JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech

Dan Lim, Sunghee Jung, Eesung Kim

Kakao Enterprise Corporation, Seongnam, Republic of Korea

{satoshi.2020, ronda.jung, chris.ekim}@kakaoenterprise.com

Abstract

In neural text-to-speech (TTS), two-stage system or a cascade of separately learned models have shown synthesis quality close to human speech. For example, FastSpeech2 transforms an input text to a mel-spectrogram and then HiFi-GAN generates a raw waveform from a mel-spectogram where they are called an acoustic feature generator and a neural vocoder respectively. However, their training pipeline is somewhat cumbersome in that it requires a fine-tuning and an accurate speech-text alignment for optimal performance. In this work, we present end-to-end text-to-speech (E2E-TTS) model which has a simplified training pipeline and outperforms a cascade of separately learned models. Specifically, our proposed model is jointly trained FastSpeech2 and HiFi-GAN with an alignment module. Since there is no acoustic feature mismatch between training and inference, it does not requires fine-tuning. Furthermore, we remove dependency on an external speech-text alignment tool by adopting an alignment learning objective in our joint training framework. Experiments on LJSpeech corpus shows that the proposed model outperforms publicly available, state-of-the-art implementations of ESPNet2-TTS on subjective evaluation (MOS) and some objective evaluations.

Index Terms: end to end text to speech, joint training, espnet

1. Introduction

Text-to-speech (TTS) based on the neural network has significantly improved synthesized speech quality over the past years. Generally, a task of neural TTS is divided into more manageable sub-tasks using an acoustic feature generator and a neural vocoder. In this two-stage system, an acoustic feature generator generates an acoustic feature from an input text first and then a neural vocoder synthesizes a raw waveform from an acoustic feature. Those models are trained separately and then joined for inference. An acoustic feature generator can be autoregressive and attention-based for implicit speech-text alignments [1], [2] or it can be non-autoregressive for efficient parallel inference and duration informed for robustness on synthesis error [3], [4], [5]. There are lots of research on neural vocoder as well and some of the famous, widely used include [6], [7], normalizing flow based one [8] and generative adversarial network (GAN) based ones [9], [10], [11], [12].

Although the two-stage system is the dominant approach for TTS, training two models separately may result in degradation of synthesis quality due to an acoustic feature mismatch. Note that a neural vocoder takes the ground-truth acoustic features for training and the predicted ones from an acoustic feature generator for inference. For optimal performance, we can further train a pre-trained neural vocoder with predicted acoustic features, which is called fine-tuning [12], [13]. Or we can train a neural vocoder with predicted acoustic feature from the beginning [1]. However, both methods make the training pipeline

somewhat complicated in that the former requires additional training steps and the latter requires completion of training of an acoustic feature generator prior to vocoder training stage.

On the other hand, end-to-end text-to-speech (E2E-TTS) [5], [13], [14], [15], [16] is a recent research trend in which a speech waveform is directly generated from an input text in a single stage without distinction between an acoustic feature generator and a neural vocoder. Although there is no intermediate conversion to human-designed acoustic features such as mel-spectrogram, it has shown comparable performance to the two-stage TTS systems. Since E2E-TTS doesn't have a problem of an acoustic feature mismatch, it usually doesn't require fine-tuning or sequential training. Moreover, some works [13], [14] further simplify the training pipeline by incorporating an alignment learning module so that the model can be trained without dependency on external speech-text alignments tools.

In this work, we propose E2E-TTS with a simplified training pipeline and high-quality speech synthesis. Our work is similar to [17] in that joint training of an acoustic feature generator and a neural vocoder is researched and the experiments are based on the ESPNet2 toolkit. However, our proposed model directly synthesizes raw waveform from an input text without an intermediate mel-spectrogram. Moreover, we incorporate an alignment learning objective so that the proposed model can be trained in single-stage training without dependency on external alignments models. The contributions of our work can be summarized as follows.

- We make the E2E-TTS model by jointly training an acoustic feature generator and a neural vocoder, which are FastSpeech2 and HiFi-GAN respectively. It does not require pre-training or fine-tuning and it synthesizes high-quality speech without an intermediate mel-spectrogram.

- We leverage an alignment learning framework [18] to obtain token duration on the fly during the training. Thus the training of our proposed model does not require external speech-text alignments models.

- The proposed model outperforms state-of-the-art implementations of ESPNet2-TTS [17] on both subjective and objective evaluations.

2. Related work

There are several E2E-TTS research that directly generates speech waveform from an input text. For examples, FastSpeech2s [5] is similar to our work in that it uses FastSpeech2 and GAN-based vocoder; Parallel WaveGAN [10]. However, it requires an auxiliary mel-spectrogram decoder and a preparation of speech-text alignments to train the model. Although LiteTTS [19] also combines an acoustic feature generator with
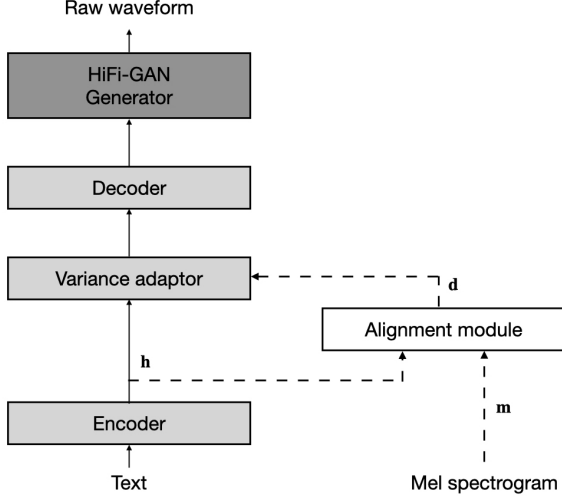
Figure 1: *An architecture of proposed model (discriminators are omitted for brevity)*

Figure 2: *Variance adaptor*

HiFi-GAN, it still depends on external alignments models and focuses more on lightweight structures for on-device uses.

On the other hand, EATS [14] integrates alignment learning into its adversarial training framework and it improves alignment learning stability by employing soft dynamic time warping to spectrogram prediction loss. VITS [13] also learns alignments during the training in the process of maximizing the likelihood of data and it improves expressiveness by utilizing variational inference and normalizing flow in an adversarial training framework. In EFTS-Wav [15], they adopt MelGAN and devise a novel monotonic alignment strategy with mel-spectrogram decoder for alignment learning. Wave-Tacotron [16] adopts an attention-based Tacotron [1] with the normalizing flow and it is optimized to simply maximize the likelihood of the training data.

In [17], joint training of an acoustic feature generator and a neural vocoder was conducted and it proved its effectiveness at solving the problem of acoustic feature mismatch by showing significant improvement compared to the separately learned model. However, the performance of the jointly trained model could not match that of a separately learned, fine-tuned model.

3. Model description

The proposed model is E2E-TTS which is jointly trained Fast-Speech2 and HiFi-GAN with an alignment module. In this section, we describe each component in order.

3.1. FastSpeech2

We adopt FastSpeech2 [5] as one of the components of the proposed model. It is a non-autoregressive acoustic feature generator with fast and high-quality speech synthesis. By explicitly modeling token duration with a duration predictor, it improves robustness on synthesis errors such as phoneme repeat and skips. Compared to its previous work; FastSpeech [3], it achieves significant improvement in speech quality by employing additional variance information which is pitch and energy. For our proposed model, We follow the structure of [5], which is a feed-forward Transformer-based [20] encoder, decoder, and
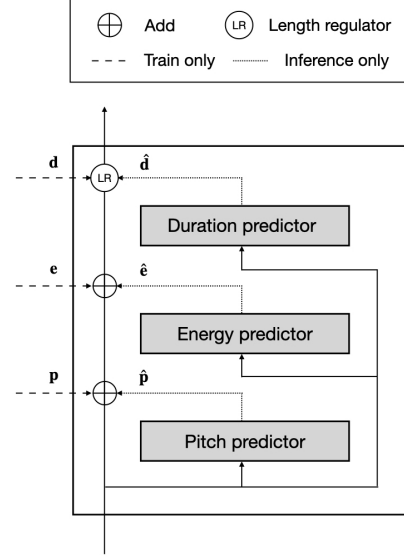
1D convolution-based variance adaptor. Figure 1 depicts each module in the proposed model. Specifically, the encoder encodes an input text as text embeddings $\mathbf{h}$, and the variance adaptor adds variance information to the text embeddings and expands according to each token duration for the decoder.

Figure 2 depicts the structure of the variance adaptor which consists of pitch, energy, and duration predictor. Pitch and energy predictors are trained to minimize token-wise pitch and energy respectively following the FastSpeech2 implementation of ESPNet2-TTS [17] or FastPitch [21] instead of frame-wise as in [5]. During training, required token-wise pitch and energy $\mathbf{p}, \mathbf{e}$ is computed on the fly by averaging frame-wise ground-truth pitch and energy according to token duration $\mathbf{d}$. The token duration is defined as the number of mel-frame assigned to each input text token and is obtained from the alignment module which will be explained later. After text embeddings are added with pitch and energy, it is expanded by a length regulator (LR) according to the token duration. We use gaussian upsampling with fixed temperature, also known as softmax-based aligner [14], instead of vanilla upsampling by repetition [3].

Note that although we adopt FastSpeech2 for our joint training, we exclude its mel-spectrogram loss so that the proposed model is trained to synthesize raw waveform directly from an input text without intermediate mel-spectrogram. Thus there remains a variance loss that minimizes each variance with L_2 loss.

$$L_{var} = ||\mathbf{d} - \hat{\mathbf{d}}||_2 + ||\mathbf{p} - \hat{\mathbf{p}}||_2 + ||\mathbf{e} - \hat{\mathbf{e}}||_2 \qquad (1)$$

where $\mathbf{d}, \mathbf{p}, \mathbf{e}$ are ground-truth duration, pitch and energy feature sequences respectively whereas $\hat{\mathbf{d}}, \hat{\mathbf{p}}, \hat{\mathbf{e}}$ are predicted ones from the model respectively.

3.2. HiFi-GAN

HiFi-GAN [11] is one of the most famous, GAN-based neural vocoders with fast and efficient parallel synthesis. In the GAN training framework, a model is trained by adversarial feedback where a generator is trained to fake a discriminator, and

a discriminator is trained to discriminate between the ground-truth sample and the predicted sample of the generator alternately. Discriminators of HiFi-GAN are designed to improve fidelity by considering a property of speech waveform, which are multi period discriminator (MPD) and multi scale discriminator (MSD). MPD handles diverse periodic patterns of speech waveform whereas MSD operates on the consecutive waveform at different scales with a wide receptive field.

As depicted in figure 1, we adopt the HiFi-GAN generator for synthesizing raw waveform from the output of the decoder. HiFi-GAN generator upsamples the output of the decoder through transposed convolution to match the length of the raw waveform where an output of the decoder has the same length as mel-spectrogram of the ground-truth waveform. It has not only adversarial loss but also auxiliary losses which are feature matching loss [9] and mel-spectrogram loss for the improvement of speech quality and training stability. Note that auxiliary mel-spectrogram loss here is L_1 loss between mel-spectrogram of synthesized waveform and that of the ground-truth waveform, which is devised and used for training HiFi-GAN [11]. The auxiliary mel-spectrogram loss is different from the mel-spectrogram loss of FastSpeech2 [5]. The training objective of HiFi-GAN follows LSGAN [22] and the generator loss consists of an adversarial loss and auxiliary losses as follows.

$$L_g = L_{g,adv} + \lambda_{fm}L_{fm} + \lambda_{mel}L_{mel} \quad (2)$$

where $L_{g,adv}$ is adversarial loss based on least-squares loss function and $\lambda_{fm}, \lambda_{mel}$ is scaling factor for auxiliary feature matching and mel-spectrogram loss respectively.

3.3. Alignment Learning Framework

Speech-text alignment is crucial in duration informed networks [3], [4], [5] where the TTS model has a separate duration model and requires explicit duration for its model training as in FastSpeech2. In our proposed model, each token duration $\mathbf{d}$ is used for training a duration predictor, for computing token-averaged pitch, energy from frame-wised ones, and for upsampling the text embeddings. The token duration can be obtained from a pre-trained autoregressive TTS model [2] as in [3] or from speech-text alignment tool such as montreal forced aligner (MFA) as in [4], [5]. Moreover, the training pipeline can be more simplified by incorporating alignment learning so that the required token duration is obtained during the model training on the fly [15], [18], [23], [24].

In this work, we incorporate an alignment learning framework [18] into our joint training framework for obtaining the required token duration $\mathbf{d}$ during the training on the fly. An alignment learning framework has shown an improved speech quality as well as fast alignment convergence by devising an alignment learning objective, which can be applied to both autoregressive and non-autoregressive TTS models. An alignment learning objective can be computed efficiently using a forward-sum algorithm. An alignment module in figure 1 represents the proposed module of an alignment learning framework [18], from which an alignment learning objective as well as each token duration are obtained.

Specifically, an alignment module encodes the text embeddings $\mathbf{h}$ and mel-spectrogram $\mathbf{m}$ as $\mathbf{h}^{enc}, \mathbf{m}^{enc}$ with 2 and 3 1D convolution layers respectively. After that, it computes soft alignment distribution $\mathcal{A}_{soft}$ which is softmax normalized across text domain based on the learned pairwise affinity between all text tokens and mel-frames.

$$D_{i,j} = dist_{L2}(\mathbf{h}_i^{enc}, \mathbf{m}_j^{enc}) \quad (3)$$

$$\mathcal{A}_{soft} = \texttt{softmax}(-D, dim = 0) \quad (4)$$

where $\mathbf{h}_i^{enc}, \mathbf{m}_j^{enc}$ is the encoded text embeddings and mel-spectrogram at timestep i, j respectively.

From soft alignment distribution $\mathcal{A}_{soft}$, we can compute the likelihood of all valid monotonic alignments which is the alignment learning objective to be maximized.

$$P(S(\mathbf{h})|\mathbf{m}) = \sum_{\mathbf{s} \in S(\mathbf{h})} \prod_{t=1}^{T} P(s_t|m_t) \quad (5)$$

where $\mathbf{s}$ is a specific alignment between a text and mel-spectrogram (e.g., $s_1 = h_1, s_2 = h_2, ..., s_T = h_N$), $S(\mathbf{h})$ is the set of all valid monotonic alignments and T, N is the length of mel-spectrogram and text token respectively. A forward-sum algorithm is used for computing the alignment learning objective and we define negative of it as forward sum loss $L_{forward\_sum}$. Notably it can be efficiently trained with off-the-shelf CTC [25] loss implementation.

To obtain token duration $\mathbf{d}$, the monotonic alignment search (MAS) [24] is used to convert soft alignment $\mathcal{A}_{soft}$ to monotonic, binarized hard alignment $\mathcal{A}_{hard}$ wherein $\sum_{j=1}^{T} \mathcal{A}_{hard,i,j}$ represents each token duration. Thus each token duration is the number of mel-frames assigned to each input text tokens and the sum of duration equals the length of mel-spectrogram. There are additional binarization loss L_{bin} which enforces $\mathcal{A}_{soft}$ matches $\mathcal{A}_{hard}$ by minimizing their KL-divergence. Note that we also apply beta-binomial alignment prior as in [18], [26] which multiplies 2d static prior to $\mathcal{A}_{soft}$ to accelerate the alignment learning by making the near-diagonal path more probable.

$$L_{bin} = -\mathcal{A}_{hard} \odot \log\mathcal{A}_{soft} \quad (6)$$

$$L_{align} = L_{forward\_sum} + L_{bin} \quad (7)$$

where $\odot$ is Hadamard product and L_{align} is final loss for alignments.

3.4. Final Loss

As depicted in figure 1, the proposed model consists of the encoder, variance adaptor, decoder, HiFi-GAN generator, and alignment module where the alignment module is used for training only. It is trained to directly synthesize raw waveform from an input text without intermediate mel-spectrogram loss in the GAN training framework. Note that we use discriminators of HiFi-GAN for the training of the proposed model though it is omitted from figure 1. Consequently, the loss of the proposed model is GAN training loss integrated with the variance loss and the alignment loss as follows.

$$L = L_g + \lambda_{var}L_{var} + \lambda_{align}L_{align} \quad (8)$$

where we used 1 for λ_{var} and 2 for λ_{align} as scaling factor of the variance and alignments loss respectively.

4. Experiments

For reproducible research, we conducted all experiments including data preparation, model training, and evaluation using ESPNet2-TTS [17] toolkit. The ESPNet2-TTS is a famous, open-sourced speech processing toolkit and it provides various recipes for reproducing state-of-the-art TTS results.

4.1. Dataset

We experimented with LJSpeech corpus [27] which is an English single female speaker dataset. It consists of 24 hours of speech recorded with a 22.05kHz sampling rate and 16bits. Following the recipe in `egs2/ljspeech/tts1` in the toolkit, we used 12,600 utterances for training, 250 for validation and 250 for evaluation.

Mel-spectrogram, which is used as an auxiliary loss and an input for an alignment module in the proposed model, was computed with 80 dimensions, 1024 fft size, and 256 hop size. For a fair comparison, g2p-en [1] without word separators was used as a G2P function, which is the same configuration as the baseline models of ESPNet2-TTS which will be explained later.

4.2. Model configuration

We implemented the proposed model using the ESPNet2-TTS toolkit following the configurations and training methods of `train_joint_conformer_fastspeech2_hifigan` in the same recipe of the toolkit used for data preparation. The differences are that the transformer was used for an encoder and decoder type instead of a conformer. And we used 256 for the attention dimension and 1024 for the number of the encoder and decoder ff units respectively. In the case of an alignment module, we simply followed the proposed structure in [18]. Note that generally a neural vocoder is trained to generate only part of the speech waveform from the corresponding portion of an input sequence for training efficiency. A related hyper-parameter in the toolkit is called segment size which determines the length of the randomly sliced output sequence of the decoder. We used 64 for this hyper-parameter.

For the comparative experiment, we prepared a conventional two-stage, cascaded TTS model as well as another E2E-TTS model. Specifically, we compared the proposed model with state-of-the-art implementations of ESPNet2-TTS. It provides the pre-trained models for public use including CF2 (+joint-ft), CF2 (+joint-tr), and VITS. CF2 (+joint-ft) is Conformer-based [28] FastSpeech2 with HiFi-GAN vocoder which are separately trained and jointly fine-tuned. CF2 (+joint-tr) is also Conformer-based FastSpeech2 with HiFi-GAN but it is jointly trained from scratch. VITS is E2E-TTS implementation of the paper [13].

4.3. Evaluation

We evaluated the performance of TTS models in objective and subjective metrics. For objective evaluations, mel-cepstral distortion (MCD), log-F_0 root mean square error (F_0 RMSE), and character error rate (CER) were computed using evaluation scripts provided by the ESPNet2-TTS toolkit. We computed CER using the same pre-trained ESPNet2-ASR model [2] which was used in [17]. For subjective evaluation, we conducted a crowdsourced Mean Opinion Score (MOS) test via Amazon Mechanical Turk where each participant, located in the United States, scored each audio sample from different models (including ground-truth audio sample) for naturalness on 5 point scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. Randomly selected 20 utterances from the evaluation set were used for the MOS test and each utterance was listened to by 20 different participants. Audio samples are available online [3]

---

[1] https://github.com/Kyubyong/g2p

[2] https://zenodo.org/record/4030677

[3] https://imdanboy.github.io/interspeech2022

Table 1: *Results on LJSpeech corpus, where "STD" represents standard deviation and "CI" represents 95% confidence intervals.*

Method	MCD ± STD	F_0 RMSE ± STD	CER	MOS ± CI
GT	N/A	N/A	1.0	4.08 ± 0.07
CF2 (+joint-ft)	**6.73 ± 0.62**	0.219 ± 0.034	1.5	3.96 ± 0.08
CF2 (+joint-tr)	6.80 ± 0.54	0.218 ± 0.035	1.5	3.93 ± 0.08
VITS	6.99 ± 0.63	0.234 ± 0.037	3.6	3.82 ± 0.09
Proposed model	7.16 ± 0.55	**0.215 ± 0.034**	**1.3**	**4.02 ± 0.07**

Table 1 shows the results on GT (ground-truth recordings), baseline models, and the proposed model. We obtained consistent outcomes as the previous work [17] in that the baseline models achieved high MOS values in the order of CF2 (+joint-ft), CF2 (+joint-tr), and VITS. Interestingly, our proposed model outperformed all of the baselines on MOS as well as objective metrics; F_0 RMSE, CER.

When it comes to the acoustic feature mismatch, the proposed model addressed the problem by the E2E approach which trains the model to generate raw waveform directly from an input text without an intermediate mel-spectrogram. Whereas CF2 (+joint-ft) and CF2 (+joint-tr) solved the problem by jointly fine-tuning and jointly training from scratch respectively. Thus we conjecture that the E2E approach was more effective for improvement than joint fine-tuning or simply joint training of an acoustic feature generator with a vocoder. Another difference compared to CF2 (+joint-ft) and CF2 (+joint-tr) was that the proposed model incorporates alignment learning in its joint training framework. It seems that those factors not only have simplified the training pipeline but also may improve the synthesized speech quality although we didn't investigate how they are related to the model performance thoroughly in this paper.

In the case of VITS, which is also an E2E model with alignment learning capability, it achieved the worst results in our experiment. One of the reasons other than the weakness on the g2p error as reported in [17], could be its training difficulty due to the somewhat complicated model structure compared to our proposed model. Note that VITS utilizes variational autoencoder and normalizing flow [13].

5. Conclusions

In this paper, we proposed an end-to-end text-to-speech model which is the jointly trained FastSpeech2 and HiFi-GAN with an alignment module. The proposed model directly generates speech waveform from an input text without intermediate conversion to an explicit human-designed acoustic feature. The training of the proposed model does not have fine-tuning which is required in the two-stage, separately learned text-to-speech models due to the problem of an acoustic feature mismatch. Moreover, we adopt an alignment learning framework so that the proposed model does not depend on external alignment tools for training. Consequently, the proposed model has a simplified training pipeline that is jointly trained in a single stage. For evaluations, we compared the proposed model with publicly available implementations of the ESPNet2-TTS toolkit on the English LJSpeech corpus, and the proposed model achieved state-of-the-art results. It would be interesting for future works to investigate other combinations of joint training other than FastSpeech2 and HiFi-GAN, or to evaluate on multi-speaker dataset.

6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for speech synthesis." in *INTERSPEECH*, 2020, pp. 2027–2031.

[5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[6] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[7] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[8] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[9] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[10] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[11] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[12] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Proc. Interspeech 2021*, 2021, pp. 2207–2211.

[13] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 5530–5540.

[14] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *International Conference on Learning Representations*, 2021.

[15] C. Miao, L. Shuang, Z. Liu, C. Minchuan, J. Ma, S. Wang, and J. Xiao, "Efficienttts: An efficient and high-quality text-to-speech architecture," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7700–7709.

[16] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5679–5683.

[17] 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 preprint arXiv:2110.07840*, 2021.

[18] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One tts alignment to rule them all," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6092–6096.

[19] H.-K. Nguyen, K. Jeong, S. Um, M.-J. Hwang, E. Song, and H.-G. Kang, "LiteTTS: A Lightweight Mel-Spectrogram-Free Text-to-Wave Synthesizer Based on Generative Adversarial Networks," in *Proc. Interspeech 2021*, 2021, pp. 3595–3599.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[21] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[22] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2813–2821.

[23] D. Lim, W. Jang, G. O, H. Park, B. Kim, and J. Yoon, "JDI-T: Jointly Trained Duration Informed Transformer for Text-To-Speech without Explicit Alignment," in *Proc. Interspeech 2020*, 2020, pp. 4004–4008.

[24] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 8067–8077.

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, 2006, p. 369–376.

[26] K. J. Shih, R. Valle, R. Badlani, A. Lancucki, W. Ping, and B. Catanzaro, "RAD-TTS: Parallel flow-based TTS with robust alignment learning and diverse synthesis," in *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

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

[28] 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 2020*, 2020, pp. 5036–5040.