Foley sound synthesis at the DCASE 2023 challenge

Keunwoo Choi, Jaekwon Im, Laurie Heller, Brian Mcfee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, Shinnosuke Takamichi

FOLEY SOUND SYNTHESIS AT THE DCASE 2023 CHALLENGE

Keunwoo Choi[1*], *Jaekwon Im*[1,2*], *Laurie M. Heller*[3*], *Brian McFee*[4],
Keisuke Imoto[5], *Yuki Okamoto*[6], *Mathieu Lagrange*[7], *Shinnosuke Takamichi*[8]

[1] Gaudio Lab, Inc., Seoul, South Korea, {keunwoo, jaekwon}@gaudiolab.com
[2] KAIST, Daejeon, South Korea
[3] Carnegie Mellon University, USA, laurieheller@cmu.edu
[4] New York University, USA, brian.mcfee@nyu.edu
[5] Doshisha University, Japan, keisuke.imoto@ieee.org
[6] Ritsumeikan University, Japan, y-okamoto@ieee.org
[7] CNRS, Ecole Centrale Nantes, Nantes Université, France, mathieu.lagrange@ls2n.fr
[8] The University of Tokyo, Japan, shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

ABSTRACT

The addition of Foley sound effects during post-production is a common technique used to enhance the perceived acoustic properties of multimedia content. Traditionally, Foley sound has been produced by human Foley artists, which involves manual recording and mixing of sound. However, recent advances in sound synthesis and generative models have generated interest in machine-assisted or automatic Foley synthesis techniques. To promote further research in this area, we have organized a challenge in DCASE 2023: Task 7 - Foley Sound Synthesis. Our challenge aims to provide a standardized evaluation framework that is both rigorous and efficient, allowing for the evaluation of different Foley synthesis systems. We received 17 submissions, and performed both objective and subjective evaluation to rank them according to three criteria: audio quality, fit-to-category, and diversity. Through this challenge, we hope to encourage active participation from the research community and advance the state-of-the-art in automatic Foley synthesis. In this paper, we provide a detailed overview of the Foley sound synthesis challenge, including task definition, dataset, baseline, evaluation scheme and criteria, challenge result, and discussion.

Index Terms— Generative models, DCASE, sound synthesis

1. INTRODUCTION

Recent years have seen remarkable progress in generative models, with applications in a variety of fields including image generation [1], text generation [2], music generation [3, 4, 5], and sound generation [6, 7]. Models like these are capable of generating high-quality and diverse samples, and have been widely adopted in both academia and industry. In particular, sound generation has gained increased attention in recent years, with advances in sound synthesis and generative models enabling the creation of realistic and diverse audio content.

Sound synthesis plays a crucial role in enhancing the auditory perception of multimedia content, such as movies, music, and videos. Automatic or machine-assisted Foley synthesis has the potential to greatly streamline the process of creating these sound effects, freeing up time and resources for multimedia content creators.

To encourage further research and development in the field of automatic Foley synthesis, we proposed a challenge that aims to provide a standardized evaluation framework for different systems. Challenges have been shown to be an effective way to motivate the development of machine learning models, particularly in the early stages of a research area. We believe that the proposed Foley sound synthesis challenge can play a critical role in advancing the state-of-the-art in automatic Foley synthesis. This challenge was held as part of the international workshop on Detection and Classification of Acoustic Scenes and Events 2023 Workshop. The topics discussed in this introduction are also well covered in a proposal document [8].

2. PROBLEM AND TASK DEFINITION

We defined the problem of this challenge as 'category-to-sound' generation. The category is chosen in one of the selected seven categories - *dog bark*, *footstep*, *gunshot*, *keyboard*, *moving motor vehicle*, *rain*, and *sneeze/cough*. The sound is specified as a 4-second mono audio snippet with a sampling rate of 22,050 Hz.

As this was the first year of this challenge, we chose the input of the system to be a sound category rather than text input with natural language. This simplification was made to ease the organizing effort such as defining the problem and the evaluation scheme, collection of dataset, etc. We also intended this to lower the bar for participation, especially from academia, as category-based systems would require less data and computational resources than free text inputs. Similarly, limiting the problem to the seven categories clarified the subjective evaluation criteria. The seven categories were chosen so that i) the categories are useful for media creation, ii) it is feasible to collect a reasonable quantity of training/evaluation sounds with manual review, and iii) the generated sounds are easy to assess for the evaluators.

Despite this simplification, our intention for this challenge is to build towards generalizable and potentially useful approaches in the real world. In this regard, we specified the submitted systems should not simply copy-paste the existing sound, i.e., the systems should be generative, not retrieving.

Our goal is to motivate the development of new methods for Foley synthesis. At the same time, we recognize that volume of

*Equal contribution

data can be instrumental in qualitative improvements across many areas of ML. We therefore created two challenge tracks: one where participants are free to augment their training data with external sources (Track A) and the other where only the provided development dataset is allowed (Track B). To enhance the efficiency of the challenge, we also provided two pre-trained models, HiFi-GAN [9] and VQ-VAE [10], for Track B. These models were trained using the official dataset.

For a fair and correct evaluation, we required the participants to submit their model embedded in a Google Colab notebook template[1]. This provided an easy, familiar, and verifiable way for participants to share models while resolving any dependency issue for the organizers, at least within the time frame of the challenge.

3. OFFICIAL DATASET AND BASELINE

The development dataset used in this task consists of 6.1 hours of audio excerpts, each annotated with one of seven distinct sound classes: footstep, sneeze/cough, rain, dog bark, moving motor vehicle, gun shot, and keyboard. We selected the categories by considering an urban sound taxonomy [11]. The seven sound categories were selected evenly from each top-level group ('human', 'nature', 'mechanical'), except for 'music.' There is no overlap in the low-level groups between the sound categories.

We collected the data from UrbanSound8K [11], FSD50K [12], and BBC Sound Effects.[2] To select the appropriate audio clips for our challenge, we followed a two-step process. First, we gathered audio samples that were annotated with labels closely related to one of the seven sound categories. Second, to ensure consistency in the challenge, we pre-processed the audio to mono 16-bit 22,050 Hz and either zero-padded or segmented it to a length of 4 seconds, a duration found sufficient for human recognition of class and audio quality. This pre-processing step was applied before selection, as the audio events comprise only a small portion of the total audio length.

To ensure the quality of the dataset, we carefully selected the audio clips for each category based on their relevance, variety, and clarity. One organizer manually selected the collection of excerpts, each of which was verified by a different organizer to ensure accuracy and clarity. Overall, we selected 5,550 labeled sound excerpts, with the number of sounds per category ranging from 681 to 900.

We divided the dataset into a development dataset and an evaluation dataset. Although the number of audio samples varies across sound classes, we ensured that the evaluation set had a consistent number of 100 audio samples per category. This decision was made to ensure that the evaluation set had a diverse range of sounds and was not too small. We also made sure that the partitions were stratified, so no source recording provided clips in both the development and evaluation sets, even if there were multiple excerpts from the same longer recording.

As a baseline system, we implemented a model [13] composed of three independently trained modules: PixelSNAIL [14], VQ-VAE [10], and HiFi-GAN [9]. The first module, PixelSNAIL, is an autoregressive model that maps a sound category input to a time-frequency representation. The second module, VQ-VAE, transforms the PixelSNAIL output into a Mel spectrogram through a compressed, latent vector encoding. The final module, HiFi-GAN, transforms the VQ-VAE output (Mel spectrogram) into a time-domain digital audio signal.

We selected the model as our baseline system for the following reasons. First, the modules were assigned the reconstruction task and the generation task separately, enhancing the whole architecture's explainability. Second, the participants were allowed to reuse some of the modules. Since each module was trained independently, improving the performance of the system can be achieved by modifying the structure or scheme of specific modules while keeping the remaining modules unchanged.

4. EVALUATION

Even for objective tasks such as classification and detection tasks, it is challenging to provide unambiguous annotations and unbiased evaluation metrics. Multiple evaluation metrics may be necessary, but it can complicate the ranking of participants. [15]. With generative tasks such as the one considered in this challenge, the problem is even more difficult, as the produced data is not a set of labels, but audio, whose qualities must be assessed. This matter is far from being solved and is currently undergoing active research [16]. Recognizing this as a challenge, we opted for a pragmatic combination of objective and subjective evaluation protocols as proposed in [8].

In detail, we chose a two-step procedure. The first step considers objective metrics to get a first ranking of the proposed systems. Due to the constraints on human listening time for subjective ratings, in each track, only the top four entries were then considered for the second step with a subjective evaluation.

We decided to measure the following qualities:

1. **Perceptual Audio Quality**: The degree of clarity of sound, free from any artifacts, fuzziness, degradation, distortion, and noise.
2. **Fit-to-category**: The degree to which a sound is recognized as belonging in the intended category.
3. **Diversity**: The degree to which a system is able to produce a diverse set of sounds.

Evaluation of the above qualities typically involves high-level perceptual and cognitive processing by humans and thus cannot be evaluated by simple computational means. For this reason, we chose to complement the objective evaluation with subjective metrics. Although essential, subjective evaluation comes with some constraints. Humans can give different ratings depending upon the context of a sound they hear, and can experience fatigue. For the latter reason, only a subset of audio samples can be presented for subjective rating. To make the sure the context is similar across raters (and potentially, across future contests), the audio samples should include some "anchors," i.e. sounds which clearly have a very low and/or high quality; anchors help to psychologically anchor the ratings and also serve as a check on the quality of the rater [17].

4.1. Step 1: Objective Evaluation

We adopted Fréchet Audio Distance (FAD) [18], a reference-free, lower-the-better, evaluation metric. FAD calculations were performed for each category. Systems were then ranked based on the average FAD across seven categories, and only the 4 top-performing systems per track were considered for the second step, due to time limitations of the subjective evaluation.

[1] https://colab.research.google.com

[2] https://sound-effects.bbcrewind.co.uk/

4.2. Step 2 : Subjective Evaluation

The subjective evaluation was operated in two steps. The first was an online survey that measured the fit-to-category and perceptual audio quality. The fit-to-category asked the listener to use their general notion of the sound category and was not restricted to referencing the exact sounds in the development set, nor was it based on the number of sound events in a file. These tasks were performed on 20 sounds from each category, along with a set of anchors taken from the development set and baseline system.

The selection of the 20 representative sounds was done as follows. OpenL3 embeddings of all the samples were computed and a k-means clustering with $k = 20$ was conducted on them [19]. The 20 "medoid" representative sounds are selected as the ones with the smallest Euclidean distance to the centroid in the embedding space.

For each representative sound, the rater was asked to rate two categories, perceptual audio quality and fit-to-category, as defined in Section 4. Raters selected among 11 levels from 0 to 10. For each category, interpretation guidance was provided for minimal and maximal rating, with 10 being the top of an absolute scale (the best possible, as opposed to the best of this contest). The rater could listen to the sound as much as needed, and one response on each scale was required before evaluating the next sound.

Before rating a category, the rater listened to 6 representative sounds of the category selected from the development set. To provide some normalization across listeners, the 'anchor' sounds for each category were embedded in the main test at random locations. The high and low quality/fit anchors, respectively, were hand-picked from the evaluation set and our baseline system. These sounds were not identified as anchors in the survey. Entries from Track A and B were intermixed so that their relative quality would be apparent, even though the competition rankings are separated within each track. The order of trials was counterbalanced across test conditions.

4.3. Execution

All of the challenge participants performed the ratings on perceptual audio quality and fit-to-category for 4-7 categories, for a total duration of about 3-6 hours. After each category, the listener could take a break.

All participants listened to the same sounds. Thus, participants who submitted one of the finalist systems actually rated sounds from their own systems but their ratings were removed by the organizers before computing results. This allowed us to streamline the rating system and will allow us to check for potential rating bias.

Rating at least 4 categories was required to be eligible for a prize. This requirement ensured that we had a fair distribution of teams doing ratings and enough ratings per sound. Some organizers did ratings as well. This resulted in 10-15 ratings per sound. 93 separate category ratings were completed, which totals to at least 47 hours of time. Two of the 93 ratings were omitted at the start of the data analysis because they mis-rated 5 or more of the 12 quality-check trials (in both cases, giving a rating of low quality & good fit to an anchor sound that had a high quality & poor fit, indicating that they had confused the two scales). The anchors that had low quality tended to get a poor fit rating, so we did not use those as an exclusion criterion. Appropriate ratings were given for anchors that had high quality & low fit, high quality & good fit, and low quality & poor fit (4 of each type).

To validate the protocol as well as the software stack, a pilot study was carried out with the outputs of the baseline system in which the listeners were the organizers. During the evaluation phase, the test was advertised to relevant mailing lists. In this version, only one category was proposed to the listener, yielding a manageable duration of about 30 minutes, and using a scheme based on last names to distribute the ratings across categories.

Finally, we also performed a subjective test on Diversity. Diversity is a "set-based" quality, meaning that a set of generated audio files are mandatory for measuring it. For this reason, Diversity could not be evaluated within the above discussed listening test, whose stimuli are considered independently.

For each system and each category, an organizer who did not participate in the ratings generated a continuous audio file sequencing the 20 representative sounds per system. Each file was given a name specifying the category and an obfuscated version of the system id. The diversity rating task took about 1.5 hours. Four other organizers, blind as to which systems they were rating, rated the diversity of the sounds per file from 0 (all the sounds appear to be identical) to 10 (Extremely large range of sounds).

Considering that 1) diversity may be less important than quality and fitness and 2) this quality has been not as rigorously tested in this edition of the challenge as the two other qualities, organizers decided in advance that the diversity ratings were weighted half as much as each of the audio quality and category fit ratings.

5. RESULTS

We provided to the participants a colab notebook as a starting point to implement their submission and we received 42 systems in total, including 11 systems in Track A [20, 21, 22, 23, 24] and 31 systems in Track B [25, 26, 27, 28, 29, 30, 31, 22, 32, 33, 34, 35, 36]. We removed disqualified submissions that failed to run on standard colab instances in a reasonable period of time. Before disqualification, we had a 4 days review period where we exchanged with the participants for potential fix. We made sure that the changes were trivial bug fixes and not changes in parametrization of the submitted systems.

With the remaining systems, we generated 700 audio samples from 9 and 27 systems for tracks A and B, respectively. The audio samples are available online[3]. A total of 36 working systems were submitted by 17 teams.

As all the details (FAD scores per category, subjective test results on audio quality, fit-to-class, and diversity) were released on the DCASE official website,[4] we focus on analyzing the evaluation results in this section.

In Fig. 1, the FAD scores of 17 systems are plotted. The (x, y) position represents the average FAD score computed on the development set (FAD-Dev) and the evaluation set (FAD-Eval), respectively. The width and height of each rectangle represents the (scaled) standard deviation over 7 categories for both sets, respectively.

First, most of the systems show better (lower) FAD-Dev than FAD-Eval, with the exception of [20]. This is expected, as the training would be at least partially based on the development set. Second, it turns out that FAD-Dev is a noisy measure to predict FAD-Eval. This is neither surprising nor negative as the organizers did not want the actual objective measure (FAD-Eval) to be precisely predictable. However, this could be discouraging for the partici-

[3] https://zenodo.org/record/8091972

[4] https://dcase.community/challenge2023/task-foley-sound-synthesis-results

pants; the discouragement could be reduced via future use of a public leader board. Third, comparing the top systems of track A and B, several systems in track B showed better performance on FAD-Dev, but not in FAD-Eval. This shows the difficulty of training a system with the limited amount of data permitted in track B.

In Fig. 2, the top 8 systems and the baseline system are plotted by their final ranking determined by a listening test as well as FAD-Eval and FAD-Dev. On the left, the scatterplot shows the importance of subjective tests. The Spearman's rank correlation coefficient of the ranking by FAD-Eval and the final ranking is only '0.238'. On the right, with FAD-Dev, the coefficient is somewhat higher, '0.524'.

We established that subjective perceptual sound qualities were not entirely predicted by objective FAD scores. In addition, we established that the three perceptual metrics were interrelated, but each had a unique contribution. Within each category, the correlations between average rating scores of finalist systems of audio quality and category fit were very strong (average across all categories was $r = 0.98$); however, when quality & fit ratings from individual trials were correlated within each category, the average correlation was less extreme ($r = .75$), showing that raters were not giving identical answers to both questions. Our anchor trials showed that the raters did know how to distinguish the two qualities, because they appropriately rated the category-inappropriate sounds with good audio quality. On the other hand, we also found that raters gave all-around low ratings to the category-appropriate sounds with poor audio quality. Because sound recognition was essential for judging category fit, it is plausible that good audio quality was required before being able to give a high category fit rating. Average diversity (within each category, across finalist systems) had a strong relationship to category fit ($r = 0.70$); nonetheless, half of the variance in diversity ratings was independent of quality/fit.

The perceptual ratings of the quality/fit of all the systems were plausible, with the highest average ratings obtained for the sounds from the development set, and the lowest for our baseline system. The submitted systems had intermediate ratings, showing that there is room for improvement in this challenge.

To summarize, there are important mismatches between the objective evaluation done by the participants (FAD-Dev) or by the organizers (FAD-Eval) as well as the listening test (final ranking). This justifies two of our choices for the evaluation scheme: i) receive submissions in the form of a system (code) instead of sounds, and ii) run a subjective evaluation.

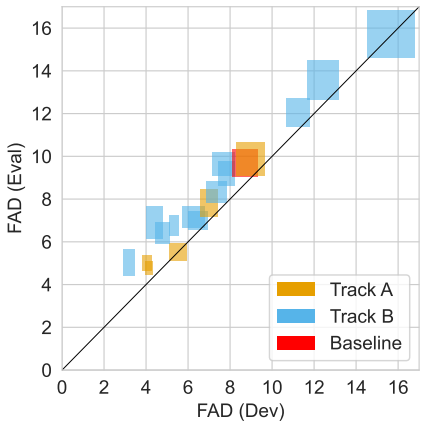

Figure 1: FAD Scores on the development set vs the evaluation set, computed on the 17 submitted systems and the baseline system.

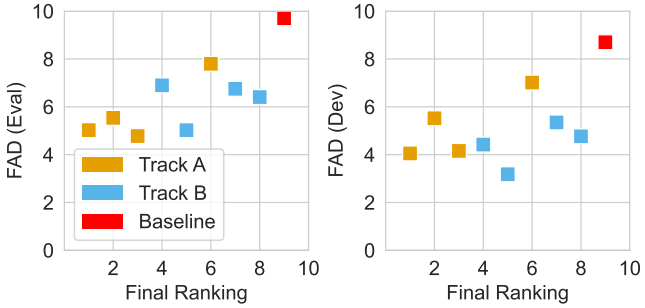

Figure 2: FAD scores on the development set and the evaluation set vs. the final ranking determined by the listening tests.

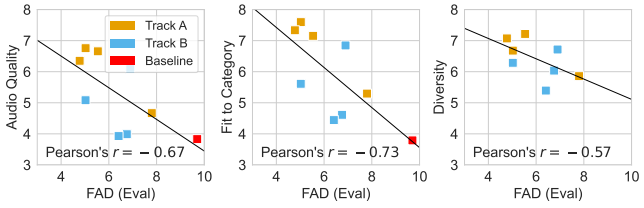

Figure 3: Relationships between objective measure (FAD-Eval) and subjective tests.

6. CONCLUSION

In this paper, we have presented a challenge for automatic Foley sound synthesis aimed at promoting further research and development in generative AI for sound. We have provided a detailed overview of the challenge, including task definition, dataset requirements, evaluation criteria, a baseline implementation, and analysis of the results. Through this challenge, we believe we have achieved our goal — to encourage active participation from the research community and advance the state-of-the-art in automatic Foley synthesis. Although it was the first year of the challenge, we received substantial submissions in both of the tracks. We also performed the generation and evaluation of the submitted systems successfully.

In both tracks, best performing systems were based on deep learning, with a sequence of a diffusion model for spectrogram generation and HIFI-Gan [9] for phase reconstruction.

There have been difficulties as well. Our analysis showed the necessity of performing a subjective evaluation and running inference by ourselves. Unfortunately, both are costly; in total, about 47 hours were spent for the evaluation of 8 systems and about 471 A100 GPU hours for the inference. We plan to release all the generated sounds as well as their subjective/objective scores soon, hoping to enable more analysis and even subjective quality prediction models based on the data.

In the future, we hope that the standardized evaluation framework provided by this challenge will help to facilitate comparisons between different Foley synthesis systems. It already seems apparent that a more complicated Foley sound synthesis can be possible in the near future with text-input, video-input, etc. We hope our challenge will ultimately lead to the development of more effective and efficient techniques.

7. ACKNOWLEDGEMENTS

We appreciate the support by Gaudio Lab, inc. for covering the computing resources to run the submitted systems. We are also thankful for all the raters who did the subjective evaluation.

8. REFERENCES

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[2] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] M. Pasini and J. Schlüter, "Musika! fast infinite waveform music generation," in *ISMIR 2022 Hybrid Conference*, 2022.

[4] 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.

[5] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations (ICLR)*, 2019.

[6] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[7] 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.

[8] K. Choi, S. Oh, M. Kang, and B. McFee, "A proposal for foley sound synthesis challenge," *arXiv preprint arXiv:2207.10760*, 2022.

[9] 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.

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

[11] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[12] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

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

[14] 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.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[16] J. Vadillo and R. Santana, "On the human evaluation of universal audio adversarial perturbations," *Computers & Security*, vol. 112, p. 102495, 2022.

[17] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[18] 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.

[19] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[20] M. Kang, S. Oh, H. Moon, K. Lee, and B. S. Chon, "FALL-E: Gaudio foley synthesis system," Tech. Rep., June 2023.

[21] S. Fan, Q. Zhu, F. Xiao, H. Lan, W. Wang, and J. Guan1, "Foley sound synthesis with AudioLDM for dcase2023 task 7," Tech. Rep., June 2023.

[22] J. Lee1, H. Nam, and Y.-H. Park, "VIFS: An end-to-end variational inference for foley sound synthesis," Tech. Rep., June 2023.

[23] R. Scheibler, T. Hasumi, Y. Fujita, T. Komatsu, R. Yamamoto, and K. Tachibana, "Class-conditioned latent diffusion model for DCASE 2023 foley sound synthesis challenge," Tech. Rep., June 2023.

[24] Y. Yuan, H. Liu, X. Liu, X. Kang, M. D.Plumbley, and W. Wang, "Latent diffusion model based foley sound generation system for dcase challenge 2023 task 7," Tech. Rep., June 2023.

[25] S. Huang, J. Bai, Y. Jia, and J. Chen, "Jless submission to dcase2023 task7: Foley sound synthesis using non-autoagressive generative model," Tech. Rep., June 2023.

[26] W.-G. Choi and J.-H. Chang, "HYU submission for the dcase 2023 task 7: Diffusion probabilistic model with adversarial training for foley sound synthesis," Tech. Rep., June 2023.

[27] C.-W. Bang, N. K. Kim, and C. Chun, "High-quality foley sound synthesis using monte carlo dropout," Tech. Rep., June 2023.

[28] Y. Chung, J. Lee, and J. Nam, "Foley sound synthesis in waveform domain with diffusion model," Tech. Rep., June 2023.

[29] H. C. Chung, Y. Lee, and J. H. Jung, "Foley sound synthesis based on GAN using contrastive learning without label information," Tech. Rep., June 2023.

[30] P. Kamath, T. N. Islam, C. Gupta, L. Wyse, and S. Nanayakkara, "Dcase task-7: StyleGAN2-based foley sound synthesis," Tech. Rep., June 2023.

[31] K. Kim, J. Lee, H. Kim, and K. Lee, "Conditional foley sound synthesis with limited data: Two-stage data augmentation approach with stylegan2-ada," Tech. Rep., June 2023.

[32] A. Pillay, S. Betko, A. Liloia, H. Chen, and A. Shah, "DCASE task 7: Foley sound synthesis," Tech. Rep., June 2023.

[33] A. Qi, "Auto-bit for DCASE2023 task7 technical reports: Assemble system of bitdiffusion and PixelSNAIL," Tech. Rep., June 2023.

[34] H. Zhang, K. Qian, L. Shen, L. Li, K. Xu, and B. Hu, "From noise to sound: Audio synthesis via diffusion models," Tech. Rep., June 2023.

[35] T. Wendner, P. Hu, T. Jadidi, and A. Neuhauser, "Audio diffusion for foley sound synthesis," Tech. Rep., June 2023.

[36] Z. Xie, X. Xu, B. Li, M. Wu, and K. Yu, "The X-LANCE system for DCASE2023 challenge task 7: Foley sound synthesis track b," Tech. Rep., June 2023.