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

GAN Vocoder: Multi-Resolution Discriminator Is All You Need

Jaeseong You, Dalhyun Kim, Gyuhyeon Nam, Geumbyeol Hwang, Gyeongsu Chae

MoneyBrain Inc., Seoul, Korea
{jaeseongyou, torch, ngh3053, goldstar, gc}@moneybrain.ai

Abstract

Several of the latest GAN-based vocoders show remarkable achievements, outperforming autoregressive and flow-based competitors in both qualitative and quantitative measures while synthesizing orders of magnitude faster. In this work, we hypothesize that the common factor underlying their success is the multi-resolution discriminating framework, not the minute details in architecture, loss function, or training strategy. We experimentally test the hypothesis by evaluating six different generators paired with one shared multi-resolution discriminating framework. For all evaluative measures with respect to text-to-speech syntheses and for all perceptual metrics, their performances are not distinguishable from one another, which supports our hypothesis.

Index Terms: vocoder, text-to-speech, GAN

1. Introduction

Neural vocoder (i.e. neural network models that transform from speech features to raw waveform) has been one of the most active fields of study in the domain of speech synthesis. Modeling raw audio is essential in vocoding, and it poses many difficulties: (1) a digital speech signal tends to be a sequence of extended length, which brings in complications of computation and memory cost, (2) a speech signal is extremely shift-variant, rendering a straightforward reconstruction loss difficult, (3) the receptive field of a model should be extensive to catch long dependencies in audio that are essential in generating low frequency signals, (4) reconstructing high frequency components, albeit being a key to fidelity of synthesized speech, is not prone to modeling due to their noise-like nature, and (5) the temporal upsampling from mel spectrogram to waveform results in patternized artifacts, which are particularly audible.

Regardless of the hurdles, many effective vocoder models have been proposed. The initial landmark success was achieved by the autoregressive models [1, 2]. Their causal inductive bias is elegantly in sync with the nature of speech signals, and they minimize the data log likelihood directly. With the powerful model assumption and the solid theoretical ground, autoregressive models had been the best-performing vocoders for a long time. Their subsequent knowledge-distilled variants are mostly geared to lessen the computational overhead for training and inference [3, 4].

To avoid the sample-by-sample causal inference or the use of teacher models, parallelizable statistical models have been devised. Particularly notable are the flow-based vocoders [5, 6] that shape a simple prior distribution into a complex data distribution conditioned on a mel spectrogram using a chain of invertible operations. Although the constraint of invertibility allows direct modeling of the data log likelihood, it also forces the use of element-wise affine transformation, which in compensation for its simplicity, necessitates a massive WaveNet-like module repeating many times. More recently, diffusion-based models [7, 8] achieve impressive negative log-likelihood and mean opinion score (MOS) by letting the network learn how to denoise from a Gaussian noise to a target signal in an interactive synthetic process.

On the other hand, Generative Adversarial Networks (GANs) [9], although its strong potential had been demonstrated in the field of image synthesis, did not show a viable performance as a vocoder until recently. Early attempts [10, 11] suffered from severe upsampling artifacts. MelGAN [12], the first effective GAN vocoder, too could not reach the MOS values of the flow-based models and the autoregressive models.

However, recent GAN vocoders made breakthrough improvements [13, 14, 15, 16, 17], often exceeding the previous state-of-the-art WaveNet on various metrics both qualitatively (e.g. MOS) and quantitatively (e.g. mel cepstral distortion) while being orders of magnitude faster to train and to infer. Their performances are so remarkable that, given a single-speaker dataset of sufficient quantity and consistent quality, one might even claim that the problem of vocoding has been solved.

How do the latest GAN vocoders perform so eminently? For each of the models, researchers underline many different key factors for its success: generator architecture, discriminator architecture, loss function composition, and various training tricks. In this paper, we question their validity as decisive elements. We instead hypothesize that the most important underlying common factor behind the recent successes of GAN vocoders is the multi-resolution discriminating framework, not necessarily the aforementioned details. To demonstrate this experimentally, we pair each of different generators of recent GAN vocoders as well as our proposed generator with the discriminator of HiFi-GAN [17] and report the outcomes in detail. We encourage readers to listen to audio samples on the companion webpage[1], where all the synthesized audios we use for the MOS test are available.

2. Related Works

We take the multi-resolution discriminating framework of HiFi-GAN as the baseline for three reasons: (1) structural simplicity: since its multi-resolution discriminating framework requires none but the final waveform prediction from the generator (no additional intermediate generator features are needed unlike in [16]), it is compatible with many different forms of generator, (2) performance: given the reported metrics in [17], it is currently one of the best vocoder models, outperforming WaveNet and WaveGlow in both ground truth (GT) reconstruction and text-to-speech (TTS) synthesis, and (3) official repository: HiFi-GAN publicly shares a well-maintained official implementation [18].

The HiFi-GAN generator takes a mel spectrogram input and expands it while reducing the number of channels by alternat-

[1] https://moneybrain-research.github.io/gan-vocoder.

http://dx.doi.org/10.21437/Interspeech.2021-41

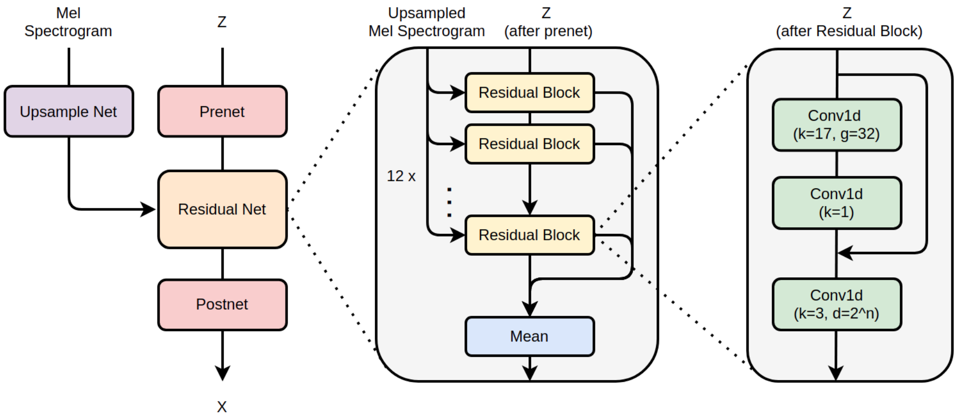

Figure 1: *Proposed architecture: overall generator (left), residual net (middle), and residual block (right). With the first convolution of a large kernel size and the third convolution of exponentially increasing dilation, the receptive field grows to 8369 samples by the 12th block. The residual net wires WaveNet-like skip connections from every residual block to the 1×1 convolution postnet.*

ing transposed-convolution-based upsamplings and multi-path residual blocks called multi-receptive field fusion (MRF). MRF processes multiple patterns of various lengths from the input to the residual block at once in a parallelized manner. As to both generator and discriminator implementations for our experiment, we refer to the official project repository [18].

MelGAN was the first demonstration that a GAN vocoder can achieve a performance that comes close to that of WaveGlow [5]. For every level in the upsampling process, it pairs a residual stack of three dilated convolution layers and a transposed convolution. This design now serves as a standard baseline for subsequent GAN vocoders due to its simplicity and effectiveness. For our experiment, the MelGAN generator of the official project repository is utilized [19].

Parallel WaveGAN (PWGAN) and its enhanced variation that applies a perceptually weighted mask to spectrogram when computing losses [15] generate audio by taking a Gaussian noise. This is a critical difference from four existing GAN vocoders in Table 1 (except for ours) where the output results from upsampling a mel spectrogram without employing a latent variable. In addition, PWGAN employs a strictly WaveNet-like but non-causal architecture as a generator, contrasting against more MelGAN-like vocoders such as VocGAN and Universal MelGAN (UMGAN). We use an unofficial but well-maintained implementation [20].

The overall architecture of the UMGAN generator is similar to that of MelGAN, consisting of residual stacks and transposed convolution upsampling layers. But within each of the residual blocks, UMGAN adopts a gated activation unit (GAU), similarly to WaveNet and PWGAN. There is no available implementation of UMGAN so we construct its generator ourselves as reported in [14]. For missing details, we refer to MelGAN and MultiBand MelGAN [21], which are the baseline models of UMGAN.

VocGAN is different from the other state-of-the-art GAN vocoders in that it is originally designed to utilize both the final output and the intermediate outputs of the generator as the target of discrimination. It has a longer upsampling sequence of 4×, 4×, 2×, 2×, 2×, 2× than the usual MelGAN style of 8×, 8×, 2×, 2× and requires auxiliary conditioning from a mel spectrogram for the last four residual blocks. Since there is no available official repository, an unofficial implementation [22] is adopted for our experiment.

3. Proposed Architecture

We propose a generator architecture of our own. Each residual block is constructed similarly to [23], which efficiently widens the receptive field by conjoining two convolution layers that process patterns over temporal axis and frequency axis respectively. In our proposed method, the first layer performs groupwise convolution, enabling a very large kernel size with a small number of parameters. The second convolution layer has a standard kernel size of 3 but with an exponentially increasing dilation. In between the two layers, we introduce 1×1 convolution that mixes up the information between channels. Every convolution layer output is followed by the non-linearity of LeakyReLU with slope of 0.1. The residual connection is made between the block input and the output of the in-between 1×1 convolution layer. The architecture is illustrated in Figure 1.

Mel spectrogram upsampling is done by alternating nearest-neighbor upsampling and 1D convolution layer with the kernel size of 3. We use nearest-neighbor upsampling over transposed convolution as [24, 25] report various advantages (e.g. less distortion in the frequency domain, less checkerboard artifacts, and better preservation of information from low resolution). Upsampling is done in 8×, 8×, 2×, 2× following the MelGAN procedure, and the upsampled mel is added element-wise to the input to every block.

The purpose behind the proposed architecture is to reinforce the diversity of the six generators so as to further test the generality of the multi-resolution discriminating framework. The proposed generator can be interpreted to be a mixture of varying generators; it takes a MelGAN-like simple residual architecture, WaveNet-like skip connection from all the intermediate layers to the ultimate 1×1 convolution layer, and a PWGAN-like latent-variable-dependent process with mel spectrogram as a guiding condition.

4. Experiments

We use the LJSpeech dataset [26], consisting of 13,100 English utterances of a single female speaker with a total duration of approximately 24 hours. It is split into a training set of 12,950 files and a validation set of 150 files. The latter is not used for training, preserved only for evaluation. The audio format is 16-bit PCM with a sample rate of 22.05 kHz. The dataset is employed as is without any modification.

Aside from the varying generator, we strictly follow the V2

Table 1: *Summary of the six models employed for our experiment. The reported numbers of parameters concern only the generators, and the training speeds are recorded when training each generator paired with the HiFi-GAN V2 discriminating framework. We average the elapsed time from the beginning to the stoppage at 700k steps per batch. The inference speed is computed by averaging the elapsed time of synthesizing 150 evaluation GT mel spectrograms.*

	Parameter	Training speed	Inference speed	Input type
HiFi-GAN	928,514	.61 s / batch	.0058 s / sample	Mel spectrogram
MelGAN	4,266,050	.58 s / batch	.0033 s / sample	Mel spectrogram
PWGAN	1,320,442	.82 s / batch	.0098 s / sample	Gaussian noise (+ mel)
UMGAN	93,077,506	1.09 s / batch	.0035 s / sample	Mel spectrogram
VocGAN	4,715,698	.61 s / batch	.0085 s / sample	Mel spectrogram
Ours	1,209,590	.78 s / batch	.0048 s / sample	Gaussian noise (+ mel)

configuration of HiFi-GAN with respect to the data processing procedure, the discriminating mechanism, and the loss composition, as proposed in [17] and as specified in [18]. We adopt the V2 configuration over V1 since its reported performance is not much different from that of the optimal V1, it is a straightforwardly smaller version of V1 (V3 introducing more fundamental changes), and it trains more quickly than V1.

80-band mel spectrograms are computed with the FFT size, the window size, and the hop size set to 1024, 1024, and 256, respectively. We use the AdamW optimizer [27] with β_1 = 0.8, β_2 = 0.99, and the weight decay of λ = 0.01. The learning rate decays by 0.999 every epoch, starting from 2e-4. The batch size is set to 16. Since the five existing GAN vocoders [12, 13, 14, 16, 17] originally report a wide range of numbers of training steps from 400k to 2.5M, we set the stoppage to an approximate midpoint (700k steps) for all the six models of our experiment.

To evaluate the audio quality, 20 sentences are randomly selected out of the 150 evaluation samples disjoint from training. We then compute the mel spectrograms of the chosen utterances and have each of the vocoders synthesize them back to raw waveforms. In addition to the GT mel spectrogram-based vocoding, we use Tacotron 2 [28] to test the performances of the six generators against the TTS output. To this end, the Pytorch example implementation of Tacotron 2 [29] and its pretrained weights [30] are employed.

Given 6 vocoders, 20 GT-based syntheses, and 20 TTS-based syntheses, 240 synthesized utterances are obtained in total. We then add 100 of the 150 evaluation samples (these are original human utterances with no vocoding involved) to the set; the samples corresponding to the 20 sentences used for GT- and TTS-based syntheses are first included, and the remaining 80 are randomly chosen. The resulting final test set thus contains 340 utterances in a random order. All the audio clips are normalized to the equal loudness of -21 dB to prevent the influence of audio volume differences.

Each of the randomly shuffled 340 samples is assigned to 30 unique subjects via Amazon Mechanical Turk (AMT) [31]. The participants are asked to evaluate the audio files regarding their naturalness (i.e. human-sounding) on a five-point Likert scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). The total number of raters who participated is 189. To avoid any response bias, we use the AMT template of audio naturalness as is with no modification and accept all the collected responses without rejection on a first-come-first-serve basis.

In addition to the qualitative method, the mel cepstral distortion (MCD) [32] is utilized to evaluate the models by a quantitative measure. For all the pairs of the GT and the corresponding synthesized utterances, we represent them as mel-frequency cepstral coefficients (MFCC). Additionally, dynamic time warping is applied to align the MFCC sequences of TTS-based syntheses to their GT pairs (the GT-based syntheses are aligned to their correspondences by default) as in [33]. We use 13 coefficients excluding the first coefficient. Before computing MCD, all audio clips are normalized to have mean of 0 and standard deviation of 1 to minimize the volume differences.

5. Results

In Table 2, the MOS scores of the six generators are distributed in an extreme vicinity of one another for both TTS-based and GT-based syntheses. For either case, the confidence interval of the best performing model (ours) overlaps with that of the worst performing model (PWGAN in MOS_{GT} and MelGAN in MOS_{TTS}) despite a large number of collected answers (600) per model; no single model is therefore significantly outperforming or underperforming.

Table 2: *Comparison of generators by MOS and MCD. The results are synthesized from both GT mel spectrograms (left) and Tacotron2 TTS outputs (right). GT* denotes MOS tests on original (not synthesized) audio files.*

	MOS_{GT}	MOS_{TTS}
HiFi-GAN	3.985 ± .065	3.657 ± .082
MelGAN	3.995 ± .065	3.580 ± .079
UMGAN	3.997 ± .067	3.615 ± .077
VocGAN	3.980 ± .066	3.700 ± .076
PWGAN	3.938 ± .074	3.650 ± .079
Ours	**4.028** ± .065	**3.708** ± .075
GT*	4.132 ± .028	
	MCD_{GT}	MCD_{TTS}
HiFi-GAN	6.425 ± .050	12.937 ± .145
MelGAN	6.160 ± .050	12.850 ± .138
UMGAN	**5.365** ± .059	**12.824** ± .138
VocGAN	6.500 ± .050	13.023 ± .140
PWGAN	6.196 ± .055	12.882 ± .137
Ours	6.680 ± .052	13.013 ± .136

In the case of the TTS-based syntheses, the MCD values in Table 2 show a similar pattern; all the models are once again within the confidence intervals of one another. Therefore, in an actual TTS application where utterances are inferred based on unseen texts, the synthesis quality will be similar both quantitatively and qualitatively amongst the six vocoders. The MCD scores of the GT-based syntheses, however, are the only case of the four types of measures, in which the models show statisti-

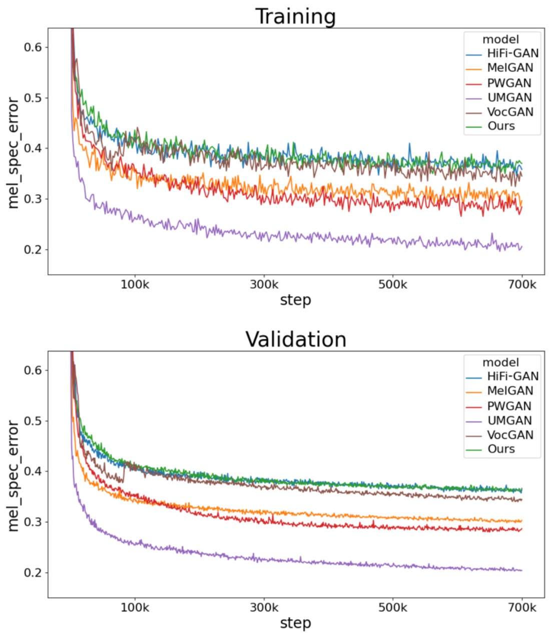

Figure 2: *L1 reconstruction loss on mel spectrograms over time; training (top) and validation (bottom). X-axis unit is batch step.*

cally significant differences.

The differences in the quantitative performance of the GT-based synthesis amongst the six generators are better illustrated in Figure 2 (since the mel spectrogram error is computed as an L1 loss between a GT sample and the corresponding reconstruction, one can interpret it as the model performance on GT mel spectrogram over time). Three groups are observed: (1) UMGAN does consistently best from early on for both training and validation errors, (2) MelGAN and PWGAN follow, and (3) HiFi-GAN, VocGAN, and our proposed model constitute the last group. Although the heaviest UMGAN shows the best performance, the subsequent models' performances are not in the order of the numbers of parameters. We thus observe no clear factor that explains the differences in MCD_{GT}.

6. Discussion

Architecture-wise, the six generators differ in many aspects: MRF of HiFi-GAN, axial residual block and nearest-neighbor upsampling of our proposed model, UMGAN's GAU residual block adapted to MelGAN-like architecture, and a longer upsampling sequence and auxiliary mel conditioning of VocGAN, to mention a few. Input-wise, PWGAN and our proposed model take a noise latent and shape it into a waveform while others upsample directly from a mel spectrogram. The capacities of the six generators greatly vary as well. Given these considerable differences, it is unexpected that the vocoders achieve highly similar performances, failing to demonstrate distinguishable scores for three out of the four metrics. Noticeable is the original HiFi-GAN generator neither performing prominently nor training/inferring quickly. It shows no sign of advantage from being paired with its intended discriminator pair.

Our conjecture is that the important common traits amongst the six generators are (1) effective capturing of long dependencies in the audio and (2) dense wiring that exposes every layer closely to the gradients from the final loss. As long as these primary conditions are met, other architectural details do not seem to affect the vocoded outcome significantly. While recent studies closely link generators and discriminators and explicate their interactions to minute details, the experimental results suggest a different possibility: given a well-established multi-resolution discriminating framework and an operating generator, the architectural details might not be critical factors of the vocoder performance.

Note that the original discriminator of UMGAN determines real/fake based on spectral representations of GT utterances and predictions as well as their raw waveforms. VocGAN requires intermediate and final waveform representations from its generator for discrimination. Although paired with a discriminating system that examines only the final output waveform here, both UMGAN and VocGAN generators achieve qualitative and quantitative scores just as high as other generators. The implication here is that the multi-resolution discriminating framework is applicable even across various forms of representations: raw waveform, spectral representation, and intermediate network features.

From the experiment, we can see that the light generators of HiFI-GAN, PWGAN, and our proposed one show comparable performances to those of heavier models such as MelGAN, VocGAN and even UMGAN that is 9 to 10 times larger. In addition, the MelGAN generator, despite being an earlier and simpler instance, does just as well as more modern designs. This observation implies the fact that the capacity and the wiring of the generator have already been sufficiently powerful since MelGAN. Therefore, increasing the number of channels, the number of layers, or the number of blocks might not be a key to further advancing the vocoder performance.

We believe that the findings above are applicable beyond vocoding. Given the powerful generality of the multi-resolution discriminating framework, it is highly likely to improve various GAN-based speech tasks such as speech synthesis, voice conversion, speech enhancement, and singing synthesis among others.

7. Conclusions

In this work, we pair the HiFi-GAN multi-resolution discriminating framework with five different generators of the latest GAN vocoders. Our proposed generator is added to the set to enhance the overall variety of the six generators' architectural characteristics. MOS and MCD evaluations are performed on GT-based and TTS-based syntheses of the vocoders, and in three of the four metrics, we observe no sign of statistical difference in their performances despite their widely different structures. This finding questions what really matters in creating a high-fidelity speech signal from a mel spectrogram in a GAN setting and supports the hypothesis that the deciding factor lies in the multi-resolution discriminating framework. From the experiment, we make three noteworthy observations: (1) the multi-resolution discriminating framework can be generalized across different generators of varying design philosophy, (2) the multi-resolution discriminating framework can be generalized across various forms of targets (i.e. spectrogram, raw waveform, or intermediate representation) to be discriminated, and (3) the capacity and the wiring of the generator have already been sufficiently powerful since MelGAN. Observing the strong generality of the multi-resolution discriminating framework, we expect it to be effective in various other GAN-based speech tasks.

8. References

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

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *Proceedings of Machine Learning Research (PMLR)*, pp. 2410–2419, 2018.

[3] W. Ping, K. Peng, and J. Chen, "Clarinet: parallel wave generation in end-to-end text-to-speech," *International Conference on Learning Representations (ICLR)*, 2019.

[4] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel wavenet: fast high-fidelity speech synthesis," *International Conference on Machine Learning (ICML)*, pp. 3915–3923, 2018.

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

[6] S. Kim, S. Lee, J. Song, J. Kim, and S. Yoon, "Flowavenet : a generative flow for raw audio," *Proceedings of Machine Learning Research (PMLR)*, vol. 97, pp. 3370–3378, 2019.

[7] N. Chen, Y. Zhang, H. zen, R. Weiss, M. Norouzi, and W. Chan, "Wavegrad: estimating gradients for waveform generation," *International Conference on Learning Representations (ICLR)*, 2021.

[8] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: a versatile diffusion model for audio synthesis," *International Conference on Learning Representations (ICLR)*, 2021.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[10] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *International Conference on Learning Representations (ICLR)*, 2019.

[11] C. Lee, A. Toffy, G. Jung, and W. Han, "Conditional wavegan," *arXiv preprint arXiv:1809.10636*, 2018.

[12] K. Kumar, R. Kumar, T. Boissiere, L. Gestin, W. Teoh, J. Sotelo, A. Brebisson, Y. Bengio, and A. Courville, "Melgan: generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14 881–14 892, 2019.

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

[14] 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.

[15] E. Song, R. Yamamoto, M. Hwang, J. Kim, O. Kwon, and J. Kim, "Improved parallel wavegan vocoder with perceptually weighted spectrogram loss," *arXiv preprint arXiv:2101.07412*, 2021.

[16] J. Yang, J. Lee, Y. Kim, H. Cho, and J. Kim, "Vocgan: a high-fidelity real-time vocoder with a hierarchically-nested adversarial network," *INTERSPEECH*, pp. 200–204, 2020.

[17] 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 (NeurIPS)*, pp. 1268–1272, 2020.

[18] J. Kong, "Official repository of hifi-gan," https://github.com/jik876/hifi-gan/, accessed: 2021-03-01.

[19] W. Teoh, "Official repository of melgan," https://github.com/descriptinc/melgan-neurips/, accessed: 2021-03-01.

[20] T. Hayashi, "Unofficial repository of parallel wavegan," https://github.com/kan-bayashi/ParallelWaveGAN/, accessed: 2021-03-01.

[21] 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," *arXiv preprint arXiv:2005.05106*, 2020.

[22] Rishikesh, "Unofficial repository of vocgan," https://github.com/rishikksh20/VocGAN/, accessed: 2021-03-01.

[23] J. You, G. Nam, D. Kim, and G. Chae, "Axial residual networks for cyclegan-based voice conversion," *arXiv preprint arXiv:2102.08075*, 2021.

[24] J. Pons, S. Pascual, G. Cengarle, and J. Serra, "Upsampling artifacts in neural audio synthesis," *arXiv preprint arXiv:2010.14356*, 2020.

[25] R. Child, "Very deep vaes generalize autoregressive models and can outperform them on images," *arXiv preprint arXiv:2011.10650*, 2020.

[26] K. Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, accessed: 2021-03-01.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[28] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.

[29] G. Karch, "The tacotron2 repository," https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/Tacotron2/, accessed: 2021-03-01.

[30] "Tacotron2 pytorch checkpoint (fp32)," https://ngc.nvidia.com/catalog/models/nvidia:tacotron2_pyt_ckpt_fp32/, accessed: 2021-03-01.

[31] "Amazon mechanical turks," https://www.mturk.com/, accessed: 2021-03-01.

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

[33] E. Battenberg, S. Mariooryad, D. Stanton, R. Skerry-Ryan, M. Shannon, D. Kao, and T. Bagby, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," *International Conference on Learning Representations (ICLR)*, 2020.