Classifying Audio Timbre Without Audio Using Text-only Training

Peter McCabe, Patrick J. Donnelly

Oregon State University

Text-only Training

Text-only Training (ToT) refers to training machine learning models on natural language descriptions of data rather than the data itself.

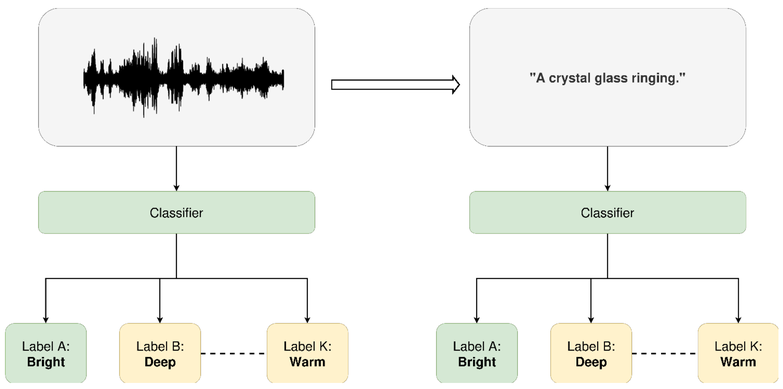


Figure 1. Audio timbre classification with text data.

Audio timbre classification refers to matching audio to the most fitting subjective sound qualia. In other words, how would a person describe a sound regardless of source, context, or content?

Modality Gap

We use a multimodal audio-language model to embed text descriptions into a shared vector space with audio. However, we cannot directly substitute the audio with text due to the **modality gap.** In practice, data from different modalities are embedded in separate subspaces.

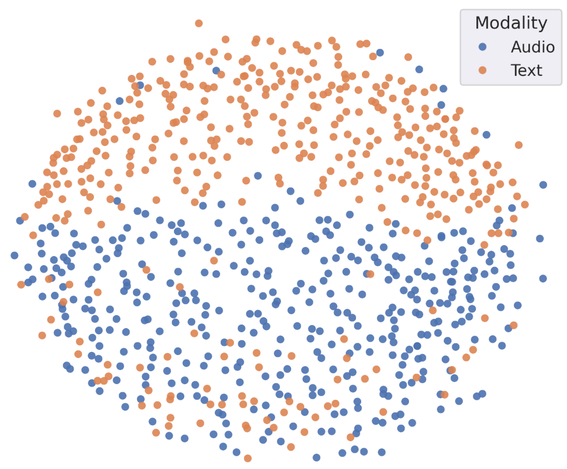
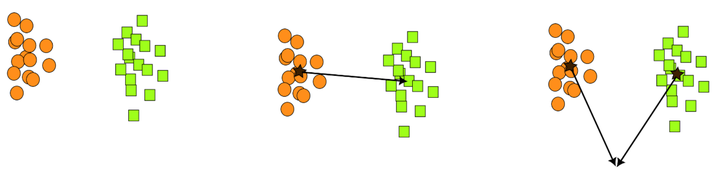


Figure 2. An example of the modality gap in CLAP [1] with samples drawn from the Clotho [2] dataset.

Augmentation Strategies

We employ **six augmentation methods** to improve approximation under the modality gap:

Clean	Nothing	Mean Shifting[3]	Connect-Collapse (C2)[4]
Noisy	Gaussian Noise[5]	Shift + Noise[6]	C2 + Corrupt (C3)[7]



No augmentation

Mean shifting pushes one modality towards the other

C2 centers data at 0

Text Datasets

We create four text datasets using three methods. First, we create two datasets by comparing the text-to-text semantic similarity between captions and labels with CLAP and SBERT [8].

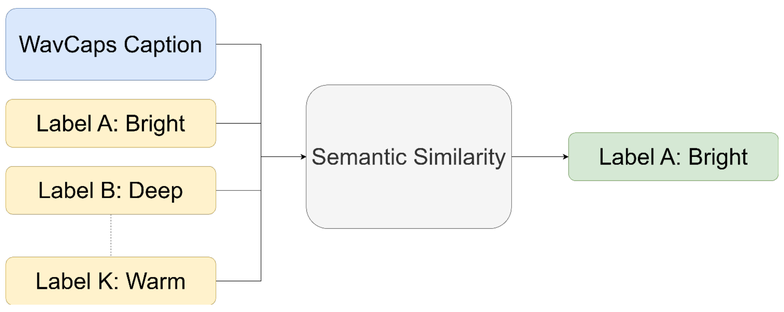


Figure 3. WavCaps labeling process. Each caption and label (as full sentences) are embedding into CLAP or SBERT [8], then the label sentence with the highest semantic similarity to the caption that is selected.

Second, we create a dataset entirely from Gaussian noise centered at each label in the CLAP embedding space. We make six versions of each label sentence to improve dataset diversity.

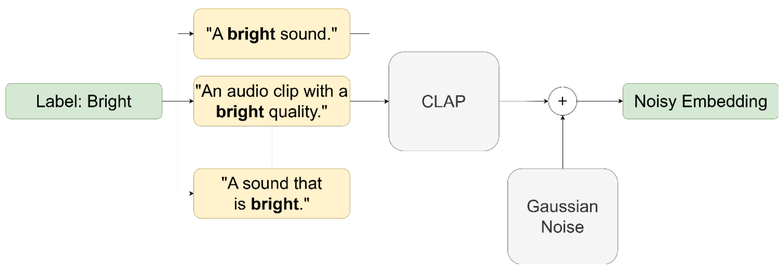


Figure 4. Label-Centered Noise dataset creation process.

Last, we use an LLM to create a caption dataset based on a given prompt and timbre label.

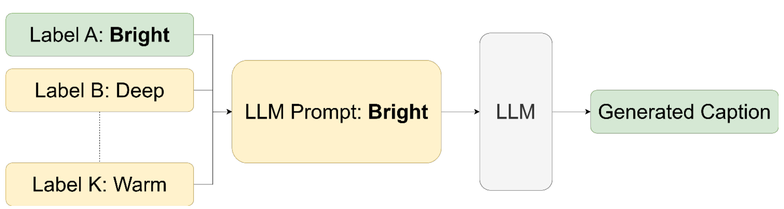


Figure 5. LLM-Generated process. The LLM [9] is prompted to write pseudo-captions for a sound that a person might describe with the provided timbre label.

References

[1] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. **Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation.** In *Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023.

[2] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *Conference on Acoustics, Speech and Signal Processing*, pages 736–740. IEEE, 2020.

[3] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Conference on Neural Information Processing Systems*, pages 17612–17625, 2022.

[4] Yuhui Zhang, Jeff Z. HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and Rectifying Vision Models using Language. 2022.

[5] Soham Deshmukh, Benjamin Elizalde, Dimitra Emmanouilidou, Bhiksha Raj, Rita Singh, and Huaming Wang. Training Audio Captioning Models without Audio. In *Conference on Acoustics, Speech and Signal Processing*, pages 371–375. IEEE, 2024.

[6] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! : Learning Visual Tasks Using Only Language Supervision. In *Conference on Computer Vision*, pages 2672–2683, 2023.

[7] Yuhui Zhang, Elaine Sui, and Serena Yeung. Connect, Collapse, Corrupt: Learning Cross-Modal Tasks with Uni-Modal Data. In *Conference on Learning Representations*, 2024.

[8] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

[9] Aaron Grattafiori et al. The Llama 3 herd of models, 2024.

[10] Andy Pearce, Saeid Safavi, Brookes Tim, Wenwu Mason, Russell Wang, and Mark Plumbley. D5.8: Release of timbral characterisation tools for semantically annotating non-musical content, 2019.

Results: Text Datasets

We train 10 classifiers on each combination of text dataset, modality gap augmentation, and test dataset. We test on real audio data from a binarized version of the AudioCommons datasets [10]. Classification results are presented as the mean and standard deviation of the F_1-scores.



Figure 6. Average F_1-scores grouped by text dataset creation method. Datasets that take advantage of CLAP's domain knowledge led to higher average test scores.

Results: Augmentations

We find that applying augmentation improves average score: C2 and Mean Shift are generally comparable, but applying no augmentation leads to strictly lower scores. We also find that adding noise has very little impact on test scores, despite common practice throughout the literature.

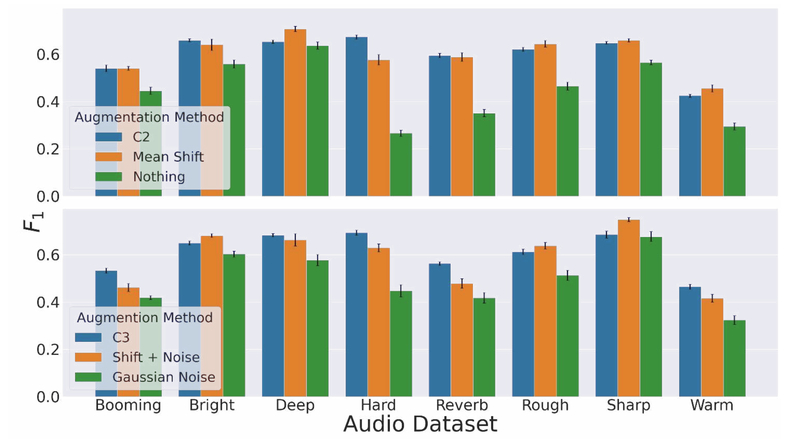


Figure 7. **Top:** Average scores for Clean augmentation methods. **Bottom:** Average scores for Noisy methods.

Final Results

Dataset	Booming	Bright	Deep	Hard	Reverb	Rough	Sharp	Warm
Average Model	0.52	0.66	0.68	0.64	0.56	0.63	0.68	0.44
CLAP Zero-Shot	0.58	0.76	**0.85**	0.25	0.57	0.62	0.75	0.47
Best Model	**0.67**	**0.79**	**0.85**	**0.81**	**0.66**	**0.78**	**0.80**	**0.62**
Audio Training	0.72	0.79	0.89	0.85	0.82	0.81	0.78	0.77

Overall we find:

- Our average model **trained entirely on text** is competitive with CLAP zero-shot
- Our best **ToT** model meets or exceeds CLAP's zero-shot scores in all test datasets
- Our best model can compete with the average model trained on real audio
- Our best model outperforms the average audio model in 37.5% of datasets.

Oregon State University

soundbendor lab