



DelightfulTTS 2: End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders

Yanqing Liu^{1*}, Ruiqing Xue^{1*}, Lei He¹, Xu Tan², Sheng Zhao¹

¹Microsoft Azure Speech

²Microsoft Research Asia

{yanqliu, ruiqingxue, helei, xuta, szhao}@microsoft.com

Abstract

This paper describes DelightfulTTS 2, a new end-to-end architecture for speech synthesis jointly optimizing acoustic model and vocoder modules with phoneme/text and audio data pairs. Current TTS systems usually leverage a cascaded acoustic model and vocoder pipeline with mel-spectrograms as the intermediate representations, which suffer from two limitations: first, the acoustic model and vocoder are separately trained instead of jointly optimized, which incurs cascaded errors; second, the intermediate speech representations (e.g., mel-spectrogram) are predesigned and lose phase information, which are sub-optimal. To solve these problems, in this paper, we develop DelightfulTTS 2, a new end-to-end speech synthesis system with automatically learned speech representations and jointly optimized acoustic model and vocoder. Specifically, 1) We propose a new codec network based on vector-quantized auto-encoders with adversarial training (VQ-GAN) to extract intermediate frame-level speech representations (instead of traditional representations like mel-spectrograms) and reconstruct speech waveform. 2) We jointly optimize the acoustic model (based on DelightfulTTS) and the vocoder (the decoder of VQ-GAN), with an auxiliary loss on the acoustic model to predict intermediate speech representations. Experiments show that DelightfulTTS 2 achieves a CMOS gain +0.14 over DelightfulTTS, and more method analyses further verify the effectiveness of the developed system.

Index Terms: DelightfulTTS, End-to-End Training, Vector-Quantization, VQ-GAN

1. Introduction

Popular TTS models [1] are based on a typical two-stage system consisting of an acoustic model and a vocoder. The phonemes or linguistic features from input text are transformed into intermediate acoustic representations like mel-spectrograms by an acoustic model and the predicted acoustic representations are then converted to waveform with a vocoder. Although two-stage TTS systems [2, 3, 4] have been showing good quality in terms of prosody and audio fidelity in the past years, they still suffer several issues: 1) Mel-spectrograms are extracted by Fourier transformation where phase information are lost, which are not an optimal representation to cascade acoustic models and vocoders together [5]. 2) Since vocoders are trained with ground-truth mel-spectrograms while mel-spectrograms predicted by acoustic models are used in inference, the inaccurate predictions from acoustic models would cause training-inference mismatch and thus inferior audio quality.

*These authors contributed equally to this work. Corresponding author: Yanqing Liu, yanqliu@microsoft.com

To solve the issues in two-stage cascaded TTS systems, fully end-to-end TTS models [6, 3, 7, 8] are developed recently. Although ideally having advantages over two-stage systems, different end-to-end models face their own challenges and limitations: 1) no significant improvement over two-stage models in terms of voice quality, such as FastSpeech 2s [3], EATS [6], and WAVE-TACOTRON [8]; 2) slow inference speed due to autoregressive generation, such as WAVE-TACOTRON; 3) complicated training pipeline, such as VITS [7] and ClariNet [9]; 4) relying on Fourier transform for representation extraction, such as mel-spectrogram FastSpeech 2s [3] and linear spectrogram in VITS [7], while mel-spectrogram or linear spectrogram is already a compact representation of speech it still lacks key speech details [5].

In this paper, we develop DelightfulTTS 2, a new end-to-end speech synthesis system with automatically learned frame-level speech representations and jointly optimized acoustic model and vocoder. Specifically, we introduce the designs in DelightfulTTS 2 as follows.

- We propose a new codec network based on vector-quantized auto-encoders with adversarial training (VQ-GAN) to automatically learn intermediate frame-level speech representations, instead of using mel-spectrograms or other pre-designed features. Specifically, we use the encoder in VQ-GAN to extract speech representations, and quantize them with multi-stage vector quantizers, and then use the decoder to reconstruct waveform with adversarial training.
- After the VQ-GAN is trained, we jointly optimize the vocoder (i.e., the decoder of VQ-GAN) with an acoustic model based on DelightfulTTS [4], with an auxiliary loss on the acoustic model to predict the learned intermediate speech representations extracted by the encoder of VQ-GAN.

We conduct experiments on an internal English dataset. Experiment results show that DelightfulTTS 2 achieves +0.14 CMOS gain over the baseline DelightfulTTS and further objective evaluations also verify its effectiveness for TTS modelling.

2. DelightfulTTS 2

As shown in Figure 1, our proposed DelightfulTTS 2 consists of two components: (1) a codec network based on vector-quantized auto-encoders with adversarial training (VQ-GAN) that encodes raw waveform into frame-level feature embeddings with its encoder and quantizer and reconstructs waveform with encoded features with its decoder, and (2) an acoustic model based on DelightfulTTS [4] that predicts encoded features from phoneme sequence. A joint training strategy for acoustic model and codec is proposed for better voice quality.

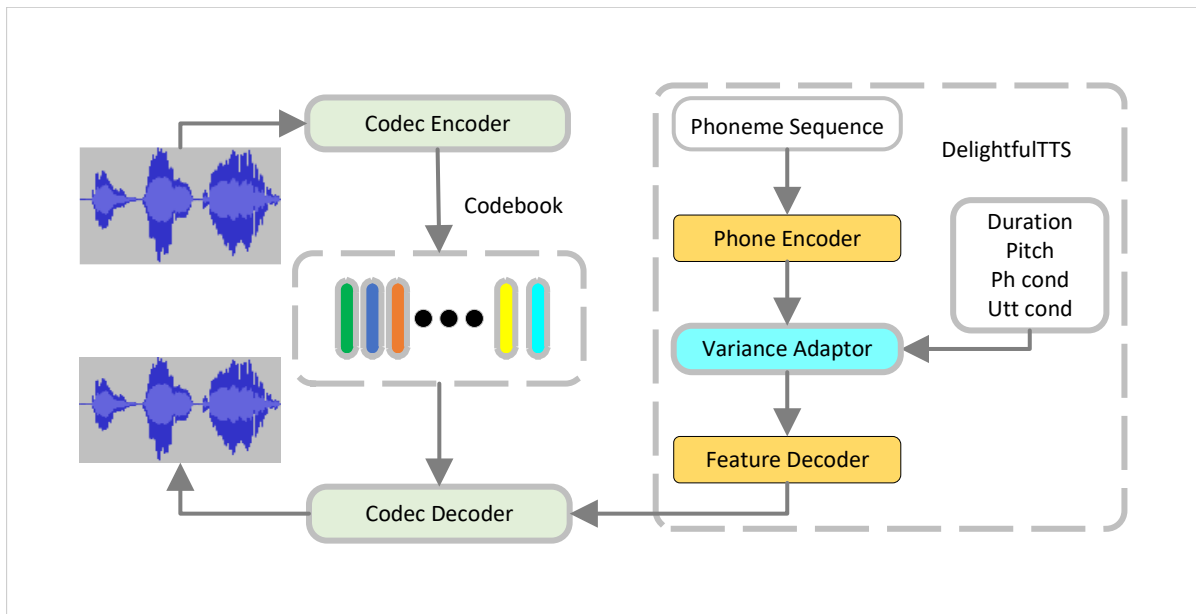


Figure 1: Overview of our proposed DelightfulTTS 2. The left is a codec network for speech representation learning and waveform reconstruction, whose encoder acts like a feature extractor and decoder acts like a vocoder. The right is an acoustic model based on DelightfulTTS that maps phoneme sequence into learnt speech representations, Ph cond is phoneme-level acoustic condition while Utt cond is utterance-level acoustic condition. The vocoder and acoustic model are jointly optimized.

2.1. Speech Representation Learning with VQ-GAN

To learn better speech representations instead of mel-spectrograms, we design a new codec network which learns frame-level speech representations based on vector-quantized auto-encoders with adversarial training (VQ-GAN). Specifically, it consists of a symmetric encoder-decoder network [10, 11, 12, 13] with skip-connections between bottom and top layers, and a multi-stage vector quantizer [14] in the middle as feature bottleneck. The overall structure of VQ-GAN based codec network is illustrated in Figure 2.

The decoder adopts the same architecture as the generator in HiFi-GAN [15], with a bidirectional Long Expressive Memory (LEM) [16] layer at upsampling stage for a stable learning of long-term sequential dependencies. The encoder mirrors the decoder in its layout. Skip-connections are added between the first three encoder blocks and its mirrored decoder blocks during training, which we found is critical to stabilize joint training and help model convergence. Multi-stage vector quantization [14] is applied on top of the codec encoder, to quantize each frame of encoded features in multiple stages.

For adversarial training, we combine the same multi-scale and multi-period discriminators as [17, 15]. Three discriminators with the same structure are applied to input audio at different resolutions: original, 2x down-sampled, and 4x down-sampled. Discrete wavelet transform is used to replace average sampling in the discriminators as [18] to reproduce high-frequency components accurately.

2.2. Acoustic Model based on DelightfulTTS

The acoustic model predicts the quantized speech representations with phoneme sequence as input. The network is based on DelightfulTTS, which consists of an encoder and a decoder with improved conformer [4] blocks, and a variance adaptor to provide multiple variance information to ease the one-to-

many mapping between text and speech. The encoder converts a phoneme sequence into its hidden representations, and the variance adaptor predict the variance information including utterance-level acoustic condition, phone-level acoustic conditions, and phoneme-level pitch and duration, and then the decoder predicts the frame-level speech representations with variance information and phoneme hidden as input.

2.3. Joint Training of Acoustic Model and Vocoder

Previous two-stage cascaded TTS systems consist of an acoustic model and a vocoder that are trained independently. Although these models can synthesize speech with good quality, they have a few drawbacks: Firstly, there is a feature mismatch between training and inference phase for the input of vocoder, i.e., ground-truth mel-spectrograms in training while predicted ones in inference. Secondly, pre-designed features like mel-spectrogram limit the performance of waveform reconstruction, which loses phase information and high-frequency details. In DelightfulTTS 2, we jointly train the acoustic model and vocoder in an end-to-end way to improve TTS performance, with a specifically designed scheduled sampling mechanism in acoustic model.

- The acoustic model [4] has four variance information modules: duration predictor, pitch predictor, utterance-level acoustic predictor and phone-level acoustic predictor. During training, the ground-truth pitch, utterance-level acoustic embedding, and phone-level acoustic embedding are extracted from ground-truth mel-spectrograms and added to phoneme hidden as decoder input to predict speech representations. This poses a mismatch between training and inference: predicted variance information in inference stage has gap compared to ground-truth counterparts. Instead of feeding all ground-truth features during training, we leverage a schedule sampling mechanism [19] for pitch, utterance-level acous-

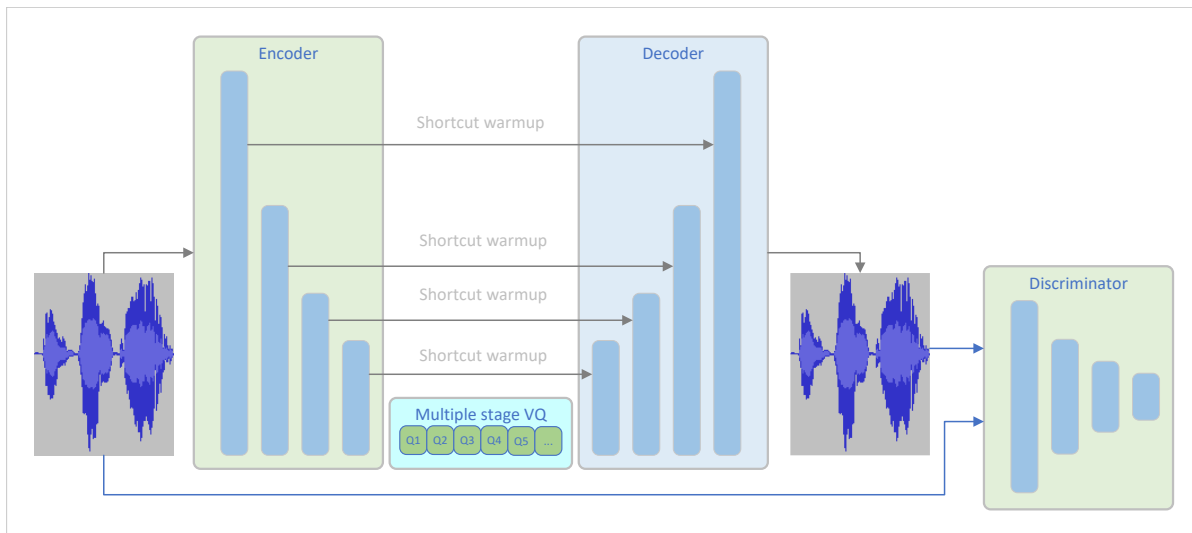


Figure 2: The structure of codec network based on VQ-GAN with multi-stage quantizer. The encoder and multi-stage quantizer convert audio into frame-level speech representations, and the decoder reconstructs audio from the quantized frame-level speech representations with adversarial training.

tic condition, and phone-level acoustic conditions, to reduce training and inference gap, and improve the performance of end-to-end model in test phase.

- During training phase, acoustic model output is directly used as vocoder input, with an auxiliary L1 loss between acoustic model output and quantized speech representations to stabilize training, and then a random segmentation process is applied on top of the predicted representations.

2.4. Training Objectives

Discriminator Loss The adversarial objectives for VQ-GAN and end-to-end training follow [15], with multi-period discriminator loss and multi-scale discriminator loss. To alleviate high-frequency loss, we further replace problematic average pooling method with discrete wavelet transform [20, 18] as the down sampling method, which is an effective way of down sampling non-stationary signals into several frequency sub-bands.

Codec Decoder Loss Codec decoder(vocoder) in end-to-end training has a multi-resolution spectrogram loss L_{mrs} , adversarial loss L_{Adv} and feature match loss L_{fm} [21, 17], which helps to generate realistic results when jointly optimizing with adversarial loss functions. L_{vq} is vector quantization loss [22] for all vector quantizers.

$$L_G = L_{Adv} + L_{vq} + L_{fm} + L_{mrs} \quad (1)$$

Acoustic Model Loss Following [4], we introduce phoneme-level pitch and duration loss, utterance-level, and phoneme-level acoustic condition loss, where L_{utt}/L_{phone} is the L1 loss between predicted utterance-level/phoneme-level acoustic condition vector and the vector extracted from utterance-level/phoneme-level reference encoder; L_{pitch}/L_{dur} is the L1 loss between predicted pitch/duration and the ground-truth pitch/duration. To improve audio fidelity, we use SSIM [23] to measure the similarity between predicted by the acoustic model and ground-truth quantized speech representations by the codec encoder, denoted as L_{ssim} . L_{feat} is the L1 loss between pre-

dicted speech representations and quantized speech representations.

$$L_{AM} = L_{pitch} + L_{dur} + L_{utt} + L_{phone} + L_{SSIM} + L_{feat} \quad (2)$$

The overall joint training loss function for DelightfulTTS 2 is a combination of acoustic model and audio codec decoder loss, W_G and W_{AM} are the loss weights of different components.

$$L_{joint} = W_G * L_G + W_{AM} * L_{AM} \quad (3)$$

3. Experiments and Results

In this section, we conduct experiments to evaluate the effectiveness of the proposed DelightfulTTS 2. We first describe the experimental setup and then introduce the results of DelightfulTTS 2. Audio samples are available at ¹.

3.1. Experimental Setup

Datasets We conduct experiments on an internal dataset with 40-hour professional English speech-script pairs. This dataset is divided into train set, dev set, and test set respectively. All text in our datasets are first processed by text normalization (TN) with a rule-based TN module and then converted into phonemes by a grapheme-to-phoneme module. The duration target is extracted by an internal force alignment model. The frame-level speech representation is obtained by downsampling audio samples by the codec encoder, which has a 12.5 ms hop size like mel-spectrogram [4]. The mel spectrogram used in our system is extracted from audio downsampled to 16 kHz and computed through a short time Fourier transform (STFT) using a 50 ms frame size, 12.5 ms frame hop. Frame-level pitch is extracted on 16KHz speech too and then averaged to phone-level pitch with phoneme alignment [4]. For codec training, the original 48 kHz audios were downsampled to 24 kHz.

Training Setting Our training process involves first training the codec network on audio samples only, followed by jointly

¹Audio samples: <https://cognitivespeech.github.io/delightfultts2>

training an acoustic model with the pretrained codec decoder (vocoder) model. To train the codec network, we apply the standard exponential moving average with a batch size of 16 on eight 32GB V100 GPUs, with a random audio segment of 24000 waveform points. The shortcut between codec encoder and decoder will be removed after the first 1k warm up steps to accelerate convergence. Adam optimizer [24] is used, and a learning rate of 0.0001 with exponentially decaying to 0.0001 starting from 100,00 iterations.

Evaluation Setting We keep an evaluation set randomly preserved from the training set and evaluate the audio quality with both subjective and objective metrics. Subjective metrics include mean opinion score (MOS) and comparative mean option score (CMOS). Audio generated on this set is sent to a human rating system where each sample is rated by at least 20 raters on a scale from 1 to 5 with 0.5-point increments, to better verify TTS system performance, we filtered the judges with same or better TTS score over recording. For each pair of utterances with a random order in CMOS, 20 raters are asked to give a score ranging from -3 (new system is much worse than baseline) to 3 (new system is much better than baseline) with intervals of 1. For objective metrics, we report results measured by means of ViSQOL [25] in ablation studies.

3.2. Results

Speech Quality Table 1 shows a comparison of our method against baselines. DelightfulTTS and FastSpeech 2 use phoneme sequence as input and mel-spectrogram as target, and [26] synthesizes waveform with mel-spectrogram. We find that the proposed system outperforms the baseline systems.

System	MOS
Ground Truth	4.39 ± 0.08
FastSpeech 2 [3]	4.08 ± 0.09
DelightfulTTS [4]	4.16 ± 0.09
DelightfulTTS 2	4.26 ± 0.09

Table 1: MOS evaluations of different systems.

We also conduct a side-by-side evaluation in Table 2 between audio synthesized by DelightfulTTS 2 and other systems, from which a comparison mean option score (CMOS) is calculated. The overall mean scores of +0.14 over DelightfulTTS and +0.13 over FastSpeech 2 show that raters have a statistically significant preference towards our system over other TTS systems; mean score of -0.06 over recording indicates that voice quality of DelightfulTTS 2 has slight regression compared to recording in general domain speaking of naturalness.

Baseline	CMOS
FastSpeech 2	+0.13
DelightfulTTS	+0.14
Recording	-0.06

Table 2: CMOS Comparison of DelightfulTTS 2 (new system) vs different TTS systems and recording.

Inference Latency Acoustic model has about 65.3M param-

eters and vocoder has about 3.3M parameters. We evaluate the inference latency of DelightfulTTS 2 on NVIDIA V100 GPU, the end-to-end RTF is about 0.008.

3.3. Analyses on Codec Network

To test the reconstruction performance of the proposed audio codec, both subjective and objective metrics are involved. For subjective test, a side-by-side CMOS evaluation with 20 raters between audio synthesized by audio codec (new system) and ground truth (baseline systems) is shown in Table 3. The overall mean score of -0.03 shows that the waveform reconstructed by our proposed codec network has no obvious significance with the ground-truth waveform.

System	CMOS
Codec Reconstruction vs Ground Truth(baseline)	-0.03

Table 3: CMOS for codec reconstruction.

For objective quality metrics, Table 4 shows the rate-quality curve of proposed codec network over a wide range of bitrates, from 3.2 kbps to 12.8 kbps. As measured by means of ViSQOL [25], we observe that the reconstructed speech quality by codec decoder decreases as the bitrate is reduced; on the other hand, the number of the intermediate speech frame also decreases as the bitrate is reduced, which may be easier for the acoustic model to predict and faster for runtime inference. Our proposed codec operates at constant bitrate 12.8 kbps, resulting in the same frame length as traditional mel-spectrogram, to balance efficiency of end-to-end TTS modelling and reconstructed speech quality. Given bitrate 1.1 kbps, Table 5 shows the ViSQOL for different frame-level speech representation dimension extracted by codec encoder, which indicates that larger dimension has slightly better quality.

Bitrate	3.2	6.4	12.8
ViSQOL	3.80	3.90	4.28

Table 4: Comparison of ViSQOL for different bitrate (kbps).

Dim	256	512	1024	2048
ViSQOL	3.39	3.41	3.51	3.55

Table 5: Comparison of ViSQOL for different frame-level speech representation dimension under 1.1 kbps.

4. Conclusions

This paper describes DelightfulTTS 2, an end-to-end TTS system that combines a convolutional codec network with adversarial vector-quantized auto-encoders and an acoustic model based on DelightfulTTS. DelightfulTTS 2 can be trained jointly from paired text/audio data, without suffering from cascaded errors in two-stage TTS models and sub-optimal pre-designed acoustic features like mel-spectrograms. The synthesized speech of DelightfulTTS 2 achieves better quality than baseline systems and is of similar quality to recordings.

5. References

- [1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.
- [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, 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 *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=piLPYqxtWuA>
- [4] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He, and S. Zhao, "Delightfultts: The Microsoft speech synthesis system for blizzard challenge 2021," *arXiv preprint arXiv:2110.12612*, 2021.
- [5] J. Cong, S. Yang, L. Xie, and D. Su, "Glow-wavegan: Learning speech representations from gan-based variational auto-encoder for high fidelity flow-based speech synthesis," *arXiv preprint arXiv:2106.10831*, 2021.
- [6] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," *arXiv preprint arXiv:2006.03575*, 2020.
- [7] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *arXiv preprint arXiv:2106.06103*, 2021.
- [8] 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.
- [9] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.
- [10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-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.
- [11] A. Mustafa, J. Bütche, S. Korse, K. Gupta, G. Fuchs, and N. Pia, "A streamwise gan vocoder for wideband speech coding at very low bit rate," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 66–70.
- [12] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [13] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "Seanet: A multi-modal speech enhancement network," *arXiv preprint arXiv:2009.02095*, 2020.
- [14] B.-H. Juang and A. Gray, "Multiple stage vector quantization for speech coding," in *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7. IEEE, 1982, pp. 597–600.
- [15] 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, 2020.
- [16] T. K. Rusch, S. Mishra, N. B. Erichson, and M. W. Mahoney, "Long expressive memory for sequence modeling," *arXiv preprint arXiv:2110.04744*, 2021.
- [17] 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," *arXiv preprint arXiv:1910.06711*, 2019.
- [18] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-gan: Adversarial frequency-consistent audio synthesis," *arXiv preprint arXiv:2106.02297*, 2021.
- [19] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based TTS," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 6274–6278.
- [20] G. P. Nason and B. W. Silverman, "The discrete wavelet transform in s," *Journal of Computational and Graphical statistics*, vol. 3, no. 2, pp. 163–191, 1994.
- [21] 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.
- [22] T. Srikotr and K. Mano, "Vector quantization of speech spectrum based on the vq-vae embedding space learning by gan technique," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2021.
- [23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [25] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.
- [26] Microsoft, "Azure neural TTS upgraded with hifinet, achieving higher audio fidelity and faster synthesis speed," p. 0, Nov 2020, hiFiNet. [Online]. Available: <https://techcommunity.microsoft.com/t5/azure-ai/azure-neural-tts-upgraded-with-hifinet-achieving-higher-audio/ba-p/1847860>