JLESS SUBMISSION TO DCASE2023 TASK7: FOLEY SOUND SYNTHESIS USING NON-AUTOAGRESSIVE GENERATIVE MODEL

Technical Report

Siwei Huang, Jisheng Bai, Yafei Jia, Jianfeng Chen

Joint Laboratory of Environmental Sound Sensing,
School of Marine Science and Technology,
Northwestern Polytechnical University, Xi'an, China
LianFeng Acoustic Technologies Co., Ltd. Xi'an, China
{hsw838866721, baijs, jyf2020260709}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

This technical report describes our proposed system for DCASE2023 task7: Foley Sound Synthesis. In our approach, we propose a GAN-based mel-spectrogram synthesis system. we take a Conditional Variational auto-encoder (CVAE) as the generator, which consists of densely-connected dilated convolution blocks, and a simple CNN as the discriminator. The decoder of CVAE synthesizes fake mel-spectrogram resampling from prior noise and class, and the discriminator determines whether it is real or not. Furthermore, we also train a classifier to help CVAE keep class-wise distribution. Finally, the audio is wrapped using the HiFiGAN vocoder.

Index Terms— Foley sound synthesis, non-autoregressive, GAN, CVAE

1. INTRODUCTION

Foley sound, in general, refers to sound effects that are created to convey (and sometimes enhance) the sounds produced by events occurring in a narrative (e.g. radio or film). Foley sounds are commonly added to multimedia to enhance the perceptual audio experience. This sound synthesis challenge requires the generation of original audio clips that represent a category of sound, such as footsteps.By generating sound that belongs to a target sound category, Foley sound synthesis can make the workflow much more time and cost-effective. With the rise of virtual environments such as the metaverse, we expect a growing need for the automated generation of more and more complex and creative sound environments. Second, it can be utilized for dataset synthesis or augmentation for a wide variety of DCASE tasks including sound event detection (SED). SED has drawn great attention and synthesized datasets have been used already, e.g., URBAN-SED dataset. A high-quality Foley sound synthesis model could lead to development of better SED models. In 2023, Foley Sound Synthesis (FSS) is introduced as a novel task in DCASE [1]. In this challenge, 7 sound classes are defined, which are Dogbark, Footstep, GunShot, Keyboard, Moving-MotorVehicle, Rain, and Sneeze Cough. As to the baseline system, Liu e.t. [2] proposed a cascaded model. Firstly, they train a VQ-VAE to get the mapping from spectrograms to the latent codebook. With the codebook, they secondly train a prior net for conditionally spectrogram generation, then rebuild the audio waveform using a vocoder.

In this report, we proposed a non-autoregressive FSS system, which consists of a conditional variational auto-encoder [3], GAN algorithm [4]. The generative procedure is completed by the decoder of the CVAE sampled from the prior noise and class-wise condition. And the adversarial training helps the CVAE to reconstruct high-fidelity spectrograms.

2. PROPOSED METHOD

In this section, we introduce the overview of the non-autoregressive model. Then, we introduce the architecture of the generator and the discriminator in the GAN-based framework.

2.1. Overview

The CVAE-GAN-based [5] FSS system is shown in Fig 1. The conditional VAE serves as the generator of GAN, a straightforward CNN serves as the discriminator of GAN and a classifier reveals the quality of class-wise discrepancy. Through adversarial training, the generator synthesizes high-fidelity spectrograms that confuse the discriminator. Finally, the waveforms are wrapped using the HiFiGAN vocoder. We use the officially pre-trained HiFiGAN wrapper, hence our work mainly focuses on the building of CVAE-GAN.

2.2. CVAE for learning presentation and reconstruction

We employ a CVAE, which consists of an encoder, a resampling module, and a decoder. The encoder learns a non-linear mapping from the fixed-shape spectrogram $x \in \mathbb{R}^{H\times W\times 1}$ onto multiple characters of Gaussian distribution: mean μ and standard deviation σ. Then, the decoder reconstructs the synthesized spectrogram according to the latent representations z resampled from Gaussian noise and conditions (one-hot class embedding c). The CVAE loss L_{cvae} consists of reconstruct loss L_{recon} and KLD loss, which can be trained by minimizing:

$$L_{cvae} = L_{recon} + D_{KL}(q(z \mid x, c) \| p(z \mid c)), \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence and $p(z \mid c)$ is the prior distribution based on different categories.

In implementing CVAE, we incorporate D3net symmetrically into the encoder and the decoder. D3net [6], a fully-connected dilated convolutional backbone, is widely used in semantic segmen-

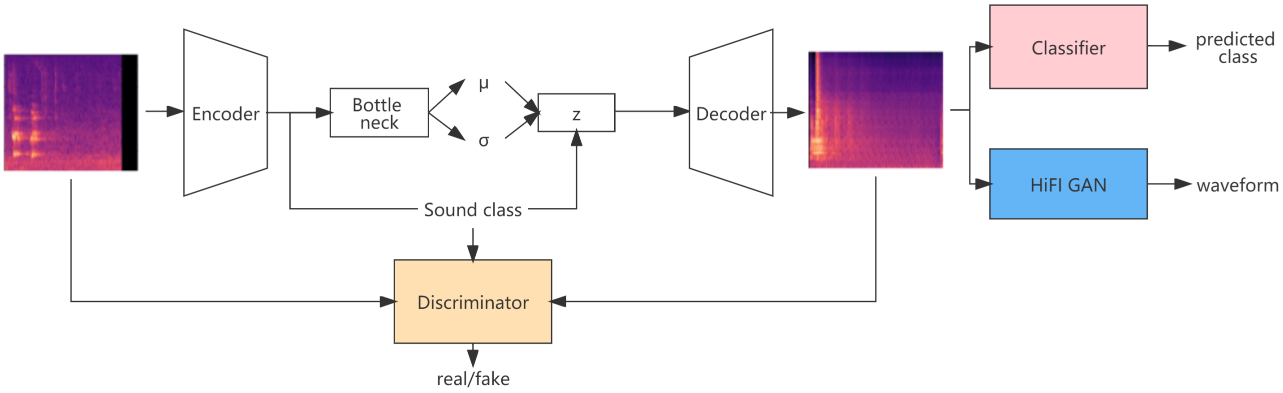

Figure 1: The sketch of the proposed CVAE-GAN.

tation, source separation and speech enhancement. In the proposed CVAE, we use 2 branches of d3net with different kernel sizes. In the decoder, we replace the transposed convolution layer with an interpolate upsampling layer and a convolution layer.

2.3. GAN for high-quality synthesis

We employ a GAN-based framework as an extra constraint to enhance the details of spectrograms reconstructed by the auto-encoder. The proposed CVAE is treated as the generator of GAN. Hence, the encoder of the CVAE firstly maps real spectrograms and the according categories into hidden z.Given the categorycondition, the decoder produces fake spectrograms sampled from z of $q(z|x,c)$. The discriminator of GAN judges the difference between the real spectrogram and the fake one. In the adversarial training, the GAN losses for the generator $\mathcal{L}_G$ and the discriminator $\mathcal{L}_D$ are defined as:

$$\mathcal{L}_G = -\log(D(G(z))), \tag{2}$$

$$\mathcal{L}_D = -\log(D(x)) - \log(1 - D(G(z))), \tag{3}$$

where G is the encoding and the decoding parts of CVAE, D represents the discriminator implemented by a straightforward CNN consisting of Conv2d layer, Instance normalization, and Leaky ReLU activation function.

As to the distinct categorydiscrepancy, we train an additional classifier using the real and the fake spectrogram. The loss of the classifier can be presented as $\mathcal{L}_C$, a multi-class classification loss.

With the above CVAE, GAN and the classifier, we optimize the CVAE-GAN with the full objective:

$$\mathcal{L} = \lambda_1\mathcal{L}_{recon} + \lambda_2\mathcal{L}_{KL} + \lambda_3\mathcal{L}_G + \lambda_4\mathcal{L}_D + \lambda_5\mathcal{L}_C, \tag{4}$$

where the $\mathcal{L}_{KL}$ corresponds the Kullback-Leibler divergence in Eq. 1, and we set $\lambda_1 = 0.1$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 1$, $\lambda_5 = 1$.

3. EXPERIMENT

3.1. Setups

We conduct experiments on track B (no external data). The characters of audio and features follow the setting of the pre-trained HiFi-GAN vocoder: the sample rate is 22050Hz, the length of each audio is set to 4 seconds, the number of FFT is 1024, the hop length is 256, and the number of Mel bins is 80. As to the CVAE, for the encoder, the growth rate of d3net is [16, 18, 20, 22] with a kernel size of 3, the growth rate of the other branch is [16, 22] with a kernel size of 5, yet the parameters are inversely set in the decoder. The discriminator is a straightforward CNN, with 3 convolution blocks, each block includes Conv2d layer, Instance normalization, and Leaky ReLU activation function. The number of kernels is [64, 128, 256].

Table 1: The FAD score of systems, * denotes ensemble results, † denotes the result of output from a single model.

FAD	baseline	ours†	ours*	submitted*
Dogbark	13.411	12.718	**10.651**	11.793
Footstep	8.109	11.168	**7.842**	8.448
Gunshot	**7.951**	16.027	10.965	14.587
Keyboard	**5.230**	10.019	6.725	8.551
Movingvehicle	16.108	15.781	15.673	**15.637**
Rain	13.337	16.885	**8.270**	9.286
Sneeze/Cough	3.770	6.738	**3.043**	3.043
Average	9.702	10.057	**9.024**	10.192

For the adversarial training, the optimizers for all models are Adam with a learning rate of 0.0001. In each iteration, we update the discriminator (D), the generator (G), and the classifier sequentially. To avoid model collapse, we pause the update of the D/G once one of their losses is greater than a loss bound manually set in the iteration. The system is trained in about 300,000 iterations. During training, we apply augments such as gain, pitch shifting, time shifting, and peak normalization. The quality of synthesized samples is evaluated using Frechet Audio Distance (FAD) [7] between 30 audio clips for each class in the development set. In the evaluation step, we define the system which outputs the lowest average FAD score as the best FAD system.

3.2. Results

As shown in the table 1, the FAD scores of each class and the average class are computed. Since they share various data when computing FAD scores, the final result might be not accurate. The single model is slightly inferior to the baseline system, what is worse, the

FAD score of each class is over-smoothed. We also evaluate the audio from specific models, 7 classes to 7 models in total. The ensemble system outperforms the baseline system slightly. For the submission, we choose 7 models for each class according to the FAD scores and audio fidelity.

3.3. Discussion

In our experiments, we notice that the final FAD scores are either smooth or prefer to several sound classes. And the synthesized audio can hardly fit the corresponding category, or the spectrogram contains the ghost from other sound classes. The synthesis can not satisfy the best quality at the same time for three types of sounds: constant noise (vehicle, rain), continual pulse (footstep, keyboard), and pulse (dogbark, gunshot, sneeze/cough). It means we are not making full use of the prior noise and condition. Furthermore, the audio samples are out of diversity because of CVAE. Most of the samples synthesized come from fixed Gaussian noises and just vary in the details of the spectrogram, which means they sound much the same.

4. CONCLUSION

In DCASE2023 task 7: Foley Sound Synthesis, we submit a CVAE-GAN-based system. The submitted system consists of a conditional VAE as the non-autoregressive generator and a CNN as the discriminator. Finally, a vocoder is used to convert the spectrogram to the waveform. The FSS system can output high-fidelity foley sounds according to the categorycondition.

5. REFERENCES

[1] K. Choi, J. Im, L. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the dcase 2023 challenge," *In arXiv e-prints: 2304.12521*, 2023.

[2] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2021.

[3] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[5] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2745–2754.

[6] N. Takahashi and Y. Mitsufuji, "D3net: Densely connected multidilated densenet for music source separation," *arXiv preprint arXiv:2010.01733*, 2020.

[7] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms," in *Proc. Interspeech 2019*, 2019, pp. 2350–2354. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2219