An Audio Dataset And Hierarchical Taxonomy For Post-Disaster Scenes

Tuan-Anh Pham (1), Seong-Hun Park (1), Jong-Hoon Lee (1), Sang-Jin Han (1), Jungyu Choi (2), and Sungbin Im (2)

(1) Department of AI Laboratory, MOADATA, Gyeonggi, South Korea
(2) Soongsil University, Seoul, South Korea

Abstract - In the aftermath of disasters, timely response and rescue operations are crucial for saving lives. However, the post-disaster environment often presents challenges such as structural damage, debris, and hazardous conditions, where visual cues are obstructed or unavailable. In such situations, effective analysis of audio data can provide critical insights and situational awareness to guide response efforts. However, existing audio datasets for disaster scenarios are limited and lack a comprehensive taxonomy to represent the diverse sound events encountered in these environments. In this paper, we present an audio dataset specifically curated for disaster scenes along with a hierarchical taxonomy to categorize possible related sound events. The audio dataset comprises audio clips organized into 31 event categories. The first 30 categories each contain 50 audio clips, while the remaining category contains 1,000 clips. To augment this dataset, we leverage a generative model AudioLDM to synthesize an additional 1,000 audio samples for each event category. Instead of relying on text prompts, we employ 50 reference clips from each category as the condition input to generate target audio samples, resulting in a total of 30,000 samples. To select high-fidelity generated samples, we propose a simple two-stage process that combines the Frechet Audio Distance (FAD) metric and a finetuned pretrained audio model PANNs. To verify the effectiveness of the audio selection process, we randomly selected 100 audio clips from the generated data, finetuned PANNs, and classified the reference audio clips. The finetuned PANNs achieves 89.27% in top-1 accuracy and 95.99% in top-2 accuracy.

1 INTRODUCTION

For the survivor localization and rescue in post-disaster scenes, it is challenging whether the disaster is natural or man-made. Normally, survivors might be trapped under piles of debris or rubble once buildings or construction collapses when disasters occur. Unfortunately, visual information is often unavailable and obstructed in these situations, thus it is impossible to localize survivors using images. Without using visual cues, acoustic signals could be great sources as alternatives for rescue teams to search and localize potential victim locations. Fundamentally, acoustic signals are sound waves that can penetrate through and bounce on a variety of materials, the sound waves might carry essential information to support the communication between survivors and rescue teams.

To rescue survivors in the post-disaster, many technologies have been studied and employed, each technology has its own strengths and weaknesses. Radar sensors (Nezirovic, 2010; Uzunidis et al., 2023) and mobile robots (Latif et al., 2016; Bhondve et al., 2016) can detect trapped victims, but their effectiveness is much affected by the density and composition of debris. UAVs and aerial images (Antoniou, Potsiou, 2020; Song et al., 2022; Amit et al., 2016) provides overhead views, but they struggle to retrieve information inside the debris. By analyzing changes in oxygen and carbon dioxide, gas sensors (Guntner et al., 2018; Imlauer et al., 2014) can identify human breath, but their accuracies can be degraded if the environmental conditions change. Optical sensors (Mäyrä et al., 2013) or infrared images (Dong et al., 2021) can be used to detect invisible lights from survivors, but the line of sight may be obstructed because of collapsed structures. Unlike previous technologies, acoustic signals (Wilson, Makris, 2006; Ekpezu et al., 2021; Rouet-Leduc et al., 2017; Hoshiba et al., 2017; Ascione et al., 2012) do not

rely on line of sight or visible cues, they can propagate or penetrate through various materials and reach acoustic sensors or rescue teams. Several information from survivors might be conveyed when analyzing the sound waves such as voices, movements, emotions, condition of survivors, or breathing patterns.

To the best of our knowledge, an audio dataset that capture diverge range of disaster-related sound events is not available. Therefore, in this paper, we propose a novel audio dataset for post-disaster scenes, this dataset contains both real-world and synthetic audio clips. We expect that this dataset could be a foundation for developing automated audio-based systems, in which tasks like survivor localization, sound event detection, and disaster monitoring are employed. Furthermore, to provide a comprehensive analysis of disaster events, we present a hierachical taxonomy that categorizes and organizes post-disaster sound events based on three main groups.

The dataset is organized into 31 sound event categories, in which the first 30 categories contain 50 audio clips each and the remaining category contains 1,000 audio samples. As the number of samples in each category is limited, we explore an audio augmentation method using a state-of-the-art audio generative model, AudioLDM (Liu et al., 2023). We propose a simple yet effective two-stage process for selecting high-fidelity generated audio samples. The filtering process is primarily built based on the combination of an audio evaluation metric Frechet Audio Distance (Kilgour et al., 2019), and the pretrained audio model PANNs (Kong et al., 2020).

Our contributions can be summarized as follows: (1) a post-disaster audio dataset containing possible related events, (2) a hierarchical taxonomy for post-disaster scenes, (3) an audio augmentation method combining FAD and large-scale pretrained audio model PANNs. We introduce the details of the dataset, taxonomy, data collection, and filtering process in section 2. Section 3 explores the effectiveness of a two-stage audio selection process by classification results of real-world audio. Finally, valuable insights of this paper are summarized in section 4.

2 TAXONOMY AND DATASET

2.1 Post-disaster Sound Taxonamy

To support future research on post-disaster scenes, we propose a hierarchical taxonomy that organizes and classifies possible related events, as shown in Figure 1. This taxonomy is designed with a multi-level hierarchy, the top level contains three groups: survivors, animals, and the surrounding environment. The taxonomy is built from a total of 33 sound events presented as the leave nodes in the hierarchy. Among the three groups, the survivor group is the main focus, this group is about sound events directly related to victims, expressing all possible activities and actions of victims such as emotions, voices, movements to ask for help, or using things like cellphone bells to signify signals to rescue teams.

Apart from the survivor's group, animals such as dogs and cats might present in accident sites, we chose dogs and cats as the representatives because of their popularity, thus other animals can be added to this group. For the surrounding environment, this group indicates events occurring in and out of collapsed areas, providing contextual information about disaster scenes. Some events in this group could be considered as noises such as safety alarms or sounds of equipment, but other events may be useful for disaster scene analysis like water flowing or wind blowing.

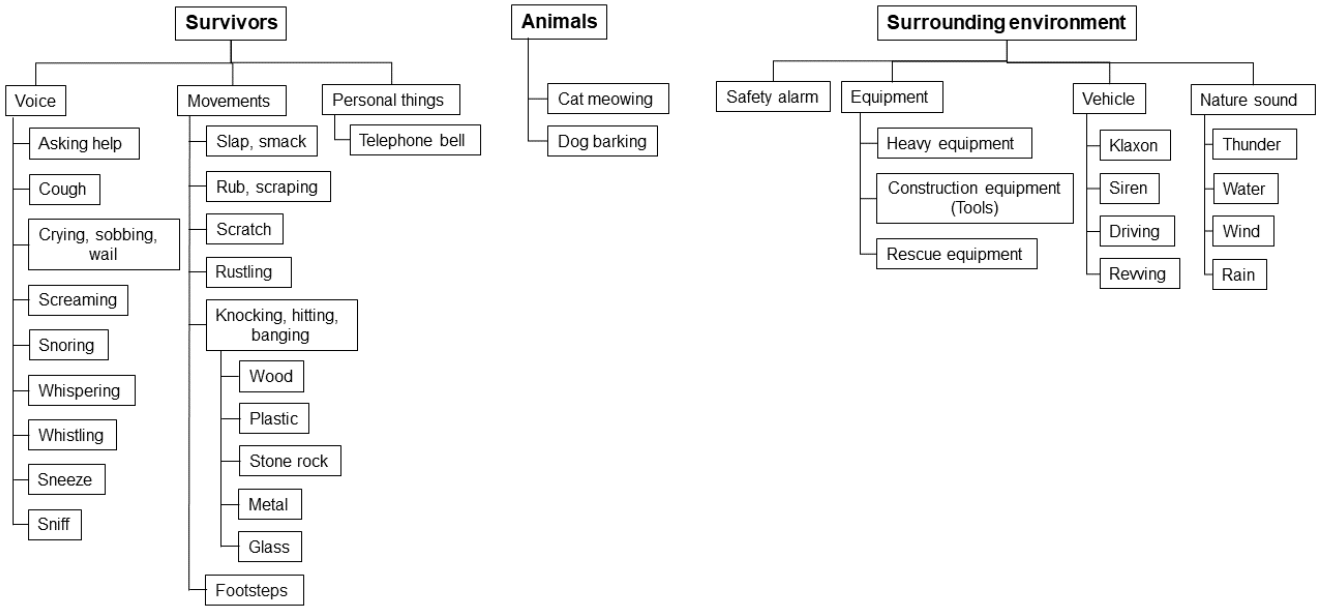

Figure 1 – Hierarchical taxonomy of post-disaster events

2.2 Dataset collection

In this dataset, we collect only 31 post-disaster events which can be referred to in the taxonomy except for "Rescue equipment" and "Footsteps". We first downloaded all of the related audios using target categories or their synonyms as keywords from:

- AudioSet (https://research.google.com/audioset/)
- FreeSound (https://freesound.org/)
- EpidemicSound (https://www.epidemicsound.com/)
- AI-Hub (https://www.aihub.or.kr/)

We use AudioSet as the primary source to collect the interest data. To filter out audio clips containing noise and unrelated events, we check if an audio sample matches the three following conditions: (1) the target event must exist in the audio; (2) the target event must not be overlapped with other events. If they overlap, they should happen naturally in real world; (3) the event duration must not be shorter than 2 seconds. We applied the pretrained PANNs model without tuning to check the event presence and duration. Then, we listen to the audio to assess the quality and overlapping events. We chose PANNs because its performance and it has been trained on large-scale datasets with 527 sound classes the same number of classes in AudioSet.

If any categories do not have enough 50 audio clips, we will collect more data from other sources. For the category "Asking help", this category was collected from AI-Hub and is about survivor speech with different emotions and recorded in Korean. Some sentences such as "도와주세요", "살려주세요" mean "Help me" or "Help me, please". We only kept speech events while excluding other events that occurred in the audio clips. Moreover, some events in the group "Knocking, hitting, banging" are not available in the above sound sources, thus we recorded them ourselves using different materials like "wood", "plastic", "stone, rock", "metal", and "glass".

In cases of the number of clips in any category is still less than 50 after searching from the above sound sources, we employed AudioLDM (Liu et al., 2023) using audio samples as condition inputs and (Liu et al., 2024) using texts as inputs to generate the remaining number of clips. Note that those generated clips were selected manually,

making sure they are realistic and relevant to the given categories. We have separated and marked the labels of audio clips are collected from real-world sound sources and generated ones in the dataset for supporting many options for the data usage of readers.

In summary, this dataset has a total of 2,500 audio clips across 31 sound events, in which 1,000 clips are from "Asking help" category and 1,500 clips are from the remaining 30 events. The sampling rate of each clip is 16,000 Hz, the shortest and longest duration of audio records are 1.93s and 27.33s, respectively. We did not count the audio duration of recordings from "Asking help" because of the very short duration in speech events. The total duration of 2,500 recordings is approximately about 4.5 hours.

2.3 Audio Augmentation

As the size of each category is 50 clips which is limited, we augmented more audio data from the reference dataset to expand the diversity and the size of the dataset. There are several traditional audio augmentation techniques that can be used to generate audio data such as time stretching, time shifting, cropping, resampling, adding noise, etc. However, applying these techniques may come with unexpected results like quality degradation, low diversity, and unnatural sounds. Additionally, generating audio data using text-to-audio (Liu et al., 2023; Liu et al., 2024; Agostinelli et al., 2023; Kreuk et al., 2023; Huang et al., 2023; Yang et al., 2022) or audio-to-audio (Liu et al., 2023) models becomes more popular in recent years, audio types could be speech, music, or specific sound effects. In this paper, we chose the audio-to-audio model AudioLDM to generate synthetic data because it is impossible to describe fully the details of an audio by just using texts. Therefore, all reference audio clips will be used as inputs to synthesize audio samples.

In this section, we propose a simple two-stage process for selecting high-fidelity generated audio samples automatically, as shown in Figure 2. In the first stage, individual FAD scores for every generated sample are calculated, in which an individual score is computed by using an audio and its corresponding set of reference clips (same category) (Gui et al., 2023). Then, for each category, individual scores of all generated samples in the category are collected and the 95th percentile is chosen as the threshold. Next, any samples whose FAD scores are larger than their corresponding thresholds are removed. The 30 FAD thresholds will be used in case we generate additional audio samples. In the data augmentation process, we excluded "Asking help" category for generation because the linguistics of audio speech can not be reproduced using the AudioLDM model. In the second stage, the pretrained audio model PANNs will be finetuned *N* times resulting *N* classifiers, *N* was selected as 3 in this experiment. If *N* is too small, the quality of generated samples might be insufficient. Otherwise, if *N* is too large, it will be too strict to select samples and we need to generate more audios. Note that PANNs models are finetuned using reference audio clips, that make the generated data more realistic and close to what we expect.

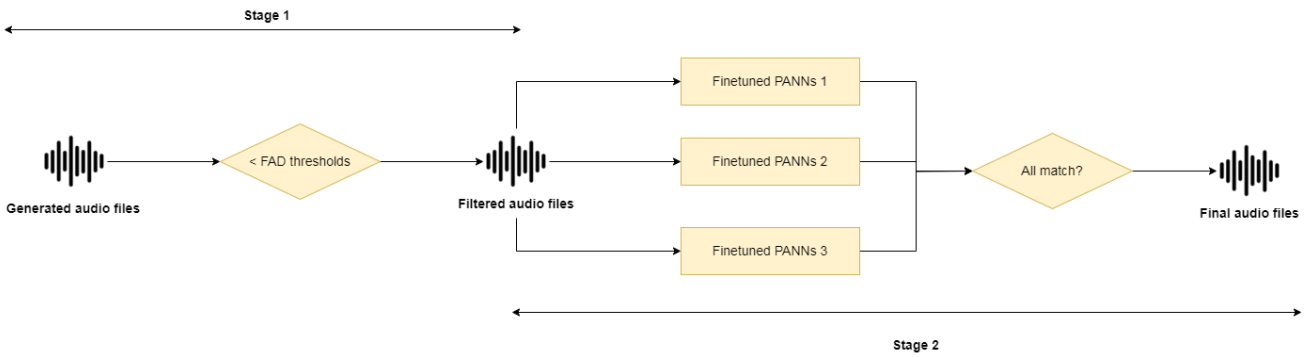

Figure 2 – Two-stage process for high-fidelity audio filtering

In AudioLDM model, two parameters affects significantly in the diversity of generated clips, which are "guidance scale" and "random seed". "Guidance scale" determines the weight of the condition information, the higher the more relevant to the condition input of the given audio. However, the diversity will be reduced when the "guidance scale" increases. We tested with both two parameters and we found that "guidance scale" did not contribute much to the diversity content of generated audios, while "random seed" had much impact on generating varieties of the given audio. To achieve 1,000 high-fidelity samples, we generated 2,000 to 3,000 clips, then applied the two-stage audio selection process to filter the desired number of audios.

3 SOUND EVENT CLASSIFICATION

To explore the quality of generated samples using the two-stage selection process, we finetuned a classification model from the pretrained PANNs with the generated data and then classifying reference audios. In this experiment, we only evaluate 30 categories except for "Asking help". For each category, 100 samples will be randomly selected, resulting in 3,000 data points for finetuning. This evaluation step aims to verify both the diversity and fidelity of generated audio samples. We use accuracy as the evaluation metric and explore the difference in performance evaluation between top-1 and top-2 accuracy. The evaluation results are interpreted as follows: the lower the score of a category the less effectiveness that generated samples capture various sound patterns from the real-world data, and vice versa.

The experiments are conducted on a server with Ubuntu OS 22.04.3, NVIDIA GeForce RTX 4090 24GB, and 256GB of RAM. The environment is set up with Anaconda 2023.09-0, Python 3.8.19, Torch 2.2.2. The code for finetuning PANNs model is available at https://github.com/qiuqiangkong/panns_transfer_to_gtzan. The classifier is finetuned in 10,000 steps, learning rate is 1e-4, and batch size is 32.

Table 1 presents the top-1 accuracy, top-2 accuracy, and false classification (false negatives) of 30 event types in post-disaster scenes. We sorted event categories by top-1 accuracy in descending order. In the false classification column, we analyze and list the categories in which most generated samples are falsely classified using Top-1 accuracy.

Table 1 – Classification results of 30 post-disaster events of reference dataset

Category	Top-1 accuracy	Top-2 accuracy	False classification
whistling	100	100	N/A
klaxon	100	100	N/A
telephone bell	100	100	N/A
wood	100	100	N/A
slap smack	98	98	metal
wind	98	100	driving
rain	98	100	wind
metal	98	100	stone rock
siren	96	100	engine, heavy equipment
sneeze	96	96	stone rock
cat	94	100	screaming
whispering	94	98	scratch
snoring	94	98	driving
driving	94	98	heavy equipment
dog	92	96	screaming
safety alarm	92	100	telephone bell
sniff	92	94	snoring
screaming	92	96	siren

heavy equipment	92	96	engine
engine	92	96	rain
scratch	86	96	rub
thunder	84	98	wind
plastic	84	100	wood
crying	82	100	screaming
rub	80	92	scratch
tools	78	82	engine, heavy equipment
stone rock	76	86	metal
water	72	98	rain
cough	72	88	sneeze
glass	52	72	metal, stone rock
Average	**89.26**	**95.93**	**N/A**

In table 1, the average top-1 accuracy achieves 89.26% which is 6.67% lower than the average top-2 accuracy 95.93%. The difference between the two accuracies is quite significant, it might show the similarity between events in top-2. In top-1 accuracy column, 20 over 30 events are classified with accuracies of more than 90% and this number increases to 26 events in the top-2 accuracy. It can be seen that the accuracies of material categories are quite low such as "glass", "stone rock", and "plastic". There are some reasons that might cause these problems: (1) we recorded these events ourselves and AudioLDM has not been trained with these events; (2) the sound of "glass" and "metal" or "glass" and "stone rock" when knocking are difficult to distinguish even with human ears.

There are some confusions among events which can be analyzed in the False classification column, and these false classifications are quite reasonable in real-world. For example, "safety alarm" and "telephone bell", "heavy equipment" and "engine", "scratch" and "rub", "crying" and "screaming", "stone rock" and "metal", "cough" and "sneeze", "glass" and "metal". The details of event classifications can be referred to in the confusion matrix, precision, recall, and F1 score in our data repository.

4 CONCLUSION

This paper presents a novel audio dataset for post-disaster sound events and a comprehensive hierarchical taxonomy. To explore the effectiveness of using the generative audio model for audio augmentation, we employed AudioLDM using realistic audio samples as condition inputs. The accuracy scores from classification experiments demonstrate the promising of using the generative model for data augmentation. Although the two-stage process for filtering high-fidelity is simple, the effectiveness is shown in the classification experiment. We observed that if AudioLDM model cannot generate high-quality audio if the sound events have not been trained in the model. Moreover, AudioLDM cannot reproduce the linguistic in the conditioning audio inputs. We publish our audio dataset and classification results on Github repository https://github.com/tuananhphamds/PostDisasterDataset.

ACKNOWLEDGEMENTS

This work was supported by the Technology Innovation Program (20025005, Development of abnormal sound-voice detection system for victims at accident sites) funded by the Ministry of the Interior and Safety (MOIS, Korea).

REFERENCES

Agostinelli, A., Denk, T., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Frank, C. (2023). MusicLM: Generating Music From Text. *arXiv preprint arXiv:2301.11325*.

Amit, S., Shiraishi, S., Inoshita, T., & Aoki, Y. (2016). Analysis of satellite images for disaster detection. *International Geoscience and Remote Sensing Symposium* (pp. 5189-5192). Beijing: IEEE.

Antoniou, V., & Potsiou, C. (2020). A Deep Learning Method to Accelerate the Disaster Response Process. *Remote Sensing*, 544.

Ascione, M., Buonanno, A., D'Urso, M., Vinetti, P., Angrisani, L., & Moriello, R. (2012). A method based on passive acoustic sensors for detection of vital signs in closed structures. *IEEE Instrumentation and Measurement Technology Conference* (pp. 1764-1769). Graz: IEEE.

Bhondve, T. B., Satyanarayan, R., & Mukhedkar, M. (2014). Mobile Rescue Robot for Human Body Detection in Rescue Operation of Disaster. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 9876-9882.

Dong, J., Ota, K., & Dong, M. (2021). UAV-Based Real-Time Survivor Detection System in Post-Disaster Search and Rescue Operations. *IEEE Journal on Miniaturization for Air and Space Systems*, 209-2019.

Ekpezu, A., Wiafe, I., Katsriku, F., & Yaokumah, W. (2021). Using deep learning for acoustic event classification: The case of natural disasters. *The Jounal of the Acoustical Society of America*, 2926-2935.

Gui, A., Gamper, H., Braun, S., & Emmanouilidou, D. (2023). Adapting frechet audio distance for generative music evaluation. *arXiv preprint arXiv:2311.01616*.

Güntner, A., Pineau, N., Mochalski, P., Wiesenhofer, H., Agapiou, A., Mayhew, C., & Pratsinis, S. (2018). Sniffing Entrapped Humans with Sensor Arrays. *analytical chemistry*, 4940-4945.

Hoshiba, K., Washizaki, K., Wakabayashi, M., Ishiki, T., Kumon, M., Bando, Y., Okuno, H. (2017). Design of UAV-Embedded Microphone, Array System for Sound Source Localization in Outdoor Environments. *Sensors*, 2535.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Zhao, Z. (2023). Make-An-Audio: Text-to-Audio Generation with Prompt-Enhanced Diffuusion Models. *International Conference on Machine Learning*, (pp. 13916-13932). New York.

Imlauer, S., Lassnig, K., Maurer, J., & Steinbauer, G. (2014). Life Sign Detection Based on Sound and Gas Measurements. *Austrian Robotics Workshop.* Linz.

Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019). Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms. *INTERSPEECH*, (pp. 2350-2354).

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks, for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2880-2894.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Adi, Y. (2023). Audiogen: Textually guided audio generation. *International Conference on Learning Representations.* Kigali .

Latif, T., Whitmire, E., Novak, T., & Bozkurt, A. (2016). Sound Localization Sensors for Search and Rescue Biobots. *IEEE Sensors Journal*, 3444-3453.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Plumbley, M. (2023). AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *International Conference on Machine Learning*, (pp. 21450-21474).

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Plumbley, M. (2024). AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining. *arXiv preprint arXiv:2308.05734*. Retrieved from https://arxiv.org/abs/2308.05734

Mäyrä, A., Känsälä, K., Ojala, K., Aitta, P., Hietavalkama, T., Fernandez, F., Bussion, J. (2013). Optical sensors and algorithms for life-sign detection in USaR-operations. *AIP* (pp. 41-46). AIP Publishing.

Nezirovic, A. (2010). Trapped-Victim Detection in Post-Disaster Scenarios Using Ultra-Wideband Radar. Delft, Netherlands.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C., & Johnson, P. (2017). Machine Learning Predicts Laboratory Earthquakes. *Geophysical Research Letters*, 9276-9282.

Song, H., Yu, J., Qiu, J., Sun, Z., Lang, K., Luo, Q., Wang, Y. (2022). Multi-UAV Disaster Environment Coverage Planning with Limited-Endurance. *International Conference on Robotics and Automation (ICRA)* (pp. 10760-10766). Philadelphia: IEEE.

Uzunidis, D., Mitilineos, S. A., Ponti, C., Schettini, G., & Patrikakis, C. Z. (2023). Detection of trapped victims behind large obstacles using radar sensors: a review on available technologies and candidate solutions. *IEEE Conference on Antenna Measurements and Applications* (pp. 1025-1030). Italy: IEEE.

Wilson, J., & Nicholas C., M. (2006). Ocean acoustic hurricane classification. *The Jounal of the Acoustical Society of America*, 168-181.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., & Yu, D. (2024). Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.0993X*.