

Oregon State University
Cascades

Bridging the Timbre Gap

A Generative Approach to Audio Labeling

Marc Vannucci & Patrick Donnelly

soundbendor lab

College of Engineering | Computer Science

Introduction

Oregon State University Cascades

- Music information retrieval (MIR) models classify genre, pitch, and instrumentation with high accuracy — but timbre, the psycho-acoustic “tone color” of a sound, remains difficult to model.
- Timbre is what makes a guitar and a piano sound distinct on the same note; it is described with subjective tags such as warm, hollow, metallic, and brittle.
- MIR lacks public, large-scale datasets labeled with these descriptors; labeling existing audio by hand is costly and inconsistent across listeners.
- Our contribution: a scalable, controllable pipeline that uses generative AI to synthesize a timbre-labeled audio dataset, addressing the data-scarcity problem directly.

Related Work

Oregon State University Cascades

- Three perceptual taxonomies inform our tag set:
  - Reymore & Huron (2020) map cognitive-linguistic dimensions of musical-instrument timbre using auditory-imagery tasks.
  - Carron et al. (2017) develop a lexicon and tool for communicating about sound features in design.
  - Pearce, Safavi, Brookes et al. (2019) release tools to semantically annotate timbre in non-musical content.
- We invert this: rather than labeling existing audio, we generate audio directly from these descriptors.

Data Set

Oregon State University Cascades

- Tags drawn from established perceptual taxonomies [1, 2, 3].
- Six physical domains broaden source variety:
  - musical, domestic, environmental, organic, electronic, mechanical
- 119 tags, stratified by domain
  - reduced from 194 raw entries via deduplication + normalization

119 Timbre tags	**6** Physical domains
62,394 Audio clips	**2 s** Clip length (mp3)

Summary of the generated dataset

Approach: Tag-Audio Pipeline

- Gemini generates diverse scenarios for each tag across domains, constrained to causal physical events (not atmospheric or spectral language).
- A validator removes cross-tag semantic collisions (near-duplicate scenarios) before synthesis.
- Cleaned scenarios fill a fixed prompt template: “A sound effect of [scenario] with a [timbre tag] quality.”
- ElevenLabs synthesizes a 2-second clip per prompt; each clip is tagged with its descriptor and a unique scenario ID, stored as a tag–scenario–clip triple.

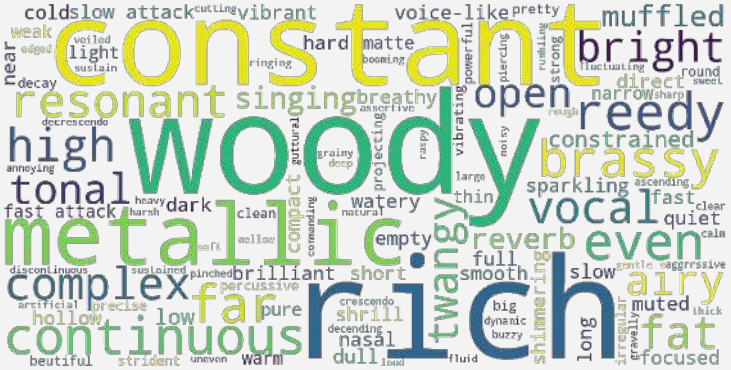


Phase 1: generating audio from timbre tags

Dataset Characterization

Oregon State University Cascades

- 62,394 clips generated across all 119 tags and 6 domains.
- Tag frequency is uneven — constant, woody, rich, and metallic dominate the generated scenarios.
- A validator removes near-duplicate scenarios; imperfect LLM adherence still produces some junk data, which is filtered out.
- Saturation ceiling: Gemini yields ~17–20 unique scenarios per tag per domain before collisions appear.



Tag frequency across the dataset

Discussion & Applications

Oregon State University Cascades

- A scalable, controllable framework for generating perceptually grounded audio in data-scarce domains; stratifying across six domains reduces bias toward any single sound source.
- Demonstrates that generative AI can synthesize audio aligned with timbre qualities humans recognize.

- Applications span music recommendation systems, audio forensics, and automated sound design.
- Limitation: imperfect LLM adherence to the prompt introduces junk data, which we filter — an inherent trade-off of synthetic generation.

Next Steps: Classification

- Train a classifier to predict the timbre tag directly from audio (Phase 2).
- Extract spectral features and learned embeddings from each clip as model input.
- Measure disagreement between the original tag and the predicted tag.
- Use that disagreement to refine perceptual modeling of subjective timbre.



Phase 2: classifying timbre tags from audio

References

[1] Maxime Carron, Thomas Rotureau, and Françoise Dubois et al. “**Speaking about sounds: a tool for communication on sound features**.” Journal of Design Research, 15(2):85, 2017

[2] Andy Pearce, Saeid Safavi, and Tim Brookes et al. **D5.8: Release of timbral characterisation tools for semantically annotating non-musical content**, 2019

[3] Lindsey Reymore and David Huron. “**Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia.**” Psychomusicology, 30(3), 2020



Oregon State University