A Demand-Driven Perspective on Generative Audio AI

Sangshin Oh Minsung Kang Hyeongi Moon Keunwoo Choi Ben Sangbae Chon
Gaudio Lab, Inc., Seoul, South Korea

We asked real users of audio gen AI.
We defined the task, challenges, and proposed solutions.

To foster deployable research on audio gen AI,

- Insights for **industry-side demands** from a survey with individuals in the movie sound production
- Summary of related **challenges** and a **proposal** on potential solutions

are presented in this poster.

Link for Full Paper

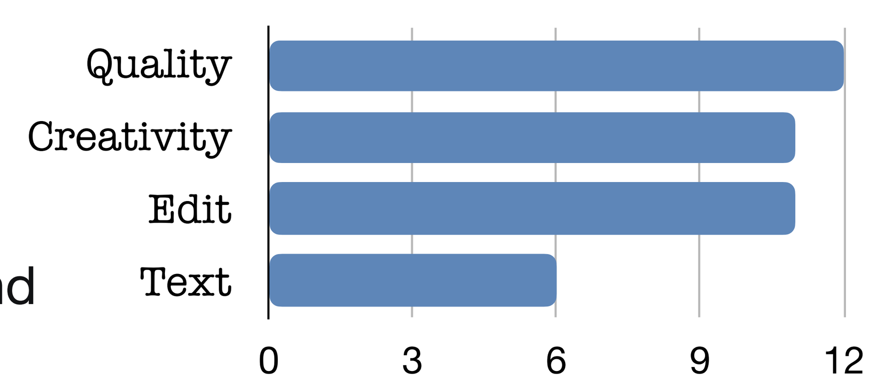
Motivations

- While essential, creating foley effects lacks reproducibility, scalability, and reusability, and the advent of audio gen AI offers a promising solution to these problems.
- We want to encourage more industry-oriented research and bridge the gap between industry and the research community.

Insights from the survey

- **What are the challenges faced in foley recording?**
  - The biggest challenge is **time synchronization** with the corresponding visual contents.
  - **Consistency in tone** with other sources or synchronous recordings is also a big challenge.
- **What is the (expected) limitation(s) of the current text-conditioned audio generation as a product?**
  - Before the questionnaire, we presented a demo page for AudioLDM, a state-of-the-art system for audio gen AI.
  - Most of the concerns are about the **sound quality**.
  - Other concerns include **controllability for detail** and a **lack of creativity or art**.

Figure 1. Bar chart for Question 2. ***Edit*** indicates 'detailed audio editing.' And ***Text*** refers to 'audio-text alignment.'



- **How would you like to condition the audio generation?**
  - First: **Video**, as they are in movie sound production.
  - Second: **Reference audio**, an example of audio excerpts to offer desired tone or mood.

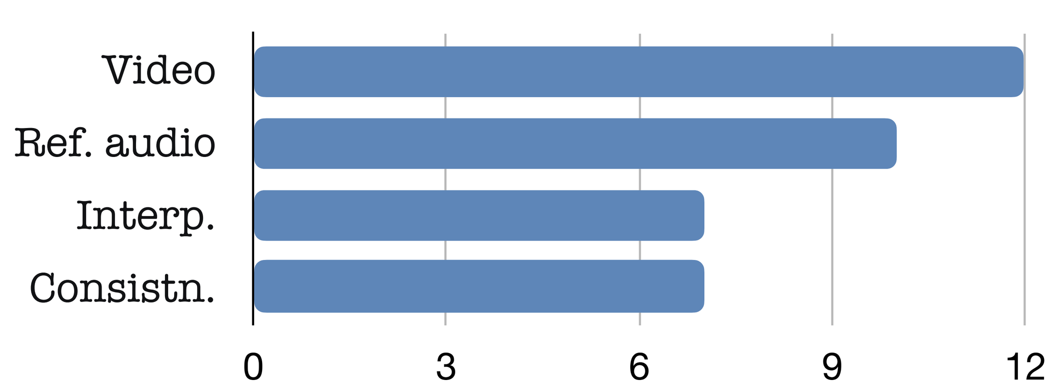


Figure 2. Bar chart for Question 3. ***Interp.*** Refers to 'interpolation,' and ***Consistn.*** is to 'consistency with other sources.'

Challenges and Solutions

- **Dataset improvement for audio quality**
  - Data scarcity deteriorates the model training and resulting audio quality.
    - There are **fewer datasets** compared to image datasets.
    - Most of the datasets suffer from **noise and interference** signals.
    - Audio datasets are often **weak-labeled**. Their labels often lack temporal information about the event.
  - Proposed ***quality-aware training (QAT)*** provides a remedy for these problems.
    - The model is trained with an additional label for dataset quality and can control cleanness in the inference phase.

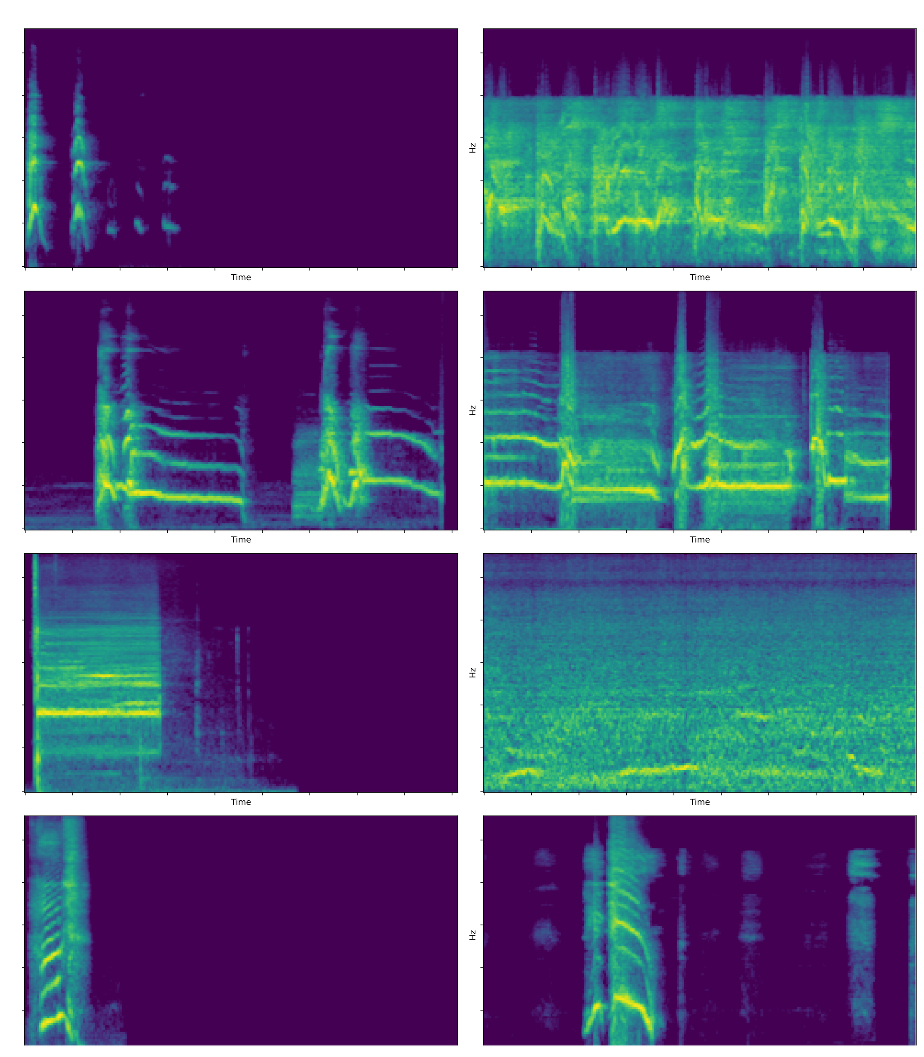


Figure 3. Spectrograms of generated audio samples. (*left*) samples with 'clean' embedding, and (*right*) samples with 'noisy' embedding.

- **Methodological improvement for audio quality**
  - Controllability is another major concern in our survey, and it is crucial to deliver sound engineers' intent
  - Classifier-free guidance is a widely adopted solution across diffusion-based and transformer-based models.
  - Research for various exemplar-based audio generation is required. We plan to explore this direction as our future direction.