Interspeech 2024
1-5 September 2024, Kos, Greece

Challenge of Singing Voice Synthesis Using Only Text-To-Speech Corpus With FIRNet Source-Filter Neural Vocoder

Takuma Okamoto[1], *Yamato Ohtani*[1], *Sota Shimizu*[2,1], *Tomoki Toda*[3,1], *Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Japan
[2]Kobe University, Japan [3]Nagoya University, Japan

{okamoto, yamato.ohtani, hisashi.kawai}@nict.go.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

Singing voice synthesis (SVS) corpora are more costly to collect than TTS corpora. SVS using only a TTS corpus is challenging because the ranges of fundamental frequency (f_o) and phoneme duration in SVS are wider than those in TTS. Although a melody-unsupervised method prototyped SVS using only a TTS corpus, some problems remain. To improve duration and f_o controllability, this paper proposes a unified TTS and SVS framework. It is based on the FastSpeech-2-based duration-expansion-robust TTS acoustic model with phoneme embedding skip connection (PESC) and FIRNet source-filter neural vocoder with source-filter acoustic features. In the inference for SVS, the input text, f_o, and phoneme duration are obtained from lyrics and notes in a musical score. Additionally, input f_o shift is proposed. Experiments using the JSUT corpus confirm that the PESC-based acoustic model using input f_o shift and FIRNet can improve the SVS quality compared with that using HiFi-GAN.

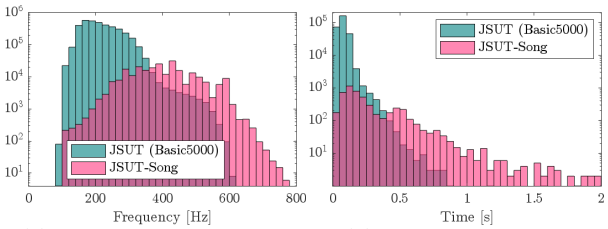
Index Terms: FastSpeech 2, FIRNet, fundamental frequency control, text-to-speech, singing voice synthesis

1. Introduction

Text-to-speech (TTS) [1–13] can synthesize speech waveforms from input text sequences. Similarly, high-fidelity singing voice synthesis (SVS), which can synthesize singing voice waveforms from input musical scores (lyrics and notes), can be achieved by recent developments in neural network technologies [14–20]. For many speech applications, it is useful to perform SVS with the same speaker voice as TTS. To construct high-quality SVS models, a large SVS corpus is required. However, SVS corpora are more expensive to collect than TTS corpora. Additionally, some speakers are not good at singing. To perform SVS with the same speaker voice as TTS in the absence of an SVS corpus that includes the speaker, the following two approaches have been investigated.

The first approach is UnySyn, a unified end-to-end TTS and SVS model that uses multi-speaker TTS and SVS corpora [19]. In UnySyn, speaker timbre and speaking style are disentangled using a multi-conditional variational autoencoder, and SVS can be performed for a speaker whose singing data are not included in the training set [19]. However, the collection of multi-speaker TTS and SVS corpora is costly.

The second approach is to train an SVS model using only a TTS corpus. Figure 1 shows the histograms of fundamental frequency (f_o) [22] and phoneme duration for the JSUT corpus [21] (Basic5000) for TTS and the JSUT-Song corpus[1] for SVS of a Japanese female speaker. The histograms indicate that performing SVS using only a TTS corpus is a challenging task because the ranges of f_o and phoneme duration in SVS are wider than those in TTS. Although a melody-unsupervised SVS method [23] prototyped SVS using only the LJSpeech corpus [24] (a TTS corpus), the following problems remain with this approach:



(a) Fundamental frequency (b) Phoneme duration

Figure 1: *Histograms of JSUT (Basic5000) corpus [21] for text-to-speech (TTS) and JSUT-Song corpus for singing voice synthesis (SVS).*

P1) Although synthesized demo samples have been published[2], no experiments have been conducted.

P2) Although mel-spectrograms [1] and HiFi-GAN [5] are used for the acoustic features and neural vocoder, respectively, it is difficult to extrapolate f_o outside the range of the training data.

P3) The f_o trajectory of the synthesized voice tends to be stair-stepped because f_o is input in each subframe and the same value is repeated.

P4) Although the model is trained using only a TTS corpus, it can only perform SVS. TTS cannot be performed because f_o and phoneme duration are required to be input.

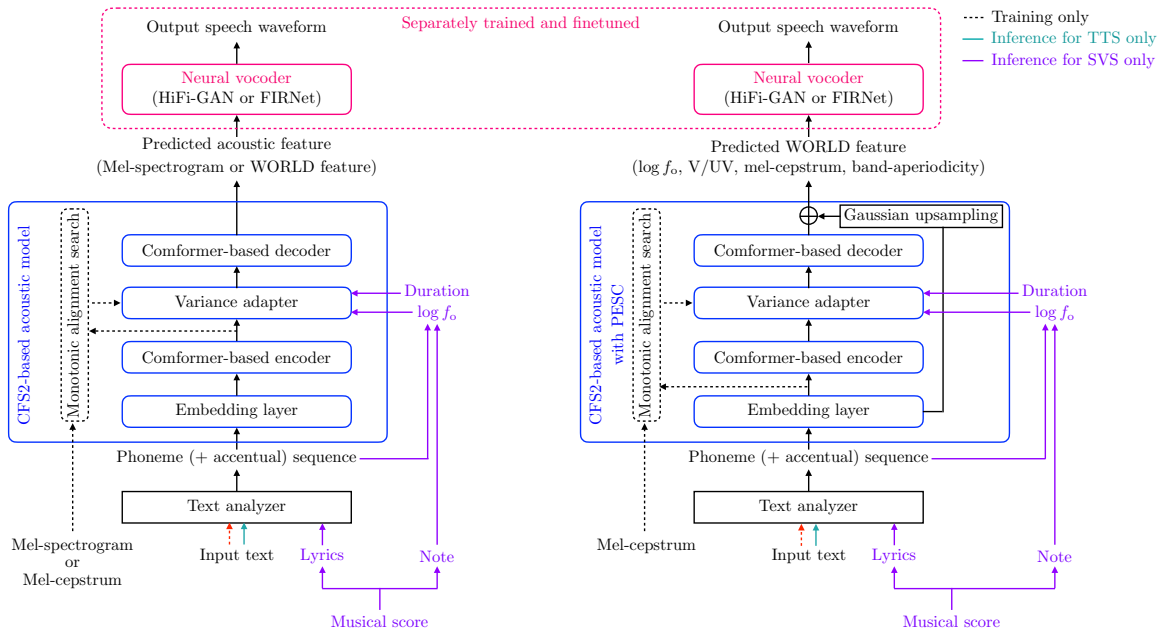
P5) It is difficult to implement the model because the detailed experimental conditions, such as the number of encoder layers, have not been described.

To tackle the challenging SVS task using only a TTS corpus and mitigate the above five problems of the previous method, this paper proposes a unified TTS and SVS framework. It is based on the duration-expansion-robust TTS acoustic model, which is based on Conformer-FastSpeech 2 (CFS2) [25], with phoneme embedding skip connection (PESC) and FIRNet source-filter neural vocoder [26]. It uses source-filter acoustic features instead of HiFi-GAN with mel-spectrograms to improve duration and f_o controllability. In the inference for SVS, the input text, f_o, and phoneme duration are obtained from lyrics and notes in a musical score. Additionally, input f_o shift

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

[2]https://soonbeomchoi.github.io/melody-unsupervised-blog/

10.21437/Interspeech.2024-2504



(a) CFS2 (b) PESC

Figure 2: *Unified TTS and SVS models using only a TTS corpus. (a) Conformer-FastSpeech-2-based (CFS2) acoustic model with monotonic alignment search (MAS). (b) CFS2-based acoustic model with MAS and phoneme embedding skip connection (PESC).*

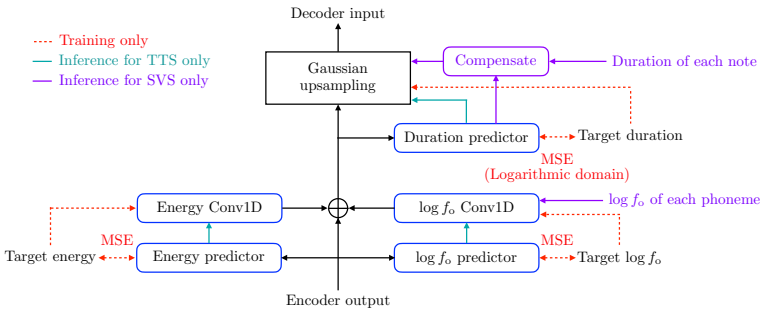


Figure 3: *Variance adapter for unified TTS and SVS models using only a TTS corpus.*

is proposed. The results of experiments using the JSUT corpus confirm that the PESC-based acoustic model using input f_o shift and FIRNet can improve the SVS quality compared with that using HiFi-GAN. The limitation is that the quality of the synthesized singing voices using only the TTS corpus is still lower than that of the original singing voices, and many problems remain. Some of the speech samples used in the experiments are available on the demo page[3].

2. Proposed approach

2.1. Baseline: CFS2 + HiFi-GAN with mel-spectrogram

To construct a unified TTS and SVS framework that can be trained using only a TTS corpus, a CFS2-based TTS acoustic model [25] is used, as illustrated in Fig. 2(a). CFS2 is a model that extends FastSpeech 2 [4] by using an encoder and a decoder [27] that are Conformer-based instead of Transformer-based. In contrast to the original FastSpeech 2, which predicts the log f_o and energy of each frame after the length regulator in the variance adapter, CFS2 predicts the log f_o and energy of each phoneme before the length regulator, in a similar manner to FastPitch [3] (Fig. 3). The latter is particularly important for SVS because it can prevent the synthesized f_o trajectory from becoming stair-stepped, as explained in P3). Monotonic alignment search (MAS) [6] is used with an alignment training framework [28] to obtain phoneme durations, in a similar manner to JETS [9], and Gaussian upsampling [2] is used as a length regulator. CFS2 with MAS can be trained with L1 loss to predict mel-spectrograms, variance loss, and forward-sum loss and bin loss for MAS. A HiFi-GAN neural vocoder is trained using only a TTS corpus and speech waveforms are synthesized from the predicted mel-spectrograms.

In the inference for SVS, the input text, f_o of each phoneme, and phoneme duration are obtained from lyrics and notes in a musical score. Figure 4 shows an example of the input sequence for SVS (the first phrase of JSUT-Song 003), where the tempo of the song is 120 beats per minute. The phoneme sequence is input to the embedding layer, as in TTS. The f_o of each phoneme is then obtained from each note. Additionally,

[3] https://ast-astrec.nict.go.jp/demo_samples/svs_using_only_tts_corpus/index.html

Phoneme	d	e	N	d	e	N	m	u	sh	i	m	u	sh	i	k	a	t	a	ts	u	m	u	r	i
Note	G4		G4	G4		E4	C4		C4		C4		D4		E4		E4		D4		C4		D4	
Number of frames	37.5		12.5	25		25	37.5		12.5		25		25		37.5		12.5		25		25		50	
Number of phonemes	2		1	2		1	2		2		2		2		2		2		2		2		2	

Figure 4: *Example of input sequence for SVS (JSUT-Song 003).*

the duration predicted by the duration predictor is modified according to the duration and number of phonemes for each note. For example, when the predicted durations of the phonemes "d" and "e" are 6 and 10 frames, respectively, the modified durations are $37.5\times6/(6+10)$ and $37.5\times10/(6+10)$, respectively. It should be noted that Gaussian upsampling can deal with float values of durations. The above procedure enables SVS to be performed by a model that was trained using only a TTS corpus.

2.2. CFS2 + HiFi-GAN / FIRNet with WORLD features

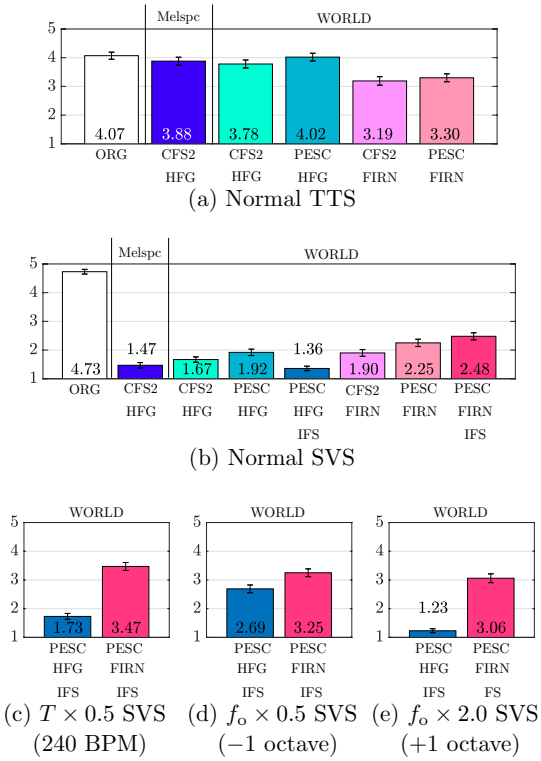
As shown in Fig. 1, the ranges of f_o and phoneme duration in SVS are wider than those in TTS. Therefore, it is necessary to extrapolate f_o and phoneme duration outside the ranges of the training data. Neural vocoders with an excitation signal input corresponding to f_o, such as PeriodNet [29], Harmonic-Net+ [30], SiFi-GAN [31], and FIRNet [26], can extrapolate the f_o and phoneme duration outside the ranges of the training data, as reported in [30]. To use these neural vocoders, the output acoustic features must be changed from mel-spectrograms to source-filter acoustic features constructed from f_o, voiced/unvoiced binary flag, mel-cepstrum, and band aperiodicity. These features can be analyzed by WORLD [32] and are known as WORLD features. As reported in [30, 33], HiFi-GAN conditioned on WORLD features without an excitation signal input can also be trained. According to the results of preliminary experiments for f_o controllability, in which FIRNet was compared with Harmonic-Net+ and SiFi-GAN, FIRNet is used as a neural vocoder with excitation signal input.

2.3. PESC + HiFi-GAN / FIRNet with WORLD features

To improve long-duration expansion, a PESC-based acoustic model is used, as shown in Fig. 2(b). PESC-based acoustic models are used to improve the robustness of duration expansion in TTS. In contrast to CFS2, MAS is calculated between phoneme embedding and acoustic features, and an additional PESC is used. This modification improves the synthesis quality in the presence of duration expansion. This study investigated the effectiveness of a PESC-based acoustic model for performing SVS using only a TTS corpus.

2.4. Proposed input f_o shift in PESC

Although f_o outside the range of the training data can be extrapolated by a FIRNet neural vocoder, f_o outside the range of the training data cannot be directly predicted by PESC. To mitigate this problem, the f_o obtained from a note is first shifted by a fixed value $-f_{\mathrm{o,shift}}$ and PESC predicts the shifted value $\log(f_\mathrm{o} - f_{\mathrm{o,shift}})$ within the range of the training data. Finally, the predicted value is shifted by $+f_{\mathrm{o,shift}}$, and a neural vocoder can synthesize the target f_o corresponding to the note. $f_{\mathrm{o,shift}}$ is the difference between the mean f_o in the training set and the mean f_o used in the synthesized songs.



(a) Normal TTS

(b) Normal SVS

(c) $T\times0.5$ SVS (240 BPM) (d) $f_\mathrm{o}\times0.5$ SVS (-1 octave) (e) $f_\mathrm{o}\times2.0$ SVS ($+1$ octave)

Figure 5: *Results of mean opinion score (MOS) tests with 20 listening subjects. The confidence level is 95%. ORG, Melspc, HFG, FIRN, and IFS represent original, mel-spectrogram, HiFi-GAN, FIRNet, and input f_o shift, respectively.*

2.5. Direct f_o input to neural vocoder

Instead of the f_o predicted by CFS2 or PESC, the f_o obtained from a note can be input to a neural vocoder. However, in this case, the synthesized f_o is stair-stepped because f_o is input in every frame and the same value is repeated. Therefore, although this method is available on the demo page, it was not used in the experiments.

3. Experiments

To evaluate the proposed unified TTS and SVS framework for the challenging task of performing SVS using only a TTS corpus, experiments were conducted for both the TTS and SVS tasks with a sampling frequency of 24 kHz. All the neural network models were implemented by modifying ESPnet2-TTS [8] on PyTorch 2.1.0 and were trained using an NVIDIA Tesla A100 GPU with 40 GB of memory.

3.1. Experimental conditions

Dataset: The experiments were conducted using the JSUT corpus (Basic5000) [21] for training the unified TTS and SVS mod-

Table 1: *Results of objective evaluations. The values in the columns for mel-cepstral distortion (MCD) and* $\log f_{\mathrm{o}}$ *root-mean-square error (RMSE) are the means and standard deviations. AM and NV represent acoustic model and neural vocoder, respectively.*

				TTS		SVS	
AM	NV	Input f_{o} shift	Acoustic feature	MCD [dB]	$\log f_{\mathrm{o}}$ RMSE	MCD [dB]	$\log f_{\mathrm{o}}$ RMSE
CFS2	HiFi-GAN		mel-spectrogram	$\mathbf{6.20 \pm 0.49}$	$\mathbf{0.22 \pm 0.06}$	11.4 ± 0.83	0.41 ± 0.14
CFS2	HiFi-GAN		WORLD	6.39 ± 0.51	$\mathbf{0.22 \pm 0.06}$	12.0 ± 1.08	0.41 ± 0.10
PESC	HiFi-GAN		WORLD	6.35 ± 0.51	$\mathbf{0.22 \pm 0.06}$	11.5 ± 1.17	0.37 ± 0.09
PESC	HiFi-GAN	$\surd$	WORLD	-	-	10.9 ± 0.92	0.57 ± 0.15
CFS2	FIRNet		WORLD	6.41 ± 0.50	0.23 ± 0.06	12.2 ± 1.02	0.35 ± 0.10
PESC	FIRNet		WORLD	6.39 ± 0.49	0.23 ± 0.06	11.6 ± 1.18	0.34 ± 0.09
PESC	FIRNet	$\surd$	WORLD	-	-	$\mathbf{10.5 \pm 1.00}$	$\mathbf{0.30 \pm 0.11}$

els. The JSUT-Song corpus was used only for evaluating the SVS model. Following the procedure established for ESPnet2-TTS [8], 4,500 utterances, 250 utterances, and 250 utterances were used for the training set, validation set, and test set for TTS, respectively. For TTS in Japanese, the G2P function based on pyopenjtalk and enhanced with prosody symbols [34] was used, following [8, 11]. For the subjective evaluations of TTS, 10 utterances were randomly selected. For the objective and subjective evaluations of SVS, the first phrases of 10 songs from JSUT-Song were used. The 10 musical scores used in the experiments were handcrafted by the first author.[4] In the experiments, the mean f_{o} calculated from the training set was approximately 206 Hz, and the minimum and maximum notes were C4 (261.6 Hz) and C5 (523.3 Hz). Therefore, for the input f_{o} shift, $f_{\mathrm{o,shift}}$ was calculated as $261.6 + (523.3 - 261.6)/2 - 206 = 186.5$ Hz.

Eighty-dimensional mel-spectrograms bandlimited to 7600 Hz were calculated and used as the acoustic features for HiFi-GAN. The WORLD features were 50-dimensional mel-cepstrum coefficients with warping coefficient $\alpha = 0.455$, three-dimensional band aperiodicity, and log-scaled continuous f_{o}. These features were extracted by cheaptrick [35], D4C [36], and Harvest [37] (based on WORLD [32]), respectively, following [30]. The STFT length and shift length were 1024 and 240 samples, respectively.

Model setting: For all the acoustic models, training and inference were conducted by modifying the FastSpeech 2 model with MAS implemented in ESPnet2-TTS [8] (`https://is.gd/9LFZtH`). The model configuration was based on that described in `https://is.gd/zcoB1Y`, where adim, eunits, dunits, and max_epoch were changed to 256, 1024, 1024, and 200, respectively. The configuration of HiFi-GAN was the same as that used in [30]. The configuration of FIRNet was the same as that used in [26] except for the frame shift of 10 ms. HiFi-GAN and FIRNet were fine-tuned using the acoustic features predicted by CFS2 and PESC. The iteration times of HiFi-GAN and FIRNet for fine-tuning were 200,000 and 100,000, respectively. The real-time factors on an AMD EPYC 7542 CPU (1 core) using PyTorch 2.1.0 of CFS2, PESC, HiFi-GAN, and FIRNet were 0.03, 0.03, 0.80, and 0.10, respectively. These results demonstrate that fast unified TTS and SVS can be performed by FIRNet combined with CFS2 or PESC.

Evaluation criteria: Mel-cepstral distortion (MCD) and $\log f_{\mathrm{o}}$ root-mean-square error (RMSE) were used as the objective evaluation criteria. The MCD and $\log f_{\mathrm{o}}$ RMSE were calculated by the ESPnet2-TTS toolkit [8, 9]. To evaluate the synthesized TTS and SVS speech subjectively, mean opinion score (MOS) tests [38] were conducted for (a) normal TTS, (b) normal SVS, (c) $T \times 0.5$ SVS (240 beats per minute), (d) $f_{\mathrm{o}} \times 0.5$ SVS (-1 octave), and (e) $f_{\mathrm{o}} \times 2.0$ SVS ($+1$ octave). In (d) and (e), f_{o} was controlled before the neural vocoders. Each subject evaluated 200 samples: 10 utterances $\times$ 20 models. The naturalness of each sample was rated on a five-point scale: (1) bad, (2) poor, (3) fair, (4) good, and (5) excellent. Twenty adult Japanese native speakers without hearing loss participated using headphones.

[4] These are also available on the demo page.

3.2. Results of experiments

Table 1 and Fig. 5 show the results of the objective and subjective experiments, respectively. HiFi-GAN-based models outperformed FIRNet-based models on the TTS task. However, PESC for duration-expansion-robust TTS, FIRNet for extrapolating f_{o} outside the range of the training data, and the proposed input f_{o} shift improved the SVS quality. PESC with FIRNet and input f_{o} shift significantly outperformed the other models. Additionally, PESC with FIRNet and input f_{o} shift outperformed PESC with HiFi-GAN and input f_{o} shift for $T \times 0.5$, $f_{\mathrm{o}} \times 0.5$, and $f_{\mathrm{o}} \times 2.0$ SVS. HiFi-GAN with PESC could not improve the synthesis quality because HiFi-GAN without an excitation signal input corresponding to f_{o} cannot synthesize f_{o} outside the range of the training data.

The experimental results outlined above validate the effectiveness of the proposed framework. Future work includes further improvement of the SVS quality; the quality of the synthesized singing voices is still lower than that of the original singing voices.

4. Conclusion

To tackle the challenging task of performing SVS using only a TTS corpus, this paper proposed a unified TTS and SVS framework. It is based on the FastSpeech-2-based duration-expansion-robust TTS acoustic model with PESC and FIRNet with source-filter acoustic features for improving duration and f_{o} controllability. In the inference for SVS, the input text, f_{o}, and phoneme duration are obtained from lyrics and notes in a musical score. Additionally, an input f_{o} shift was proposed. Experiments using the JSUT corpus for TTS confirmed that the PESC-based acoustic model using input f_{o} shift and FIRNet can improve the SVS quality compared with that using HiFi-GAN.

5. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Apr. 2018, pp. 4779–4783.

[2] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *Proc. ICLR*, May 2021.

[3] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*, June 2021, pp. 6573–6577.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, May 2021.

[5] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17 022–17 033.

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

[7] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, July 2021, pp. 5530–5540.

[8] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv:2110.07840*, 2021.

[9] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, Sept. 2022, pp. 21–25.

[10] H. Yamashita, T. Okamoto, R. Takashima, Y. Ohtani, T. Takiguchi, T. Toda, and H. Kawai, "Fast neural speech waveform generative models with fully-connected layer-based upsampling," *IEEE Access*, vol. 12, pp. 31409–31421, 2024.

[11] T. Okamoto, Y. Ohtani, T. Toda, and H. Kawai, "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, pp. 12456–12460.

[12] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. ICLR*, May 2024.

[13] T. Okamoto, Y. Ohtani, and H. Kawai, "Mobile PresenTra: NICT fast neural text-to-speech system on smartphones with incremental inference of MS-FC-HiFi-GAN for low-latency synthesis," in *Proc. Interspeech*, Sept. 2024.

[14] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "XiaoiceSing: A high-quality and integrated singing voice synthesis System," in *Proc. Interspeech*, Oct. 2020, pp. 1306–1310.

[15] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Sinsy: A deep neural network-based singing voice synthesis system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2803–2815, 2021.

[16] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "DiffSinger: Singing voice synthesis via shallow diffusion mechanism," in *Proc. AAAI*, Feb. 2022, pp. 11 020–11 028.

[17] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "VISinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *Proc. ICASSP*, May 2022, pp. 7237–7241.

[18] Y. Zhang, H. Xue, H. Li, L. Xie, T. Guo, R. Zhang, and C. Gong, "VISinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer," in *Proc. Interspeech*, Aug. 2023, pp. 4444–4448.

[19] Y. Lei, S. Yang, X. Wang, Q. Xie, J. Yao, L. Xie, and D. Su, "UniSyn: An end-to-end unified model for text-to-speech and singing voice synthesis," in *Proc. AAAI*, Feb. 2023, pp. 13 025–13 033.

[20] R. Yamamoto, R. Yoneyama, and T. Toda, "NNSVS: A neural network-based singing voice synthesis toolkit," in *Proc. ICASSP*, June 2023.

[21] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoust. Sci. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[22] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, J. Kreiman, M. Kob, A. Löfqvist, S. McCoy, D. G. Miller, H. Noé, R. C. Scherer, J. R. Smith, B. H. Story, J. G. Švec, S. Ternström, and J. Wolfe, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *J. Acoust. Soc. Am.*, vol. 137, no. 5, pp. 3005–3007, May 2015.

[23] S. Choi and J. Nam, "A melody-unsupervision model for singing voice synthesis," in *Proc. ICASSP*, May 2022, pp. 7242–7246.

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

[25] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent developments on ESPnet toolkit boosted by Conformer," in *Proc. ICASSP*, June 2021, pp. 5874–5878.

[26] Y. Ohtani, T. Okamoto, T. Toda, and H. Kawai, "FIRNet: Fundamental frequency controllable fast neural vocoder with trainable finite impulse response filter," in *Proc. ICASSP*, Apr. 2024, pp. 10871–10875.

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 5036–5040.

[28] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One TTS alignment to rule them all," in *Proc. ICASSP*, May 2022, pp. 6092–6096.

[29] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "PeriodNet: A non-autoregressive raw waveform generative model with a structure separating periodic and aperiodic components," *IEEE Access*, vol. 9, pp. 137 599–137 612, 2021.

[30] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, "Harmonic-Net: Fundamental frequency and speech rate controllable fast neural vocoder," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1902–1915, 2023.

[31] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-Filter HiFi-GAN: Fast and pitch controllable high-fidelity neural vocoder," in *Proc. ICASSP*, June 2023.

[32] M. Morise, F. Yokomori, and K. Ozawa, "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.

[33] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, "Comparison of real-time multi-speaker neural vocoders on CPUs," *Acoust. Sci. Tech.*, vol. 43, no. 2, pp. 121–124, Mar. 2022.

[34] K. Kurihara, N. Seiyama, and T. Kumano, "Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS," *IEICE trans. Inf. Syst.*, vol. E104-D, no. 2, pp. 302–311, Feb. 2021.

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

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

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

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