Text2EQ

Human-in-the-loop Co-Creation Interface for EQ

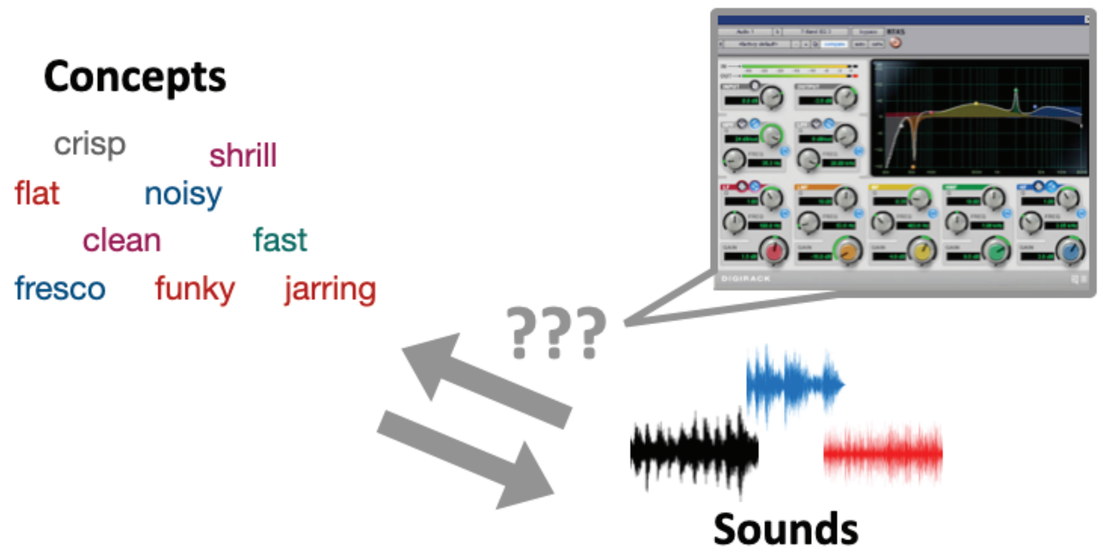
Annie Chu, Hugo Flores García, Patrick O'Reilly, Bryan Pardo
Interactive Audio Lab, Northwestern University, USA

Motivation

>Applying **audio effects (FX)**, typically within a DAW, is unintuitive and complex [1], requiring users to adjust multiple (>10+) technical DSP-based knobs

>**Natural language** could offer a more intuitive way to describe and apply sound transformations, streamlining the process of adjusting multiple controls

>**Key question:** Can we build a human-centered interface that allows users to apply DSP-based audio FX by simply describing their goal in text?

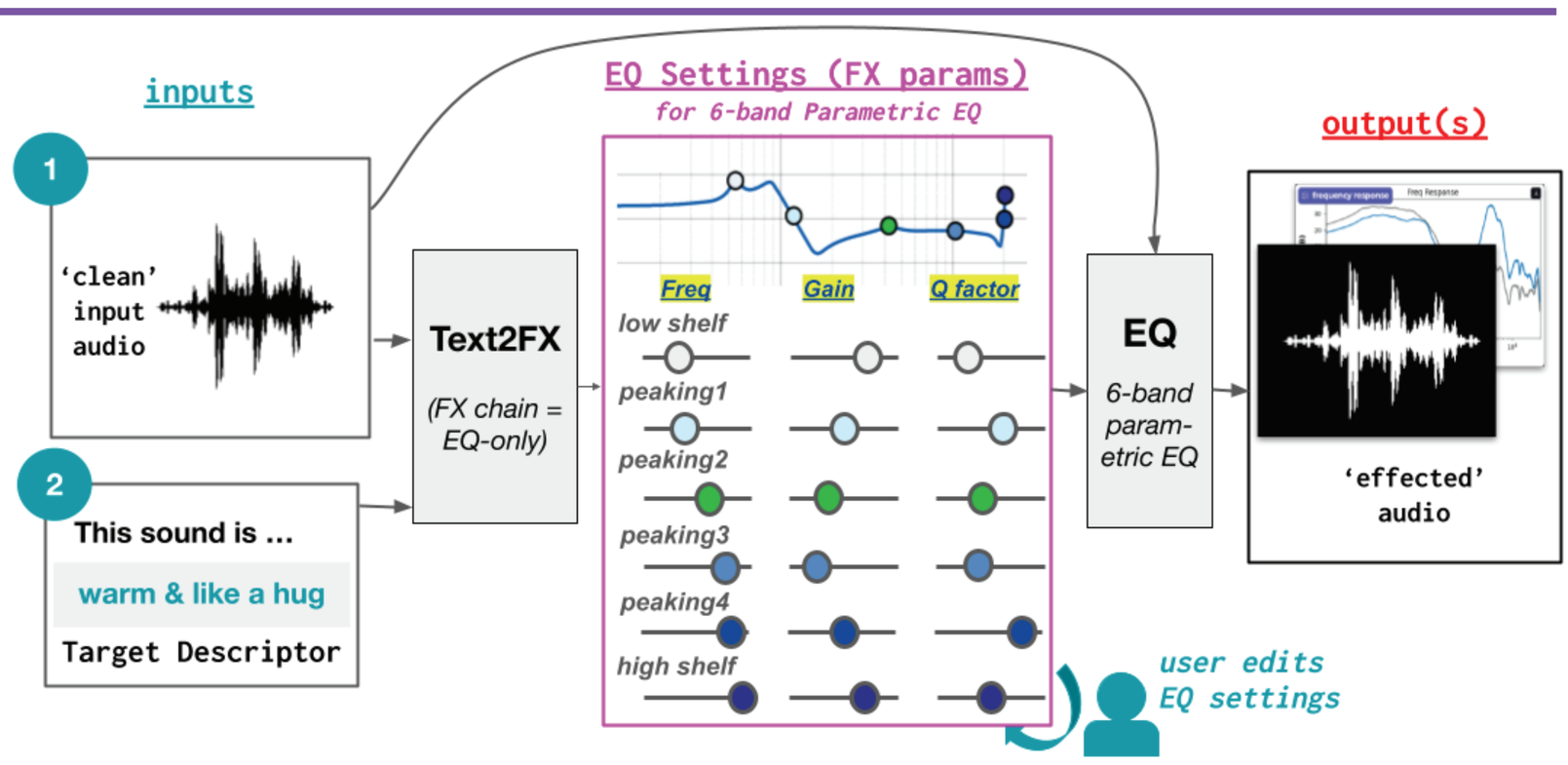


Key Guiding Design Principles [3]

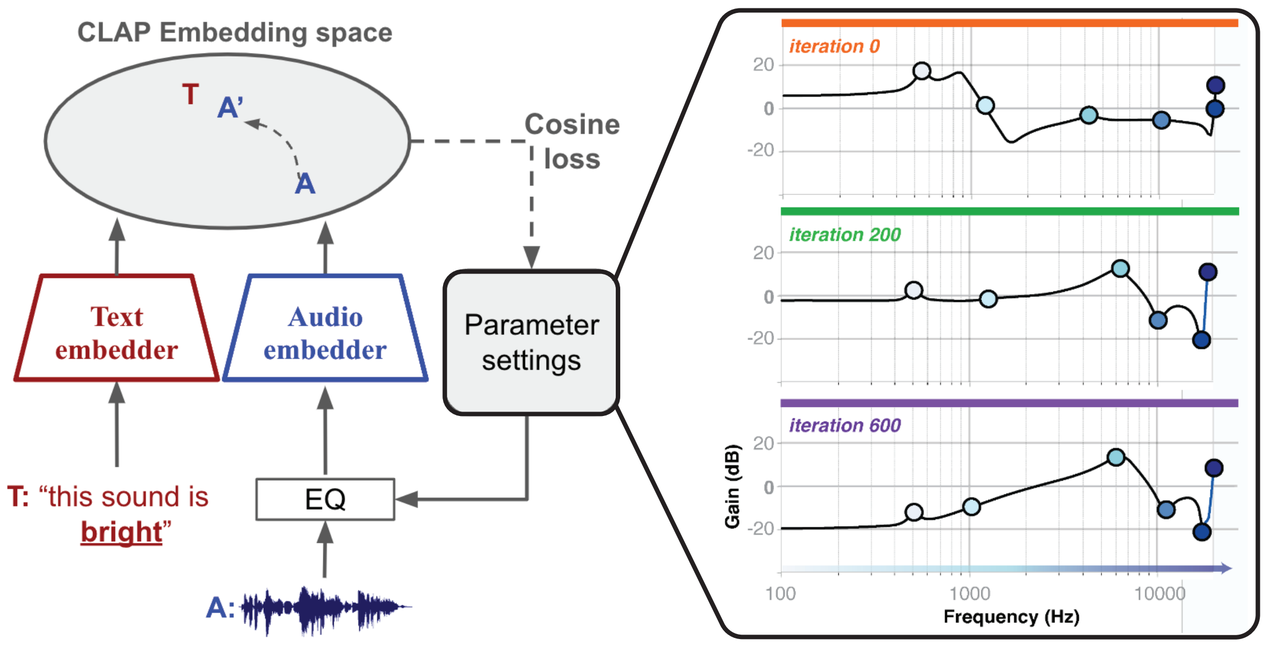
Concept	*Rationale*	*Leads to...*
Human-in-the-loop	Audio production is an inherently iterative process, human judgement is key	Adjustable EQ sliders, separate button to apply EQ parameters — *Interface*
Multimodal feedback	To build intuition about system functionality and confirm successful actions	Returns both (1) 'effected' **audio;** (2) **visual** plot of frequency response pre/post-FX — *Interface*
Customization & Flexibility	Allowing users to disregard or modify outputs promotes user agency & ownership	Core functionality of system model -- Text2FX is audio-to-FXparameters, not (only) audio-to-audio — *Model*
Balanced Unpredictability	Sound-to-descriptor matching is a one-to-many problem, yet overly random outputs create distrust in the system	Adaptation & selection of Text2FX model variant with most consistent performance (via hyper-parameter sweep & small batch subjective validation) — *Model*

Text2EQ: Interaction Paradigm

A human-in-the-loop interface that optimizes and maps natural language descriptors to suggested EQ settings, allowing users to refine iteratively, combining manual control with ML-driven suggestions



System Backend: Text2FX Model



Single-instance optimization of audio FX parameters in the shared text-audio CLAP [2] embedding space

arxiv

Key Future Considerations

01 / Human-aligned parameter prediction: ML system should be designed such that the distribution of FX parameter predictions align with those chosen by human experts

02 / Low-Latency: currently takes ~40s which diminishes the user experience, how can we optimize for real-time usage, thus allowing for immediate feedback?

03 / FX chain generalization: modular architecture for toggling individual FX on/off within a larger FX chain

04 / Integration: should fit easily into existing DAW software

References

[1] B. Pardo, M. Cartwright, P. Seetharaman, and B. Kim, "Learning to Build Natural Audio Production Interfaces," Arts, vol. 8, no. 3, pp. 110, 2019.

[2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," in ICASSP 2023, pp. 1–5, IEEE, 2023.

[3] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models," in CHI 2020, pp. 1–13, 2020.