HIFIDENOISE: HIGH-FIDELITY DENOISING TEXT TO SPEECH WITH ADVERSARIAL NETWORKS

Lichao Zhang[⋆] *Yi Ren*[⋆] *Liqun Deng*[†] *Zhou Zhao*[⋆]

[⋆]Zhejiang University
[†]Huawei Noah's Ark Lab

ABSTRACT

Building a high-fidelity speech synthesis system with noisy speech data is a challenging but valuable task, which could significantly reduce the cost of data collection. Existing methods usually train speech synthesis systems based on the speech denoised with an enhancement model or feed noise information as a condition into the system. These methods certainly have some effect on inhibiting noise, but the quality and the prosody of their synthesized speech are still far away from natural speech. In this paper, we propose HiFiDenoise, a speech synthesis system with adversarial networks that can synthesize high-fidelity speech with low-quality and noisy speech data. Specifically, 1) to tackle the difficulty of noise modeling, we introduce multi-length adversarial training in the noise condition module. 2) To handle the problem of inaccurate pitch extraction caused by noise, we remove the pitch predictor in the acoustic model and also add discriminators on the mel-spectrogram generator. 3) In addition, we also apply HiFiDenoise to singing voice synthesis with a noisy singing dataset. Experiments show that our model outperforms the baseline by 0.36 and 0.44 in terms of MOS on speech and singing respectively.

***Index Terms*—** text to speech, singing voice synthesis, noisy audio, denoise, generative adversarial network

1. INTRODUCTION

A high-quality text to speech (TTS) system usually requires a large amount of clean speech data (e.g., LJSpeech [1] and VCTK [2]) for training, most of which is recorded in a recording studio by professionals to ensure cleanliness. A lot of manpower and material resources are often required for data collection. There being a similar situation, high-quality singing voice synthesis (SVS) system also suffers this problem. On the contrary, in an uncontrolled environment, data is very easy to collect and people can take out the device (mobile phone, etc.) to record it at any time. However, these data usually contain a lot of noise. How to use these low-quality and noisy data to synthesize high-quality and clean speech and singing voices is a problem worth studying. It may greatly reduce the cost and the difficulty to collect data and make the speech synthesis technology better popularized.

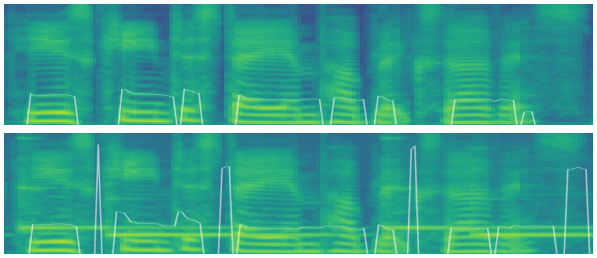
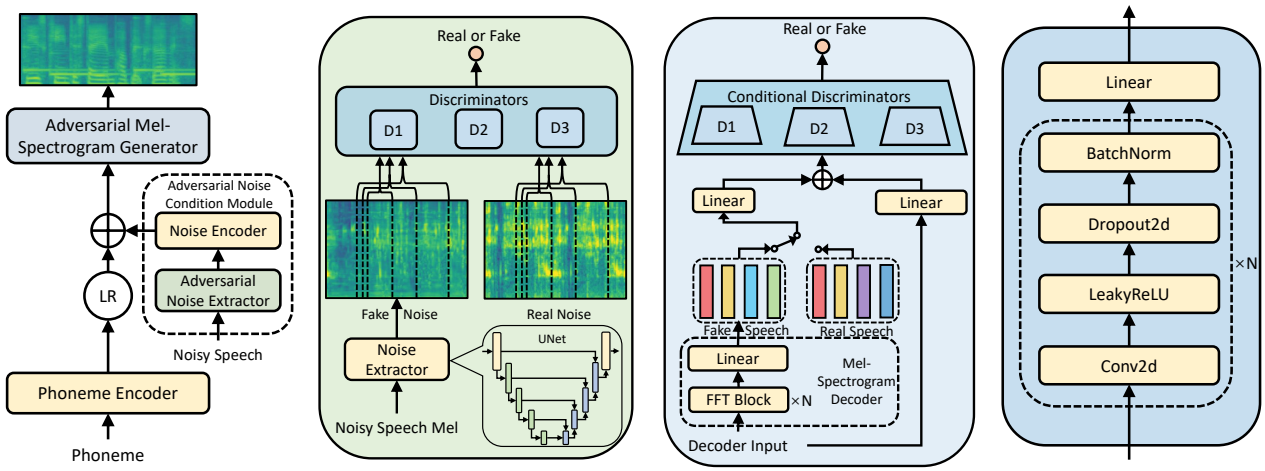


Fig. 1. The fundamental frequency comparisons of clean speech and noisy speech. It can be seen that the fundamental frequency of the added noise could affect the pitch extraction of the original audio.

Previous approaches [3, 4, 5, 6] utilize the method based on data enhancement, which first use a speech enhancement model to process the noisy speech and then feed it into the TTS model for training. And some methods [7, 8] are based on utterance-level noise embedding: as a condition, the noise is fed into the speech synthesis system and the noise condition is removed in the inference to obtain the clean speech. Recently, in DenoiSpeech [9], the utterance-level noise is changed to the frame-level noise with a noise condition module. This fine-grained condition better solves the complicated noise in the real environment. However, there is still a large gap between the voices synthesized by the model and the real voices in terms of the quality and prosody. We dive into the task of training TTS with noisy data and find two problems in the previous work [9]:

- It only uses the simple sum of mean absolute error (MAE) and mean structural similarity (MSSIM) [10] losses to optimize the noise extractor. These losses work under strong assumptions on how our output distribution is shaped and impose important modeling limitations (like not allowing multi-modal distribution and biasing the predictions towards an average of all the possible predictions) [11]. In our experiments, the extracted noise can be over-smoothing and contains



(a) HiFiDenoise (b) Adversarial Noise Extractor (c) Adversarial Mel Generator (d) Discriminator

Fig. 2. The overall architecture of HiFiDenoise.

unrealistic high-frequency details, resulting in noise information losses. When taking this lossy noise information as the condition, the acoustic model may not fully distinguish noise from the target speech, leading to some electrical noise in the synthesized speech.

- The noise could affect the extraction of the pitch. As shown in Fig. 1, noise can interfere with the pitch extraction of speech, making the extracted pitch incorrect and causing voiced/unvoiced errors. Thus, if we train the pitch predictor and the acoustic model using the pitch with some errors they can easily fit the noise, and therefore hurt the prosody and naturalness of the synthesized speech.

In this paper, we propose HiFiDenoise to generate high-fidelity speech and singing voices using noisy speech and singing datasets. Specifically, 1) to address the problem of lossy noise condition, we leverage multi-length adversarial training on noise extractor, which allows multi-modal target distributions modeling and increases the degree of disentanglement between speech and noise from noisy audio. 2) To address the problem of incorrect pitch, we remove the pitch predictor and the pitch condition in the acoustic model to eliminate the influence of errors in pitch. However, the removal of the pitch information leads to severe one-to-many prediction problem and over-smoothing mel-spectrogram outputs [12]. Therefore, we also add discriminators to the mel-spectrogram generator to make up for the performance degradation.

We conduct experiments on speech dataset VCTK [2] and singing dataset OpenSinger [13] and both of them are mixed with NonSpeech100 [14] to simulate the noisy audio. Experimental results show that HiFiDenoise outperforms the previous work [9] on both speech and singing datasets (by +0.36 and +0.44 in terms of MOS). Audio samples generated by HiFiDenoise can be found at https://hifidenoise.github.io/.

2. METHOD

2.1. Overview

As illustrated in Fig. 2(a), the acoustic model is based on FastSpeech 2 [12]. HiFiDenoise consists of a speech synthesis generator with a noise condition module and two discriminators for noise extractor and mel-spectrogram generator respectively. It is worth noting that we remove the pitch predictor in FastSpeech 2 because its ground truth which is extracted from the audio is greatly affected by noise. The generators and discriminators are trained adversarially. During training, the noise encoder converts the extracted noise into the noise embedding as a condition and is added to the hidden sequence which is then directly fed into the mel-spectrogram generator. Correspondingly, the target of the mel-spectrogram generator is noisy speech so that the model can learn both speech information and noise information at the same time. When inference, we input silence audio to the noise encoder and expect the mel-spectrogram generator to synthesize speech without noise. We describe the designs of the adversarial noise condition module and adversarial mel-spectrogram generator in detail in the following subsections.

2.2. Adversarial Noise Condition Module

The noise condition module [9] aims to capture the noise information in noisy speech. It mainly contains two parts: the noise extractor and the noise encoder. The background noise audio is extracted by the noise extractor, and then it is converted into the noise condition by the noise encoder. The noise extractor is based on UNet [15] with MAE and

MSSIM losses. In our experiments, this approach makes the extracted noise over-smoothing, which does not well restore high-frequency parts of the original noise, so we attempt to leverage a discriminator on the noise extractor to solve this problem. Considering that noise has a large difference and complicated fluctuations in time dimension, using a single discriminator to distinguish the entire audio cannot well model the noise with various fluctuations in different time ranges, so we use multiple discriminators with different time lengths. Moreover, there are various of noise but the types of training set only contain limited kinds of noise, so it is easy to cause over-fitting problem. It has been demonstrated that using random windows of different sizes has a data augmentation effect [16], and we consider it may alleviate over-fitting problem. Therefore, we leverage a multi-length GAN (ML-GAN) [17] on noise extractor for adversarial training on mel-spectrograms, as shown in Fig. 2(b), which uses multiple discriminators to distinguish the mel-spectrogram clips in different lengths. The formulation of ML-GAN is shown in Equation 1 and 2:

$$\min_{G_{ne}} \mathbb{E}_x[\sum_{t\in(0,len(y))} (1 - D_t(G_{ne}(x)))^2], \tag{1}$$

$$\min_{D_t} \mathbb{E}_y[(1-D_t(y))^2]+\mathbb{E}_x[(D_t(G_{ne}(x)))^2], \forall t \in (0, len(y)), \tag{2}$$

where x and y represent mel-spectrogram of noisy audio input and noise audio output respectively, G_{ne} represents the noise extractor and D_t represents the discriminator for different time length t. By modeling different lengths of mel-spectrogram clips, ML-GAN can better extract diverse types of noise and better construct noise with obvious high-frequency parts to provide acoustic model with richer noise information which could be helpful to the training process (During the training process, the mel-spectrogram generator receives the outputs of the phoneme encoder and the noise encoder at the same time, which are expected to provide as complete speech and noise information respectively as possible to synthesize mel with specific noise).

2.3. Adversarial Mel-Spectrogram Generator

Mel-spectrogram generator is the decoder of acoustic model, which is used to convert the linguistic features and variation information (pitch, etc.) of speech from encoder into mel-spectrogram. In FastSpeech 2 [12], pitch predictor is proposed to better predict the variations in pitch of speech. The ground truth of the pitch predictor is extracted from the audio. But due to the influence of noise, the extracted pitch information is usually inaccurate. Therefore, we remove the pitch predictor, and in order not to affect the rhythm of the synthesized audio we apply conditional ML-GAN on the generator to model mel-spectrogram of different lengths and make the synthesized audio more natural, as shown in Fig. 2(c). The conditional discriminators additionally consider the linguistic features when distinguish acoustic features. We concatenate the expanded linguistic features (phoneme encoder outputs) and the mel-spectrograms, and feed them into discriminators.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

Datasets. We conducted experiments on both speech and singing datasets. VCTK corpus [2], OpenSinger [13] are used as the dataset for speech and singing respectively. And both datasets are mixed with Nonspeech100 [14] for background noise. The VCTK corpus contains 44 hours clean English speech from 109 speakers, and OpenSinger consists of 50 hours of Chinese singing voices recorded in a professional recording studio. We randomly mix the recordings with Nonspeech100 with an SNR 1 - 10 dB to simulate the noisy audio. For both speech and singing, we split the speakers in half: half are the clean speakers who use the original clean audio, and the other half are the noisy speakers who only have noisy audio and use the artificial noisy audio for the training of the acoustic model. We also divide noisy speakers into seen and unseen parts with the ratio of 8:2. For the seen speakers, we feed the ground truth noise of the corresponding noisy audio to the noise encoder during training, and feed the noise extracted by the noise extractor for the unseen speakers. And in inference, we feed the silence audio to the noise encoder to get synthesized clean audio of the unseen speakers.

As for the text, we first convert it into phoneme sequence by open source tools: apply phonemizer[1] to convert English text into corresponding phonemes, and use pypinyin[2] to convert Chinese lyrics into phonemes. Then the MFA [18] is used to align the phoneme sequence to the speech and singing audio frames.

In addition, we extract pitch from the original singing audio with Parselmouth[3] to provide music score (lyrics, note pitch and note duration) for singing voice synthesis. Besides, we use the following parameters to process speech and singing respectively to convert them into mel-spectrogram: 1) VCTK (sample rate:24000, fft size:2048, hop size:300, window size:1200); 2) OpenSinger (sample rate:22050, fft size:512, hop size:128, window size:512).

Configuration Detail. We choose FastSpeech 2 [12] as the basic acoustic model and Parallel WaveGAN (PWG) [19] as the vocoder, both of which are non-autoregressive generation models and can cooperate to synthesize high-quality speech with high speed. And our method can also be easily ported to other TTS and SVS models trained by noisy speech and singing voices. We stack 4 feed-forward Transformer (FFT) blocks in both the encoder and decoder of the

[1]https://github.com/bootphon/phonemizer
[2]https://github.com/mozillazg/python-pinyin
[3]https://github.com/YannickJadoul/Parselmouth

acoustic model and set the hidden size to 256. As shown in Fig. 2(d), all discriminators share the similar model structure but different model parameters, which consists of three 2D-convolution layers with Leaky ReLU activation, each followed by the dropout and the batch normalization layer, and a linear projection for final output. The number of channels is set to 128 and the kernel size is set to 5. The window sizes of discriminators are set to [32, 64, 128]. In addition, we reserve the pitch condition for the SVS system, which is used to feed the pitch information of the music score. For the training process, we first train the noise extractor for 40k steps and set the batch size to 12 sentences. After that, speech/singing acoustic models are trained for 100k/200k steps respectively.

3.2. Results

To verify the effectiveness of HiFiDenoise, we conduct evaluation both in subjective and objective aspects to measure the quality of the synthesized speech and singing voices. We mainly compare HiFiDenoise with the following systems: 1) Clean GT, the original clean recordings; 2) Clean GT (PWG), where we first convert the original clean recordings into mel-spectrograms, and then convert the mel-spectrograms to audio using PWG; 3) Noisy GT, the clean recordings mixed with background noise; 4) Enhancement-Based, FastSpeech 2 trained on enhanced audio denoised with a pre-trained enhancement model rnnoise[4]; 5) Denoispeech, which adds frame-level noise conditions to the speech synthesis system.

Table 1. MCD and MOS with 95% confidence intervals of speech and singing.

Method	VCTK		OpenSinger	
	MCD	MOS	MCD	MOS
Clean GT	-	4.32	-	4.49
Clean GT (PWG)	4.43	4.13	1.48	4.31
Noisy GT	7.31	2.32	3.78	1.54
Enhancement-Based	7.09	2.89	4.05	2.62
DenoiSpeech	5.15	3.68	3.09	3.79
HiFiDenoise	4.50	4.04	2.82	4.23

Objectively, as the original audio recorded by human is of high naturalness, low noise and good rhythm, a lower difference between the synthesized audio and original audio not only indicates higher naturalness but also demonstrates better noise suppression effect and better rhythm. Mel-cepstral distortion (MCD) is adopted to measure the difference between synthesized and original audio, where lower MCD means higher similarity. In our experiments, 50 utterances are randomly selected for MCD calculation.

Subjectively, we use the mean opinion score (MOS) to measure the perceptual quality of synthesized speech and singing voices: each synthesized audio sample is judged by 5 native speakers. As shown in Table 1, it is obvious that the MCD of HiFiDenoise is lower than the two baselines both on VCTK and OpenSinger datasets. Besides, HiFidenoise achieves higher MOS score than the Denoispeech by 0.36 and 0.44, and only has 0.09 and 0.08 MOS gap to clean GT (PWG) upper bound on the speech and singing respectively, which verifies the high-fidelity voices synthesized by HiFidenoise.

[4]https://jmvalin.ca/demo/rnnoise

3.3. Ablation Studies

We conduct ablation studies to verify the effectiveness of the components in HiFiDenoise, including noise extraction discriminators (NED) and mel-spectrogram generation discriminators (MGD). Table 2 shows the CMOS measured by human and MCD score objectively. We find that removing the NED and the MGD (Row 3 and Row 4) can both result in a decrease CMOS and an increase MCD, indicating their effectiveness in synthesizing high-fidelity speech. The baseline, which removes both NED and MGD (Row 5), makes the voice quality further drop, indicating that both NED and MGD can help improve the performance of HiFiDenoise.

Table 2. MCD and CMOS of speech.

Method	MCD	CMOS
HiFiDenoise	4.50	0
w/o NED	4.75	-0.29
w/o MGD	5.03	-0.80
BaseLine (Denoispeech)	5.15	-1.05

4. CONCLUSION

In this paper, we proposed HiFiDenoise, a high-fidelity TTS system trained by noisy speech data with adversarial networks to address the issues in previous work and improve sound quality, also, refine prosody at the same time: 1) we leverage multi-length adversarial training on noise extractor to solve the problem of sound quality degradation caused by lossy noise condition; meanwhile 2) we remove the pitch predictor and also add discriminators on mel-spectrogram generator to address the problem of prosody degradation caused by the inaccurate pitch extraction. Moreover, based on HiFiDenoise, we also build a SVS system which can generate high-fidelity singing voices with a noisy singing dataset. Experiments demonstrate that HiFiDenoise outperforms previous works on both speech and singing datasets. In the future, we will continue to close the quality gap between the synthesized voices and recordings with noisy training data, and will also make improvements to adapt our model to more languages and more complex scenarios.

5. REFERENCES

[1] Keith Ito and Linda Johnson, "The lj speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017. 1

[2] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016. 1, 1, 3.1

[3] Cassia Valentini Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Interspeech 2016*, 2016, pp. 352–356. 1

[4] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *SSW*, 2016, pp. 146–152. 1

[5] Cassia Valentini-Botinhao and Junichi Yamagishi, "Speech enhancement of noisy and reverberant speech for text-to-speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1420–1433, 2018. 1

[6] Dongyang Dai, Li Chen, Yuping Wang, Mu Wang, Rui Xia, Xuchen Song, Zhiyong Wu, and Yuxuan Wang, "Noise robust tts for low resource speakers using pre-trained model and speech enhancement," *arXiv preprint arXiv:2005.12531*, 2020. 1

[7] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al., "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018. 1

[8] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5901–5905. 1

[9] Chen Zhang, Yi Ren, Xu Tan, Jinglin Liu, Kejun Zhang, Tao Qin, Sheng Zhao, and Tie-Yan Liu, "Denoispeech: Denoising text to speech with frame-level noise modeling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7063–7067. 1, 1, 2.2

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 1

[11] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017. 1

[12] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020. 1, 2.1, 2.3, 3.1

[13] Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao, "Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3945–3954. 1, 3.1

[14] Guoning Hu and DeLiang Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010. 1, 3.1

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 2.2

[16] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan, "High fidelity speech synthesis with adversarial networks," *arXiv preprint arXiv:1909.11646*, 2019. 2.2

[17] Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu, "Hifisinger: Towards high-fidelity neural singing voice synthesis," *arXiv preprint arXiv:2009.01776*, 2020. 2.2

[18] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502. 3.1

[19] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203. 3.1