FRE-GAN 2: FAST AND EFFICIENT FREQUENCY-CONSISTENT AUDIO SYNTHESIS

Sang-Hoon Lee[1], *Ji-Hoon Kim*[2], *Kang-Eun Lee*[2], *Seong-Whan Lee*[1,2]

[1]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
[2]Department of Artificial Intelligence, Korea University, Seoul, Korea

ABSTRACT

Although recent advances in neural vocoder have shown significant improvement, most of these models have a trade-off between audio quality and computational complexity. Since the large model has a limitation on the low-resource devices, a more efficient neural vocoder should synthesize high-quality audio for practical applicability. In this paper, we present Fre-GAN 2, a fast and efficient high-quality audio synthesis model. For fast synthesis, Fre-GAN 2 only synthesizes low and high-frequency parts of the audio, and we leverage the inverse discrete wavelet transform to reproduce the target-resolution audio in the generator. Additionally, we also introduce adversarial periodic feature distillation, which makes the model synthesize high-quality audio with only a small parameter. The experimental results show the superiority of Fre-GAN 2 in audio quality. Furthermore, Fre-GAN 2 has a 10.91× generation acceleration, and the parameters are compressed by 21.23× than Fre-GAN.

Index Terms— audio synthesis, neural vocoder, generative adversarial networks, speech synthesis, test-to-speech

1. INTRODUCTION

Recently, deep generative models have shown significant improvement in audio synthesis models [1, 2, 3]. An audio synthesis model, called *"Vocoder"*, converts a low-resolution acoustic feature such as mel-spectrogram into a high-resolution waveform. Furthermore, the recent end-to-end text-to-speech model also needs an audio synthesis model to synthesize a high-resolution waveform from text without any intermediate features [4, 5]. Although current systems synthesize almost realistic audio, these systems suffer from the increased model complexity to generate higher fidelity high-resolution audio. For practical applicability [6, 7], these models have a limitation on low-resource environments such as mobile device.

Especially, WaveNet [8] has shown significant improvement in audio quality. Due to an autoregressive manner, WaveNet has limitations in slow inference speed. To overcome this limitation, parallel waveform synthesis models are introduced. For parallel audio synthesis, Parallel WaveNet [9] uses the knowledge distillation from a pre-trained WaveNet by an inverse autoregressive flow. WaveGlow [10] uses a sequence of invertible flow operations to synthesize audio in parallel. However, these models have high computational complexity.

There are many generative adversarial networks (GAN) based parallel audio synthesis studies by modeling various representations of audio, such as MelGAN [11], Parallel WaveGAN [12], and HiFi-GAN [13]. Among them, HiFi-GAN achieves both high-quality audio synthesis and fast audio generation by modeling the various periodic patterns. Similar to StyleMelGAN [14] using filter-bank discriminators, UnivNet [15] employs the multi-resolution spectrogram discriminator to alleviate the over-smoothing problem on spectra. In addition, Fre-GAN [16] adopt a resolution-connected generator and resolution-wise discriminator to capture the various scales of spectral distribution. However, all models still have limitations to be implemented in low-resource devices due to their computational complexity.

In this paper, we present Fre-GAN 2, a fast and efficient frequency-consistent audio synthesis model. For fast audio synthesis, we do not synthesize target-resolution audio, but we still train the model with target-resolution audio. To do this, Fre-GAN 2 only synthesizes low and high components of audio, and we introduce the inverse discrete wavelet transform (iDWT) to reproduce the target resolution audio in the generator. We also adopt the resolution-wise discriminator of Fre-GAN [16], which uses the discrete wavelet transform (DWT) as a dowmsampling method to reproduce all components without losing information. By utilizing DWT in the discriminator, Fre-GAN 2 can optimize in the sub-audio domain. Furthermore, we also introduce adversarial periodic feature distillation to increase the audio quality with a small parameter. The results show that Fre-GAN 2 can achieve comparable performance with fewer parameters than other models. Especially, Fre-GAN 2 using multi-level iDWT has a 10.91× generation acceleration, and the parameters are compressed by 21.23× than Fre-GAN.

2. FRE-GAN 2

2.1. Generator

To alleviate model complexity in high-resolution audio synthesis, Fre-GAN 2 only synthesizes low and high-frequency sub-audio parts of the target resolution audio from a mel-spectrogram as illustrated in Fig.1, and reproduces the target resolution audio by applying inverse discrete wavelet transform (iDWT) as:

$$\hat{x} = \phi^{-1}(\hat{x}_{low}, \hat{x}_{high}) \tag{1}$$

where $\hat{x}$, $\hat{x}_{low}$, and $\hat{x}_{high}$ denote generated audio reproduced by iDWT, generated low-frequency component and high-frequency component of audio, and ϕ^{-1} represents iDWT. Note that DWT is invertible, so Fre-GAN 2 can reproduce the target-resolution audio from sub-audio sets.

For model compression, we also simplify the Fre-GAN [16] by removing the resolution-connected generator (RCG) and upsampled mel-spectrogram conditioning. While Fre-GAN adopts the RCG to progressively capture various levels of spectral distributions by summing multiple waveforms at the different resolution, Fre-GAN 2 is able to capture different frequency domain representations by sub-audio synthesis and iDWT. To explore the multi-level sub-audio synthesis, we extend Fre-GAN 2 to reproduce the target resolution audio by the multi-level iDWT as illustrated in Fig.1. This allows the model to synthesize the audio much faster and compress the model parameter.

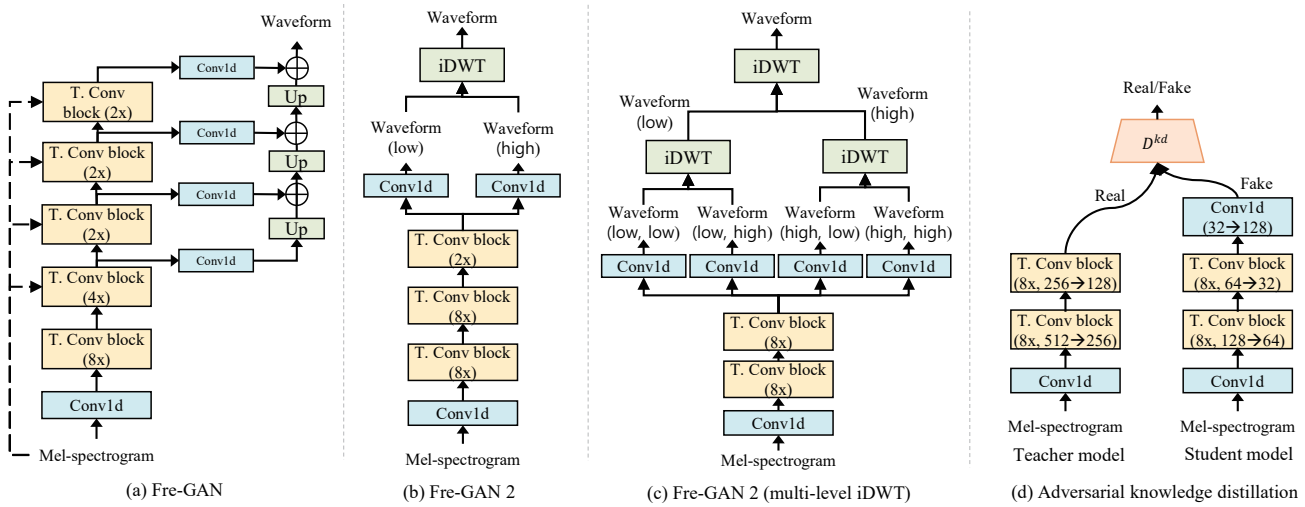

Fig. 1: The generator framework of Fre-GAN 2. (a) Fre-GAN generator architecture. (b) Fre-GAN 2 generator architecture. Fre-GAN 2 only synthesizes low and high components of audio, not target-resolution audio. The target-resolution audio is reproduced by iDWT of sub-audio. (c) Fre-GAN 2 with multi-level iDWT. (d) Adversarial Periodic Feature Distillation. We use the Fre-GAN 2 with a large parameter as a teacher Fre-GAN 2, use the Fre-GAN 2 with a small parameter as a student Fre-GAN 2.

2.2. Discriminator

To train the target-resolution waveform reproduced from sub-audio, we use the resolution-wise discriminators adopted from Fre-GAN. Because the resolution-wise discriminators use the DWT instead of average pooling as a downsampling method, they disentangle the target resolution audio into sub-audio sets, which enable the reproduced audio from sub-audio sets to be trained for each frequency domain without any information loss. It is worth noting again that DWT is invertible [17], so the resolution-wise discriminators make our generator learn to synthesize each sub-audio. Specifically, the resolution-wise discriminators consist of resolution-wise multi-scale discriminator (RSD) and resolution-wise multi-period discriminator (RPD). RSD consists of three sub-discriminators that operate on different audio scales: target resolution audio, stacked sub-audio sets with a DWT (2× downsampled audio), stacked sub-audio sets with a multi-level DWT (4× downsampled audio). RPD consists of five sub-discriminators to capture the different periodic information from audio; We use the same period p of [13, 16] which is $p \in \{2, 3, 5, 7, 11\}$. Due to the DWT in discriminators, the generator learns the consecutive pattern and periodic patterns of audio at the sub-audio domain without synthesizing the target resolution audio.

We use the least-squares GAN objective [18] for the discriminators and generator, and the feature matching loss for generator as following:

$$\begin{aligned}\mathcal{L}_{adv}(D) = &\sum_{n=0}^{4}\mathbb{E}\Big[(D_n^P(x)-1)^2 + (D_n^P(G(s)))^2\Big] \\ &+ \sum_{m=0}^{2}\mathbb{E}\Big[(D_m^S(\phi^m(x)-1))^2 + (D_m^S(\phi^m(G(s))))^2\Big]\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{adv}(G) = &\sum_{n=0}^{4}\mathbb{E}\Big[(D_n^P(G(s))-1)^2\Big] \\ &+ \sum_{m=0}^{2}\mathbb{E}\Big[(D_m^S(\phi^m(G(s)))-1)^2\Big]\end{aligned} \tag{3}$$

$$\mathcal{L}_{fm}(G) = \mathbb{E}\Big[\sum_{i=0}^{T-1}\frac{1}{N_i}\|D^{(i)}(x) - D^{(i)}(G(s))\|_1\Big] \tag{4}$$

where x denotes ground-truth audio and s denotes the input mel-spectrogram of the ground-truth audio. The output of $G(s)$ is generated audio which is reproduced by iDWT. D denotes discriminator which consists of D^P and D^S. D^P and D^S indicate RPD and RSD, respectively. ϕ^m represents m-level DWT. T denotes the number of layers in the discriminator. $D^{(i)}$ is the i^{th} layer feature map of the discriminator, and N_i is the number of units in each layer.

2.3. Adversarial Periodic Feature Distillation

To improve the audio quality of Fre-GAN 2 with a small parameter (student Fre-GAN 2), we use the knowledge distillation [9, 19] from teacher Fre-GAN 2 with a large parameter to student Fre-GAN 2. We adopt the adversarial feature map distillation (AFD) [20, 21], and modified the discriminator as a multi-period discriminator which is able to capture the periodic information of features [13]. The periodic feature discriminators for adversarial periodic feature distillation (APFD) use the features of final transposed convolutional block output from teacher Fre-GAN 2 and the transformed feature of that from student Fre-GAN 2 as an input. The student Fre-GAN 2 mimics the feature of teacher Fre-GAN 2 to fool the periodic feature discriminator. To stabilize model training, spectral normalization [22] is applied to all periodic feature discriminators. We use LSGAN objective for periodic feature discriminator and generator of student model G, and we also use the feature matching loss for student generator as:

$$\mathcal{L}_{kd}(D^{kd}) = \sum_{n=0}^{3}\mathbb{E}\Big[\|D_n^{kd}(f_t)-1\|_2 + \|D_n^{kd}(T(f_s))\|_2\Big] \tag{5}$$

$$\mathcal{L}_{adv}^{kd}(G) = \sum_{n=0}^{3}\mathbb{E}\Big[(D_n^{kd}(T(f_s))-1)^2\Big] \tag{6}$$

$$L_{fm}^{kd}(G) = \mathbb{E}\Big[\sum_{i=0}^{T-1}\frac{1}{N_i}\|D^{kd,(i)}(f_t) - D^{kd,(i)}(T(f_s))\|_1\Big] \tag{7}$$

Table 1: Objective and subjective evaluation results. The MOS is presented with 95% confidence intervals. Higher is better for MOS, PESQ, and speed, and lower is better for the other metrics. Speed of n kHz means that the model can synthesize $n \times 1000$ audio samples per second. The numbers in () denote the synthesis speed over real-time. V1 and V2 denote the large and small parameter model, respectively. m denotes using multi-level iDWT in the generator. Fre-GAN 2* indicates the Fre-GAN 2 model with knowledge distillation.

Model	MOS (↑)	MCD_{13}	RMSE_{f0}	PESQ	Param	Speed on CPU		Speed on GPU	
Ground Truth	4.01 ± 0.04	–	–	–	–	–		–	
WaveNet	3.95 ± 0.04	2.14	42.39	2.96	24.73M	–		0.10	($\times$0.004)
HiFi-GAN (V1)	3.99 ± 0.04	1.07	39.64	3.64	13.92M	60.80	($\times$2.75)	2,249	($\times$102.03)
Fre-GAN (V1)	4.00 ± 0.04	1.00	39.56	**3.76**	18.69M	55.72	($\times$2.52)	2,315	($\times$104.98)
Fre-GAN 2 (V1)	**4.01 ± 0.04**	**0.89**	**39.19**	3.75	13.78M	85.10	($\times$3.86)	2,762	($\times$125.26)
Fre-GAN 2 (V1, m)	4.00 ± 0.04	0.92	39.62	3.72	**13.24M**	**143.84**	($\times$**6.52**)	**4,127**	($\times$**187.17**)
HiFi-GAN (V2)	3.94 ± 0.04	1.55	40.52	3.25	0.93M	218.46	($\times$9.91)	11,763	($\times$533.47)
Fre-GAN (V2)	3.96 ± 0.04	**1.25**	40.05	**3.52**	1.47M	193.21	($\times$8.76)	10,930	($\times$495.70)
Fre-GAN 2 (V2)	3.96 ± 0.04	1.35	40.31	3.37	0.91M	342.66	($\times$15.54)	16,418	($\times$744,61)
Fre-GAN 2 (V2, m)	3.95 ± 0.04	1.45	**39.82**	3.31	**0.88M**	**570.71**	($\times$**25.88**)	**25,258**	($\times$**1,145.49**)
Fre-GAN 2* (V2, m)	**3.98 ± 0.04**	1.56	40.90	3.23	0.91M	501.14	($\times$22.72)	24,180	($\times$1,096.63)

where D^{kd} is periodic feature discriminator, f_t and f_s denote the feature maps of final MRF output from teacher model and student model, respectively. T is a transformation function to match the feature map size between the teacher and student model. Note that using L1 distance loss between f_t and $T(f_s)$ makes it harder to optimize the feature matching objective, so we do not use L1 distance loss between features for feature distillation.

2.4. Final Loss

With adversarial periodic feature distillation, the total loss for the student Fre-GAN 2 model is describe as following:

$$\begin{aligned}\mathcal{L}_{total}(G) = \mathcal{L}_{adv}(G) + \lambda_{fm}\mathcal{L}_{fm}(G) + \lambda_{mel}\mathcal{L}_{mel}(G) \\ + \lambda_{adv}^{kd}\mathcal{L}_{adv}^{kd}(G) + \lambda_{fm}^{kd}\mathcal{L}_{fm}^{kd}(G)\end{aligned} \tag{8}$$

where we set $\lambda_{fm} = 2$, $\lambda_{mel} = 45$, $\lambda_{adv}^{kd} = 1$, and $\lambda_{fm}^{kd} = 2$, and $\mathcal{L}_{mel}$ is defined as L_1 loss between the target mel-spectrogram and predicted mel-spectrogram which are converted from waveform by the STFT function.

3. EXPERIMENTS

3.1. Training Setup

We conducted experiments on the LJSpeech dataset[1] which is a single English speaker dataset. We use the dataset at a sampling rate of $22,050$ Hz. The dataset contains $13,100$ audio samples, and we randomly split the dataset into train (80%), validation (10%), and test (10%) sets. Fre-GAN 2 was compared against several neural vocoders trained on the same dataset: the open-source implementation of a mixture of logistics WaveNet[2], the official implementation of HiFi-GAN[3], and our Fre-GAN implementation. We train all of the models with a large parameter (V1, initial channel of 512) and a small parameter (V2, initial channel of 128). We train the models with a batch size of 16 for the 200M steps. We use 80 bands mel-spectrogram which is transformed with 1024 of window size, 256 of hop size, and 1024 points of Fourier transform. We used AdamW optimizer [23] with $\beta_1 = 0.8$, $\beta_2 = 0.999$, and followed the same learning rate schedule in [13].

[1] https://keithito.com/LJ-Speech-Dataset

[2] https://github.com/r9y9/wavenet_vocoder

[3] https://github.com/jik876/hifi-gan

3.2. Implementation Details

We conducted experiments based on two variations of the generator: V1, V2 with the same discriminator configuration. We simplify the Fre-GAN by removing the stacked audio layer and mel-conditional layer, and we reduce the transposed convolutional blocks by sub-audio synthesis. For Fre-GAN 2 with a single iDWT, we set the kernel sizes of transposed convolutions to [16,16,4], the upsampling sizes to [8,8,2], and the dilation rates of MRF to $[[1,1],[3,1],[5,1]] \times 3$, and two audio components are projected from the block output. For Fre-GAN with multi-level iDWT, we change the kernel sizes of transposed convolutions to [16,16], the upsampling sizes to [8,8], and four audio components are projected from the block output. For RSD and RPD, we use the same architecture of Fre-GAN. The periodic feature discriminator consists of the MPD with periods [2,3,5] and the single-scale discriminator which operates on raw waveforms, not multi-scale discriminator. We decrease the stride size of the first layer to 3 in MPD. We also decrease the kernel size of [7,15,15,41,41,5] and the stride size of [1,2,2,4,4,1]. The source code with specific hyperparameters and audio samples are available on the demo page.[4]

3.3. Audio Quality and Inference Speed

For subjective evaluation, we conduct the naturalness mean opinion score (MOS) test. We randomly select 100 sentences from the test set. For fair evaluation, we normalize all audio samples. The samples are evaluated by at least 20 raters on a scale of 1 to 5. For comparison of models with large parameter, the MOS results show that both Fre-GAN 2 (V1) and Fre-GAN 2 (V1, multi-level iDWT) has comparable performance with Fre-GAN, and higher performance than HiFi-GAN and WaveNet as indicated in Table 1. For models with a small parameter, Fre-GAN (V2) has higher MOS than Fre-GAN 2 (V2). However, Fre-GAN 2 (V2) with knowledge distillation has a higher performance than Fre-GAN (V2).

For objective evaluation, we conduct 3 objective metrics; the mel-cepstral distortion (MCD) [24], $f0$ root mean square error (RMSE_{f0}), and the perceptual evaluation of speech quality (PESQ) [25]. We randomly select 200 sentences from the test set. For MCD, we use the first 13 mel-frequency cepstral coefficients (MFCCs). The results show Fre-GAN 2 has better performance in MCD and RMSE_{f0}. The PESQ results also show Fre-GAN 2 has comparable performance

[4] https://prml-lab-speech-team.github.io/demo/FreGAN2

Table 2: Subjective preference scores and computational complexity comparison between Fre-GAN 2 (V1, multi-level iDWT (m)) and other models. Positive preference scores indicate that Fre-GAN 2 (V1, m) was rated better than the other model. Inference speedup is indicated that how much Fre-GAN 2 (V1, m) achieves generation acceleration than the other model. Parameter reduction (P. reduction) is indicated that how much Fre-GAN 2 (V1, M) is compressed than the other model.

Model	Preference	Speedup	P. reduction
Fre-GAN 2 (V1, m)		Reference	
Ground Truth	-0.01	-	-
HiFi-GAN (V1)	0.01	$1.84\times$	$1.05\times$
Fre-GAN (V1)	-0.02	$1.78\times$	$1.41\times$
Fre-GAN 2 (V1)	-0.01	$1.49\times$	$1.04\times$

Table 3: Subjective preference scores and computational complexity comparison between Fre-GAN 2 (V2, m) and other models.

Model	Preference	Speedup	P. reduction
Fre-GAN 2 (V2, m)		Reference	
Ground Truth	-0.13	-	-
HiFi-GAN (V1)	-0.09	$11.23\times$	$15.81\times$
Fre-GAN (V1)	-0.10	$10.91\times$	$21.23\times$

Table 4: Subjective preference scores between Fre-GAN 2 with knowledge distillation (Fre-GAN 2*) and other models.

Model	Preference
Fre-GAN 2* (V2, m)	Reference
Ground Truth	-0.09
HiFi-GAN (V1)	-0.02
Fre-GAN (V1)	-0.07
Fre-GAN 2 (V1, m)	0.04
Fre-GAN 2 (V2, m)	0.05

with Fre-GAN. We found that feature-level distillation degrades the performance in objective metrics, but the perceptual score increases.

Although Fre-GAN 2 models have a small number of parameters, Fre-GAN 2 has comparable performance in objective and subjective metrics. We also measured the generation speed on Intel Xeon Gold 6148 2.40 GHz CPU and a single NVIDIA Titan Xp GPU. Fre-GAN 2 also shows generation acceleration in both CPU and GPU. Especially, Fre-GAN 2 with multi-level iDWT achieves 25.88 times faster than real-time on CPU and 1,145.49 times faster than real-time on GPU.

3.4. Preference Evaluation

We conducted preference evaluation between several Fre-GAN 2 with other models on a scale of -3 to 3. Although accelerating the inference speed with fewer parameters, Fre-GAN 2 with multi-level iDWT has comparable performance than Fre-GAN and Fre-GAN 2 with a single iDWT as indicated in Table 2.

Table 3 showed that Fre-GAN 2 (V2) has almost the same preference with HiFi-GAN (V1). Furthermore, Fre-GAN 2 (V2) has a $10.91\times$ generation acceleration, and the parameters are compressed by $21.23\times$ than Fre-GAN (V1).

Table 4 demonstrated knowledge distillation improved the Fre-GAN 2 performance slightly. Especially, Fre-GAN 2 (V2) with APFD has better preference than other small models (V2) .

Table 5: Subjective MOS and preference scores between Fre-GAN 2 (V2, m) with APFD and different distillation methods.

Distillation method	MOS	MCD	PESQ
APFD	$\mathbf{3.92 \pm 0.04}$	1.65	3.15
Fre-GAN 2 (V2, m)	3.91 ± 0.04	**1.60**	**3.16**
L1 distance	3.89 ± 0.04	1.70	3.15
AFD	3.90 ± 0.04	1.65	3.15

Table 6: Ablation study for sub-audio modelling

Method	MOS	MCD	PESQ
Multi-level iDWT	–	2×10^{-5}	4.5
PQMF	–	0.50	4.07
Fre-GAN 2 (V2, m)	$\mathbf{3.91 \pm 0.04}$	**1.60**	**3.16**
Fre-GAN 2 (V2, PQMF)	3.89 ± 0.04	1.64	3.12

3.5. Ablation Study

We conducted ablation study for knowledge distillation methods. First, we use the L1 distance between the feature of the teacher and student model. However, using the L1 distance loss increases metallic sound in the audio. Second, we use adversarial feature distillation (AFD) with only a single discriminator with a period of 1. In our final model, we use adversarial periodic feature distillation (APFD). The results show APFD has the highest audio quality as shown in Table 4.

We also conducted ablation study for sub-audio modelling. Durian [26] introduces multi-band WaveRNN by pseudo quadratue mirror filter bank (PQMF) [27], and multi-band MelGAN [28] also uses PQMF for efficient audio synthesis. To compare iDWT with PQMF, firstly, we evaluate the audio reproduced by multi-level iDWT and audio reproduced by synthesis filter of PQMF. We also train the Fre-GAN 2 with each method. The Table 6 shows that iDWT has better reconstruction performance than PQMF, and the Fre-GAN 2 with iDWT has better performance than Fre-GAN 2 with PQMF. For both ablation studies, we train each model for 500k steps.

4. CONCLUSION

We presented Fre-GAN 2, a fast and efficient frequency-consistent neural audio synthesis model. By adopting the inverse discrete wavelet transform in a generator, Fre-GAN 2 only synthesizes low and high components of audio, not target-resolution audio. It decreases the model parameter and accelerates the audio synthesis speed, and even the results show it also does not degrade audio quality. Furthermore, the adversarial periodic feature distillation increased the audio quality. With $21.23\times$ compression and $10.91\times$ audio generation acceleration, Fre-GAN 2 model has comparable performance on audio quality with other baselines. In this works, we focused on the effect of sub-audio synthesis with iDWT in the generator. For future work, we will apply our method to the on-device TTS model to synthesize speech from the text with a small parameter.

5. ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079 , Artificial Intelligence Graduate School Program (Korea University)) and Netmarble AI Center.

6. REFERENCES

[1] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A Survey on Neural Speech Synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] Hyun-Wook Yoon, Sang-Hoon Lee, Hyeong-Rae Noh, and Seong-Whan Lee, "Audio Dequantization for High Fidelity Audio Generation in Flow-based Neural Vocoder," in *Proc. Interspeech*, 2020, pp. 3545–3549.

[3] Sang-Hoon Lee, Hyun-Wook Yoon, Hyeong-Rae Noh, Ji-Hoon Kim, and Seong-Whan Lee, "Multi-SpectroGAN: High-Diversity and High-Fidelity Spectrogram Generation with Adversarial Style Combination for Speech Synthesis," in *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[4] Hyunseung Chung, Sang-Hoon Lee, and Seong-Whan Lee, "Reinforce-Aligner: Reinforcement Alignment Search for Robust End-to-End Text-to-Speech," in *Proc. Interspeech 2021*, 2021, pp. 3635–3639.

[5] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," *arXiv preprint arXiv:2106.06103*, 2021.

[6] Renqian Luo, Xu Tan, Rui Wang, Tao Qin, Jinzhu Li, Sheng Zhao, Enhong Chen, and Tie-Yan Liu, "Lightspeech: Lightweight and Fast Text to Speech with Neural Architecture Search," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5699–5703.

[7] Zhengxi Liu and Yanmin Qian, "Basis-MelGAN: Efficient Neural Vocoder Based on Audio Decomposition," in *Proc. Interspeech 2021*, 2021, pp. 2222–2226.

[8] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv:1609.03499*, 2016.

[9] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al., "Parallel Wavenet: Fast High-fidelity Speech Synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.

[10] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[11] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[12] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel WaveGAN: A Fast Waveform Generation Model based on Generative Adversarial Networks with Multi-resolution Spectrogram," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[13] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs, "Stylemelgan: An Efficient High-fidelity Adversarial Vocoder with Temporal Adaptive Normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6034–6038.

[15] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Proc. Interspeech 2021*, 2021, pp. 2207–2211.

[16] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee, "Fre-GAN: Adversarial Frequency-consistent Audio Synthesis," in *Proc. Interspeech 2021*, 2021, pp. 2197–2201.

[17] Ingrid Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Commun. Pure. Appl. Math.*, vol. 41, no. 7, pp. 909–996, 1988.

[18] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least Squares Generative Adversarial Networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2794–2802.

[19] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao, "Knowledge Distillation: A Survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[20] Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak, "Feature-map-level Online Adversarial Knowledge Distillation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2006–2015.

[21] Liang Gao, Haibo Mi, Boqing Zhu, Dawei Feng, Yicong Li, and Yuxing Peng, "An Adversarial Feature Distillation Method for Audio Classification," *IEEE Access*, vol. 7, pp. 105319–105330, 2019.

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, "Spectral Normalization for Generative Adversarial Networks," in *Proc. International Conference on Learning Representation (ICLR)*, 2018.

[23] Ilya Loshchilov and Frank Hutter, "Decoupled Weight Decay Regularization," in *Proc. International Conference on Learning Representation (ICLR)*, 2019.

[24] Robert Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment," in *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM)*, 1993, pp. 125–128.

[25] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual Evaluation of Speech Quality (pesq)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[26] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, et al., "Durian: Duration Informed Attention Network for Speech Synthesis.," in *INTERSPEECH*, 2020, pp. 2027–2031.

[27] Truong Q Nguyen, "Near-perfect-reconstruction Pseudo-QMF Banks," *IEEE Transactions on signal processing*, vol. 42, no. 1, pp. 65–76, 1994.

[28] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie, "Multi-band MelGAN: Faster Waveform Generation for High-quality Text-to-Speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 492–498.