである. この VOICEVOX については, 一部を除き OSS として公開されている.[*2]

これらの OSS を活用したソフトウェアとして, 筆頭著者の芦田は SHAREVOX をリリースした.[*3] 本研究では, リリース後の音声合成モデルの改善について取り扱う. 一般にソフトウェアユーザーは, 合成音声の品質と動作の快適性を求める. 品質面では, 一気通貫モデルや Source-Filter モデルの採用, 速度面では逆離散ウェーブレット変換 (iDWT) を用いて波形を生成するモデルを採用し, ソフトウェアに組み込まれている現行手法と性能を評価する.

2 関連研究

TTS 分野においては, 音響モデルとニューラルボコーダを合わせる手法と, 一気通貫で学習・推論ができる手法がある.

2.1 音響モデル

音響モデルは, メルスペクトログラムをはじめとした音響特徴量を生成するモデルである. FastSpeech 2 [1] は, 音素長と F0 の制御が可能で, Transformer を用いて非自己回帰に音響特徴量を生成できるモデルである. 現在 SHAREVOX で採用しており, 内部に音素列から音素長と F0 を推論するネットワークを持っている. SHAREVOX にはアクセントを制御できる機能があるが, これは先行研究報告 [2] に基づくアクセント埋め込みによるものである.

2.2 ニューラルボコーダ

ニューラルボコーダは音響特徴量から波形を生成するモデルである. 代表例として, WaveNet や Parallel WaveGAN, HiFi-GAN が挙げられる. HiFi-GAN は数層の転置畳み込みを用いることで, 品質と高速な動作を両立している. 品質が高いが, 厳密な F0 には従わない上, 長音に弱い特性がある.

これらの特性や問題を克服するため, 音響特徴量と F0 を入力にとって F0 の制御を可能とした, Period-HiFi-GAN や Source-Filter HiFi-GAN(SiFi-GAN) [3] が提案されている.

その他にも, HiFi-GAN をベースに識別器に離散ウェーブレット変換 (DWT) を取り入れ, 生成器の構造を変更し, 高品質化を図った Fre-GAN があり, 現在 SHAREVOX で用いている. より品質を高めて高速化を図るため, 生成器に iDWT を取り入れ, 転置畳み込み層を減らした, 後継の Fre-GAN 2 [4] も存在する.

2.3 一気通貫モデル

音響特徴量を介さず, 音響モデルとニューラルボコーダを結合して一気通貫で学習, 推論を行う方式である. VAE を用いる VITS や, JETS [5] といった手法がある. 学習時に音素アライメントも学習するため, FastSpeech 2 のようにあらかじめ外部で音素アライメントデータを用意する必要がなく, テキストと音声データがあれば学習可能である. ニューラルボコーダとして HiFi-GAN が用いられていることが多い.

3 提案手法

一気通貫モデルに近い手法として, 音素と F0, 音素長から音声波形を生成するモデルとして, 図 1 のモデルを提案する. SHAREVOX に必要な要素である, 音素とアクセントから, F0 と音素長を推論するモデルについては, 現在 SHAREVOX にて採用している先行研究の手法 [6] に従う.

3.1 一気通貫モデルと Source-Filter モデルによる品質向上

FastSpeech 2 で生成される音響特徴量は, 過剰に平滑化される傾向があり, この音響特徴量から生成した音声は品質が過度に落ちることがある. これを回避するためには, FastSpeech 2 が生成する音響特徴量でニューラルボコーダを転移学習しなければならない. この手間を減らすため, 一気通貫モデルの JETS を採用する. 加えて, 長音に弱い問題を解決するため, Decoder 部分に SiFi-GAN を採用する. また, JETS では Variance Adapter 内で埋め込まれている F0 を Decoder に直接入力するようにする. なお SHAREVOX では, F0 を音素単位で推論するため, Decoder に入力する場合にはフレームレベルまで拡張しなければならない. そのため, FastSpeech 2 の Variance Predictor を用い, 音素単位の F0 からフレームレベルの F0 を予測する. また, 事前実験において, 予測したフレー

[*] y.ashida@uec.ac.jp

[*1] https://voicevox.hiroshiba.jp

[*2] https://github.com/VOICEVOX

[*3] https://sharevox.app

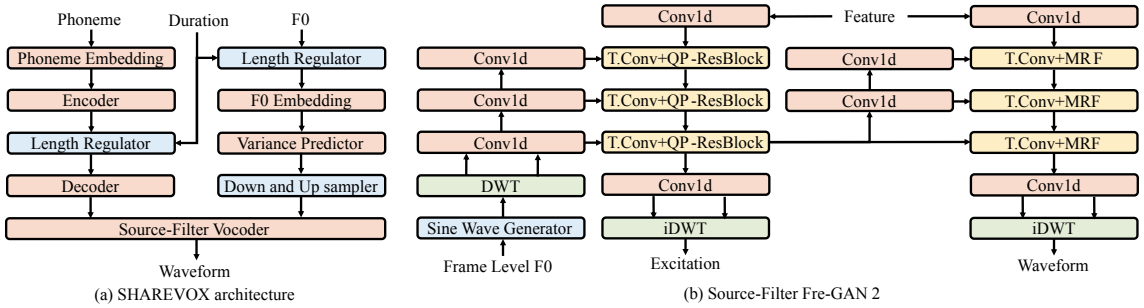

図 1　新しく提案する SHAREVOX のモデル図, (a) は全体的なアーキテクチャ, (b) は新しく提案する Source-Filter Vocoder の SFre-GAN 2 を表す

ムレベルの F0 がフレームごとにぶれるために, 出力音声が震え声になる傾向があったため, F0 を一度ダウンサンプリングし, 線形補間でアップサンプリングすることで, なめらかになるようにする.

3.2　転置畳み込み層の置き換えによる速度向上

Fre-GAN 2 の iDWT による実行速度の向上を JETS のニューラルボコーダに取り入れる. 先の品質向上の項で示した SiFi-GAN と Fre-GAN 2 を組み合わせた, Source-Filter Fre-GAN 2(SFre-GAN 2) を提案し, 組み込む. 識別器には, Fre-GAN 2 で用いられている RPD と SiFi-GAN で用いられている MRD を組み合わせて用いた.

4　実験

全ての実験は実際に SHAREVOX に用いている事前学習用データセット 1 話者 2024 文と製品用データセット 7 話者 1672 文を用いた. 従来の SHAREVOX モデルを除き, 事前学習用データセットで 10 万ステップ学習したのち, 転移学習を行い, 製品用データセットで 20 万ステップ学習させた. バッチサイズは共通して 12 とした. 音声に関しては, 量子ビット数を 16, サンプリングレートを 48kHz とした.

4.1　品質評価

合成音声の自然性を評価するため, 主観的評価として 14 人に対して Mean Opinion Score(MOS) テストを行った. 従来の SHAREVOX の手法である **FS2+Fre** と, 一気通貫モデル **JETS**, JETS のニューラルボコーダに SiFi-GAN を用いた **JETS-SiFi**, SFre-GAN 2 を用いた **JETS-SFre2**, Fre-GAN 2 で提案されている multi-level iDWT を用いた **JETS-SFre2m** の計 5 手法で合成した音声と, 自然音声をランダムに 100 文聴いてもらい, 5 段階で評価してもらった.

4.2　速度評価

一気通貫モデルや SFre-GAN 2 の採用による, 実行速度の変化を評価するため, 100 文を生成して CPU と GPU での実行時間を計測した. この評価では, モデルの大部分を ONNX 形式に変換して ONNX Runtime で実行し, モデル化できない一部の処理は CPU で行った. 評価にあたり, CPU は AMD Ryzen 5 3600X, GPU は GeForce RTX 3090 を用いた.

5　結果

表 1 に実験の結果を示す. **JETS** が RTF において最も良いスコアを出したが, **JETS+SiFi** のスコアが低いこと,

表 1　各モデルの MOS テストの結果と RTF(Real-Time Factor), MOS は 95% 信頼区間で表される

Method	MOS	RTF(CPU)	RTF(GPU)
Ground Truth	4.58 ± 0.11	-	-
FS2+Fre	**3.21 ± 0.32**	0.28	0.093
JETS	3.14 ± 0.30	**0.18**	**0.068**
JETS+SiFi	2.22 ± 0.38	0.35	0.10
JETS+SFre2	1.99 ± 0.32	0.28	0.085
JETS+SFre2m	2.67 ± 0.33	0.20	0.071

JETS+SFre2m が **JETS** に迫るスコアを出していることから, iDWT の利用に関する効果が示唆された. MOS に関しては **FS2+Fre** が最も良いスコアとなった. **FS2+Fre** は製品化のためにかなり長い時間学習したものであるため, 一概に比較することは難しいが, 提案手法は既存手法を超えることができなかった. ただ, 速度面のために採用した **JETS+SFre2m** が **JETS+SiFi** よりも高い品質という非常に興味深い結果となったと同時に, **JETS+SFre2** の品質が低くなるという興味深い結果も得られた.

6　まとめ

品質向上や速度向上のために, 最先端の手法を取り入れた. 結果として, 提案手法は既存手法を超えられなかった. 製品と同等の学習時間でのモデル制作や品質比較, SFre-GAN 2 の識別器の改善などが今後の課題だ.

参考文献

[1] Ren, Y., et al.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, in *Proc. ICLR* (2021).
[2] 藤井一貴他：韻律情報で条件付けされた非自己回帰型 End-to-End 日本語音声合成の検討, Technical report, 情報処理学会 (2021).
[3] Yoneyama, R., et al.: Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder, arXiv:2210.15533 (2022).
[4] Lee, S.-H., et al.: FRE-GAN 2: Fast and Efficient Frequency-Consistent Audio Synthesis, in *Proc. ICASSP*, pp. 6192–6196 (2022).
[5] Lim, D., et al.: JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech, in *Proc. Interspeech*, pp. 21–25 (2022).
[6] 芦田裕飛：手軽に製作可能かつ、調整可能で品質の高いテキスト音声合成システムの開発, `https://onsite.gakkai-web.net/ipsj/poster/pdf/8056.pdf` (2022).