CONDITIONAL FOLEY SOUND SYNTHESIS WITH LIMITED DATA: TWO-STAGE DATA AUGMENTATION APPROACH WITH STYLEGAN2-ADA

Technical Report

Kyungsu Kim, Jinwoo Lee, Hayoon Kim, Kyogu Lee

Seoul National University
Department of Intelligence and Information

ABSTRACT

This report introduces an audio synthesis system designed to tackle the task of Foley Sound Synthesis in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 challenge [1]. Our proposed system is an ensemble system composed of a baseline model and StyleGAN2-ADA [2]. To optimize the system with limited data without relying on external datasets and pretrained systems, we propose a two-stage data augmentation strategy. This approach involves augmenting input waveforms to expand the size of the training dataset, as well as employing adaptive discriminator augmentation (ADA) to alleviate the overfitting of the discriminator and ensure stable training. Experimental results demonstrate that our proposed ensemble system achieves an FAD (Fréchet Audio Distance) [3] score of 5.84 on the evaluation dataset.

Index Terms— Sound synthesis, generative networks, adaptive discriminator augmentation

1. INTRODUCTION

In recent years, the field of sound synthesis has seen significant advancements, especially in the context of text-conditioned sound generation [4, 5, 6, 7]. Despite these advancements, the synthesis of realistic and high-quality sounds from limited data remains a challenge. The field faces the dual obstacles of preserving the quality of generated sounds while maintaining data efficiency.

This report presents our approach to addressing this problem, undertaken in the context of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 challenge [8]. Specifically, we developed a system for conditional Foley sound synthesis tasked with producing high-quality synthetic sounds corresponding to seven different classes from a small dataset, consisting of less than a thousand samples per class.

We leveraged the capabilities of the StyleGAN2-ADA, known for its data efficiency, and implemented a two-stage data augmentation strategy. This strategy involved not only augmenting input waveforms to expand the size of the training dataset, but also utilizing adaptive discriminator augmentation (ADA) to prevent the discriminator from overfitting and ensure stable training.

The performance of our system, characterized by a better FAD score, surpassed that of the baseline, which highlights the effectiveness of our approach. Our findings indicate that StyleGAN2-ADA, combined with a thoughtful data augmentation strategy, can serve as an effective solution for audio synthesis tasks even when faced with limited data.

The implications of this finding extend beyond the Foley sound synthesis task, potentially contributing to other audio domains like music and speech synthesis. Thus, this report offers a valuable contribution to the broader field of audio synthesis, particularly in scenarios constrained by data limitations.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions)

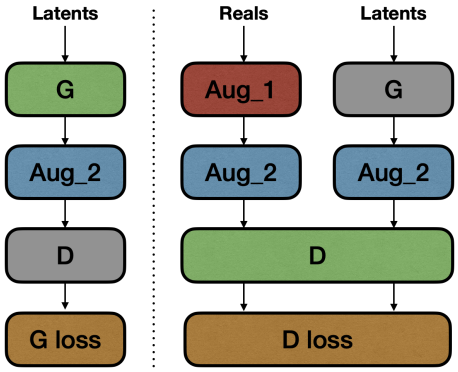

Figure 1: A flowchart of the StyleGAN2-ADA model, featuring the two-stage data augmentation procedure. The first stage, labeled as Aug_1, represents the audio augmentation pipeline. The modifications introduced at this stage are deemed acceptable to 'leak' into the learning process of the generator. Conversely, the second stage, Aug_2, depicts the augmentation process applied to the spectrograms. This latter stage of augmentation is designed to avoid 'leakage', thereby ensuring that these augmentations do not influence the generator while preventing the overfitting of the discriminator.

2. SYSTEM DESCRIPTION

2.1. Audio Augmentation

We utilized various audio augmentations due to the limited quantity of available data for the audio generation task. We observed that the seven classes of training data exhibited unique sound characteristics. Considering this diversity, we applied distinct augmentation techniques and specific parameters tailored to each class.

For each class, we applied three distinct types of augmentations, conducting each augmentation six times. This resulted in a total of

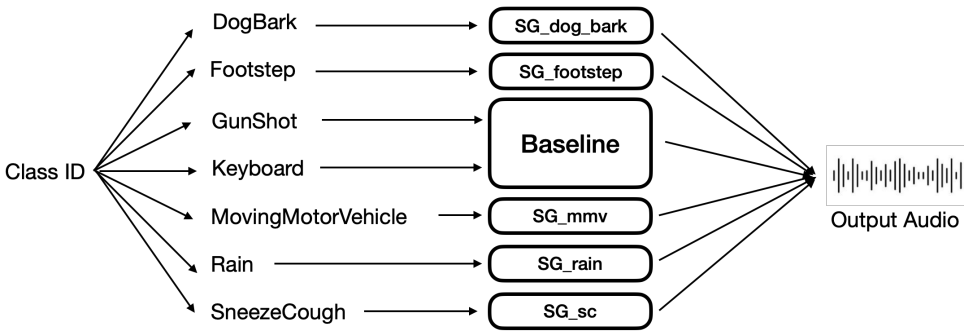

Figure 2: Schematic representation of the proposed ensemble system. The baseline model is utilized for the generation of the 'GunShot' and 'Keyboard' classes. For all other classes, individual StyleGAN2-ADA models, trained separately, are used. SG_[class_name] represents the StyleGAN2-ADA model for the corresponding classes.

	DogBark	Footstep	GunShot	Keyboard	MovingMotorVehicle	Rain	SneezeCough
Aug1	TimeShift RoomSimulator	PitchShift Shift	TanhDistortion PitchShift TimeShift	PitchShift TimeShift	PitchShift TimeShift Resample	PitchShift TimeShift	PitchShift TimeShift
Aug2	PitchShift TimeShift	TimeShift RoomSimulator	PitchShift TimeShift	TimeShift RoomSimulator	PitchShift TanhDistortion TimeShift	TimeShift RoomSimulator	TimeShift RoomSimulator
Aug3	TimeShift TimeMask RoomSimulator	TimeShift TimeMask	TimeShift TimeStretch	TimeShift TimeStretch	TimeShift TimeMask RoomSimulator	PitchShift TimeShift RoomSimulator	TimeShift TimeStretch RoomSimulator

Table 1: Augmentation techniques employed for each class. Each class was subjected to three different augmentation techniques, with each technique applied six times to augment the data.

19 augmentations per class, including the original data and the six repetitions for each of the three augmentation types ($1 + 3 \times 6$). Table 1 shows each augmentation pipeline applied to each class. We used `audiomentation`[1] library for audio augmentation.

The selection of augmentation types and parameters was carefully determined to preserve the unique sound characteristics of each class. This was evaluated by comparing the FAD scores between the original and augmented data. The objective of this evaluation was to maintain the integrity of each class's sound characteristics while expanding the volume of training data.

2.2. StyleGAN2-ADA

Our primary approach centered around the use of StyleGAN2-ADA, a model tailored for training generative adversarial networks, especially when data availability is limited. The core concept of StyleGAN2-ADA lies in its innovative approach to data augmentation, specifically designed to tackle the challenges of training generative adversarial networks (GANs) with limited data. This advanced architecture addresses the problem of overfitting, a common pitfall when working with sparse datasets.

In standard GAN training, data augmentation is applied to both the generator and discriminator. However, this often leads to 'leakage', where the generator learns to mimic the augmentation rather than the underlying data distribution, resulting in artificial and distorted outputs.

[1] `https://github.com/iver56/audiomentations`

StyleGAN2-ADA circumvents this issue with its concept of non-leaking augmentation. In this setup, data augmentation is solely applied to the discriminator, effectively increasing the diversity of the data it sees without influencing the data the generator is trained to reproduce. For example, this augmentation pipeline includes geometric and color (intensity) transformations of log-magnitude spectrograms. The generator, nevertheless, is exposed only to the original data, allowing it to focus on learning the true underlying data distribution.

Furthermore, StyleGAN2-ADA introduces an adaptive mechanism, adjusting the intensity of augmentation applied to the discriminator based on how well the training process is progressing. This dynamic adaptation provides a balance between preventing overfitting (with high augmentation) and preserving the quality of the generated samples (with low augmentation). Overall, this adaptive, non-leaking augmentation strategy is the key component that enables StyleGAN2-ADA to effectively learn and generate high-quality samples, even when trained on limited data.

2.2.1. Class-Specific Model Training

However, we noticed that the version of StyleGAN2-ADA conditioned on classes exhibited lower-than-expected performance. Therefore, we decided to adopt an alternative strategy that involved training separate models for each class. This methodology showed better performance, thus highlighting the benefits of utilizing individual models for each class, rather than a single model conditioned

	DogBark	Footstep	GunShot	Keyboard	MovingMotorVehicle	Rain	SneezeCough	Average
StyleGAN2-ADA	**6.92**	**4.45**	9.56	7.24	**11.53**	**5.20**	**1.49**	6.63
Baseline (reproduced)	11.20	7.47	**7.17**	**4.14**	13.76	12.51	2.65	8.41
Ensemble System	6.92	4.45	7.17	4.14	11.53	5.20	1.49	5.84

Table 2: Comparison of FAD scores of the StyleGAN2-ADA, the baseline, and the ensemble system for each class and the average across all classes. Given that we selected the best-performing model for each class in the ensemble system, the FAD score for the ensemble system corresponds to the best FAD score observed for each individual class.

on class ids.

2.2.2. Spectrogram Transformation and Audio Synthesis

In the initial stages, we transformed the waveforms into linear spectrograms exhibiting log magnitudes. This transformation was accomplished using the short-time Fourier transform (STFT) with the parameters `n_fft=512` and `hop_length=256`. Subsequently, these log-magnitude linear spectrograms were cropped along the time axis to establish the final data format of [256, 256]. We based our strategy on the official implementation of StyleGAN2-ADA, and followed its default training configuration.

During the generative phase, our first step was to sample log-magnitude linear spectrograms. These were then transformed to have a linear magnitude scale. Finally, the spectrograms were converted into waveforms using Griffin-Lim [9] algorithm.

2.3. Baseline System

The baseline system is composed of three distinct networks, each being separately trained using only the training data allowed to use in Task 7-B. These networks include PixelSNAIL [10], VQ-VAE [11], and HiFi-GAN [12].

PixelSNAIL, constituting the first network, functions as an autoregressive generative model that takes a class id, subsequently yielding a corresponding discrete time-frequency representation (DTFR). The second network, VQ-VAE, serves to transform the compressed DTFR into a Mel-spectrogram. Finally, the third network, HiFi-GAN, converts the Mel-spectrogram into a time-domain waveform.

2.4. Ensemble System

As instructed in the task description, we constructed an ensemble system that combines the StyleGAN2-ADA with the baseline system. The system with the best FAD score for each sound class is selected as the representative system for that particular class. The ensemble system consists of six subsystems, including the baseline system that generates two classes (GunShot and Keyboard), and the five different StyleGAN2-ADA systems that generate the other five classes independently.

3. RESULTS

In this section, we present our experimental results for the Foley sound synthesis task. We tested our ensemble system, comprising StyleGAN2-ADA and the baseline model, on the evaluation dataset. The FAD scores were computed using the official GitHub repository. [2]

As shown in Table 2, our ensemble system achieved an average FAD score of 5.84, demonstrating its superior performance over the baseline model (FAD score of 8.41) and individual StyleGAN2-ADA models (average FAD score of 6.63). This result highlights the effectiveness of our ensemble system, particularly when faced with the challenge of synthesizing high-quality sound from limited data.

The FAD scores for individual classes in Table 1 show that our StyleGAN2-ADA model achieved the best FAD scores across almost all classes, except for the class 'GunShot' and 'Keyboard.' For these two classes, the baseline model outperformed the StyleGAN2-ADA. This suggests that certain classes might be better suited to different models, reinforcing the importance of our ensemble approach.

The success of our system can largely be attributed to our two-stage data augmentation strategy, which successfully expanded the size of the training dataset and ensured stable training by preventing the overfitting of the discriminator. Moreover, the use of individual StyleGAN2-ADA models for each class, rather than a single class-conditioned model, also contributed to our system's superior performance.

4. CONCLUSION

Our research aimed to tackle the task of Foley sound synthesis, particularly focusing on the challenge of working with limited data. In this context, we put forward an ensemble system combining a baseline model with StyleGAN2-ADA. The experiments demonstrate the promising performance of our proposed system.

The results, as reported in the previous section, highlight the benefits of the ensemble system. Our system achieved a competitive average FAD score of 5.84 on the DCASE 2023 evaluation dataset.

We also noticed variations in the performance of the different models across classes. While the StyleGAN2-ADA model achieved superior FAD scores in most categories, the baseline model outperformed in the 'GunShot' and 'Keyboard' classes. This suggests that distinct classes might respond better to different models, hence underscoring the value of our ensemble approach in accommodating these variations.

A major factor contributing to the success of our ensemble system was our two-stage data augmentation strategy. This approach allowed us to effectively expand the training dataset, while also preventing the overfitting of the discriminator and ensuring more stable training of GAN. The strategy of using individual StyleGAN2-ADA models for each class, rather than a single class-conditioned model,

[2] https://github.com/DCASE2023-Task7-Foley-Sound-synthesis/dcase2023_task7_eval_fad

also played a significant role in the superior performance of our system.

In conclusion, this work has made substantial contributions to the field of Foley sound synthesis, especially under conditions of data scarcity. It has highlighted the viability of an ensemble approach that combines different models for different classes and has emphasized the importance of effective data augmentation strategies. Our findings offer valuable insights for future work in audio synthesis, paving the way for improvements in sound quality and model efficiency, which are particularly crucial when dealing with scenarios with limited data.

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] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in neural information processing systems*, vol. 33, pp. 12 104–12 114, 2020.

[3] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms." in *INTERSPEECH*, 2019, pp. 2350–2354.

[4] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[6] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[7] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[8] http://dcase.community/challenge2023/.

[9] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[10] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," in *International Conference on Machine Learning*. PMLR, 2018, pp. 864–872.

[11] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[12] 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*, vol. 33, pp. 17 022–17 033, 2020.