Glow-WaveGAN: Learning Speech Representations from GAN-based Variational Auto-Encoder For High Fidelity Flow-based Speech Synthesis

Jian Cong[1†], Shan Yang[2], Lei Xie[1], Dan Su[2]*

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2] Tencent AI Lab, China

npujcong@mail.nwpu.edu.cn, {shaanyang,dansu}@tencent.com
lxie@nwpu.edu.cn

Abstract

Current two-stage TTS framework typically integrates an acoustic model with a vocoder – the acoustic model predicts a low resolution intermediate representation such as Mel-spectrum while the vocoder generates waveform from the intermediate representation. Although the intermediate representation is served as a bridge, there still exists critical mismatch between the acoustic model and the vocoder as they are commonly separately learned and work on different distributions of representation, leading to inevitable artifacts in the synthesized speech. In this work, different from using pre-designed intermediate representation in most previous studies, we propose to use VAE combining with GAN to learn a latent representation directly from speech and then utilize a flow-based acoustic model to model the distribution of the latent representation from text. In this way, the mismatch problem is migrated as the two stages work on the same distribution. Results demonstrate that the flow-based acoustic model can exactly model the distribution of our learned speech representation and the proposed TTS framework, namely Glow-WaveGAN, can produce high fidelity speech outperforming the state-of-the-art GAN-based model.

Index Terms: Speech synthesis, Speech representations, Variational auto-encoder, Generative adversarial network

1. Introduction

With the continuous development of deep learning, text-to-speech (TTS) has made tremendous progress. Due to the huge difference in temporal resolution between text and audio samples, most approaches divide the TTS procedure into two stages, namely acoustic model and vocoder. The acoustic model first generates frame-level intermediate representations, such as Mel-spectrum given character or phoneme sequences, while the vocoder produces audio samples conditioned on the intermediate representation.

Attention-based auto-regressive (AR) model, such as Tacotron [1] and Tacotron2 [2], is a typical framework of acoustic modeling, which takes character or phoneme sequences as input and generates Mel-spectrum frame by frame. The AR models have achieved reasonably naturalness with human-parity speech in closed domains. The other family is the non-AR acoustic model that can synthesize speech much faster in a parallel manner. The key to design a non-AR acoustic model is alignment prediction. FastSpeech [3] was proposed to learn alignment from pre-trained AR teacher model, while Fast-Speech2 [4] bypassed the requirement of knowledge from the

teacher model through an external force aligner to extract duration label. Glow-TTS [5] was a flow-based generative model for parallel generation, which obtains the alignment by the properties of flows and dynamic programming.

For the vocoder that generates audio samples from intermediate representation, most high-fidelity neural vocoders are auto-regressive [6, 7, 8], which produce outputs relying on the previous sample. However, because of the AR structure, it's prohibitively slow in synthesizing high temporal resolution audio. The neural vocoders [9, 10, 11, 12] based on GAN (generative adversarial networks) are proposed to address this problem. They all rely on an adversarial game of two networks, a generator network that produces samples conditioned with the intermediate speech representation and a discriminator or several discriminators to differentiate between real and generated samples. And to stabilize the training process, many studies [9, 10, 12] adopted auxiliary loss, i.e. multi-resolution STFT loss and feature matching loss.

Despite the recent success on acoustic models and neural vocoders, there still exists critical mismatch between the two-stages, which may cause artifacts in the synthesized speech during inference. We believe a decisive reason is that the generated acoustic features from text lie in a different distribution with those extracted directly from speech. Note that the conventional acoustic models often use mean square error (MSE) to estimate model parameters, which only considers the numerical differences and may lead to over-smoothing problems [13]. To overcome the above issue, generative adversarial networks are utilized to model the distribution of target speech [13, 14]. Another intuitive way to alleviate the mismatch is to fine-tune the neural vocoder with the generated speech representation, which is widely used in the practical sequence-to-sequence based acoustic models [2, 15]. But in the auto-regressive manner, the ground-truth alignment mode is still different from the inference mode. Integrating acoustic model and vocoder into an end-to-end model is another possible solution [4, 15, 16, 17], but this kind of models are notoriously difficult to train and extra Mel-spectrum based loss has to be adopted to stabilize the convergence process.

In our work, we propose Glow-WaveGAN, which utilizes a flow-based acoustic model to learn the distribution of hidden representations from a variational auto-encoder (VAE) [18]. Different from all the existing methods, we do not directly use any predesigned acoustic features such as Mel-spectrum. This is because the estimated approximations of the Mel distribution from the acoustic model are still different from those inputs to a neural vocoder, which may cause artifacts i.e., metallic effects in synthesized audios. To make sure the target distribu-

---

tion of the acoustic model is exactly from the modeled distribution from the vocoder, we propose to learn a hidden representation directly from speech through variational auto-encoder (VAE) [18], where we further apply the generative adversarial network (GAN) [19] on the outputs to improve the quality of reconstructed speech.

Specifically, our proposed Glow-WaveGAN consists of a WaveGAN and a Flow-based acoustic model. The proposed WaveGAN utilizes GAN-based variational auto-encoder to learn speech latent representation. The encoder in VAE acts as an extractor to extract speech representation, while the decoder works as a vocoder to reconstruct waveform from the latent representation. With the extracted latent representation from speech via the well-trained WaveGAN, we then employ a flow-based acoustic model to learn the distribution of the latent representation from text. Experiments show that our proposed approach can produce high-fidelity speech, obviously outperforming the popular GAN-based models in audio quality.

2. Proposed method

Fig. 1 illustrates the proposed Glow-WaveGAN for high-fidelity speech synthesis. The architecture consists of a WaveGAN module and a Glow-TTS [5] module. The WaveGAN module is intended to learn latent speech representation by conducting waveform reconstruction, while the Glow-TTS model is designed to map input text to speech representation learned by the WaveGAN module. To learn a robust distribution of the latent speech representation, we utilize a variational auto-encoder (VAE) [18] to conduct feature extraction and reconstruction. To improve the quality of reconstructed speech, we further introduce pitch prediction as an auxiliary task with the latent representation as input and then employ the generative adversarial network [19] to model the distribution of target speech. Finally, we utilize the flow-based acoustic model to learn the distribution of the latent representation extracted from the encoder of the VAE. In this way, the main idea of our proposed Glow-WaveGAN is to make both the acoustic model and the waveform reconstruction module follow the same distribution, which successfully produces high quality natural speech compared to the state-of-the-art models.
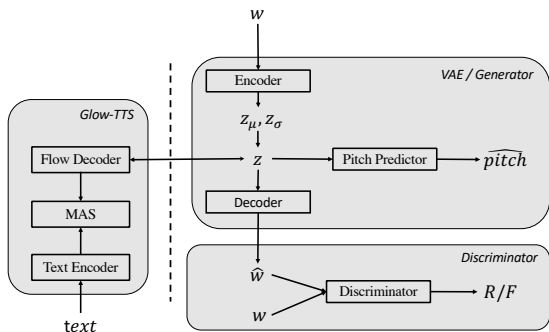

Figure 1: *Model architecture of proposed Glow-WaveGAN.*

2.1. VAE for learning representation and reconstruction

To relieve the mismatch between the acoustic model and the vocoder in conventional two-stage speech synthesis, we propose to force the acoustic model to directly learn the distribution of a latent feature from the vocoder. To achieve this goal, we first use VAE to learn speech representation in an unsupervised manner.

Conventional VAE typically contains an encoder and a decoder, where the encoder encodes the inputs w into hidden features z and the decoder tries to reconstruct w from z:

$$z = Enc(w) \sim q(z|w) \qquad (1)$$

$$\hat{w} = Dec(z) \sim p(w|z) \qquad (2)$$

where w means the input waveform. $q(z|w)$ is the distribution of hidden features, which is utilized to reconstruct waveform $\hat{w}$ that lies in $p(w|z)$.

Conventional VAE usually contains a reconstruction objective L_{recon} and a prior regularization term as

$$L_{vae} = L_{recon} + D_{KL}(q(z|w)||p(z)), \qquad (3)$$

where D_{KL} is the Kullback-Leibler divergence and $p(z)$ is the prior distribution. As for the proposed VAE for speech reconstruction, we adopt multi-resolution STFT loss [10, 20, 21] to measure the distance between natural speech w and generated speech $\hat{w}$ in frequency domain. Carrying out the multiple STFT losses with different analysis parameters is proved to be effective to learn the time-frequency characteristics of speech [9].

The detailed architecture of the proposed VAE for learning speech representation is shown in Fig. 2. The encoder contains a stack of down-sampling convolution layers followed by a residual block to summarize the abstract information in speech. Since it is hard for the acoustic model to learn long-time dependency in sample-level, in the residual block, we adopt the dilated convolutions to increase the receptive field to model the long-range context [9]. After the encoding process, we treat the produced mean and variance as the statistics of the learned latent distribution $q(z|w) = \mathcal{N}(Z_\mu, Z_\sigma)$. The structure of the decoder is mirror symmetry of the encoder, except that we employ a stack of transposed convolution with residual blocks to up-sample the sampled z to speech $\hat{w}$.

Since pitch information is important for natural speech synthesis and z is treated as the target of the acoustic model, we further propose to enhance the latent representation z with pitch information. To achieve this, we introduce a pitch predictor to predict pitch information from the sampled z with a frame-level pitch reconstruction objective as:

$$L_{pitch} = \|f(z) - p\|_2 \qquad (4)$$

where p represents the extracted pitch in log-scale, $f(\cdot)$ is the predictor with two convolution layers followed by a linear output layer. The frame-shift in pitch extraction depends on the total down-sampling factor.

2.2. GAN for high-fidelity speech reconstruction

Since speech reconstructed by the above auto-encoder may exist audio artifacts, we further propose to use GAN to improve the quality of the reconstructed speech, in viewing of the recent success of GAN-based vocoders [9, 10, 22]. As shown in Fig. 1, we treat the above VAE with pitch predictor as the generator of GAN. Hence, the generator firstly encodes the waveform into hidden z, and then produces speech from sampled z of $q(z|w)$. The discriminator network is optimized to classify the real samples w and the generated samples $\hat{w}$. During synthesis, only the decoder part of the generator is used to reconstruct waveform.

As for the adversarial training, we utilize the least-squares GANs objectives [23] to stabilize the training. The GAN losses for generator and discriminator are defined as:

$$L_{adv\_g} = (D(G(w)) - 1)^2 \qquad (5)$$

$$L_{adv\_d} = (D(w) - 1)^2 + D(G(w))^2 \qquad (6)$$

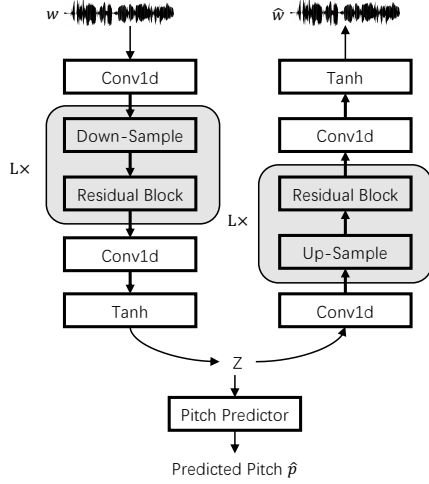where G represents both the encoding and decoding parts in the

Figure 2: *Model architecture of proposed auto-encoder for learning hidden features.*

VAE. D denotes multi-resolution spectrum discriminator [11] operating on Mel-spectrum with different analysis parameters.

Furthermore, we use the feature matching loss [9] as an additional loss to train the generator. The feature matching loss aims to minimize the L1 distance between the feature map extracted from intermediate layers in each discriminator, which is defined as:

$$L_{fm} = \sum_s \sum_i \frac{1}{N_i} \|D_s^i(w) - D_s^i(G(w))\|_1, \quad (7)$$

where $D_s^i(\cdot)$ is the output from the ith in s^{th} discriminator, and N_i denotes the number of units in each layer.

With the above VAE with pitch predictor and the GAN, we optimize the WaveGAN with the full objective

$$L = \lambda_1 L_{kl} + \lambda_2 L_{pitch} + \lambda_3 L_{recons} + \lambda_4 L_{adv\_g} + \lambda_5 L_{fm} \quad (8)$$

where the L_{kl} corresponds the Kullback-Leibler divergence in Eq.(3) and we empirically set λ_1=10, λ_2=1, λ_3=1, λ_4=1, λ_5=20. The discriminator is optimized with adversarial loss L_{adv_d}. The whole training procedure is summarized in Algorithm 1.

Note that the architecture only combining the decoder VAE and the discriminator can be treated as a GAN-based vocoder like [9, 22] when replacing the speech representation Z with Mel-spectrum. And we call the combination vocoder *Inner-GAN* for experimental comparison.

2.3. Flow based acoustic model

With the WaveGAN including VAE and GAN, we can obtain an extractor to extract latent representations and a speech reconstructor to synthesize waveform from the representations. Accordingly, acoustic model is designed to model the distribution of the latent represent z. To achieve this goal, we employ the flow-based acoustic model – Glow-TTS [5] to directly maximize the likelihood of z by applying invertible transformations.

Firstly, we extract the statistics Z_μ, Z_σ of the training data. Hence, the Glow-TTS models the conditional distribution of speech representation $P_Z(z|t)$ by transforming a conditional prior distribution $P_C(c|t)$ through the flow-based decoder $f_{dec} : c \rightarrow z$, where t denotes the text sequences. As for the prior distribution $P_C(c|t)$, we compute its statistics $C_\mu = C_{\mu_{1:M}}$ and $C_\sigma = C_{\sigma_{1:M}}$ from each token in the text $t = t_{1:M}$, where M is the sequence length.

Algorithm 1 Training the WaveGAN model
───
$\{\Theta_E, \Theta_G, \Theta_D, \Theta_{PN}\} \leftarrow$ initial encoder, decoder (generator) discriminator and pitch predictor network parameters
repeat
    $W \leftarrow$ random mini-batch audios from data-set
    $P \leftarrow$ pitch corresponding to W
    $\{Z_\mu, Z_\sigma\} \leftarrow E(W)$
    $Z \leftarrow$ sample from $\mathcal{N}(Z_\mu, Z_\sigma)$
    $L_{kl} \leftarrow KL(P_z \| \mathcal{N}(0, I))$
    $\hat{P} \leftarrow f(Z)$ // predict pitch using pitch predictor
    $\hat{W} \leftarrow G(Z)$
    $L_{pitch} \leftarrow \|\hat{P} - P\|_2$
    $L_{recons} \leftarrow$ multi-resolution STFT loss.
    $L_{adv\_g} \leftarrow (D(\hat{W}) - 1)^2$
    $L_{adv\_d} \leftarrow (D(W) - 1)^2 + D(\hat{W})^2$
    $L_{fm} \leftarrow$ feature matching loss with Eq. (7)
    // Update parameters according to gradients.
    $\Theta_E \xleftarrow{+} -\nabla_{\Theta_E}(L_{kl} + L_{pitch} + L_{fm} + L_{adv\_g} + L_{recons})$
    $\Theta_P \xleftarrow{+} -\nabla_{\Theta_P}(L_{pitch})$
    $\Theta_G \xleftarrow{+} -\nabla_{\Theta_G}(L_{fm} + L_{adv\_g} + L_{recons}))$
    $\Theta_D \xleftarrow{+} -\nabla_{\Theta_D}(L_{adv\_d})$
until done
───

We can calculate the log-likelihood of the Z using change of variables as follows:

$$\log P_Z(z|t) = \log P_C(c|t) + \log |det \frac{\partial f_{dec}^{-1}(z)}{\partial_z}| \quad (9)$$

where the prior distribution P_C follows an isotropic multivariate Gaussian distribution.

Since the sequence of target variable z and the input variable t are totally different, we need to align these two sequences to compute $\log P_C(c|t)$. Same as Glow-TTS [5], we also use the Monotonic Alignment Search (MAS) to map from z to each c, where $A(j) = i$ if $c_j \sim \mathcal{N}(c_j; C_{\mu_i}, C_{\sigma_i})$. So the log-likelihood of prior distribution is:

$$\log P_C(c|t; \Theta, A) = \sum_{j=1}^{N} \log \mathcal{N}(c_j; C_{\mu_{A(j)}}, C_{\sigma_{A(j)}}). \quad (10)$$

where N is the length of the speech representation Z from the above WaveGAN. The trainable variables in parameter set Θ and the alignment A are optimized iteratively [5].

With the Glow-TTS, we transform the explicit distribution $\mathcal{N}(Z_\mu, Z_\sigma)$ from WaveGAN into the prior conditional distribution $P(c|t)$ during training. At each training step, we sample the target z from $\mathcal{N}(Z_\mu, Z_\sigma)$ instead of using pre-sampled constant targets. At inference time, the flow-based acoustic model can produce $\hat{z}$ that follows the distribution $\mathcal{N}(Z_\mu, Z_\sigma)$ from texts, which is further fed into the decoder part of the WaveGAN to reconstruct high-fidelity synthesized speech.

3. Experiments and Results

3.1. Basic setups

To evaluate the proposed method, we conduct experiments on two English datasets. For the single speaker setting, we use LJSpeech [24], which consists of 13,100 audio clips in 22.5 kHz with a total duration of approximately 24 hours. We randomly reserve 500 samples as test set. For the multi-speaker setting, we use VCTK [25] corpus, which consists of approximately 44,200 audio clips uttered by 109 native English speakers with various accents. The total length of the audio clips is approximately 44 hours and the sample rate is 44kHz. We reduce the sample rate to 24kHz and randomly select nine speak-

ers for testing and exclude all their audio clips from the training set[1]. We evaluate our proposed approach on speech synthesis by training several models on the datasets described above:

1) The state of the art GAN-based vocoder *HiFi-GAN* [22].

2) The *Inner-GAN* vocoder which combines the VAE decoder and the discriminator of our proposed model mentioned in Section 2.1

3) Our proposed *Glow-WaveGAN*.

We use the publicly available implementation of HiFi-GAN vocoder [2]. The same original implementation of Glow-TTS[3] is adopted for all our experiments, and phonemes are chosen as input text tokens. For our proposed Glow-WaveGAN, the encoder down-samples audio to Z by a factor of 256 and the dimension of Z is 256. The down-sample factors and the corresponding channels are [2,4,4,8] and [128, 128, 256, 512]. The decoder takes 256 dimension Z as input and up-samples 256 factors to audio space, while the up-sample factors and channels are [8,4,4,2] and [256, 128, 128, 64]. For the discriminator, we use six different spectrum discriminators.

3.2. Evaluation on text-to-speech

We conduct the opinion score (MOS) tests in terms of naturalness and audio quality on the test set of LJSpeech and VCTK. The MOS scores are listed in Table 1. The results show whether reconstruction to waveform (copysyn) or text to speech, the proposed Glow-WaveGAN scores consistently higher than the other models. And the gap between the text to speech and the reconstruction to waveform of the proposed Glow-WaveGAN is only 0.08, which is significantly lower than that of HiFi-GAN (0.6). It implies that the speech representation with exact distribution has learned the true distribution of speech and the learned distribution is easier for the acoustic model to estimate. And the results of Inner-GAN further verifies that the improvement from the proposed Glow-WaveGAN compared with HiFi-GAN is not due to the difference between the model architecture or auxiliary losses. For the comparison between Inner-GAN and HiFi-GAN, the Inner-GAN scores lower than HiFi-GAN. We believe that the gap is because that the HiFi-GAN model adopts the multi-receptive field fusion module and takes more discriminators.

Table 1: *The MOS results of a multi-speaker TTS on VCTK and single-speaker TTS on LJSpeech. SR stands for speech representation (Mel and Z are extracted directly from speech, while $\hat{Mel}$ and $\hat{Z}$ are predicted from acoustic model).*

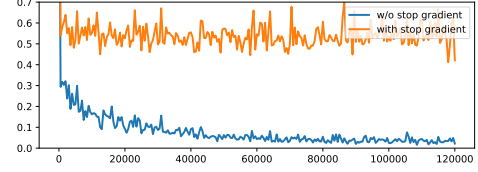| Model	SR	VCTK	LJ
Ground Truth		4.5	4.6
HiFi-GAN	Mel	3.8	3.86
Glow + HiFi-GAN	$\hat{Mel}$	3.21	3.4
Inner-GAN	Mel	3.78	3.84
Glow + Inner-GAN	$\hat{Mel}$	3.20	3.35
WaveGAN	Z	**4.13**	**4.15**
Glow-WaveGAN	$\hat{Z}$	**4.05**	**4.09**
Glow-WaveGAN (w/o pitch predictor)	$\hat{Z}$	3.87	3.93

Figure 3: *The pitch loss of our proposed model at the training procedure.*

The ablation without pitch prediction achieved substantially worse MOS, demonstrating the importance of adding pitch supervision in learning the speech representation Z. To further examine the effectiveness of the pitch predictor, we use the speech representation Z without pitch supervision to predict pitch. Specifically, we train Glow-WaveGAN and apply stop gradient operator to the input of the pitch predictor. The pitch loss curve is shown in Fig. 3. We can see that the pitch loss of the model with stop gradient even does not get converge, which demonstrates that there is no exact pitch information in the speech representation Z if we do not utilize the pitch predictor. As a result, the listeners point out many jitters in the speech synthesized by Glow-WaveGAN without the pitch predictor.

3.3. Generalization to unseen speakers

To measure the generalization ability of our model to unseen speakers, we use 100 randomly selected utterances from nine unseen speakers in the VCTK for MOS test. Note that for training of vocoder and WaveGAN, the utterances of unseen speakers are totally excluded from the training set, but for the training of Glow-TTS, only the test set for MOS test is excluded.

The results are shown in Table 2. We can see that our proposed model scores better than other models, and the score is even close to that of seen speakers, indicating that the proposed model generalizes well to unseen speakers. For the HiFi-GAN and Inner-GAN, due to the mismatch between the predicted and the ground truth Mel-spectrum, listeners can tell more artifacts like metallic sounds from the synthesized samples.

Table 2: *The MOS results for the unseen speaker in VCTK.*

Model	SR	MOS
Ground Truth		4.5
HiFi-GAN	Mel	3.76
Glow + HiFi-GAN	$\hat{Mel}$	3.14
Inner-GAN	Mel	3.70
Glow + Inner-GAN	$\hat{Mel}$	3.12
WaveGAN	Z	**4.08**
Glow-WaveGAN	$\hat{Z}$	**4.0**

4. Conclusions

In this work, we introduce Glow-WaveGAN, which can synthesize high fidelity speech from text, without using Mel-spectrum as the intermediate representation. Specifically, we propose WaveGAN which combines VAE and GAN to extract a latent speech representation by encoder and the representation can be reconstructed back to waveform by decoder. And to model the distribution of latent representation, we utilize Glow-TTS to estimate the distribution from text. Through this way, we migrate the problematic data distribution mismatch between acoustic model and vocoder in a typical two-stage TTS system. Results show that our proposed model outperforms the state-of-the-art GAN-based text-to-speech framework in audio quality.

---

[1]Samples can be found at `https://syang1993.github.io/glow_wavegan`

[2]`https://github.com/jik876/HiFi-GAN`

[3]`https://github.com/jaywalnut310/glow-tts`

5. References

[1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. of INTERSPEECH*, 2017, pp. 4006–4010.

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

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. of NeurIPS*, 2019, pp. 3165–3174.

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

[5] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Proc. of NeurIPS*, 2020.

[6] A. van den 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," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[7] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. of ICML*, 2018, pp. 2415–2424.

[8] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," in *Proc. of ICLR*, 2017.

[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," in *Proc. of NeurIPS*, 2019, pp. 14 881–14 892.

[10] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. of ICASSP*, 2020, pp. 6199–6203.

[11] W. Jang, D. Lim, and J. Yoon, "Universal melgan: A robust neural vocoder for high-fidelity waveform generation in multiple domains," *arXiv preprint arXiv:2011.09631*, 2020.

[12] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, 2021, pp. 492–498.

[13] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. of ICASSP*, 2017, pp. 4910–4914.

[14] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 685–691.

[15] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," *arXiv preprint arXiv:2006.03575*, 2020.

[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 *Proc. of ICASSP*, 2021, pp. 5679–5683.

[17] 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.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. of ICLR*, 2014.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. of NeurIPS*, 2014, pp. 2672–2680.

[20] M. Binkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *Proc. of ICLR*, 2020.

[21] R. Yamamoto, E. Song, and J. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," in *Proc. of INTERSPEECH*, 2019, pp. 699–703.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. of NeurIPS*, 2020.

[23] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. of ICCV*, 2017, pp. 2813–2821.

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

[25] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.