A Demand-Driven Perspective on Generative Audio AI

Sangshin Oh [*1] **Minsung Kang** [*1] **Hyeongi Moon** [1] **Keunwoo Choi** [1] **Ben Sangbae Chon** [1]

Abstract

To achieve successful deployment of AI research, it is crucial to understand the demands of the industry. In this paper, we present the results of a survey conducted with professional audio engineers, in order to determine research priorities and define various research tasks. We also summarize the current challenges in audio quality and controllability based on the survey. Our analysis emphasizes that the availability of datasets is currently the main bottleneck for achieving high-quality audio generation. Finally, we suggest potential solutions for some revealed issues with empirical evidence.

1. Introduction

The use of audio generative models has the potential to significantly impact a variety of industries. Although essential, the process of creating foley effects is often tedious, non-reproducible, and lacks scalability. Moreover, the utilization of pre-recorded sounds is not conducive to real-time or interactive applications, rendering it inadequate for fields like gaming, metaverse, or any domain requiring the simulation of lifelike environments. The advent of generative audio AI offers a promising solution to address these limitations, significantly impacting areas like film production, gaming, social platforms, and more.

Audio synthesis research has a long history (Dudley, 1955; Chowning, 1973), but we will focus on the data-driven approaches as they are the recent pioneers with huge potential. The current generative audio AI is still in its early stages, necessitating further advancements in various aspects. We present this paper to provide a demand-driven perspective on task definitions, challenges, and potential solutions within audio generation. Specifically, our focus is on general audio, excluding speech and music.

[*]Equal contribution [1]Gaudio Lab, Inc., Seoul, South Korea. Correspondence to: Ben Sangbae Chon <bc@gaudiolab.com>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

The key contributions of this paper include:

- A survey with individuals working in movie sound productions to share insights into the industry-side demands.
- Detailed definitions and review of distinct tasks in audio generation regarding input types and conditions.
- A summary of the related challenges towards industrial demands and a proposal on potential solutions supported by empirical evidence, including a method with which we achieved 2nd place in the foley synthesis challenges at DCASE 2023.

2. Demands from Industry

To gather insights regarding the impact of audio generative models on the industry, we first interviewed two professionals from the field of movie sound production. They highlighted that their role extends beyond that of sound technicians, as they contribute to the artistic dimension of creating immersive and captivating sound experiences. Despite the inevitable laborious nature of foley and sound effect recording, they are compelled to record new sounds since existing sounds are hardly reusable. While they have a vast library of previous sound stems, there is effectively no efficient method at hand for searching and finding suitable sounds. Even if they find a suitable sound, they have to spend time on editing the time synchronization and sound tone.

Based on this knowledge, we conducted a survey involving 18 individuals working in movie sound production, addressing the topic of AI audio generation. We first presented them with some examples of AI image generation applications and a demo page[1] of a recent text-to-audio model (Liu et al., 2023). We then asked three following primary questions with multiple-choice options.

Q1. *What are the major challenges faced in foley recording?* The most frequently selected option for this question was the time synchronization problem. Following that, respondents expressed the importance of audio quality and consistency in tone with the synchronous recording. In the additional comments, respondents emphasized again that for foley sound, audio quality, synchronization with the scene, and

[1]https://audioldm.github.io/

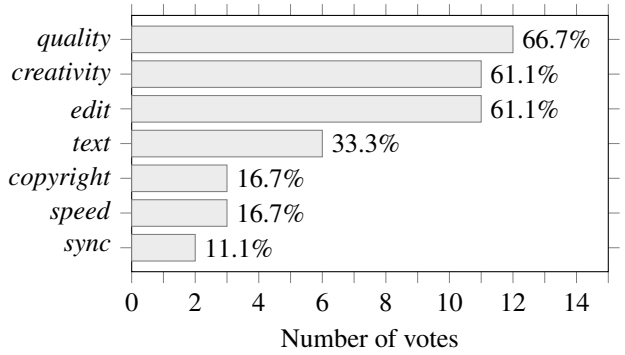

Figure 1: Answers of Q2: *What is the limitation(s) of the current text-conditioned audio generation as a product?*

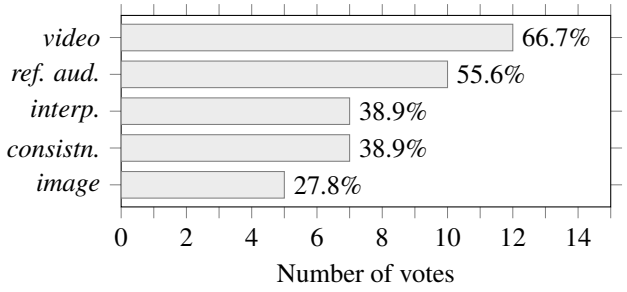

Figure 2: Answers of Q3: *How would you like to condition the audio generation by?*

consistency in tone with other sound sources are crucial – to the point that without a good synchronization, some might only consider using AI-generation for ambient sounds. This indicates that relying solely on text-based conditioning may not be sufficient for a majority of use-cases.

Q2. *What is the limitation(s) of the current text-conditioned audio generation as a product?* The survey result is plotted in Figure 1. In this question, it was found that audio quality presents the most significant challenge for practical usage. According to their comments, the concerns about *quality* encompass other aspects such as low fidelity, low sampling rate, roughness, and other related factors. A majority of respondents expressed complaints regarding the sample rate. It is noteworthy that while the industry requires full-band signals at 48kHz or higher, most of the current systems still operate within the 16kHz-24kHz range (Kreuk et al., 2022; Huang et al., 2023; Liu et al., 2023). For *creativity*, which was the second most frequently chosen category, it refers to the generation of new sounds that fulfill artistic intentions, e.g., creating "the sound of a lightsaber in Star Wars." The terms such as *edit* and *text*, which received the third and fourth highest numbers of votes, indicate the problems of controllability.

Q3. *How would you like to condition the audio generation?* As in Figure 2, the most frequently chosen option is the utilization of video for time synchronization and achieving an appropriate sound tone. More than half of the respondents were interested in generating similar sounds to reference audio samples. The third and fourth popular options, namely *interp.* and *consistn.*, are related to refining the generated audio based on reference audio samples. Here, *interp.* indicates generation via interpolation between two reference audio, and *consistn.* means generation of audio sample consistent to other tracks or sources. The respondents seemed to show their hope for a more efficient workflow in Q3, in contrast to showing their expectations in Q2.

This survey result presents important remarks on generative audio research. First, texts and videos are complementary to each other towards a more complete generative audio system. Second, sound and event synchronization is an important topic that deserves more attention. Third, although it is somewhat deviated from our topic, high-quality audio indexing, search, and separation may be also a solution for some of the problems generative audio AI aims to solve. Based on this understanding, we delve into the current state and challenges of the audio generation field in the following sections.

3. Task Definitions

In a recent proposal paper on foley sound synthesis challenge (Choi et al., 2022), the audio generative AI task is specified based on the input and output types. The authors outline three distinct input types: i) category index, ii) text description, and iii) videos. While the categorization of output types is not explicitly stated, it can be inferred as follows: i) individual foley sounds representing a single event, ii) a combination of multiple events and/or ambient sounds, and iii) a comprehensive soundtrack comprising foley sounds ambient elements, and spatially enhanced mixing. We will focus on the input types since the determination of output types is primarily governed by technical feasibility, allowing a limited scope with the current technology.

3.1. Input Types

First, a category index, that indicates a single type of audio event, would be the simplest form of input type for a sound synthesis system. This was adopted in some previous works (Ghose & Prevost, 2020; 2022) and this year's DCASE Task 7 (Choi et al., 2023). Solutions with this approach would improve foley recording processes for some popular categories such as dog barks, door slams, or footsteps.

The second type would be text descriptions as employed in recent research (Kreuk et al., 2022; Yang et al., 2023; Liu et al., 2023; Huang et al., 2023), relying on audio caption datasets. There are several promising aspects associated

with this text-to-audio approach. i) Extensive research has already been conducted on text-to-X generation (e.g., text-to-image generation studies (Ramesh et al., 2021; Saharia et al., 2022; Rombach et al., 2022)), which simplifies its adaptation for audio generation purposes. ii) The familiarity of users with UI/UX utilizing text inputs further supports the feasibility of this approach. However, there are difficulties as well. i) Compared to text-image pairs, there is a scarcity of text-audio pairs available for training models (Huang et al., 2023). For example, the number of items of AudioCaps (Kim et al., 2019), the largest audio captioning dataset, is 0.013% of (or 7561 times smaller than) that of LAION-400M, an text-image pair dataset (Schuhmann et al., 2021). ii) Text input has limitations in providing highly detailed descriptions at a professional level, as audio engineers rely on precise controls like knobs and sliders to make fine adjustments to the sound (e.g., equalizers).

Third, video input types have pros and cons. Unlike the previous input types, videos may provide the exact timings of events (Zhou et al., 2018; Ghose & Prevost, 2022; Cui et al., 2023). As its importance was discussed in Section 2, there is a huge potential for improving the workflow of video creation in this scenario by efficient time synchronization. However, the video itself does not provide complete information because it is common that not everything visible should sound, as well as not everything that sounds is visible. Additionally, there are deliberate artistic intentions involved in video creation such as muting/exaggerating certain sounds. These artistic decisions may vary significantly. Therefore, when developing video-to-sound generation methods, the ability to edit and manipulate the generated audio becomes crucial, just as it is important for text-based generation approaches as we will discuss in the following section.

3.2. Conditioning

Conditioning can be viewed as a form of input in a broader sense and is deeply related to controllability and editability. AudioLDM pioneered sound editing through text-based approaches (Liu et al., 2023), and we believe that this direction of research will continue toward more diverse, intuitive, and fine-grained conditioning. For example, users may want to control factors such as sound bandwidth, F0 contours, temporal and spectral envelopes, etc. Our exploration of these product development considerations will continue in the following sections.

4. Challenges

4.1. Dataset Improvement for Audio Quality

Recently, there have been some generative AI products successfully deployed on language and image (Touvron et al., 2023; Chowdhery et al., 2022; OpenAI, 2023; Ramesh et al., 2021). However, the current state of audio generation research does not seem mature enough to be adopted into professional sound production. As audio quality was the most prominent issue as in Figure 1, we focus on the issues and potential solutions on datasets to improve the generated audio quality in this section.

First of all, the current data scarcity deteriorates the model training and resulting audio quality. Compared to image generation datasets that go beyond a few billion pairs (Ramesh et al., 2022), there are much less text-paired audio data available (Kreuk et al., 2022; Huang et al., 2023). Moreover,

Name	AQ	Dataset Size		Modality		
		Dura.	N. Files	Lb	Cp	Vd
Audioset						
AudioSet (Gemmeke et al., 2017)	*noisy*	5420 *h*	1,951,460	✓		✓
AudioCaps (Kim et al., 2019)	*noisy*	144.9 *h*	52,904	✓	✓	✓
Freesound						
Freesound (Font et al., 2013)	*noisy*	3003 *h*	515,581		△	
UrbanSound8K (Salamon et al., 2014)	*noisy*	8.75 *h*	8,732	✓		
ESC-50 (Piczak, 2015)	*noisy*	2.78 *h*	2,000	✓		
Clotho (Drossos et al., 2020)	*noisy*	37.0 *h*	5,929		✓	
FSD50K (Fonseca et al., 2021)	*noisy*	108.3 *h*	51,197	✓		
Others						
VGG Sound (Chen et al., 2020)	*noisy*	550 *h*	≈ 200,000	✓		✓
BBC sound effects [2]	*clean*	463.5 *h*	15,973		✓	
Epidemic Sound effects [3]	*clean*	220.4 *h*	75,645	✓		
Free To Use Sounds [4]	*noisy*	175.7 *h*	6,370		✓	
Sonnis Game Effects [5]	*clean*	84.6 *h*	5,049		△	
WeSoundEffects [6]	*clean*	12.0 *h*	488		△	
Odeon Sound Effects [7]	*clean*	19.5 *h*	4,420		△	

Table 1: A list of audio datasets. AQ: audio quality, Dura.: duration, N. Files: number of files. Modality columns refer to the existence of labels, captions, and videos, respectively. *Clean* recording: Audio is recorded in well-treated environments and mastered for professional content production. *Noisy*: dataset contains environmental noises or interference signals. △: Textual information included, not necessarily captions. This table is partially from (Kreuk et al., 2022) and (Wu et al., 2023)

[2] https://sound-effects.bbcrewind.co.uk
[3] https://www.epidemicsound.com/sound-effects/
[4] https://www.freetousesounds.com/all-in-one-bundle/
[5] https://sonniss.com/gameaudiogdc
[6] https://wesoundeffects.com/we-sound-effects-bundle-2020/
[7] https://www.paramountmotion.com/odeon-sound-effects

most of such paired datasets are *weakly labeled*, i.e. their labels or captions lack time resolution. This is problematic as it is common practice to slice audio signals for ease of training and memory-related issues. Since the text in the pairs depicts audio coarsely in the time axis, there should be potential risks of mismatching when the audio signal is sliced into smaller segments for some practical reasons. Augmentation method (Kreuk et al., 2022; Huang et al., 2023) or using a contrastive embedding network (Liu et al., 2023) can help this, but not as an absolute treatment.

The characteristics of the audio itself even exacerbates the problem. It is a difficult problem to separate foreground and background audio sources, and obtaining isolated audio recording would remain to be costly. The spatial characteristics of the recording environment often have negative affects to the recording quality. Altogether, there are many factors that make it tricky to create a studio-quality audio dataset. We listed available audio datasets in Table 1. Since the largest datasets in the list are collected or curated from crowd-sourced audio (Font et al., 2013) or video (Gemmeke et al., 2017; Chen et al., 2020), their recording conditions may vary and are usually not good. Thus, the samples from those datasets often suffer from severe background noises, low recording bandwidth / bit rate, and various types of distortion. *Clean* datasets are limited to several commercial sound effect libraries.

To this trade-off problem of *more* data vs. *clean* data, we propose a solution called *quality-aware training* (QAT). This can be simply done by prompting, i.e., appending dataset labels indicating the quality of the dataset in the text input. QAT enables to utilize a broader range of datasets. During the training phase, a model can learn from both *clean* and *noisy* datasets with quality labels. As a result, the model would learn not only the concepts of different audio events but also their audio quality; i.e., the model would have *compositionality* of audio events and audio quality. During the inference phase, we can force the model to generate clean signals by conditioning the model, i.e., by appending 'clean' labels to the text input. This enabled us to use all data pairs regardless of their quality without deteriorating their output quality. In our experience, this approach let us control the audio quality, reverberation, signal bandwidth, and audio event independently and achieve 2nd place in the recent foley synthesis challenge at DCASE 2023 (Kang et al., 2023a;b). Details about experiments are provided in Appendix B.

4.2. Methodological Improvement for Controllability

Controllability was another major concern in our survey, as the audio engineers have specific intent about how the generated output should sound. Audio generation may take a long time, hence it is crucial for deployable audio AI systems to have effective controllability

Classifier-free guidance is a widely adopted solution for the problem across diffusion-based and Transformer-based generative models. At the cost of sample qualities by extrapolating intermediate features or logits, it introduces diversity, which would make exploration easier for the users of generative audio AI systems. Most of the recent text-to-audio generation research adopted this technique (Kreuk et al., 2022; Liu et al., 2023; Huang et al., 2023).

Controllability can be also attained by introducing new features or new modalities, for example, a reference audio or a conditioning video as in Figure 2. As AudioLDM demonstrated audio manipulation without fine-tuning (Liu et al., 2023), we believe text-guided audio-to-audio generation is a compelling research direction towards deployable generative audio AI. Video-based foley generation has been less popular, but it would be an interesting direction for future research along with the existing research (Zhou et al., 2018; Ghose & Prevost, 2020; 2022). Finally, conventional signal features such as F0 contour or envelopes can be a great user interface for experienced audio engineers. As those features are easy to extract from audio signals, it is plausible to use them as one of the inputs during the training phase, then build a user interface that allows control of the generated output by modifying the features.

5. Conclusion

In this paper, we presented a survey conducted with sound engineers in the movie industry. Based on the survey results, we have provided task definitions for audio generation research and identified related research challenges. Our objective was to bridge the gap between current research and industry practices, offering potential solutions to address the challenges of audio quality and controllability.

Surprisingly, there are limited opportunities for researchers to gain insights from the industry side. We believe that this work serves as a valuable starting point for understanding the difficulties faced by both researchers and potential users, ultimately aligning our efforts to solve the real-world problems.

While our perspective focuses on the movie industry, it is important to acknowledge that neighboring industries may face different challenges with varying priorities. For example, the demand for real-time generation systems may be stronger in the virtual reality or gaming industry, while the standards for audio quality or artistic intent may be lower for non-professional movie creation platforms such as YouTube. We hope that our work represents a meaningful step towards comprehending the diverse demands placed on generative audio AI and its diverse applications.

Acknowledgement

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2022 (Project Name: R&D on AI Text-to-Sound Generation, Project Number: RS-2023-00229204, Contribution Rate: 100%)

References

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Choi, K., Oh, S., Kang, M., and McFee, B. A proposal for foley sound synthesis challenge, 2022.

Choi, K., Im, J., Heller, L., McFee, B., Imoto, K., Okamoto, Y., Lagrange, M., and Takamichi, S. Foley sound synthesis at the dcase 2023 challenge, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chowning, J. M. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the audio engineering society*, 21(7):526–534, 1973.

Cui, C., Zhao, Z., Ren, Y., Liu, J., Huang, R., Chen, F., Wang, Z., Huai, B., and Wu, F. Varietysound: Timbre-controllable video to sound generation via unsupervised information disentanglement. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Dudley, H. Fundamentals of speech synthesis. *Journal of the Audio Engineering Society*, 3(4):170–185, 1955.

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

Font, F., Roma, G., and Serra, X. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 411–412, 2013.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Ghose, S. and Prevost, J. J. Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning. *IEEE Transactions on Multimedia*, 23:1895–1907, 2020.

Ghose, S. and Prevost, J. J. Foleygan: Visually guided generative adversarial network-based synchronous sound generation in silent videos. *IEEE Transactions on Multimedia*, 2022.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., and Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023.

Kang, M., Oh, S., Moon, H., Lee, K., and Chon, B. S. Fall-e: Gaudio foley synthesis system. Technical report, Gaudio Lab, Inc., Seoul, South Kore, June 2023a.

Kang, M., Oh, S., Moon, H., Lee, K., and Chon, B. S. Fall-e: A foley sound synthesis model and strategies, 2023b.

Kim, C. D., Kim, B., Lee, H., and Kim, G. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011.

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

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

OpenAI. Gpt-4 technical report, 2023.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

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

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

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

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

Zhou, Y., Wang, Z., Fang, C., Bui, T., and Berg, T. L. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3550–3558, 2018.

A. Details of survey in Section 2

A.1. Exact expression of the options in Figure 1 and Figure 2

Question	Option	Exact expression
Figure 1	*quality*	Audio quality.
	creativity	Lack of creativity in fulfilling artistic intentions (e.g. the sound of lightsabers in Star Wars).
	edit	Detailed audio editing (e.g. I like the footstep sound I've created, but I wish it was a bit lighter).
	text	Difficult to create the desired sound with just text.
	copyright	Copyright.
	speed	Speed of generation.
	sync	Time synchronization with the scene.
Figure 2	*video*	Time synchronization and incorporating tone through video.
	ref. aud.	Create sounds similar to a reference (e.g., "Create 10 sounds similar to Sound A" or"Make a slightly more light version of Sound A").
	interp.	Interpolation of two sounds (e.g., "I need a footstep sound that is a middle ground between Sound A and Sound B").
	consistn.	Generate sounds consistent with the reference audio for other tracks or other sources. (e.g., "Create a car sound that matches the 0:00:32 - 0:00:42 segment of this track").
	image	Expressing the sensation of sound that is difficult to convey in words through image.

Table 2: Exact expressions of the options.

A.2. Results on the other questionnaire

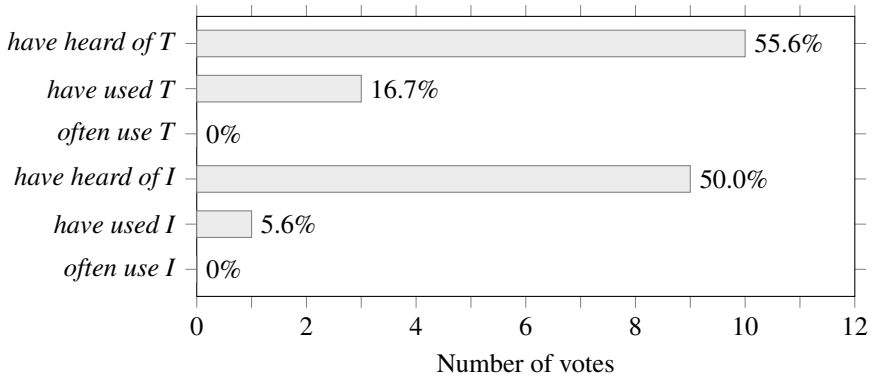

Figure 3: Answers of a multiple-choice question.

Q: Have you heard of or used text or image generation AI, such as ChatGPT, Bard, Stable Diffusion or Midjourney?

A1: *have heard of T* - Have heard of text generative model
A2: *have used T* - Have used text generative model
A3: *often use T* - Often use text generative model
A4: *have heard of I* - Have heard of image generative model
A5: *have used I* - Have used image generative model
A6: *often use I* - Often use image generative model

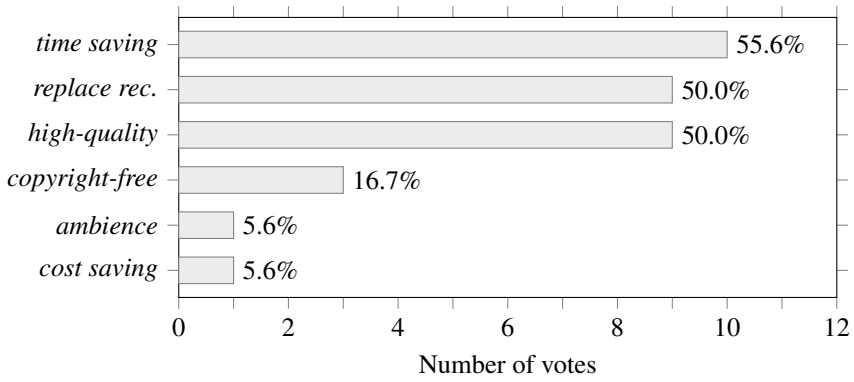

Figure 4: Answers of a multiple-choice question.

Q: What do you expect for generative audio AI product?

A1: *time saving* - Time saving due to the fast generation speed
A2: *replace rec.* - Replacing recording or sampling process with generation
A3: *high-quality* - Obtaining high-quality, well-aligned audio
A4: *copyright-free* - Obtaining copyright-free sources
A5: *ambience* - Generating ambient sound
A6: *cost saving* - Cost saving for human resources

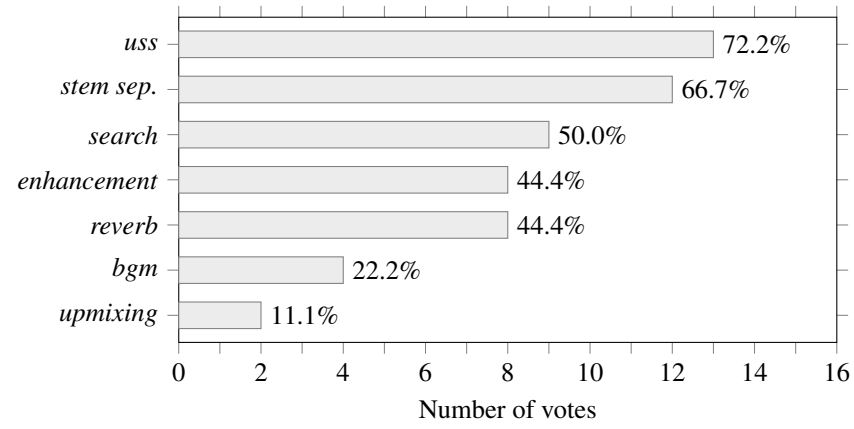

Figure 5: Answers of a multiple-choice question.

Q: Except for generative audio AI, which technology do you think would be useful?

A1: *uss* - Universal source separation with text condition
A2: *stem sep.* - Automatic separation of stems from mixed or mastered tracks
A3: *search* - Simplified and efficient search algorithms or visualization methods
A4: *enhancement* - Audio enhancement to improve audio quality, such as sample rate or fidelity
A5: *reverb* - De-reverberation or room-impulse response estimation
A6: *bgm* - Automatic rearrangement for background music
A7: *upmixing* - Automatic upmixing (e.g. mono to 5.1ch audio)

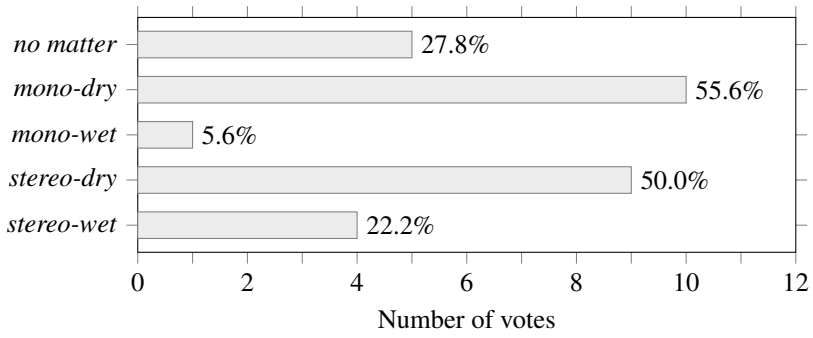

Figure 6: Answers of a multiple-choice question.

Q What audio type do you prefer for output?

A1: *no matter* - All possible types
A2: *mono-dry* - 1-channel audio signal without any reverb
A3: *mono-wet* - 1-channel audio signal with proper reverb
A4: *stereo-dry* - 2-channel audio signal without any reverb
A5: *stereo-wet* - 2-channel audio signal with proper reverb

B. Experiment Results for Section 4

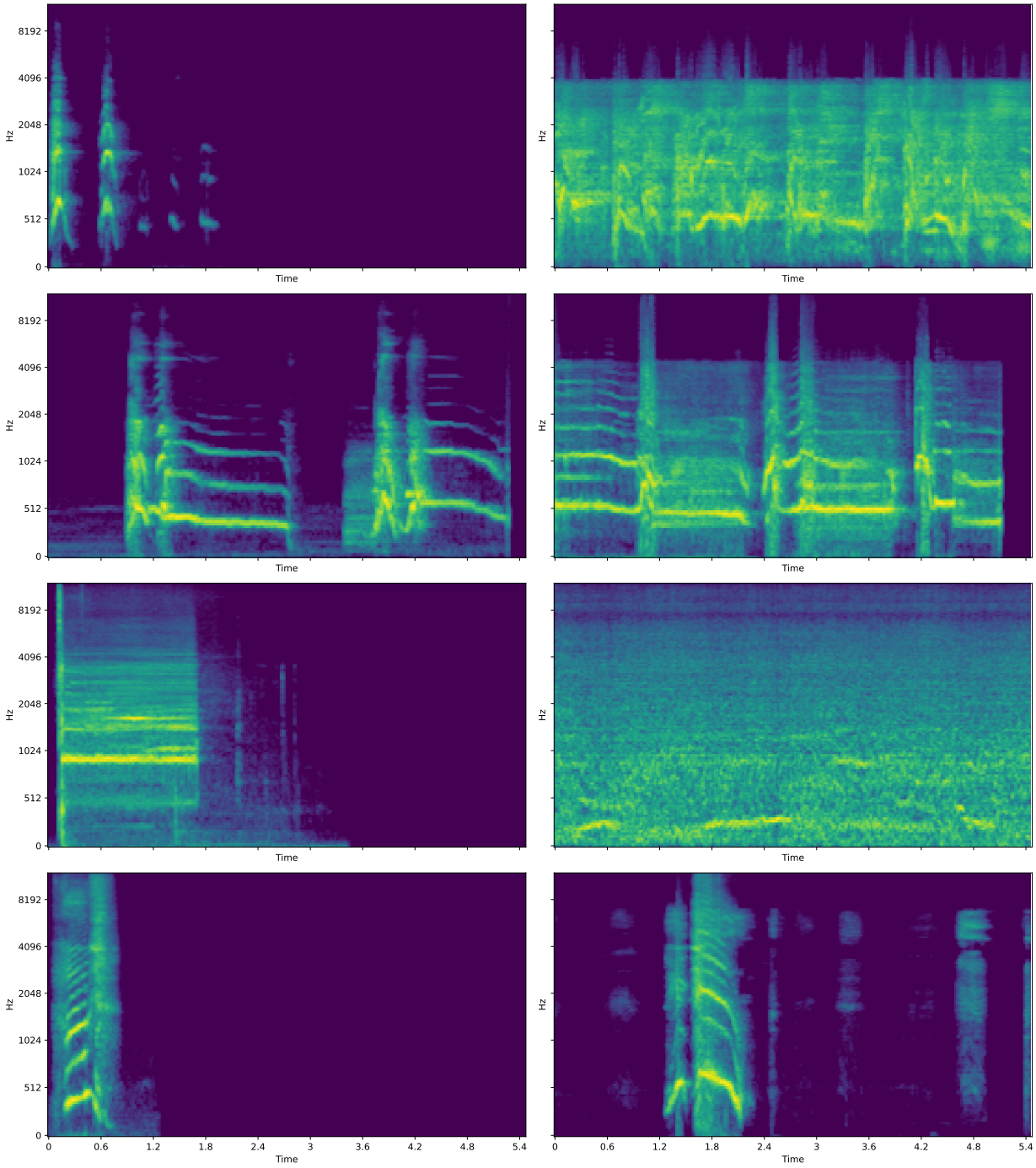

Figure 7: Mel-spectrograms of the generated audio samples with different dataset labels – *left*: with "clean" label, and *right*: with "noisy" label. Samples in the same row are generated with the same prompt. Prompts used to generate samples in each row are as follows; (i) "small dog bark", (ii) "dog howling", (iii) "the sound of starting a car engine", and (iv) "male sneeze". With the *clean* dataset label, the model generates high quality signals with less background noise and more high frequency components. Conversely, generated samples show low quality results when the model is conditioned with the *noisy* label. Specifically, most of noisy-labeled signals show underlying noise or interference and have limited bandwidth for some samples.