INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Reinforce-Aligner: Reinforcement Alignment Search for Robust End-to-End Text-to-Speech

Hyunseung Chung[1], *Sang-Hoon Lee*[2], *Seong-Whan Lee*[1,2]

[1]Department of Artificial Intelligence, Korea University, Seoul, Korea
[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
{hs_chung, sh_lee, sw.lee}@korea.ac.kr

Abstract

Text-to-speech (TTS) synthesis is the process of producing synthesized speech from text or phoneme input. Traditional TTS models contain multiple processing steps and require external aligners, which provide attention alignments of phoneme-to-frame sequences. As the complexity increases and efficiency decreases with every additional step, there is expanding demand in modern synthesis pipelines for end-to-end TTS with efficient internal aligners. In this work, we propose an end-to-end text-to-waveform network with a novel reinforcement learning based duration search method. Our proposed generator is feed-forward and the aligner trains the agent to make optimal duration predictions by receiving active feedback from actions taken to maximize cumulative reward. We demonstrate accurate alignments of phoneme-to-frame sequence generated from trained agents enhance fidelity and naturalness of synthesized audio. Experimental results also show the superiority of our proposed model compared to other state-of-the-art TTS models with internal and external aligners.
Index Terms: text to speech, reinforcement learning

1. Introduction

Rapid progress in text-to-speech (TTS) was initiated by autoregressive models. These models replaced traditional methods [1–5] and are typically sequence-to-sequence in an encoder-decoder framework with attention mechanism [6–12]. The purpose of the encoder is to extract hidden representation feature vectors from phoneme sequence, and the decoder generates mel-spectrograms from produced vectors. Despite the advantages, end-to-end attention within autoregressive models have limitations such as slow inference speed, word skipping, and reading [11, 13]. As a means of remedy to this problem, non-autoregressive models are proposed for parallel generation of mel-spectrograms from text or phoneme [13–19]. Although the new architecture alleviates some of the drawbacks from autoregressive models, the duration aligner of non-autoregressive models still require guidance from external aligners. One of the main issues of having an external aligner is an increase in complexity of the training process. Properly aligned text and speech attention maps are required from an autoregressive model before training. This delays the training process and the non-autoregressive model becomes reliant on the quality of alignments generated by the external aligner. Therefore, recent synthesis pipelines are designed with the objective of robust end-to-end internal aligners [20–24].

Most recent works on TTS related to our work are EATS [21] and HiFi-GAN [25]. EATS is an end-to-end text to waveform network with an internal aligner that approximates phoneme to mel-spectrogram sequence alignments with Gaussian kernels. Although the model proposes robust text-to-wave synthesis, the alignments are not additionally trained to ensure improved duration alignment. Hifi-GAN generates high quality audio waveforms through multi-scale and multi-period discriminators, but the model is limited to considering only mel-spectrograms as input. Our proposed text-to-waveform network represents the best of both worlds and improvement on the drawbacks.

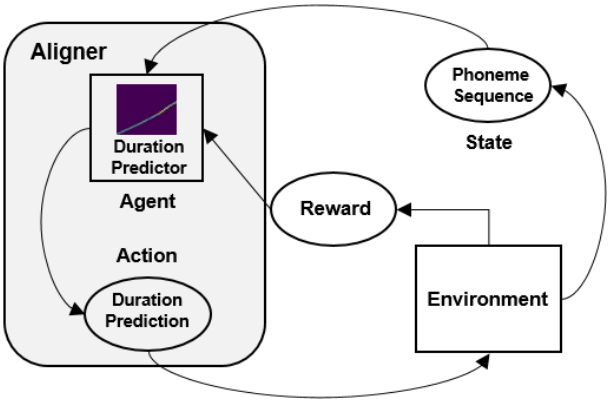


Figure 1: *Overall architecture of reinforce-aligner. The agent (duration predictor) interacts with the environment (text-to-waveform network) to receive current state (phoneme sequence) and reward feedback.*

In this paper we propose an end-to-end text-to-wave network with reinforce-aligner, which is a reinforcement learning based alignment search method for robust speech synthesis. Our agent interacts with the environment over a sequence of steps to select the best action given the current state. Then, the environment applies an update to the action and returns the reward of the action for the agent to take into account for the next step. This training process is repeated until convergence, and enables the network to internally learn its own alignment. Our experimental results show the positive impact of the reinforce-aligner on the duration alignment and quality of generated audio waveforms. The synthesized audio samples are provided on our online demo webpage.[1]

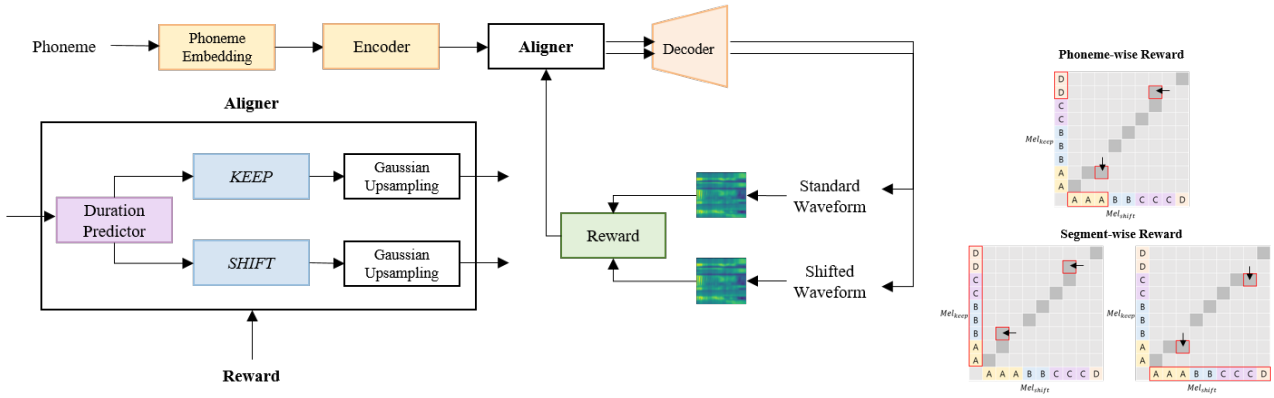
2. Model Architecture

Our proposed fully convolutional generator utilizes text or phoneme as input and generates a raw audio waveform as output. The encoder contains a modified multi-receptive field

[1]https://prml-lab-speech-team.github.io/demo/Reinforce-Aligner

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-831



(a) Generator training process (b) Reward setups

Figure 2: *Figure (a) is the overall architecture of the generator in training process. The aligner is shown in detail to describe the actions available for the agent. Figure (b) is the phoneme-wise and segment-wise rewards. Each arrow represents a shift that deviates towards the lower mel-spectrogram loss value.*

fusion module (MRF) [25]. The original implementation of MRF uses mel-spectrograms as input and upsamples the mel-spectrogram with transposed convolution to a raw waveform. Our implementation utilizes phoneme embeddings as input and outputs hidden representations without upsampling. The MRF based phoneme encoder contains multiple residual blocks with multiple kernel sizes and dilation rates, which is an essential component of our network because various receptive fields are able to extract distinct contextual features from phoneme embeddings. Then, the encoder output progresses through the reinforce-aligner to produce output frames. The output frames are randomly segmented by γ. Finally, the decoder upsamples γ by 256 to produce raw audio waveforms.

Our model contains two discriminators for different objectives during training. The first discriminator is the Multi-Scale Discriminator proposed in [26]. This discriminator effectively learns different frequency components of audio through variation in scales. Each of the three sub-discriminators contain convolution on different scales: raw audio, downsampled by factor of 2, and downsampled by factor of 4. The second discriminator is Multi-Period Discriminator [25], which captures the distinct features of audio with convolutions in periodic variations. Each of the different periodic values consider different periodic segments of the input audio.

3. Reinforce-Aligner

3.1. Reinforcement Learning Setup

The general architecture of the reinforce-aligner is shown in Figure 1. The duration predictor is the agent that interacts with the environment for each step of the training process. The environment is the overall text-to-waveform network. As shown in the generator training process of Figure 2(a), duration predictions are each upsampled to waveforms, and reward is calculated by mel-spectrogram losses of the respective mel-spectrograms. The reward feedback is given back to the agent for the final action selection.

3.2. Agent

The duration predictor returns one scalar value for each phoneme duration prediction. Predictor consists of two 1D convolutional layers, each with layer normalization, ReLU activation, and dropout following [13]. The last linear layer reshapes the convolutional output into a single scalar value. There are two actions available for the agent:

- *KEEP*: Keep phoneme duration prediction without any alterations to the prediction output.
- *SHIFT*: Shift phoneme duration prediction with shift value applied in alternating signs.

Shift applied to each phoneme duration consists of alternating signs, which is important to maintain the total sum of the duration prediction outputs. Additionally, we designed two types of shifts: Segments-wise and phoneme-wise shifts. Segment-wise shift corresponds to shift applied to the entire phoneme sequence segments. Phoneme-wise shift is shift applied to each phoneme duration of a phoneme sequence.

3.3. Environment

In this reinforcement learning based setup, the environment is the trained text-to-waveform network that outputs audio waves from phoneme inputs. There are two main objectives of the environment in the reinforce-aligner: (i) Provide input phoneme sequence information to the agent before it takes action, (ii) Give feedback to the agent after considering each of the two possible actions.

3.3.1. State

The environment produces phoneme embedding outputs for the agent in each training step. From phoneme sequence inputs, the multi-receptive field fusion based phoneme encoder generates phoneme embedding outputs through multiple residual blocks. These encoder outputs are the current state inputs for the agent to decide on an action.

3.3.2. Rewards

We have two different rewards depending on the type of shift. Both rewards consider the mel-spectrogram loss, which is

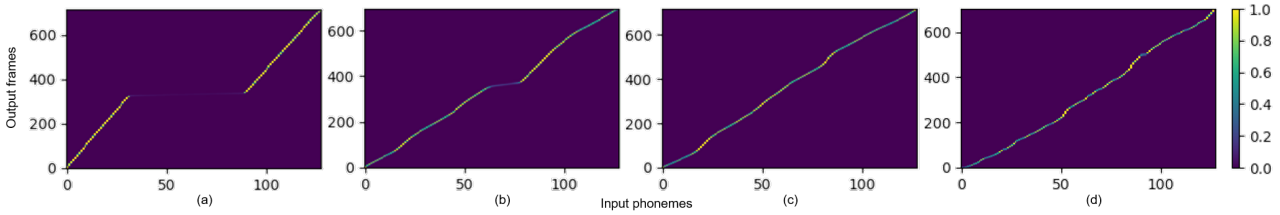


Figure 3: *Duration alignments between input phonemes and output frames in our proposed model. The three unsupervised models are compared to the last predicted alignment computed from supervision of target durations. (a) Without Soft-DTW, (b) With Soft-DTW, (c) Phoneme-wise ($\alpha = 2$) (d) With duration target*

the L1 loss between mel-spectrograms of ground-truth and generated waveforms. As shown in Figure 2(a), each mel-spectrograms are produced by the waveforms synthesized from predicted (*KEEP*) and shifted (*SHIFT*) durations. The segment-wise reward compares loss values of the entire wave segments used for training. Lower loss value implies higher similarity of the generated waveform to the ground-truth. The phoneme-wise reward considers phoneme-wise mel-spectrogram loss values. Specifically, the mel-spectrogram loss values are interpolated by downsampling into the shape of the phoneme duration sequences. Denote the phoneme duration sequences of predicted duration as $D_k = [d_1, d_2, d_3, ..., d_j]$ and shifted duration as $D_s = [d_1 + \alpha, d_2 - \alpha, d_3 + \alpha..., d_j \pm \alpha]$, where α represents the shift value. Then, our reward is formulated as:

$$r_j = \begin{cases} r_{kj} = 1 \text{ and } r_{sj} = 0, & \text{if } \mathcal{L}_{kj} \leq \mathcal{L}_{sj} \\ r_{kj} = 0 \text{ and } r_{sj} = 1, & \text{if } \mathcal{L}_{kj} > \mathcal{L}_{sj} \end{cases} \quad (1)$$

Here, $\mathcal{L}_k$ represents the L1 loss between predicted and ground-truth mel-spectrograms, and $\mathcal{L}_s$ is the L1 loss between shifted and ground-truth mel-spectrograms. r_k and r_s are the keep reward and shift reward values, respectively. Each j is a phoneme duration index value in the phoneme duration sequence for a total of N indices. For the segment-wise reward, keep reward values and shift reward values each have equal values $\forall j$. For the phoneme-wise reward, there are unique keep and shift reward values for each j.

3.4. Gaussian Upsampling

Predicted durations are scaled to the length of frame sequence outputs. Each scaled predictions are used to find cumulative sum of scaled token lengths and their center positions, as introduced in [21]. We first compute the weights:

$$w_t^i = \frac{\exp(\sigma^{-2}(t - c_i)^2)}{\sum_{j=1}^{i} \exp(\sigma^{-2}(t - c_j)^2)} \quad (2)$$

given fixed temperature parameter σ^2, scaled token center position c, and time step t. We finalize upsampling by producing weighted-sum between the encoder outputs and weights. The output features are used as decoder inputs and transposed convolutions upsample the features to a raw waveform.

3.5. Reinforced Duration Loss

We provide the agent with appropriate feedback by a loss that incorporates the rewards and duration prediction actions. For each phoneme sequence index j, the duration values are compared between the original duration prediction and the duration of the selected action. The loss is defined as:

$$\mathcal{L}_{re} = \sum_{j=1}^{N} (||D_{predj} - (D_{kj} \times r_{kj} + D_{sj} \times r_{sj})||_1) \quad (3)$$

Given N total tokens, each D_k, r_k pairs and D_s, r_s pairs represent duration and reward values for keep, shift actions, respectively. D_{pred} is equal to D_k because the keep action does not shift the predicted duration values. Therefore, the loss returns positive loss for shift action, and zero loss for keep action.

4. Auxiliary Loss

We utilize GAN loss [27] and reinforced duration loss. Additionally, auxiliary losses were used to support training of our text-to-wave network.

4.1. Total Duration Loss

The main purpose of the aligner is to produce accurate alignments of phoneme-to-frame sequences. However, the aligner does not have "correct" duration alignments to refer to during training, and therefore is not certain the duration outputs are accurate. Therefore, total duration loss is utilized as guidance for the model. We refer to the aligner length loss in [21]. Let m_{length} be the length of ground-truth mel-spectrogram, and l_j be the predicted length of jth token. The total duration loss is:

$$\mathcal{L}_{total} = (m_{length} - \sum_{j=1}^{i} l_j)^2 \quad (4)$$

4.2. Mel-Spectrogram Loss

In [25], mel-spectrogram loss is mentioned to be able to optimize the generator and improve quality of generated waveforms. Additionally, reward designs for the reinforce-aligner depends on the mel-spectrogram loss values to produce quality feedback for the agent in our model. The loss is formulated as:

$$\mathcal{L}_{mel} = \sum_{t=1}^{T} (||mel_{gt}[t] - mel_{pred}[t]||_1) \quad (5)$$

where mel_{gt}, mel_{pred} are mel-spectrograms of ground-truth and synthesized waveforms for T time steps.

Table 1: *MOS with 95% confidence intervals and duration prediction error results of the ablation study.*

Setting	MOS	Dur. Error
Ground truth	4.52 ± 0.06	-
w/ duration target	4.42 ± 0.04	1.03
w/o Soft-DTW	1.49 ± 0.08	2.85
w/ Soft-DTW	2.91 ± 0.06	2.53
Segment-wise ($\alpha = 1$)	3.48 ± 0.08	2.24
Segment-wise ($\alpha = 2$)	3.93 ± 0.07	1.84
Phoneme-wise ($\alpha = 1$)	3.78 ± 0.05	2.14
Phoneme-wise ($\alpha = 2$)	$\mathbf{4.06 \pm 0.06}$	**1.78**

4.3. Soft Dynamic Time Warping

We enable the mel-spectrograms to have room for error by iteratively finding an alignment path between ground-truth and synthesized spectrograms with dynamic time warping (DTW) [21, 28]. The main objective of this method is to alleviate the requirement that both spectrograms must be exactly aligned.

The total cost is defined as:

$$c_{total} = \sum_{l=1}^{L}(\delta_l \cdot \omega + \sum_{t=1}^{T}(||mel_{gt}[t] - mel_{pred}[t]||_1)) \quad (6)$$

where L is the mel-spectrogram length, and ω is the warp penalty that occurs for actions 2 and 3. δ is a binary indicator that is 1 when warp penalty is greater than zero. For our loss, we use the soft minimum from [29] to produce Soft-DTW.

5. Experimental Results and Analysis

5.1. Experimental Settings

For our experiments, we use the National Institute of Korean Language (NIKL) corpus [30] and the LJSpeech dataset [31]. The NIKL corpus contains about 45K audio samples of 50 native Korean speakers reading Korean text. The LJSpeech dataset contains 13,100 samples recorded by a single English female speaker. For both datasets we unified the sampling rates to 22,050 Hz. In NIKL corpus, we randomly split 10 audio samples for validation, 10 audio samples for testing, and the rest for training in all of the speakers. In LJSpeech dataset, we randomly divided 300 samples for validation, 300 samples for testing, and the rest for training. We conducted our experiments on 4 NVIDIA A100 GPUs.

5.2. Training Setup

All experiments are conducted with batch size of 64, using AdamW optimizer [32] with β_1=0.8, β_2=0.99, and weight decay of λ=0.01. Learning rate decay of 0.999 was used for each epoch, and the initial learning rate started with 0.0002. We used hidden representation of 256 , frame segment γ=128, fixed temperature parameter σ^2=10, and other hyper-parameters identical to HiFi-GAN V1, V2 [25] for all models to ensure fairness of experiments. The FFT, window size, and hop size were set to 1024, 1024, and 256. The english texts were converted to phoneme using the method of [33]. All models were trained for 350k steps.

5.3. Ablation Study

We conduct ablation study with NIKL corpus to determine the optimal reward for our model. We feed in speaker embedding alongside the encoder output to the reinforce-aligner for training in the multi-speaker dataset. All models are trained with the network proposed in this paper, with variations in each settings. "with duration target" setting represents experiments conducted with attention alignments extracted from Tacotron 2 [9]. We synthesize 100 utterances randomly sampled from NIKL corpus test dataset. Afterwards, 12 subjects rated the quality of synthesized speech with scores in the range of 1 (worst) to 5 (best). MOS score and duration error results are displayed in Table 1. "Phoneme-wise ($\alpha = 1$)" represents a phoneme-wise reward shifted by scalar value of 1 in alternating signs. In table 1, "Phoneme-wise ($\alpha = 2$)" showed the highest MOS score and lowest duration error. The duration error is the L1 loss between predicted duration and target duration extracted from attention alignments of Tacotron 2 . Additionally, Figure 3 visualizes the effectiveness of each settings with different alignment methods.

Table 2: *MOS with 95% confidence intervals, MCD_{13}, and $RMSE_{f0}$ results of proposed and state-of-the-art methods.*

Method	MOS	MCD_{13}	$RMSE_{f0}$
Ground Truth	4.32 ± 0.03	-	-
Vocoded [25]	4.10 ± 0.04	1.464	38.013
Tacotron2 [9]	4.02 ± 0.05	4.455	39.260
Glow-TTS [20]	3.93 ± 0.05	3.944	40.438
BVAE-TTS [37]	3.70 ± 0.06	4.002	41.686
Proposed (Ours)	$\mathbf{4.07 \pm 0.04}$	**2.887**	**38.988**

5.4. Comparison with Other Methods

We compare our model with other state-of-the-art methods developed for the task of TTS. In this comparison, we utilize the "Phoneme-wise ($\alpha = 2$)" setting, which represents the best MOS from our ablation study. Other methods which are not text-to-waveform networks require a vocoder to synthesize waveform from produced spectrograms. For the vocoder, HiFi-GAN [25] is used. We conducted a subjective 5 scale MOS test on Amazon Mechanical Turk [34]. At least 20 subjects rated naturalness of audio on a scale of 1 to 5. In Table 2, we display the MOS score followed by computed MCD_{13} [35] and $RMSE_{f0}$ [36] results. We synthesize 100, 200 utterances for subjective, objective evaluations, respectively. Our model has the highest MOS score, and lowest MCD_{13}, $RMSE_{f0}$ values.

6. Conclusion

We propose an end-to-end text-to-waveform network with a novel reinforcement learning based duration alignment search method. The advantage of this model is in the agent's ability to actively search for the optimal duration alignment through action based on reward feedback. Our proposed model was able to outperform other state-of-the-art methods with more accurate duration alignments and enhanced naturalness of synthesized audio.

7. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Department of Artificial Intelligence, Korea University), and the Netmarble AI Center.

8. References

[1] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Proceedings of the European Conference on Speech Communication and Technology*, 1999.

[3] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[4] H.-I. Suk and S.-W. Lee, "Subject and class specific frequency bands selection for multiclass motor imagery classification," *International Journal of Imaging Systems and Technology*, vol. 21, no. 2, pp. 123–130, 2011.

[5] M.-H. Lee, J. Williamson, D.-O. Won, S. Fazli, and S.-W. Lee, "A high performance spelling system based on eeg-eog signals with visual feedback," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 7, pp. 1443–1459, 2018.

[6] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 195–204.

[7] A. Gibiansky, S. Ö. Arik, G. F. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in Neural Information Processing Systems*, 2017.

[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of Interspeech*, 2017, pp. 4006–4010.

[9] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 4779–4783.

[10] 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*, 2019, pp. 6706–6713.

[11] N. Li, Y. Liu, Y. Wu, S. Liu, S. Zhao, and M. Liu, "Robutrans: A robust transformer-based text-to-speech model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 8228–8235.

[12] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021.

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

[14] D. Lim, W. Jang, H. Park, B. Kim, J. Yoon *et al.*, "Jdi-t: Jointly trained duration informed transformer for text-to-speech without explicit alignment," in *Proceedings of Interspeech*, 2020, pp. 4004–4008.

[15] J. Vainer and O. Dušek, "Speedyspeech: Efficient neural speech synthesis," in *Proceedings of Interspeech*, 2020, pp. 3575–3579.

[16] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021.

[17] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee, "Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[18] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021.

[19] R. Luo, X. Tan, R. Wang, T. Qin, J. Li, S. Zhao, E. Chen, and T.-Y. Liu, "Lightspeech: Lightweight and fast text to speech with neural architecture search," *arXiv preprint arXiv:2102.04040*, 2021.

[20] 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*, 2020.

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

[22] C. Miao, S. Liang, Z. Liu, M. Chen, J. Ma, S. Wang, and J. Xiao, "Efficienttts: An efficient and high-quality text-to-speech architecture," *arXiv preprint arXiv:2012.03500*, 2020.

[23] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling," *arXiv preprint arXiv:2010.04301*, 2020.

[24] P. Liu, Y. Cao, S. Liu, N. Hu, G. Li, C. Weng, and D. Su, "Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention," *arXiv preprint arXiv:2102.06431*, 2021.

[25] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, 2020.

[26] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019.

[27] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[28] H. Sakoe, "Dynamic-programming approach to continuous speech recognition," in *Proceedings of the International Congress of Acoustics*, 1971.

[29] M. Cuturi and M. Blondel, "Soft-dtw: A differentiable loss function for time-series," in *International Conference on Machine Learning*, 2017, pp. 894–903.

[30] N. I. of Korean Language, "Nikl corpus," https://www.korean.go.kr/, 2018.

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

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *Proceedings of International Conference on Learning Representations*, 2018.

[33] J. Park, Kyubyong & Kim, "g2pe," https://github.com/Kyubyong/g2p, 2019.

[34] "Amazon mechanical turk," https://www.mturk.com/, 2005.

[35] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993, pp. 125–128.

[36] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop*, 2017, pp. 712–718.

[37] Y. Lee, J. Shin, and K. Jung, "Bidirectional variational inference for non-autoregressive text-to-speech," in *International Conference on Learning Representations*, 2020.