Language-Augmented Soundsketcher

Semantic Enrichment of Perceptual Audio Representations

Konstantinos Velenis Asteris Zacharakis Emilios Cambouropoulos

Aristotle University of Thessaloniki, School of Music Studies, Faculty of Fine Arts

Introduction

Soundsketcher visualises electroacoustic music through perceptually grounded mappings: audio features are translated to visual form based on audio-visual correspondences, producing sketches that reflect *how sound is perceived* rather than merely measured.

This work introduces a **language-augmented sound library**: each sound is automatically positioned on a semantic map whose axes are derived *directly from human ratings*—not from pre-specified categories or hand-picked labels.

Methodology

1. CLAP audio embeddings. Each sound encoded as a 512-dim vector via Contrastive Language-Audio Pretraining.

2. Ridge regression + LOOCV. Three models trained to predict human ratings. Leave-One-Out Cross-Validation provides reliable out-of-sample estimates.

3. Vocabulary decoding. 380 acoustic descriptors scored by CLAP and projected onto the learned directions. Pole words *emerged from the data*.

Two-axis map: Rough and smooth are near-opposite directions ($\cos = -0.79$); sharp is independent of rough ($\cos = +0.16$). Map axes: **X = roughness**, **Y = sharpness**.

Human Semantic Study

30 electroacoustic sounds spanning diverse textures and timbres and timbres

29 participants rated each sound on three semantic dimensions (0–100 continuous sliders):

- **Smoothness**
- **Roughness**
- **Sharpness**

Key Results & Conclusions

CLAP predicts human ratings accurately:

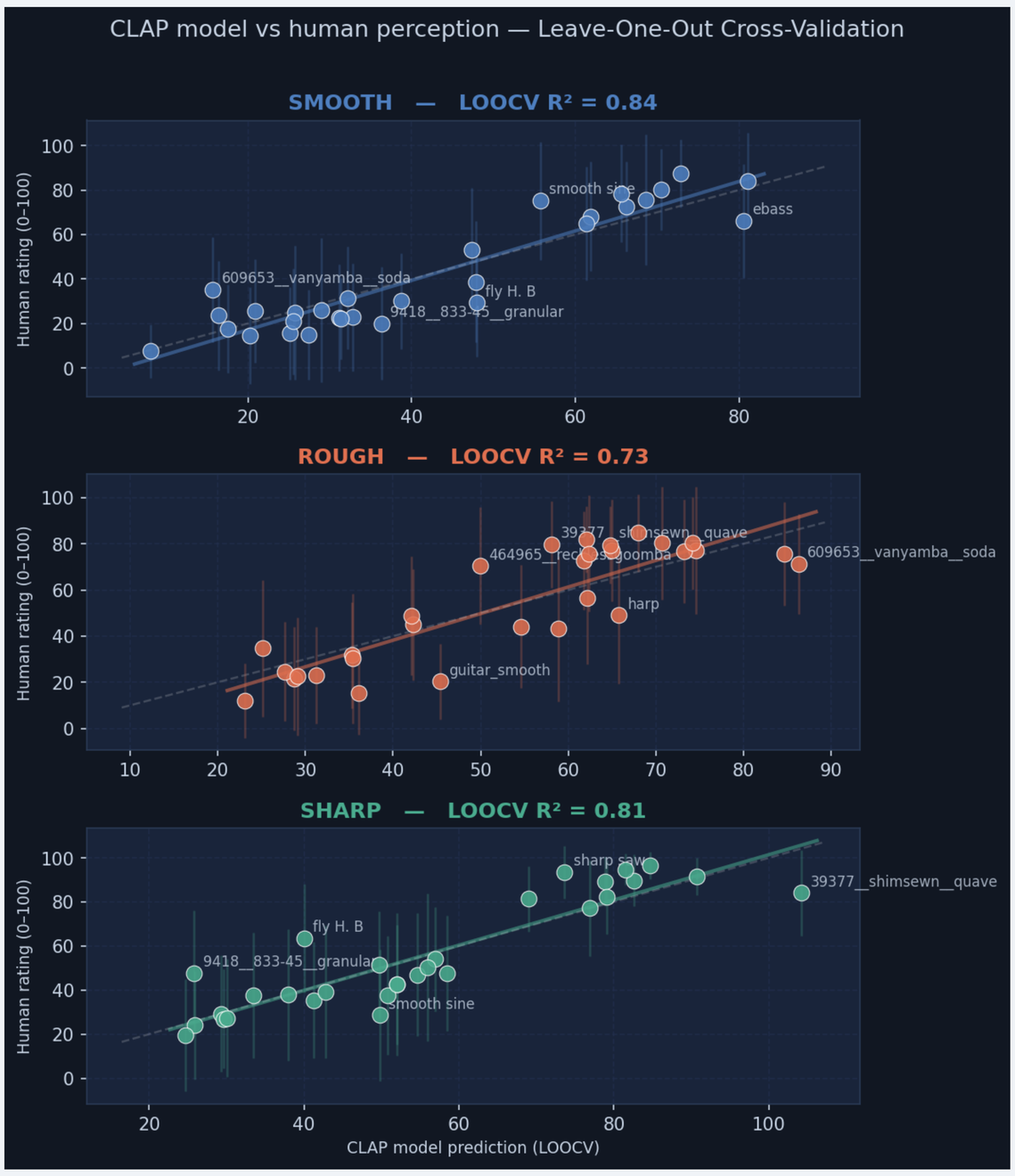
Dimension	LOOCV R^2
Smooth	**0.84**
Rough	**0.73**
Sharp	**0.81**

Pole words align with audio vocabulary: *sparkling* / *tinny* vs. *rumbling* / *dark*; *jagged* / *buzzy* vs. *pure* / *gentle*.

The semantic map grounds AI-based organisation in **human perception**

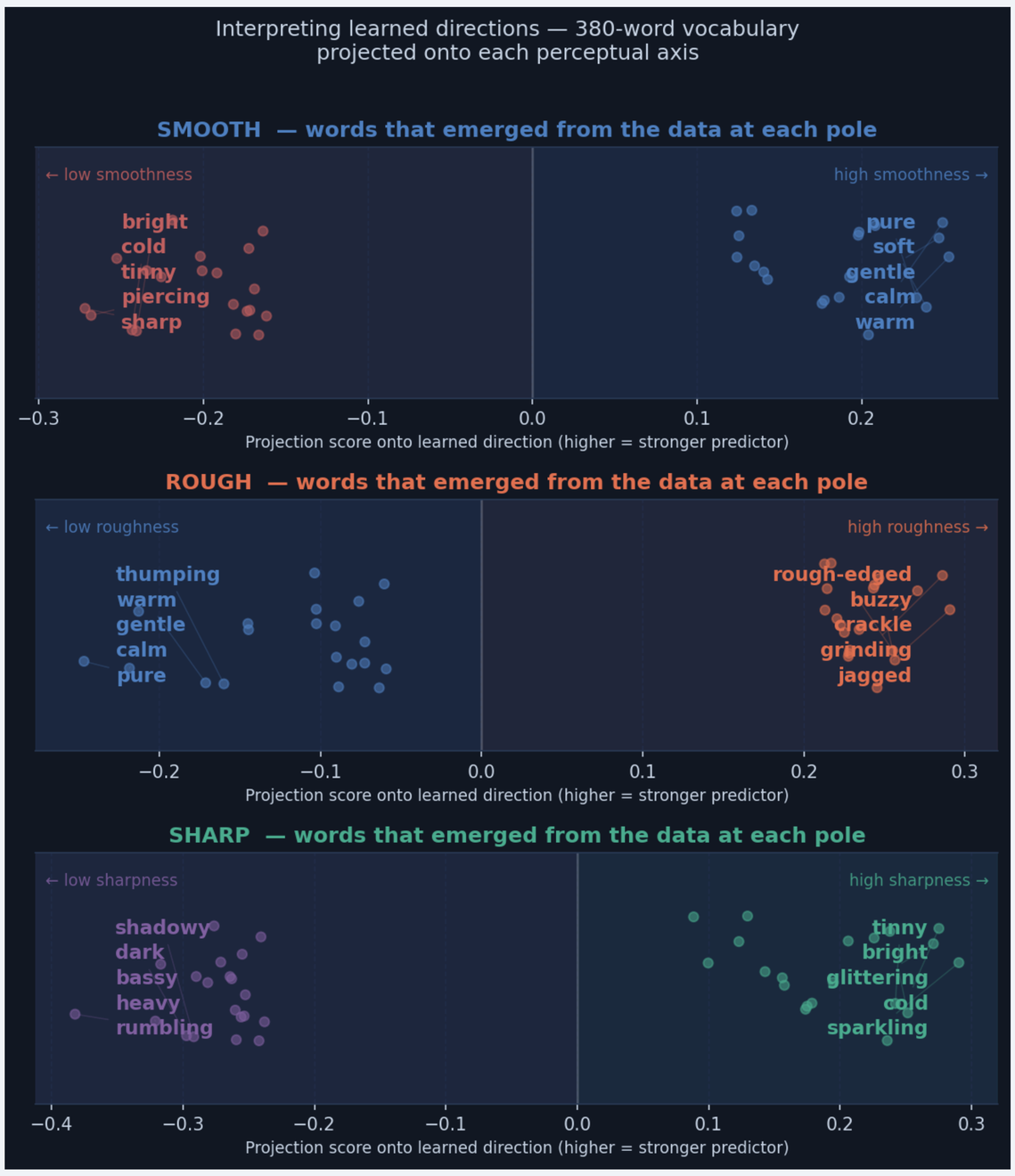
Future work: expand participant pool; add perceptual dimensions;

CLAP Predictions vs. Human Ratings — LOOCV



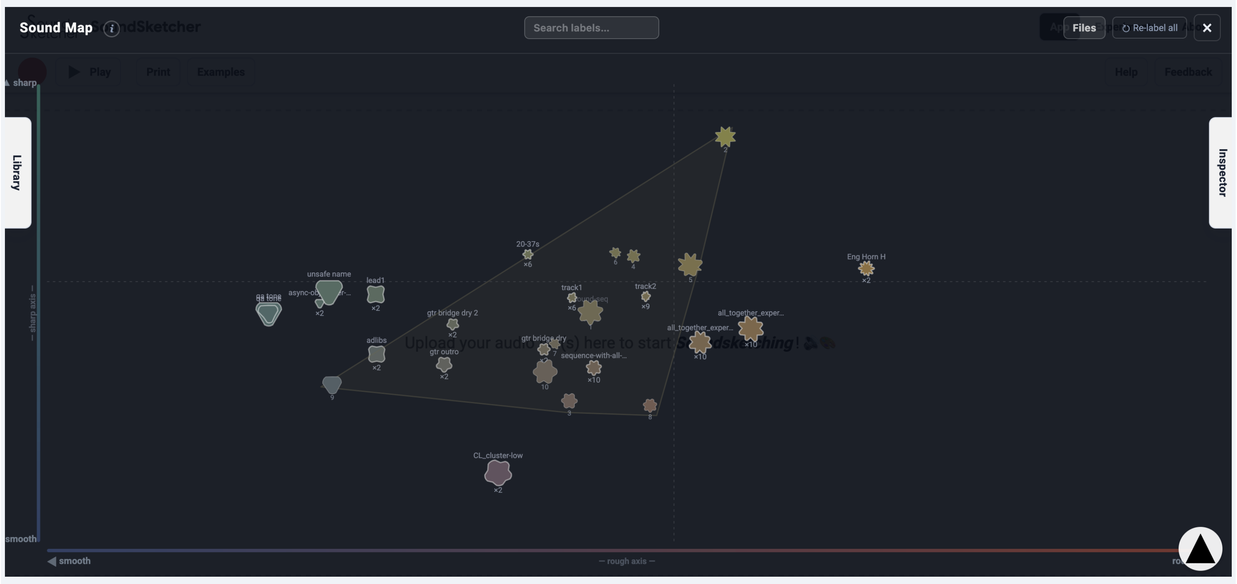
Each dot = one sound. Error bars = inter-rater spread (±1 SD). Dashed line = perfect prediction. Model never tested on the sound it predicts (Leave-One-Out Cross-Validation).

Vocabulary Decoding — What the Axes Mean



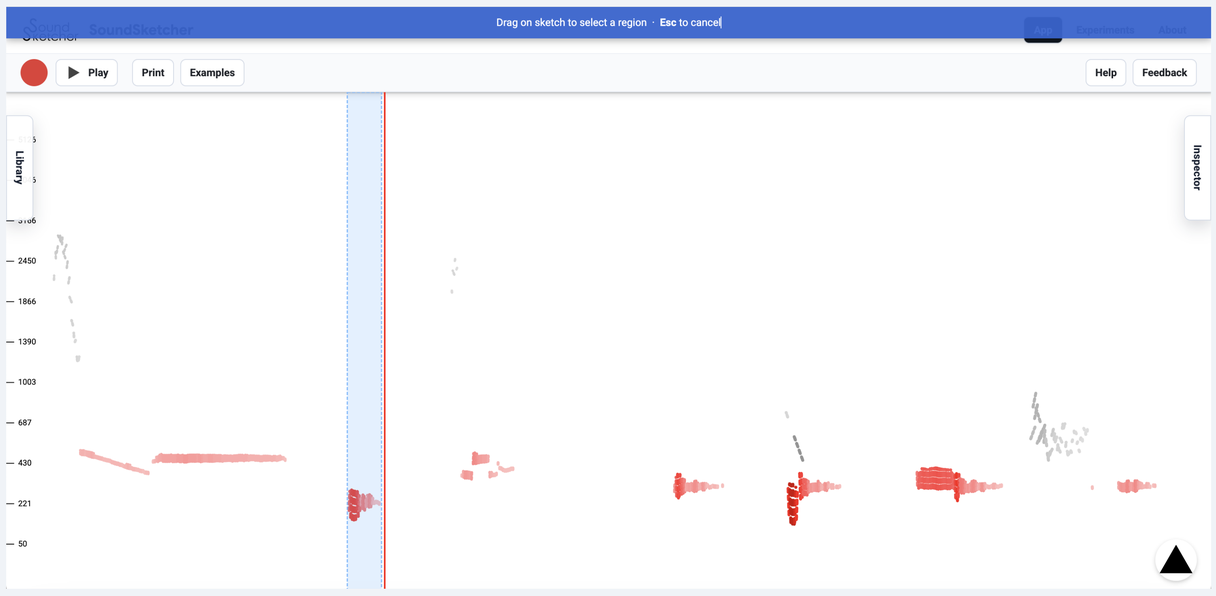
380 acoustic words projected onto each learned direction. Top-5 at each pole labelled. Words were **not selected in advance** — they surfaced from the regression.

The Semantic Sound Map in Soundsketcher



Sounds positioned by mean human ratings: **X = roughness**, **Y = sharpness**. Shape morphology encodes perception based on audio-visual correspondences: **spikiness** encodes roughness; **tooth depth** encodes sharpness; **size** encodes loudness. Composers navigate the collection from smooth, warm sounds (lower-left) to rough, sharp textures (upper-right) — guided by *how sounds feel*.

Soundsketcher — Sketch View



Spectral sketch of an electroacoustic piece in which each sound event is represented as a perceptual shape reflecting its timbral and psychoacoustic characteristics. The interface supports intuitive exploration of the sound library through semantic queries based on learned perceptual axes and content-based retrieval of acoustically similar sounds directly from the sketch.