Creative Text-to-Audio Generation via Synthesizer Programming

Nikhil Singh*, Manuel Cherep*, Jessica Shand

NEURAL INFORMATION PROCESSING SYSTEMS
ML4AUDIO WORKSHOP

Summary

Generating semantic sound sketches. We use text prompts to create abstract, sketch-like sounds that capture their meaning, rather than focusing on acoustic realism. Our approach is simple and lightweight (78 parameters), relying on a virtual modular synthesizer.

"Of course, bubbles don't make sound, but this is the magic of sound design...you can create the concept of a sound and it seems real."

— Suzanne Ciani

http://ctag.media.mit.edu

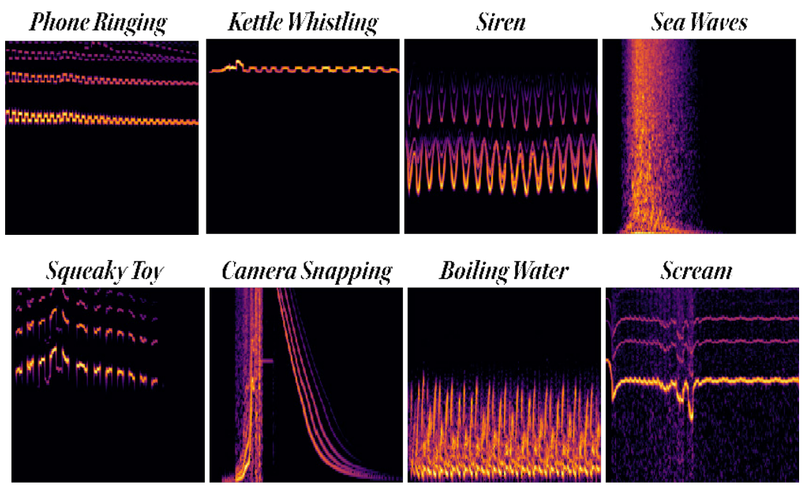


Figure 1: CTAG leverages a virtual modular synthesizer to generate sounds which capture the semantics of user-provided text prompts in a sketch-like way, rather than being acoustically literal. Spectrograms of auditory outputs corresponding to eight text prompts showcase the range of sounds.

Methods

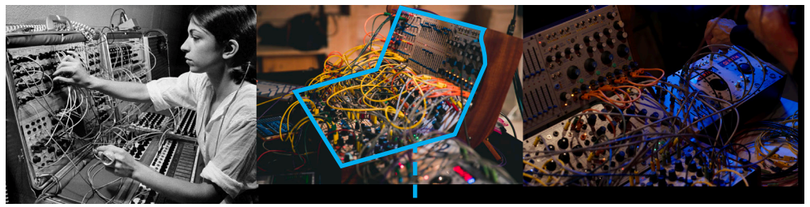
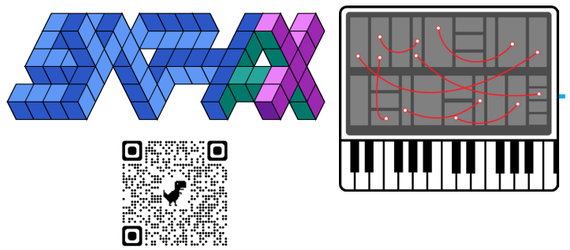


Figure 2: The modular synthesizer is a canonical sound synthesis tool, relying on components like oscillators and envelopes that produce, control, and process sound in networked configurations.

We use a virtual modular synthesizer (Cherep* and Singh* 2023) controlled by text prompts.



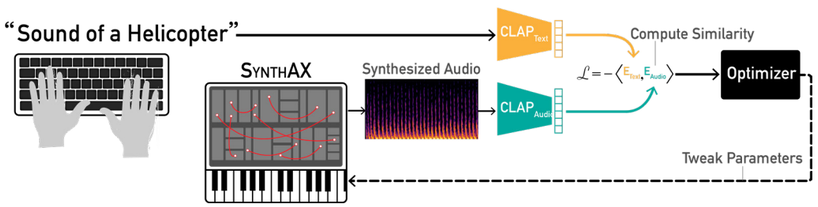


Figure 3: In our approach, we use the LAION-CLAP (Wu *et al.* 2023) model to compute the similarity between a user-provided text prompt and SYNTHAX's output. Then, we use a non-gradient optimization algorithm (Lange *et al.* 2023) to iteratively adjust parameter settings and maximize text-sound similarity.

Cherep* and Singh* AES 2023. *SynthAX: A Fast Modular Synthesizer in JAX.*
Wu *et al.* ICASSP 2023. *Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation.*
Lange *et al.* ICLR 2023. *Discovering evolution strategies via meta-black-box optimization.*

Results

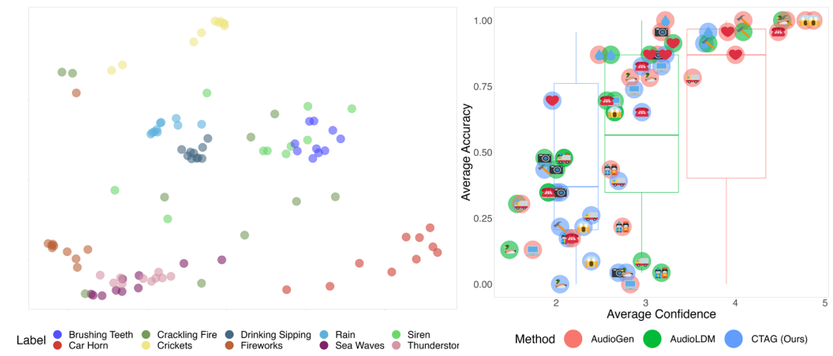


Figure 5: UMAP projection of parameters, showing how semantically-related sounds sometimes have similar parameters.

Figure 6: User accuracy and confidence, shown for 10 different prompts across CTAG, AudioGen, AudioLDM.

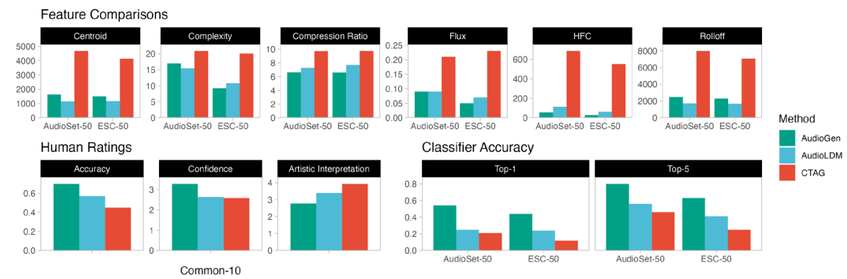


Figure 7: Quality-related feature comparisons, human ratings, and classification results comparing CTAG, AudioGen, AudioLDM. CTAG is rated as more artistically interpretive than other methods, and is qualitatively distinct as shown by several auditory descriptors. This also makes CTAG's sounds harder to classify than those from state-of-the-art generative models.

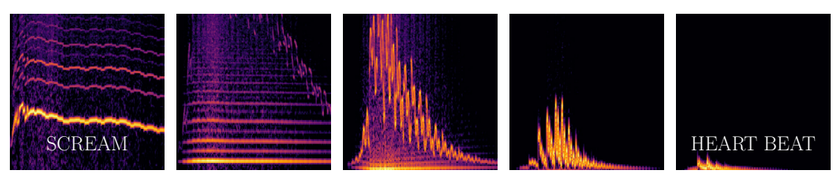


Figure 8: Spectrogram series showcasing a linear interpolation of the synthesizer parameters, from "Scream" (left) to "Heart Beat" (right). Each spectrogram in the sequence represents a step in the interpolation, highlighting the systematic shift in acoustic properties this approach can yield, accompanied by a fully interpretable and controllable parameter space.

*Equal contribution | MIT Media Lab