IteraTTA: An interface for exploring both text prompts and audio priors in generating music with text-to-audio models

Hiromu Yakura (Univ. Tsukuba), Masataka Goto (AIST)

Core question: **Can text-to-audio models help novice users show their creativity in music composition?**

Our answer: Yes, ... ***if there is*** *an interactive interface that allows them to explore and learn various text prompts*.

Underlying issue

💡 Observed in our formative study

- We invited participants who possessed no formal musical training beyond compulsory education.
- We provided them with AudioLDM on Google Colab and asked freely use it.
- We used think-around protocol and semi-structured interviews to analyze their usage.

Descriptions in the dataset used for training models

"An orchestra plays a happy melody while the strings and wind instruments are being played"

Enabling precise control of output audios

Text prompts by novice users

"a song sounds like star wars"

Lacking vocabulary to control outputs

Our proposal

a) **Augmentation of an input prompt into variations of rich descriptions.**

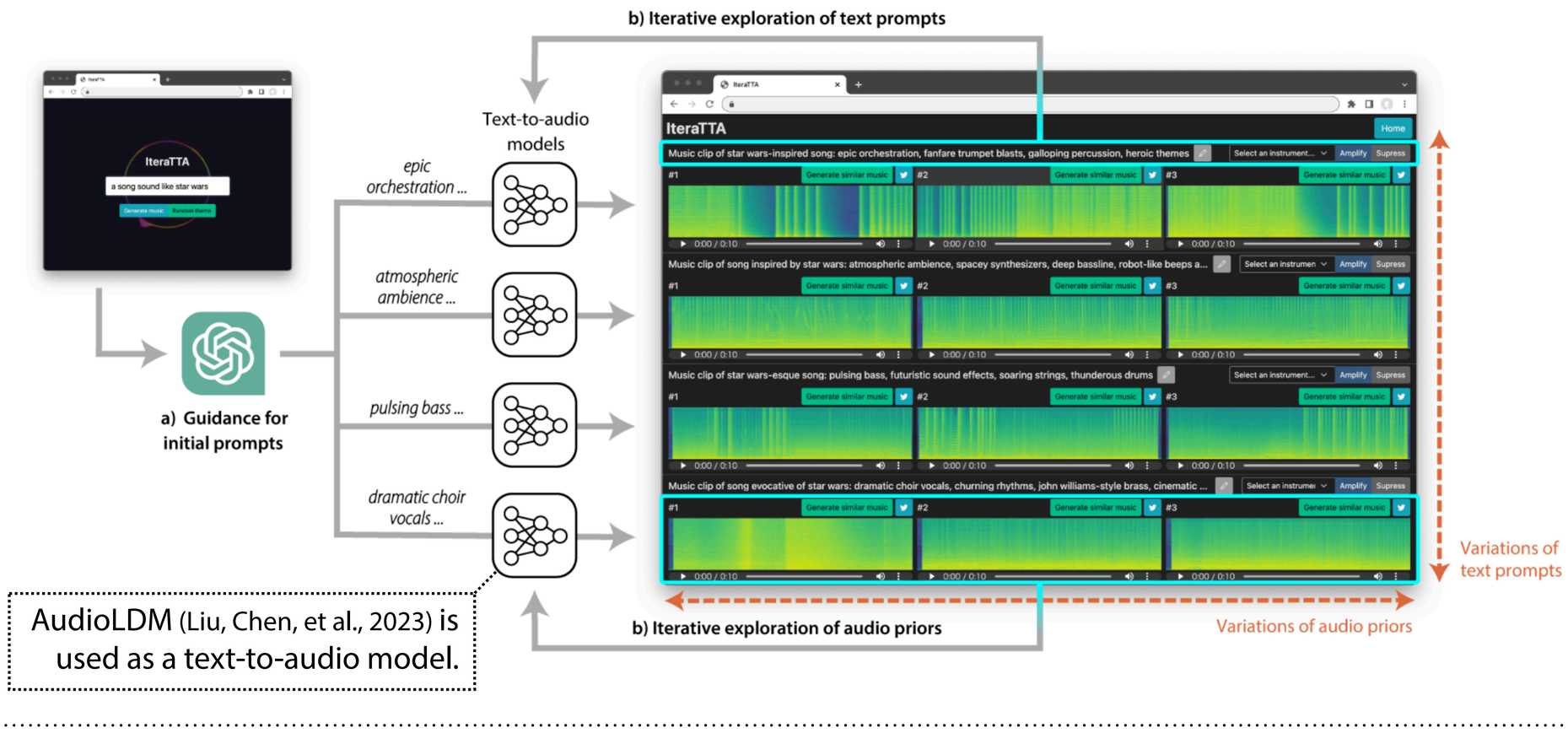
b) **Intuitive comparison of generated audios for iteratively customizing both text prompts and audios.**

👉 Demo

Video

Actual system

https://iteratta.duckdns.org/



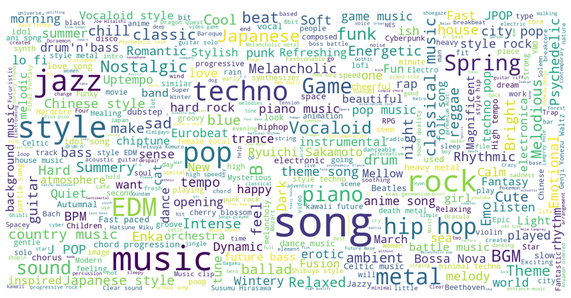
AudioLDM (Liu, Chen, et al., 2023) is used as a text-to-audio model.

Findings

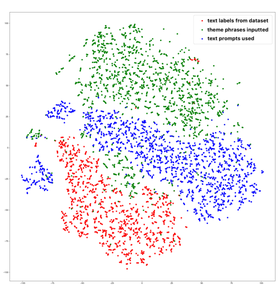
We deployed this interface as a public web service and collected the log of ...

~9,000 users *and* ***~250,000 audios***

Diversity of users' inputs and importance of filling the vocabulary gap

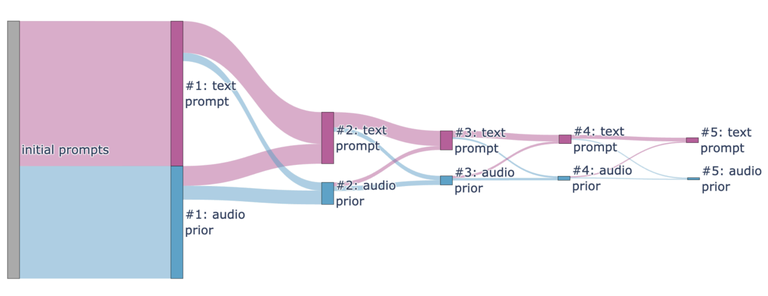


Word cloud of users' inputs



Embedding spaces of users' inputs, augmented inputs, and training data

Effectiveness of iteratively customizing text prompts and audios



Visualization of how the users utilized the iterative customization functions