Generative Adversarial Network based Neural Vocoder for Myanmar End-to-End Speech Synthesis

Aye Mya Hlaing, Win Pa Pa
Natural Language Processing Lab.,
University of Computer Studies, Yangon, Myanmar
{ayemyahlaing, winpapa}@ucsy.edu.mm

Abstract

Recently, neural vocoders have been employed in end-to-end speech synthesis for converting the intermediate spectral representations to the corresponding speech waveform. In this paper, two generative adversarial network (GAN) based vocoders, Parallel WaveGAN and HiFi-GAN are proposed for Myanmar end-to-end speech synthesis and subjective evaluations are conducted to compare the performance of the models. The subjective evaluation results show that both models trained on small Myanmar speech dataset achieve the high fidelity speech synthesis with fast inference speed, showing the ability of generalizing to the mel-spectrogram inversion of unseen speakers. Specifically, in end-to-end speech synthesis, Tacotron2 with HiFi-GAN vocoder achieves state-of-the-art performance resulting in a 4.37 mean opinion score (MOS) for Myanmar language.

1 Introduction

Text-to-speech (TTS) models focus on synthesizing intelligible and natural sounding speech which are indistinguishable from the original human speech. For the past few decades, statistical parametric speech synthesis (SPSS) has been the dominant technology for TTS (Tokuda et al., 2013; Qian et al., 2014; Wu et al., 2015; Zen and Sak, 2015). However, SPSS needs a complex pipeline for getting language dependent good linguistic features and that is time consuming and very expensive. This paper is a part of the ASEAN IVO 2023 project, "Spoof Detection for Automatic Speaker Verification", which aims to enhance the security and reliability of speaker verification by effectively detecting spoofing attacks.

In recent years, end-to-end neural TTS models, such as Tacotron (Wang et al., 2017), Tacotron2 (Shen et al., 2018), Transformer based TTS (Li et al., 2019), FastSpeech (Ren et al., 2019), FastSpeech2 (Ren et al., 2020), have emerged to simplify traditional speech synthesis pipeline and their synthesized speeches can be comparable with human recordings. The end-to-end neural TTS is typically composed of two main processing models, the spectral representation generator and the vocoder. The first one generates the spectral representation such as mel-spectrograms given the input text or phoneme, and the vocoder converts the speech waveforms from the generated mel-spectrograms. Griffin Lim algorithm (Griffin and Lim, 1984), the classic phase estimation method is generally used for speech waveform reconstruction.

Recently, in the context of end-to-end TTS synthesis, the separately trained neural vocoders such as WaveNet (Van Den Oord et al., 2016), Parallel WaveNet (Oord et al., 2018), MelGAN (Kumar et al., 2019), WaveGlow (Prenger et al., 2019), Parallel WaveGAN (Yamamoto et al., 2020) and HiFi-GAN (Kong et al., 2020) have demonstrated remarkable capabilities in generating natural-sounding synthetic speech. Inspired by this factor, in this work, the advantage of neural vocoder is combined into the Myanmar end-to-end speech synthesis to achieve both efficient and high-fidelity speech synthesis.

We trained two generative adversarial network based neural vocoders, Parallel WaveGAN and HiFi-GAN on Myanmar speech dataset because of their remarkable performance on generating waveform at fast inference speed while maintaining the quality of speech comparative to the other neural vocoders. To confirm the effectiveness of the vocoders, experiments were conducted by utilizing them in different conditions. We examined the ability of each vocoder in ground truth mel-spectrogram inversion, generalization on unseen speakers, and Myanmar end-to-end speech synthesis. Audio samples are available on this website[1].

[1] http://nlpresearch-ucsy.edu.mm/subeval-voc.html

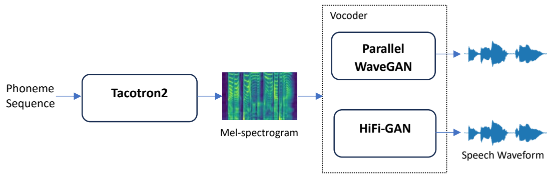

Figure 1: A proposed model of End-to-End TTS with GAN-based neural vocoders

2 Myanmar End-to-End Speech Synthesis

For Myanmar language, statistical parametric speech synthesis with different input linguistic features have been done on Myanmar speech synthesis. In HMM-based Myanmar TTS (Thu et al., 2015), CART-based Myanmar TTS (Hlaing and Pa, 2018), DNN-based Myanmar speech synthesis (Hlaing et al., 2018), LSTM-RNN-based Myanmar speech synthesis (Hlaing et al., 2019; Hlaing and Pa, 2020, Oo et al., 2020), we found that traditional speech synthesis pipeline and traditional vocoder such as WORLD vocoder (Morise et al., 2016) were used.

The first end-to-end Myanmar TTS System based on Tacotron was introduced in (Win et al., 2020) and Tacotron2 based end-to-end Myanmar TTS with phone-level embedding was found in (Qin et al., 2020). However, there is no research on the effectiveness of neural vocoder specifically trained on Myanmar speech dataset for Myanmar end-to-end TTS. To the best of our knowledge, this is the first effort to explore the advance of neural vocoder in Myanmar end-to-end TTS.

Figure 1 shows our proposed model of Myanmar end-to-end TTS with generative adversarial network based neural vocoders. In this work, a Tacotron2 (Shen et al., 2018) model was trained for the purpose of text to mel-spectrogram generation and the generated mel-spectrograms were given into our separately trained GAN-based vocoders, including Parallel WaveGAN and HiFi GAN as the input conditions. Tacotron2 uses character sequences as input, but our model was trained on phoneme sequences to alleviate the mispronunciation problems of rarely occurred words in the small training set.

3 GAN based Neural Vocoder

The first attempt of applying GAN (Goodfellow et al., 2014) to the synthesis of raw-waveform audio is WaveGAN (Donahue et al., 2018) and followed by many variants of GAN-based vocoders such as MelGAN (Kumar et al., 2019), StyleMelGAN (Mustafa et al., 2021), Multi-band MelGAN (Yang et al., 2021), Parallel WaveGAN (Yamamoto et al., 2020) and HiFi-GAN (Kong et al., 2020). GAN-based vocoders show significant performance over autoregressive models in the speed and quality of synthesized speech (AlBadawy et al., 2022). Among the different variants of GAN-based vocoders, we selected to train the vocoders using Parallel WaveGAN and HiFi-GAN for Myanmar end-to-end speech synthesis.

3.1 Parallel WaveGAN

The Parallel WaveGAN (Yamamoto et al., 2020) is a distillation-free, fast, and small-footprint waveform generation method using GAN. Though a WaveNet-based model conditioned on mel-spectrogram is used as the generator, the model is non-autoregressive at both training and inferencing. The generator is trained by jointly optimizing the multi-resolution short-time Fourier transform (STFT) auxiliary loss L_{aux} and the waveform-domain adversarial loss L_{adv}.

$$L_G = L_{aux}(G) + \lambda_{adv} L_{adv}(G, D) \quad (1)$$

where λ_{adv} represents the hyperparameter that balances the two loss terms.

Meanwhile, the discriminator is trained to correctly classify the generated sample as fake and simutaneously ground truth sample as real with the following equation:

$$L_D = \mathbb{E}_{x\sim p}[(1 - D(x))^2] + \mathbb{E}_z[D(G(z))^2] \quad (2)$$

where x denotes the target waveform, p denotes its distribution, and z denotes the input white noise.

3.2 HiFi-GAN

HiFi-GAN has been composed of one generator and two discriminators containing multi-scale discriminator (MSD) and multi-period discriminator (MPD) (Kong et al., 2020). The generator of HiFi-GAN is a fully convolutional neural network with multi-receptive field fusion (MRF) module that can perceives the various length of patterns in parallel. The final loss terms for the generator in HiFi GAN is as follows:

$$L_G = L_{Adv}(G; D) + \lambda_f L_F(G; D) + \lambda_m L_M(G) \quad (3)$$

where L_F and L_M are the feature matching loss and mel-spectrogram loss, respectively.

In the discriminator part, each sub-discriminator of MPD handles equally spaced samples of input audio and MSD was used to capture consecutive patterns and long-term dependencies. The discriminator with respect to the sub-discriminators of MPD and MSD is as follows:

$$L_D = \sum_{k=1}^{K} L_{Adv}(D_k; G) \tag{4}$$

where D_k represents k-th sub-discriminator in MPD and MSD.

4 Experiments

The dataset and the experimental setups of the models are presented in this section. The training of both GAN-based vocoders had been conducted on the open-source implementation from this site[2] and Tacotron2 model was implemented using ESPnet[3], an end-to-end speech processing toolkit. Each vocoder was trained on a single Nvidia Tesla K80 GPU and Tacotron2 model was trained on two Nvidia Tesla K80 GPUs.

4.1 Dataset

For training our proposed end-to-end pipeline including Tacotron2 model, Parallel WaveGAN and HiFi-GAN vocoders, we used a Myanmar phonetically balanced speech corpus (PBC) (Thu et al., 2015) built from Basic Travel Expression Corpus (BTEC) (Kikui et al., 2003) recorded by a female native speaker. In total, 3,800 utterances were utilized for training, 100 utterances each for validation and testing. The sampling rate of speech data was 16kHz.

4.2 Experimental setup of Parallel WaveGAN

For training the Parallel WaveGAN on Myanmar speech dataset, we used 80-band log-mel spectrograms with band-limited frequency range (80 to 7600 Hz) as the input auxiliary features for waveform generation models. The same configuration setting for the generator and the discriminator networks with the original paper (Yamamoto et al., 2020) was used in our work. Weight normalization was applied to all convolutional layers of both generator and discriminator. The hyperparameter λ_{adv} in Equation 1 was also set to 4.0. The model was trained for 200K steps and the discriminator was fixed for the first 100K steps, and then both the generator and the discriminator were trained jointly. We set the length of each audio clip to 25600 and mini-batch size to 6. The generator was set with the initial learning rate of 1×10^{-4} and the discriminator with the initial learning rate of 5×10^{-5}.

4.3 Experimental setup of HiFi-GAN

Among the variations of the generators in original source of HiFi-GAN(Kong et al., 2020), the configuration of HiFi-GAN V1 was applied to train the model on Myanmar speech dataset. We used 80-band log-mel spectrograms with band-limited frequency range (80 to 7600 Hz) as input conditions. The FFT and hop size were set to 1024 and 256, respectively. Adam (Kingma and Ba, 2014) optimizer with $\beta_1 = 0.5, \beta_2 = 0.9$ was used for training both the generator and the discriminator networks, and the initial learning rate was set to 2×10^{-4}. The batch size was 16 and the length of each audio clip was 8192. The model was trained for only 200K steps, the same steps used for training the Parallel WaveGAN model. This is very small compared to the training steps used in the original paper (2.5M steps).

4.4 Experimental setup of Tacotron2

Tacotron2 (Shen et al., 2018), a recurrent sequence-to-sequence feature prediction network with attention that maps phoneme embeddings to mel-spectrograms, was trained on the dataset mentioned in section 4.1 with a batch size of 32. The model was trained for 125K steps with Adam optimizer (Kingma and Ba, 2014) and a learning rate of 1×10^{-3}. In the training process, the guided attention loss was used to promote a fast and robust attention learning.

5 Results

To examine the performance of our trained Parallel WaveGAN and HiFi-GAN models, three mean opinion score (MOS) tests were performed for ground truth mel-spectrogram inversion, mel-spectrogram inversion for unseen speakers, and end-to-end Myanmar speech synthesis tasks. Ten native non-expert speakers participated in all MOS tests. Subjects were given the synthesized speeches of two models and ground truth audio, and they had to rate the quality of synthesized speeches on a scale of 1 to 5 where 1 is bad and 5 is excellent. The speech samples were randomly ordered.

[2] https://github.com/kan-bayashi/ParallelWaveGAN
[3] https://github.com/espnet/espnet

Model	MOS	RTF
Ground Truth	4.69 ± 0.10	-
Parallel WaveGAN	4.49 ± 0.12	0.015
HiFi-GAN	4.59 ± 0.11	0.011

Table 1: Comparison of MOS with 95% confidence intervals and inference speed (RTF) in ground truth mel-spectrogram inversion. Note that RTF is based on the average inference time of 100 utterances in evaluation set on a single Nvidia Tesla K80 GPU.

5.1 Ground Truth Mel-spectrogram Inversion

The MOS test and speed measurement with Real Time Factor (RTF) were performed to evaluate the performance of Parallel WaveGAN and HiFi-GAN models in terms of the quality of synthesized speeches and the inference speed. 10 utterances randomly selected from the testing set, were used for MOS test of mel-spectrogram inversion and the results are shown in Table 1. It can be clearly seen that both models can synthesize the high quality speech comparable to the ground truth speech. Remarkably, HiFi-GAN model achieves the highest MOS score 4.59 with a gap of 0.10 compared to the ground truth audio score 4.69 and this means that the synthesized speech is almost indistinguishable from the original speech. The RTF results indicate that both models get very small RTF values. Specifically, HiFi-GAN model gives the lowest RTF value (0.011) which implies that the model can synthesize speech 99.9 times faster than real-time on single Nvidia Tesla K80 GPU.

5.2 Generalization to Unseen Speakers

In this MOS test, 10 utterances of two unseen female speakers were utilized for investigating the ability of our trained models on generalizing to unseen speakers. However, we did not conduct an additional training for each model on multi-speaker dataset for this task. The MOS results for the mel-spectrogram inversion of the unseen speakers are shown in Table 2. The results show that Parallel WaveGAN and HiFi-GAN achieved 4.42 and 4.48 scores, respectively, indicating that both models can generalize well to unseen speakers.

5.3 End-to-End TTS

To verify the effectiveness of the Parallel WaveGAN and HiFi-GAN models in Myanmar end-to-end TTS pipeline, each model was integrated to the Tacotron2 model mentioned in section 4.4 as the vocoder. In the inferencing step, the Tacotron2 model convert the input phoneme sequences to the corresponding mel-spectrograms, and by inputting generated mel-spectrograms to vocoder models, they generate the corresponding speech waveform. To evaluate the quality of the generated speech samples, we conducted MOS test and the results are presented in Table 3. It can be observed that end-to-end TTS systems with independently trained neural vocoders can generate high quality synthesized speech. In particular, our model using Tacotron2 with Parallel WaveGAN vocoder achieves 4.33 MOS score which is comparable to the MOS results of the Parallel WaveGAN with the Transformer-based TTS (Yamamoto et al., 2020), and also the model using Tacotron2 with HiFi-GAN vocoder achieves 4.37 MOS score which is comparable to HiFi-GAN V1 model without fine-tuning (Kong et al., 2020) in the end-to-end TTS settings.

Model	MOS
Ground Truth	4.68 ± 0.12
Parallel WaveGAN	4.42 ± 0.12
HiFi-GAN	4.48 ± 0.11

Table 2: Comparison of MOS with 95% confidence intervals for generalizing on unseen speakers

Model	MOS
Ground Truth	4.68 ± 0.15
Tacotron2 + Parallel WaveGAN	4.33 ± 0.13
Tacotron2 + HiFi-GAN	4.37 ± 0.13

Table 3: Comparison of MOS with 95% confidence intervals in end-to-end Myanmar speech synthesis with neural vocoders

6 Conclusion

In conclusion, both Parallel WaveGAN and HiFi-GAN models achieve high-fidelity speech synthesis with fast inference speeds, showing the ability of generalizing to unseen speakers. By integrating these GAN-based models with Tacotron2 in the end-to-end TTS framework as the vocoders, we achieved the state-of-the-art speech quality for Myanmar language. Our work demonstrates that the GAN-based models, even trained on the small dataset with limited training steps, can achieve high quality speech for low-resource languages. Future work includes improving the mel-spectogram generator to better capture the prosody of speech and using GAN-based vocoders in various end-to-end speech synthesis settings.

Limitations

Due to the limited GPU resources, we can use the limited training steps on the models, however, more robustness of the models can be achieved by fine-tuning the hyperparemeters and training the models until an optimal point is reached. When the ability of vocoder is examined with the aim of generalizing to unseen speakers, one of the limitations is the unavailability of multi-speaker Myanmar dataset.

Acknowledgements

The ASEAN IVO (http://www.nict.go.jp/en/asean_ivo/index.html) project, "Spoof Detection for Automatic Speaker Verification", was involved in the production of the contents of this work and financially supported by NICT (http://www.nict.go.jp/en/index.html).

References

Ehab A AlBadawy, Andrew Gibiansky, Qing He, Jilong Wu, Ming-Ching Chang, and Siwei Lyu. 2022. Vocbench: A neural vocoder benchmark for speech synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 881–885. IEEE.

Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243.

Aye Mya Hlaing and Win Pa Pa. 2018. Word-based myanmar text-to-speech with clustergen. In *The 16th International Conference on Computer Applications (ICCA2018)*, pages 203–208.

Aye Mya Hlaing and Win Pa Pa. 2020. Word representations for neural network based myanmar text-to-speech s. *International Journal of Intelligent Engineering & Systems*, 13(2).

Aye Mya Hlaing, Win Pa Pa, and Ye Kyaw Thu. 2018. Dnn based myanmar speech synthesis. In *SLTU*, pages 142–146.

Aye Mya Hlaing, Win Pa Pa, and Ye Kyaw Thu. 2019. Enhancing myanmar speech synthesis with linguistic information and lstm-rnn. In *Proc. 10th ISCA Speech Synthesis Workshop*, pages 189–193.

Gen-ichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *INTERSPEECH*, pages 381–384.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884.

Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs. 2021. Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Yin May Oo, Theeraphol Wattanavekin, Chenfang Li, Pasindu De Silva, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson, and Alexander Gutkin. 2020. Burmese speech corpus, finite-state text normalization and pronunciation grammars with an application to text-to-speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6328–6339.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong. 2014. On the training aspects of deep neural network (dnn) for parametric tts synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE.

Qinglai Qin, Jian Yang, and Peiying Li. 2020. Myanmar text-to-speech synthesis using end-to-end model. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 6–11.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Ye Kyaw Thu, Win Pa Pa, Jinfu Ni, Yoshinori Shiga, Andrew M Finch, Chiori Hori, Hisashi Kawai, and Eiichiro Sumita. 2015. Hmm based myanmar text to speech system. In *INTERSPEECH*, pages 2237–2241.

Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. 2013. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Yuzana Win, Htoo Pyae Lwin, and Tomonari Masada. 2020. Myanmar text-to-speech system based on tacotron (end-to-end generative model). In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 572–577. IEEE.

Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King. 2015. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4460–4464. IEEE.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. 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)*, pages 6199–6203. IEEE.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2021. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 492–498. IEEE.

Heiga Zen and Haşim Sak. 2015. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474. IEEE.