INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

LightVoc: An Upsampling-Free GAN Vocoder Based On Conformer And Inverse Short-time Fourier Transform

Dinh Son Dang[1], *Tung Lam Nguyen*[1], *Bao Thang Ta*[1], *Tien Thanh Nguyen*[1], *Thi Ngoc Anh Nguyen*[1], *Dang Linh Le*[1], *Nhat Minh Le*[1], *Van Hai Do*[1]

[1]Viettel Cyberspace Center, Viettel Group, Hanoi, Vietnam

{sondd9,lamnt45,thangtb3,thanhnt116,anhntn51,linhld6,minhln2,haidv21}@viettel.com.vn

Abstract

Most neural vocoders based on generative adversarial networks (GANs) rely on iterative upsampling to generate audio sequences from mel-spectrograms as well as dilated convolution to expand their receptive fields. Nevertheless, iterative upsampling increases the network's complexity and thus decreases the inference speed. Moreover, convolution neural networks are geared towards extracting fine-grained local information and still struggle to capture long-term dependencies. In this work, we propose LightVoc, an efficient and high-quality GAN-based neural vocoder that replaces all upsampling blocks with a stack of Conformer blocks and uses a novel combination of discriminators to generate high-resolution waveforms over the full-band. From our experiments on LJSpeech dataset, LightVoc produces comparable audio quality while being 52.5 times faster in terms of CPU-based inference speed in comparison to HiFi-GAN V1.

Index Terms: text-to-speech, neural vocoder, generative adversarial networks, Conformer, invert short-time Fourier transform, upsampling-free

1. Introduction

Most TTS systems comprise of an acoustic model that maps input text sequence into a set of acoustic features, and a neural vocoder (NV) that synthesizes raw waveform from the these features [1–3]. In this work, we focus on developing a high-performance NV without compromising the audio quality.

NVs can be classified into two categories: autoregressive (AR) and non-autoregressive (non-AR). While AR NVs [4–7] have gained popularity due to their ability to generate human-sounding audio, their notoriously slow inference severely limits their applications in the real world. To achieve a higher level of performance, researchers have turned to non-AR NVs [8–10]. Among these works stand out GAN-based NVs [10–18] with the ability to synthesize high-fidelity audio while still be lighter and much faster than other alternatives. Their leading representative is HiFi-GAN [15], which achieves higher mean opinion score (MOS) than WaveNet [4] (a top-tier AR NV) while synthesizing audio incredibly faster.

Most GAN-based NVs consist of two adversarially trained neural networks: a generator to synthesize audio from mel-spectrogram, and a discriminator to evaluate the authenticity of the synthesized audio. Their generators usually employ a stack of convolution-based blocks to upsample the input sequence until the output sequence matches the target waveform's temporal resolution. However, this reliance on iterative upsampling increases the network's complexity and may slow down the inference process. To address this issue, *iSTFTNet* [19] proposed to remove a part of these upsampling blocks and replace some output-side layers with inverse short-time Fourier transform (iSTFT) (Fig. 1b). Using HiFi-GAN as the baseline, iSTFTNet proved to be faster and more lightweight while maintaining a comparable audio quality.

In addition to iterative upsampling, GAN-based NVs often rely on dilated convolution to increase the receptive field. In [15], HiFi-GAN proposed multi-receptive field fusion (MRF) module, consisting of multiple residual blocks with varying dilation rates and kernel sizes, to further diversify receptive field patterns (Fig. 1a). However, convolution neural networks (CNNs) are geared towards extracting fine-grained local information, and still struggle to capture long-term dependencies such as pitch and energy. On the other hand, based on self-attention [20], Transformer [21] has shown great potential to replace CNNs in various fields [22–25]. Therefore, in [26], Miao et al. proposed strategically replacing convolution with self-attention, which is computed within dilated sliding windows to expand the receptive field by constant scale factor while not increasing the computation load. Interestingly, in [27], Gulati et al. proposed *Conformer*, a novel attempt to combine convolution with self-attention to model both local and global dependencies. It achieved state-of-the-art speech recognition performances and has since been widely adopted in many tasks of speech processing [28–35]. To date, however, there has been no successful application of Conformer in building a neural vocoding model.

Inspired by these works, we propose LightVoc (Fig. 1c), an up-sampling-free GAN vocoder based on Conformer and iSTFT. Our main contribution is threefold:

- *First*, we replace all upsampling blocks of iSTFTNet with a stack of Conformer blocks to make the generator lighter and faster.
- *Second*, we propose a combination of discriminators to facilitate our model in generating high-resolution waveforms over the full-band.
- *Third*, our model outperforms all compared others in terms of inference speed while maintaining audio quality competitive with HiFi-GAN V1, according to experiments conducted on LJSpeech [36] dataset.

2. Proposed Model

Built upon GAN architecture, LightVoc consists of two adversarially trained neural networks: a generator to synthesize audio from mel-spectrogram, and a discriminator to evaluate the authenticity of the synthesized audio. The rest this section reflects the design decisions we made for these components.

10.21437/Interspeech.2023-677

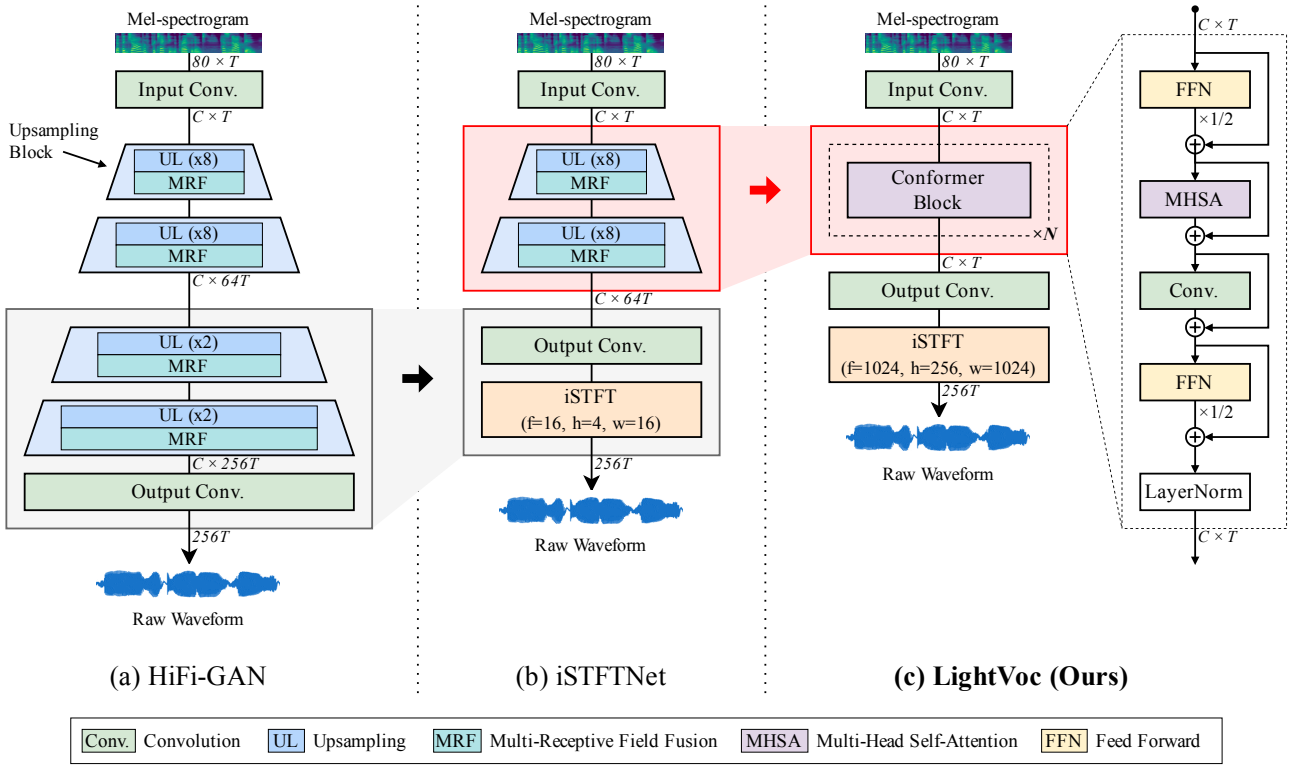

Figure 1: *Comparison between generators of HiFi-GAN, iSTFTNet and LightVoc (Ours). We replace all upsampling blocks of iSTFTNet with a stack of Conformer blocks and use a different iSTFT configuration.* $iSTFT(f, h, w)$ *denotes an invert short time Fourier transform with size of* f*, hop length of* h*, and window length of* w*.*

2.1. Generator

Our proposed generator is inspired by iSTFTNet. However, in order to reduce the model's complexity and to improve its ability to capture long-term dependencies, we substitute all upsampling blocks with a stack of Conformer blocks, as shown in Fig. 1c. The data pipeline goes as follows:

- *Stage 1*: First, a full-band log-mel spectrogram of magnitude is fed into a stack of N Conformer blocks sandwiched between two convolution layers. Their output is magnitude and phase information.
- *Stage 2*: Then, using phase and magnitude information, iSTFT layer generates the final raw waveform. The size, hop length, and window length of this iSTFT layer are identical to those of the STFT layer that is used in stage 1 for mel-spectrogram extraction.

In this pipeline, each Conformer block consists of a feed-forward module (FFN), a multi-head self-attention module (MHSA), a convolution module (Conv), and finally FFN with a normalization layer (LayerNorm). The combination of MHSA and Conv modules allow Conformer to extract both local and global context features from input sequences. Besides, skip connections allow Conformer to avoid gradient vanishing when multiple layers are stacked.

2.2. Discriminator

Our discriminator is a combination of (1) a collaborative multi-band discriminator (CoMBD) [37], (2) a sub-band discriminator (SBD) [37], and (3) a multi-resolution spectrogram discriminator (MRSD) [38].

In fact, CoMBD and SBD were recently proposed by Avocodo [37] to reduce aliasing and imaging artifacts caused by the training objective biased on the low-frequency band and naive downsampling technique. Avocodo outperformed HiFi-GAN on both single and unseen speaker synthesis tasks. Hence, CoMBD and SBD were chosen for our model.

However, in our preliminary experiments employing only CoMBD and SBD, the over-smoothing problem usually occurs in the high-frequency band of the generated audio (Fig. 2a). Fortunately, the over-smoothing issue had been resolved in UnivNet [38] with the proposal of MRSD. Their ablation study proved that it played the key role in reducing over-smoothing and increasing MOS significantly.

Therefore, we propose using MRSD in conjunction with CoMBD and SBD to facilitate LightVoc in generating high-resolution waveforms over the full-band. To the best of our knowledge, there is no prior work that uses the same combination of discriminators.

3. Experiments

3.1. Experimental setup

Our experiments are performed on LJSpeech dataset, containing 24 hours of 22,050Hz audio samples recorded by a native English-speaking female. We use 12,600 samples for training, 250 samples for validation, and 250 samples for testing.

For generator, we use two Conformer blocks as configured in [27] and apply some modifications: model dimension is set

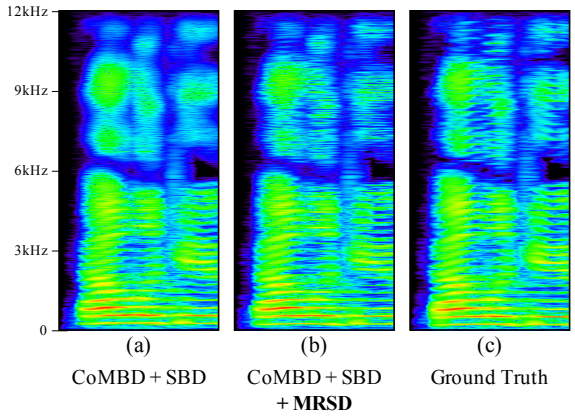

Figure 2: *Spectrograms of the generated waveforms.*

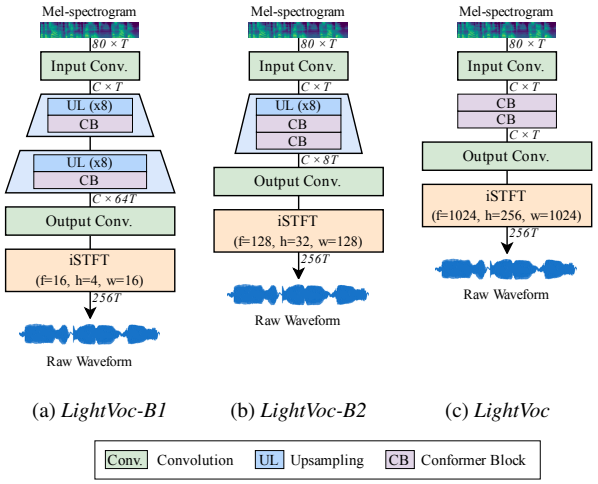

Figure 3: *The generators of our internal baselines and proposed model.*

to 256, number of attention heads is set to 8, convolution kernel size is set to 31, and dropout is set to 0.1. The input features are 80 dimensional log-mel spectrograms extracted by a STFT with size of 1024, hop length of 256, and window length of 1024. We used the AdamW optimizer to train the model for 1M iterations with an initial learning rate of 2×10^{-4}, $\beta_1 = 0.9$ and $\beta_2 = 0.98$, using multi-resolution STFT loss [14]. *For discriminator*, we use least-squares GAN loss [39] for each of discriminator blocks. The final loss is computed by taking average of the losses of these blocks.

In order to compare LightVoc with prior works on audio fidelity, inference speed and number of parameters, we use open-source implementations[1,2] of Parallel WaveGAN, Multi-band MelGAN, HiFi-GAN V1/2 (two variants of HiFi-GAN), and iSTFTNet V1/2 (based upon HiFi-GAN V1/2 respectively).

Furthermore, to gain more insights, we set up two additional baselines, namely LightVoc-B1/2 s (Fig. 3). These models share the same discriminator as proposed in Sect. 2.2, but their generators are set up different from each others: (1) For LightVoc-B1, we take iSTFTNet's generator (Fig. 1b) and replace the MRF module in each upsampling block with a Conformer block (Fig. 3.a). (2) For LightVoc-B2, we take the generator of LightVoc-B1, remove the second upsampling block, use two Conformer blocks instead, and change the configuration of iSTFT accordingly (Fig. 3.b).

3.2. Evaluation metrics

Compared models are evaluated both subjectively and objectively[3]. For subjective evaluation, we crowdsourced a MOS test with 20 native English speakers via *Amazon Mechanical Turk*. For objective assessments, we use multiple metrics including: (1) NISQA [40] and WV-MOS [41] to estimate MOS; (3) mel-cepstral distortion (MCD) [42] and perceptual evaluation speech quality (PESQ) [43] to measure the level of distortion; (4) RTFX, the inverse of real time factor (RTF), to measure the inference speed on CPU and GPU; and lastly, (5) the number of parameters.

[1]https://github.com/kan-bayashi/ParallelWaveGAN
[2]https://github.com/rishikksh20/iSTFTNet-pytorch
[3]Audio samples are available at **https://lightvoc.github.io**

4. Results and Discussion

4.1. Impact of audio length on GPU-based synthesizing speed

In order to investigate the relationship between audio length on inference speed of each model, we generate audio files of different durations: 1, 2, 3…100 seconds. For each selected audio length, each model must synthesize 1000 audio files of the same length, and the RTFX value is calculated through dividing the total audio length by the total execution time. futre These experiments were done on a single NVIDIA A100 GPU. The results are shown in Fig. 4.

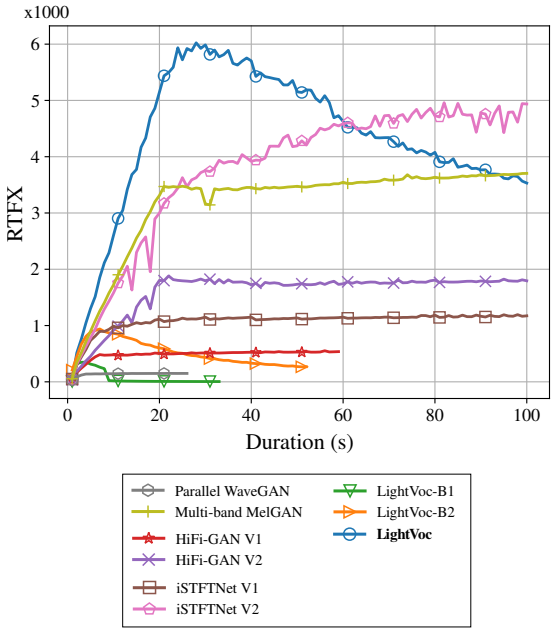

Figure 4: *Impact of audio length on GPU-based synthesizing speed. Parallel WaveGAN, HiFi-GAN V1, and LightVoc-B1/2 are unable to complete the test due to out-of-memory.*

Table 1: *Comparison of our proposed model with prior works on multiple metrics on LJSpeech. The numbers in () indicate the rates (%) comparing LightVoc with other models. A Wilcoxon signed-rank test with $\alpha = 0.05$ is conducted to check the statistical significance between the results of LightVoc and the others:* $(+)$ *= better,* $(\approx)$ *= similar, and* $(-)$ *= worse.*

Model	**MOS** ± CI↑	**WV-MOS** ± STD↑	**NISQA** ± STD↑	**PESQ** ± STD↑	**MCD** ± STD↓	**RTFX↑ @CPU**	**RTFX↑ @GPU**	**# Params↓ (M)**
Ground Truth	4.477 ± 0.601 (+)	4.222 ± 0.246 (+)	3.828 ± 0.632 (≈)	-	-	-	-	-
Parallel WaveGAN	3.772 ± 0.981 (−)	3.889 ± 0.266 (−)	3.478 ± 0.471 (−)	2.355 ± 0.145 (−)	5.456 ± 0.204 (−)	×0.48 (37)	×79.02 (34)	1.84 (13)
Multi-band MelGAN	3.795 ± 1.083 (−)	3.915 ± 0.260 (−)	3.369 ± 0.427 (−)	2.569 ± 0.135 (−)	5.313 ± 0.165 (−)	×10.74 (823)	×984.72 (428)	2.54 (18)
HiFi-GAN V1	4.372 ± 0.705 (≈)	**4.146 ± 0.251** (+)	3.759 ± 0.614 (−)	**3.622 ± 0.147** (+)	3.909 ± 0.154 (−)	×1.31 (100)	×229.89 (100)	13.94 (100)
HiFi-GAN V2	4.078 ± 0.850 (−)	4.039 ± 0.257 (−)	3.572 ± 0.590 (−)	2.947 ± 0.173 (−)	4.524 ± 0.156 (−)	×13.83 (1060)	×769.23 (335)	0.93 (7)
iSTFTNet V1	4.358 ± 0.809 (−)	4.147 ± 0.257 (+)	3.741 ± 0.591 (−)	3.530 ± 0.157 (−)	3.967 ± 0.160 (−)	×2.06 (158)	×432.90 (188)	13.26 (95)
iSTFTNet V2	4.053 ± 0.939 (−)	4.056 ± 0.252 (≈)	3.534 ± 0.606 (−)	2.873 ± 0.186 (−)	4.652 ± 0.158 (−)	×24.69 (1892)	×1293.66 (563)	**0.89** (6)
LightVoc (Ours)	**4.376 ± 0.683**	4.059 ± 0.247	**3.818 ± 0.599**	3.597 ± 0.147	**3.489 ± 0.195**	**×68.48 (5247)**	**×1430.46 (622)**	3.94 (28)

∗ Due to out-of-memory issue, LightVoc-B1/2 models couldn't be trained, and could only be initialized with random parameters for our inference speed tests, of which results are shown in Fig. 4. Therefore, LightVoc-B1/2 are absent from this table.

We observe that: *First*, LightVoc is faster than all compared others except iSTFTNet V2 when synthesizing audio files shorter than 90 seconds; and iSTFTNet V2 only outperforms our model in terms of speed when synthesizing audio files longer than 60 seconds.

Second, most compared prior works without self-attention generate audio files longer than 30 seconds with a steady speed. However, once the audio files are longer than 30 seconds, LightVoc's speed begins to slow down, due to the quadratic computational complexity of the Conformer block. This clearly demonstrates a limitation of placing the self-attention layer on top of variable-length inputs.

Third, both LightVoc-B1 and B2 can't finish the speed test because the upsampling layer significantly increase the length of inputs fed to attention layer by an order of magnitude, which exacerbates the aforementioned drawback of self-attention layer. In fact, LightVoc-B1/2 couldn't be trained due to out-of-memory and get initiated with random parameters instead.

In our future work, a possible solution to improve current self-attention module is to utilize more memory-efficient architectures such as Informer [44], Reformer [45], and Grouped Self-Attention [46].

4.2. Comparison with prior works on multiple metrics

Table 1 compares our proposed model with prior works both subjectively and objectively. Some main observations are as follows.

First, MOS and RTFX results show that LightVoc synthesizes faster than all compared other without compromising the audio quality. According to our Wilcoxon signed-rank test with $\alpha = 0.05$, our proposed model produces audio with quality similar to HiFi-GAN V1's while being 52.5 times faster in terms of inference speed on CPU. In addition, LightVoc scores 8.0% higher MOS than iSTFTNet V2 does while in terms of inference speed, our model still outperforms iSTFTNet V2, which is currently the fastest among mentioned GAN NVs', by 2.8 times on CPU and 1.1 times on GPU. This result also proves LightVoc's great efficiency as the number of parameters of our generator' is about 4 times larger than iSTFTNet V2's.

Second, our conclusions are also backed by other objective evaluation metrics: (1) our WV-MOS is higher than most compared others'; (2) our NISQA is statistically similar to the ground truth's; and (3) we achieve the lowest MCD with a 12.0-to-25.0% relative improvement to iSTFTNet V1/2, 10.7-to-22.8% relative improvement to HiFi-GAN V1/2, and 34.3-to-36.1% relative improvement to Parallel WaveGAN and Multi-band MelGAN.

Third, LightVoc achieved the aforementioned improvements while still being relatively lightweight. Our generator has only about 4 millions parameters, which is only $1/3$ the size of HiFi-GAN V1 or iSTFTNet V1.

5. Conclusions and Future Work

In this work, we propose LightVoc, an efficient and high-fidelity neural vocoder based on Conformer and invert short-time Fourier transform. Based on our experiments on the LJSpeech dataset, compared to HiFi-GAN V1, LightVoc produces competitive audio quality while being 52.5 times faster in terms of CPU-based inference speed. In the future, we will experiment with more memory-efficient self-attention architectures, as well as investigate the effectiveness of LightVoc in a fully end-to-end text-to speech system.

6. References

[1] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," *arXiv*, vol. abs/1710.07654, 2018.

[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," *arXiv*, vol. abs/1905.09263, 2019.

[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," *arXiv*, vol. abs/2006.04558, 2020.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *SSW*, 2016.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *ICASSP*, 2018.

[6] 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 *ICML*, 2018.

[7] J.-M. Valin and J. Skoglund, "LPCNET: Improving Neural Speech Synthesis through Linear Prediction," in *ICASSP*, 2019.

[8] R. J. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *ICASSP*, 2019.

[9] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," *arXiv*, vol. abs/2009.09761, 2021.

[10] C. Donahue, J. McAuley, and M. Puckette, "Adversarial Audio Synthesis," in *ICLR*, 2019.

[11] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-Excited Linear Prediction for Speech Synthesis from Mel-spectrogram," 2019.

[12] 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," *arXiv*, vol. abs/1909.11646, 2020.

[13] K. Kumar, R. Kumar, T. de Boissière, L. Gestin, W. Z. Teoh, J. M. R. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *NeurIPS*, 2019.

[14] 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.

[15] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *arXiv*, vol. abs/2010.05646, 2020.

[16] J. Yang, J. Lee, Y.-I. Kim, H. Cho, and I. Kim, "VocGAN: A High-Fidelity Real-time Vocoder with a Hierarchically-nested Adversarial Network," *arXiv*, vol. abs/2007.15256, 2020.

[17] A. A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek, and N. Kalchbrenner, "A Spectral Energy Distance for Parallel Speech Synthesis," *arXiv*, vol. abs/2008.01160, 2020.

[18] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-GAN: Adversarial Frequency-consistent Audio Synthesis," in *Interspeech*, 2021.

[19] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "ISTFTNET: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform," in *ICASSP*, 2022.

[20] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *NIPS*, 2017.

[21] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *ArXiv*, vol. abs/1910.10683, 2020.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.

[23] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context," *arXiv*, vol. abs/1901.02860, 2019.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *ICCV*, 2021.

[25] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," in *ICCV*, 2021.

[26] C. Miao, T. Chen, M. Chen, J. Ma, S. Wang, and J. Xiao, "A compact transformer-based GAN vocoder," in *INTERSPEECH*, 2022.

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech*, 2020.

[28] E. G. Ng, C.-C. Chiu, Y. Zhang, and W. Chan, "Pushing the Limits of Non-Autoregressive Speech Recognition," in *Interspeech*, 2021.

[29] S. Kim, A. Gholami, A. E. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An Efficient Transformer for Automatic Speech Recognition," *arXiv*, vol. abs/2206.00888, 2022.

[30] M. Burchi and V. Vielzeuf, "Efficient Conformer: Progressive Downsampling and Grouped Attention for Automatic Speech Recognition," in *ASRU*, 2021.

[31] D. D. Son, L. D. Linh, D. X. Vuong, D. Q. Tien, and T. B. Thang, "ASR - VLSP 2021: Conformer with Gradient Mask and Stochastic Weight Averaging for Vietnamese Automatic Speech Recognition," *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 38, no. 1, 2022. [Online]. Available: //jcsce.vnu.edu.vn/index.php/jcsce/article/view/322

[32] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. M. Meng, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," *Interspeech*, 2022.

[33] H. Wu, J. Kang, L. Meng, Y. Zhang, X. Wu, Z. Wu, H.-y. Lee, and H. M. Meng, "Tackling Spoofing-Aware Speaker Verification with Multi-Model Fusion," *arXiv*, vol. abs/2206.09131, 2022.

[34] B. T. Ta, T. L. Nguyen, D. S. Dang, D. L. Le *et al.*, "A Multi-task Conformer for Spoofing Aware Speaker Verification," in *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*. IEEE, 2022, pp. 306–310.

[35] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent Developments on Espnet Toolkit Boosted By Conformer," in *ICASSP*, 2021.

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

[37] T. Bak, J. Lee, H. Bae, J. Yang, J.-S. Bae, and Y.-S. Joo, "Avocodo: Generative adversarial network for artifact-free vocoder," *arXiv preprint arXiv:2206.13404*, 2022.

[38] W. Jang, D. C. Y. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Interspeech*, 2021.

[39] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least Squares Generative Adversarial Networks," in *ICCV*, 2017.

[40] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *arXiv preprint arXiv:2104.11673*, 2021.

[41] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HiFi++: a Unified Framework for Neural Vocoding, Bandwidth Extension and Speech Enhancement," *arXiv preprint arXiv:2203.13086*, 2022.

[42] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1. IEEE, 1993, pp. 125–128.

[43] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001.

[44] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *ArXiv*, vol. abs/2012.07436, 2020.

[45] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *ArXiv*, vol. abs/2001.04451, 2020.

[46] B. Jung, Y. Mukuta, and T. Harada, "Grouped self-attention mechanism for a memory-efficient transformer," *ArXiv*, vol. abs/2210.00440, 2022.