Bridging the Timbre Gap: *A Generative Approach to Audio Labeling*

Marc Vannucci and Patrick Donnelly

Oregon State University-Cascades

Background

In the field of music information retrieval, machine learning models have achieved high accuracy in objectively classifying audio by genre, pitch, and instrumentation. However, a significant gap remains in the analysis of timbre which is the psycho-acoustic quality or tone color of a sound. Timbre is what makes a guitar sound distinct from a piano even when playing the same note. Currently, MIR lacks public high-quality, large-scale datasets labeled with subjective, descriptive timbre tags (e.g., *hollow, warm, metallic, brittle*).

Overview

This research aims to address the data scarcity problem for automatic timbre classification by constructing a novel, synthetic dataset designed specifically for subjective timbre analysis. Rather than manually labeling thousands of existing audio clips, which is costly and inconsistent, we will automate the workflow. We will utilize state-of-the-art Generative AI audio models (specifically, ElevenLabs API) to synthesize sound snippets based on specific, subjective text prompts.

Dataset

- Timbre tags from established perceptual taxonomies [1, 2, 3]
- Consideration of six physical domains for increased variety:
  - *musical, domestic, environmental, organic, electronic, mechanical*
- Scenarios are generated across 119 tags, stratified by domain
- Audio is generated directly from textual scenario descriptors
- Synthetic audio clips are paired with a timbre descriptor label

Table 1. Summary of the generated dataset

Physical Domains	6
Timbre Tags	119
Total audio clips	62,394
Duration of clip	2 seconds
Audio Format	`mp3`

Generative Process

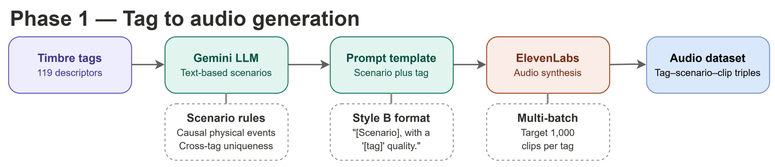

Figure 1. Diagram illustrating the process of generating audio from timbre tags

1. Gemini LLM generates diverse scenarios per tag across domains
2. Validator removes cross-tag semantic collisions (near duplicates)

Table 2. Examples of text scenarios generated by Gemini

tag	domain	source	scenario
bright	musical	piccolo flute	a piccolo flute playing a rapid trill on a mountain peak
metallic	electronic	static	the clean hiss of FM white noise on a receiver
woody	organic	hooves	horse hooves trotting over a hollow wooden bridge
reedy	environmental	roots	a tree root snapping with a reedy resonance
vocal	domestic	floor	a floorboard creaking with a vocal tone
heavy	mechanical	turbine	a heavy hydroelectric turbine whining

3. Cleaned text scenarios are inserted into audio prompt template:
   - *"A sound effect of* `[scenario]` *with a* `[timbre tag]` *quality."*
4. ElevenLabs API performs audio synthesis based on scenario

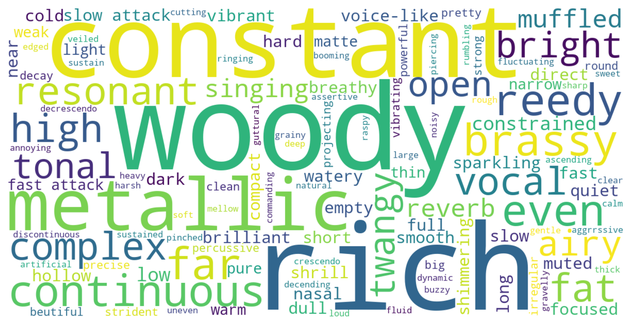

Figure 2. Word cloud illustrating frequency of the generated timbre tags

5. Audio clips are tagged with descriptor and unique scenario ID
6. Dataset is organized by scenario, tag, and audio file path

Potential Applications

- By utilizing Gen AI to populate data-scarce domains, we provide researchers with datasets to study qualitative aspects of sound
- Many broad applications ranging from music recommendation systems to audio forensics and automated sound design

Next Steps

- Train classifier on dataset to predict timbre tag from audio
- Measure disagreement between original tag and prediction
- Use results to refine perceptual modeling of subjective timbre

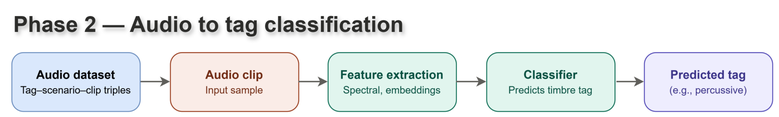

Figure 3. Diagram illustrating the next phase of classifying timbre from audio

Conclusion

- Scalable, controllable framework for generating perceptually grounded audio datasets for domains in which data is scarce
- Demonstrates that generative AI can synthesize audio aligned with timbre qualities that can be understood by humans
- Stratified scenario generation reduces bias to specific sources
- **Limitation**: imperfect LLM adherence to the prompt introduced some junk data which we filtered out from the final dataset

References

[1] Maxime Carron, Thomas Rotureau, and Françoise Dubois *et al.* **Speaking about sounds: a tool for communication on sound features.** *J. 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.