正弦波入力型ニューラルボコーダを用いたTTSモデルによる歌声合成*

☆清水聡太[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

ニューラルネットワーク技術の進展により，テキストから自然音声を生成するテキスト音声合成 (Text-To-Speech: TTS) や，ある話者の音声を発話内容を保持したまま別の話者の音声へと変換する声質変換 (Voice Conversion: VC) において，高速かつ高品質な音声波形生成が可能となっている [1,2]。近年は入力テキストや変換元話者の音響特徴量から出力音声波形を1つのニューラルネットで直接生成可能な End-to-end(E2E)TTS および VC モデルが提案され，従来のパイプライン方式を凌駕する高品質な生成を実現している [3–5]。

また，楽譜と歌詞情報を入力とし，自然な歌声を生成する歌声合成 (Singing voice synthesis: SVS) についてもニューラルネットワークに基づく方式が盛んに検討され，E2E モデルも提案されており [6–8]，SVS モデル学習ツールキット [9] も公開され，歌声変換チャレンジ [10] も実施されている。

SVS モデルを学習するためには歌声コーパスを収録する必要があるが，歌声コーパスは TTS や VC 用の音声コーパスよりもさらに収録コストが高い。これに対して，既存の TTS 用音声コーパスのみを用いて歌声合成を実現する枠組みが提案されている [11]。このモデルでは，音素系列と音声から分析した基本周波数系列を入力とし，音響特徴量 (メルスペクトログラム) 系列を出力するニューラルネットワークを学習する。学習の際は Connectionist Temporal Classification(CTC) 損失 [12] によって得られる各フレームの音素確率を入力とし，合成の際は楽譜から得られる各フレームの音素と基本周波数を入力とし，歌声のメルスペクトログラム系列を出力する。推定されたメルスペクトログラムから別途学習した HiFi-GAN [13] によって歌声波形を出力する。文献 [11] では，LJSpeech コーパス [14] を用いた英語歌声合成デモが紹介されている。

TTS コーパスのみで SVS モデルを学習する場合，大きく2つの問題を解決する必要がある。1つは基本周波数の外装，もう1つは話速の外装である。TTS コーパスは通常発話のみで構成されるため，基本周波数のレンジはそれほど広くない。それに対して SVS は基本周波数のレンジが通常発話よりも広いため，合成の際には学習データ範囲外の基本周波数を外装する必要がある。また，通常発話の話速は歌声の話速とは異なるため，各音素ごとの音素継続長や音響特徴量を音符の長さに合わせて時間方向に伸縮する必要がある。従来法 [11] ではメルスペクトログラム入力を用いた HiFi-GAN をニューラル波形生成モデルとして導入していたが，上記2つの課題を解決するために，本論文では基本周波数に対応した正弦波入力型のニューラル波形生成モデル Harmonic-Net+ [15] および SiFi-GAN [16] を導入する。Harmonic-Net+では，正弦波入力型のニューラル波形生成モデルはメルスペクトログラム入力型ニューラル波形生成モデルよりも基本周波数制御や話速制御に適していることを示している [15]。Harmonic-Net+ や SiFi-GAN はメルスペクトログラムではなくソースフィルタボコーダである WORLD 特徴量 [17] で動作するため，WORLD 特徴量を出力する SVS 音響モデルを新たに提案する。JSUT [18] コーパスを用いた SVS モデルを構築し (Fig. 1)，客観評価実験により，Harmonic-Net+や SiFi-GAN は HiFi-GAN よりも TTS コーパスを用いた SVS モデルにおいて基本周波数の制御性能が高いことを示す。

2 音声合成用コーパスを用いた歌声合成モデル

本論文では，TTS 用コーパスを用いた SVS 音響モデルとして，従来法である CTC 損失に基づく方式ではなく，非自己回帰型 TTS モデルに基づく方式を提案する。具体的には，Conformer-FastSpeech 2(CFS2) を用いた非自己回帰型 TTS

*Investigation of singing voice synthesis using TTS models with sinusoidal input-based neural vocoders. by SHIMIZU, Sota[1,2], OKAMOTO, Takuma[2], , TAKASHIMA, Ryoichi[1], OHTANI, Yamato[2], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], KAWAI, Hisashi[2] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

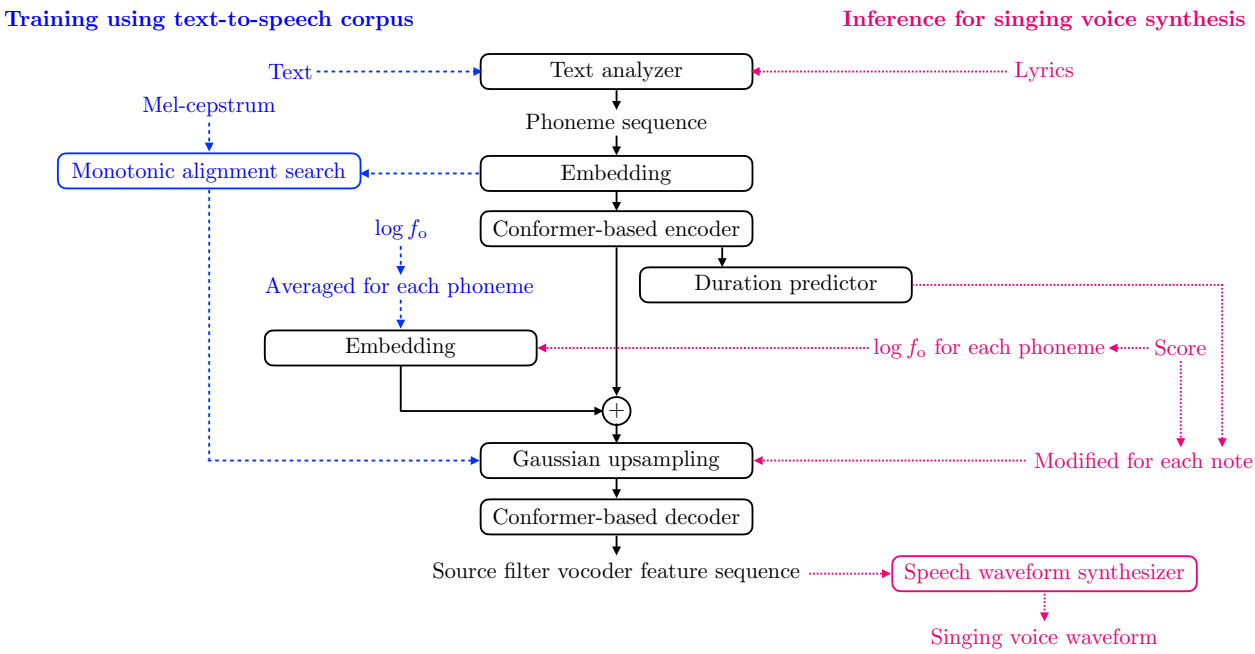

Fig. 1　Conformer-FastSpeech 2-based acoustic model for singing voice synthesis trained using text-to-speech corpus.

モデル [19] に JETS [4] で用いられているモノトニックアライメントサーチ [20] に基づく音素アライメント獲得機構 [21] を組み込んだモデルを提案する (Fig. 1)。通常の FastSpeech 2 [22] ではアップサンプリングの後フーレム単位の基本周波数 (とエネルギー) を推定し，隠れ特徴量と足し合わせてデコーダに入力するのに対して，FastPitch [23] や CFS2 ではエンコーダー出力から音素単位の基本周波数を推定し，隠れ特徴量と足し合わせてアップサンプリングを行っている。後者を用いることにより，歌声合成時に各音素の基本周波数を音符情報から陽に入力することができるところがポイントである。学習時はメルケプストラム[1]と音素埋め込みベクトル間でモノトニックアライメントサーチにより各音素の継続長を逐次更新にて推定し，その推定結果を元にガウシアンアップサンプリング [4] を行う。歌声合成時は，歌詞から音素系列を取得し，楽譜から各音符の継続長と基本周波数を取得し，それぞれ入力することにより WORLD 特徴量系列を推定する。推定された WORLD 特徴量系列を音声波形生成モデルに入力することにより，最終的な歌声波形を出力する。

[1]JETS ではメルスペクトログラムとエンコーダから出力された隠れ特徴量間でモノトニックアライメントサーチを実施しているが [4]，本論文ではメルケプストラムと音素埋め込みベクトル間でモノトニックアライメントサーチを実施した。

3　実験

3.1　実験条件

提案法の有効性を確認するために，JSUT コーパス [18] を用いた実験を行った。サンプリング周波数は 24 kHz とした。学習セットの平均基本周波数 f_o は 206 Hz であった。音響モデルは ESPnet2-TTS [2] における CFS2 の JSUT レシピを元に構築した。WORLD 音響特徴量として，基本周波数 (連続 log f_o および有声無声フラグの 2 次元)，メルケプストラム 50 次元，非周期成分 3 次元の計 55 次元をそれぞれ Harvest [26]，CheapTrick [24]，D4C [25] によりフレームシフト量 10 ms，FFT サイズ 1024 サンプルで抽出した。モノトニックアライメントサーチのための音響特徴量としてメルケプストラム 50 次元を使用した。音響モデルの学習回数は 200 epoch とした。WORLD 特徴量を用いた波形生成モデルとして，WORLD [17]，HiFi-GAN [13]，Harmonic-Net+ [15]，SiFi-GAN [16] とした。WORLD 以外は `https://github.com/kan-bayashi/ParallelWaveGAN` をベースにして 250 万回の学習を行った。今回の実験では音響モデル，音声波形生成モデル共にファインチューニングは行っていない。

歌声合成のテストセットは童謡の「どんぐりころころ」の 1 番の 4 フレーズとした。基本の楽譜条件 ($f_o \times 1.0$) に加え，1 オクターブ下げた

条件 ($f_0 \times 0.5$) および 1 オクターブ上げた条件 ($f_0 \times 2.0$) も合成した。基本周波数制御は音響モデルから推定した f_0 に対して行った。客観評価として $\log f_0$ 二乗平均平方根誤差を算出した。ここで，正解の f_0 は音符情報とした。$f_0 \times 1.0$ 条件，$f_0 \times 0.5$ 条件，$f_0 \times 2.0$ 条件における平均 f_0 はそれぞれ 352 Hz，176 Hz，704 Hz である。

3.2 実験結果

合成された音声サンプルはデモページ[2]より視聴できる。f_0 二乗平均平方根誤差の結果を Table 1 に示す。ソースフィルタボコーダである WORLD が最も精度よく基本周波数を制御できていることがわかる。HiFi-GAN は正弦波入力はしていないため，$f_0 \times 0.5$ 条件では問題なく動作するが，$f_0 \times 1.0$ では f_0 が高い音符の音は出せておらず，$f_0 \times 2.0$ 条件においてはほぼ全ての音符の音が出せていなかった。これは，正弦波入力がない場合は学習データ範囲内の f_0 しか合成できず，学習データ範囲外の f_0 を外装できなかったためである。それに対して，Harmonic-Net+や SiFi-GAN は正弦波入力があるため学習データ範囲外の f_0 を外装することができるため，HiFi-GAN よりも精度の高い合成が可能であることが確認できた。

今後は，音響モデルと正弦波入力型ニューラル波形生成モデルの同時ファインチューニングや一貫学習，歌声テストセット [27] を用いた評価，および我々が新たに提案した ConvNeXt 型音響モデル [28] や f_0 制御可能なニューラル波形生成モデル [29] の導入等を行う。

4 おわりに

TTS コーパスのみを用いて学習可能なニューラル SVS モデルを提案し，WORLD 特徴量入力型波形生成モデルを用いて歌声合成を行った。正弦波入力型波形生成モデルの方が基本周波数制御に頑健であることを示した。

参考文献

[1] 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.

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

[3] 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.

[4] 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.

[5] T. Okamoto *et al.*, “E2E-S2S-VC: End-to-end sequence-to-sequence voice conversion,” in *Proc. Interspeech*, Aug. 2023, pp. 2043–2047.

[6] Y. Zhang *et al.*, “VISinger: Variational inference with adversarial learning for end-to-end singing voice synthesis,” in *Proc. ICASSP*, May 2022, pp. 7237–7241.

[7] Y. Zhang *et al.*, “VISinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer,” in *Proc. Interspeech*, Aug. 2023, pp. 4444–4448.

[8] Y. Lei *et al.*, “UniSyn: An end-to-end unified model for text-to-speech and singing voice synthesis,” in *Proc. AAAI*, Feb. 2023, pp. 13025–13033.

[9] R. Yamamoto *et al.*, “NNSVS: a neural network based singing voice synthesis toolkit,” in *Proc. ICASSP*, June 2023.

[10] W.-C. Huang *et al.*, “The Singing Voice Conversion Challenge 2023,” in *Proc. ASRU*, Dec. 2023.

[11] S. Choi *et al.*, “A melody-unsupervision model for singing voice synthesis,” in *Proc. ICASSP*, May 2022, pp. 7242–7246.

[12] A. Graves *et al.*, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. ICML*, June 2006, pp. 369–376.

[2] https://www.okamotocamera.com/asj24s_svs/

Table 1 Results of f_o root mean square error.

	$f_o \times 1.0$	$f_o \times 0.5$	$f_o \times 2.0$
Acoustic model	0.05	N/A	N/A
WORLD	0.12	0.14	0.14
HiFi-GAN	0.58	0.15	1.25
Harmonic-Net+	0.35	0.30	0.99
SiFi-GAN	0.35	0.22	1.11

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

[14] K. Ito and L. Johnson, "The LJ Speech Dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[15] K. Matsubara *et al.*, "Harmonic-Net: Fundamental frequency and speech rate controllable fast neural vocoder," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1902–1915, 2023.

[16] R. Yoneyama *et al.*, "Source-Filter HiFi-GAN: Fast and pitch controllable high-fidelity neural vocoder," in *Proc. ICASSP*, June 2023.

[17] 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, Jul. 2016.

[18] 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.

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

[20] J. Kim *et al.*, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, Dec. 2020, pp. 8067–8077.

[21] R. Badlani *et al.*, "One TTS alignment to rule them all," in *Proc. ICASSP*, May 2022, pp. 6092–6096.

[22] Y. Ren *et al.*, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, May 2021.

[23] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*, June 2021, pp. 6588–6592.

[24] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, Mar. 2015.

[25] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 67–65, Nov. 2016.

[26] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. Interspeech*, Aug. 2017, pp. 2321–2325.

[27] `https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song`

[28] T. Okamoto *et al.*, "ConvNeXt-TTS and ConvNeXt-VC: ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion," in *Proc. ICASSP*, Apr. 2024. (accepted, in press)

[29] Y. Ohtani *et al.*, "FIRNet: Fundamental frequency controllable fast neural vocoder with trainable finite impulse response filter," in *Proc. ICASSP*, Apr. 2024. (accepted, in press)