Investigating Timbre Representations in CLAP Across Modalities via Perturbations

Devyani Hebbar DH3677@NYU.EDU
Steinhardt, New York University, USA

Brian McFee BRIAN.MCFEE@NYU.EDU
Steinhardt, Center for Data Science, Music and Audio Research Laboratory (MARL), New York University, USA

Editors: D. Herremans, K. Bhandari, A. Roy, S. Colton, M. Barthet

Abstract

The transition from feature-based language-audio representations to more high-dimensional ones from pre-trained foundation models has enabled us to map audio content to a significantly broader vocabulary of natural language. However, some interpretability of the alignment between the embedding spaces of the two modalities and their relation to psychoacoustic features is lost as a byproduct. In this study, we investigate timbre representations in CLAP in both the text embedding space and audio embedding space. We identify directions for different timbral qualities in each embedding space and use them as perturbation vectors in uni-modal and cross-modal Text-to-Music (TTM) Retrieval and Generation downstream tasks. We find that although both audio and text embeddings move monotonically along their respective timbre directions, timbral variation is more linearly distributed—and therefore more easily exploitable—in the audio embedding space. Cross-modal perturbation experiments further reveal that the audio and text embedding spaces form a geometrically aligned subspace with respect to timbre. Additionally, our analysis identifies cases where CLAP's timbre representations closely align with perceptually grounded spectral features, and cases where such alignment is limited.

Keywords: CLAP, Timbre, Embedding Space, Interpretability, Text-to-Music Retrieval, Text-to-Music Generation

1. Introduction

Early Music Information Retrieval (MIR) systems linked Natural Language to Audio by connecting descriptive words to psychoacoustic features (Slaney, 2002), (Turnbull et al., 2008). With the advent of Deep Learning, these discrete representations have been replaced by more continuous, high-dimensional ones, in the form of embeddings from pre-trained language-audio models. CLAP (Wu et al., 2022) is one of the most widely used pre-trained models in this category, in tasks such as Text-to-Music (TTM) Retrieval, Generation, and Query-Based Audio Source Separation. While such a representation opens the gateway to language-audio tasks of higher complexity by allowing the use of more diverse text descriptions, some interpretability on the alignment between language and audio is lost. This "black box" nature makes it challenging to design balanced training sets for downstream tasks. Identifying semantic directions in CLAP's latent space—and further, investigating the nature of alignment between directions across the two modalities—would

allow us to regularize downstream models more easily. This is especially beneficial for timbre, an aspect of sound that is more complex and harder to quantify as compared to pitch or rhythm (Saitis and Weinzierl, 2019). Verifying that a foundation model captures timbre coherently across modalities, and identifying circumstances under which it does not, would benefit several downstream tasks where timbral accuracy is imperative. This study aims to identify timbre directions in both the audio and text latent spaces of CLAP. Our pipeline for identifying these directions is inspired by the approach proposed in (Srivastava et al., 2022) and (Deng et al., 2025). The validity and dimensionality of these directions are then investigated by using them as perturbation vectors in both unimodal and cross-modal settings. The perturbed embeddings are used in TTM Retrieval and Generation downstream tasks and performance is compared with that of the systems that use the original embeddings.

2. Methodology

2.1. Datasets

To isolate timbre, each audio signal for a given instrument must maintain constant timbre, pitch, and loudness throughout its duration. Furthermore, for signals of the same instrument at the same pitch, any differences should arise solely from timbre. The audio data must also be paired with natural language descriptions relating to spectral content for this study. Large descriptive datasets like MusicCaps (Agostinelli et al., 2023) and SongDescriber (Manco et al., 2023) were unsuitable because each clip contains complex mixtures of multiple instruments and exhibits varying acoustic properties. We therefore use Nsynth (Engel et al., 2017) which provides isolated single-instrument notes labeled with three timbral adjectives ('bright,' 'dark,' 'distorted'). We keep only those audio samples in each instrument category that are labeled with one of the three timbral adjectives. The large scale of Nsynth provides sufficient data to support a robust investigation of CLAP's embedding space despite this filtering.

2.2. Identifying timbre directions in the audio embedding space

We use Singular Value Decomposition (SVD) to find a direction in the audio embedding space that encodes each timbral quality. Rather than treating this direction as a unique semantic axis for a timbral quality, we use it to describe how variation associated with a given timbral quality is spread within CLAP's audio embedding space. Deng et al. analyze how audio embeddings evolve as continuous audio effects are gradually applied, showing that these changes follow consistent trajectories while exhibiting structure beyond a strictly one-dimensional path. Motivated by this, we treat timbre as continuous and potentially multi-dimensional, and represent each timbral quality using a single composite direction formed as a weighted combination of multiple directions that capture the main patterns of variation associated with that quality. CLAP embeddings of audio samples from the Nsynth training set are organized into a matrix for each timbral quality. We use the LAION-CLAP implementation (Wu et al., 2022) with fusion disabled and the default pretrained checkpoint loaded. It is ensured that the selected samples only have the timbral adjective in consideration– and no other qualities related to spectral content– listed in the metadata. Since the same timbral adjective can be associated with varying spectral qualities when used

across different instruments, the selected samples are distributed evenly across instruments to ensure that the resulting direction does not strongly encode any instrument over another. The expression for the resulting matrix is given as:

$$X_a \in \mathbb{R}^{N_a \times d} \tag{1}$$

where X_a is the matrix of CLAP audio embeddings associated with timbral adjective a, N_a is the number of rows and is equal to the number of audio embeddings, and d = 512 (dimensionality of each CLAP embedding). The embeddings are mean-centered before performing the SVD. Singular values were normalized and used to weight the top-k right-singular vectors capturing ≥60% variance. To determine whether each component vector must be positive or negative, for each quality, a small set of audio embeddings is perturbed using one vector and one sign at a time. The sign that results in higher cosine similarity with the text embedding of the corresponding adjective is derived as the correct orientation for each component vector. The audio perturbation vector for each timbral quality is then computed as the signed linear combination of the scaled singular vectors:

$$P = \sum_{i=1}^{k} z_i V_{ti} \tag{2}$$

where V_{ti} is the i-th right-singular vector obtained from the SVD of X_a and z_i is the corresponding normalized singular value used to weight that component.

2.3. Identifying timbre directions in the text embedding space

Our approach to identifying a direction for a given timbral adjective in the text embedding space is inspired by the work presented in (Mikolov et al., 2013), where regularities in linear offsets between word embeddings were first observed. We create captions for each adjective in a Likert scale format:

$t1$	This recording of a musical instrument does not sound [adjective]
$t2$	This recording of a musical instrument sounds a little [adjective]
$t3$	This recording of a musical instrument sounds [adjective]
$t4$	This recording of a musical instrument sounds very [adjective]
$t5$	This recording of a musical instrument sounds extremely [adjective]

Based on the property of regularities in linear offsets between word embeddings, for a given timbral adjective, the difference vectors between adjacent prompts on the Likert scale should encode incremental changes in the timbral quality. To ensure that only consistent directional changes contribute to the final vector, we compute cosine similarities between all adjacent offsets and retain those that point most coherently in the same direction. For the "bright" prompts, it is observed that only the difference vectors $(t3-t2)$ and $(t4-t3)$ have a moderately higher cosine similarity with each other as compared to all other pairs. Thus, a mean of these two difference vectors is computed and chosen as the text perturbation vector for this adjective. For the "dark" and "distorted" prompts, there is no pair of difference

vectors that has considerably higher cosine similarity than the others, so a mean of all five difference vectors is computed as the text perturbation vector for these adjectives.

3. DOWNSTREAM TASKS

The identified text and audio directions perturbation vectors are used in two downstream tasks – Text-to-Music Retrieval and Text-to-Music generation.

3.1. Text-to-Music Retrieval

Audio and text pairs from the Nsynth test set are used for this task. It should be noted that there is no overlap of samples with the train set that was used in the matrix SVD computations. Samples are chosen in a manner similar to Section 2.1. The task is set up as a cosine similarity-based ranking problem. Given a text query, the algorithm extracts CLAP text embeddings for it and computes its cosine similarity with the audio embeddings of all the audio files in the database. The top k audio files with the highest ranking cosine similarities are then returned. The system is evaluated using four evaluation metrics: Precision, Recall, Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), all computed at K=100. A larger value of K is chosen, rather than a small pool of only the top 5 or 10 items, to ensure that the evaluation results reflect the global alignment between the text and audio embeddings.

The retrieval process is conducted using two query templates for each adjective. The query templates are: "A [adjective] sound" and "a musical instrument that sounds [adjective]". Evaluation metrics are computed separately for each adjective, and averaged across the two queries. The retrieval system without embedding perturbations of any kind serves as the baseline for each adjective. The retrieval process is then repeated by performing two types of perturbation procedures: unimodal perturbation, where the audio perturbation vector is applied to audio embeddings (AA), and the text perturbation vector is applied to text embeddings (TT); and cross-modal perturbation, where the audio perturbation vector is applied to text embeddings (TA) and the text perturbation vector is applied to audio embeddings (AT). In both settings where audio embeddings are perturbed, only those audio embeddings corresponding to the adjective in consideration are perturbed, while the rest remain unchanged. Similarly, for text perturbation experiments, all audio embeddings remain fixed and only the query embedding is perturbed. The perturbation vectors are derived as described in Sections 2.2 and 2.3 , and are scaled by a range of α values to examine the effect of increasing perturbation strength. The expression for the perturbed embeddings is given as:

$$E_{perturbed} = E + \alpha_i P \tag{3}$$

where $E_{perturbed}$ is the perturbed text or audio embedding, E is the original text or audio embedding, α_i is the scaling variable, and P is the text or audio perturbation vector.

3.2. Text-to-Music Generation

To quantitatively assess whether the identified audio and text directions correspond to perceptual timbre attributes, we generate audio using AudioLDM (Liu et al., 2023), since perturbed embeddings cannot be directly converted to sound and retrieval results alone

do not provide numerical measures of timbral change. The model leverages CLAP's cross-modal alignment by using CLAP text embeddings as a conditioning signal and generating audio in CLAP's audio latent space. The CLAP text encoder is modified such that a perturbation vector can be added to the embedding of the input text prompt. We modify the encode prompt function in the Hugging Face AudioLDMPipeline to append a custom perturbation vector to the original CLAP text embedding before it enters the diffusion pipeline. Thus, perturbation vectors are patched to the text encoder, which allows direct control over how much the conditioning prompt is perturbed. Two sets of experiments are performed for each adjective– one where the text encoder is modified using the text perturbation vector (uni-modal or TT) and another using the audio perturbation vector (cross-modal or TA). The perturbation vectors are scaled by a range of α values as expressed in Equation (1) For each adjective, two prompt templates are used for generation tests. For every prompt and each perturbation vector scaled across different values, a fixed seed ensures that changes in spectral features reflect only the perturbation and not diffusion randomness. Three distinct seeds are used per prompt, and the resulting 10-second audio clips are averaged across seeds and prompts.

4. RESULTS

4.1. Text-to-Music Retrieval

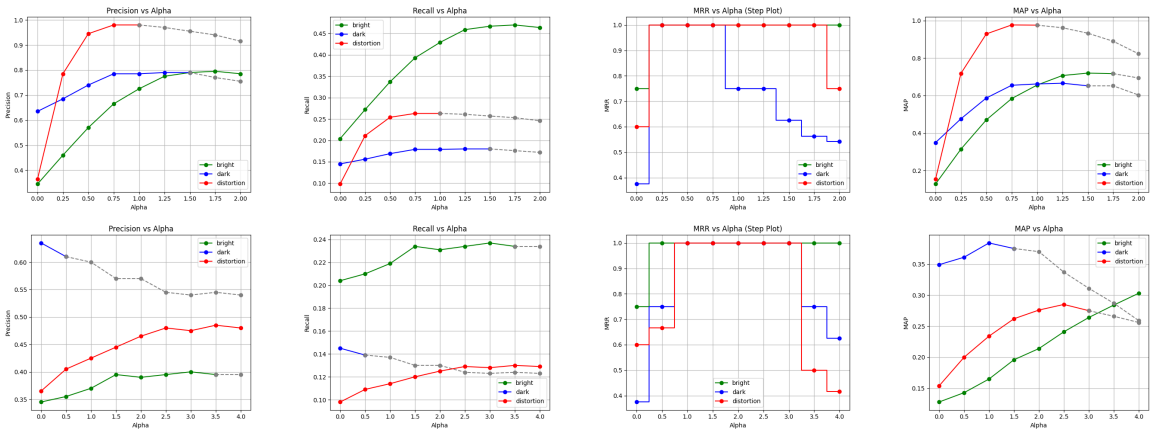


Figure 1: From left to right: Precision, Recall, MRR, and MAP vs Perturbation Strength (α) for AA (top) and TT (bottom). Gray segments mark perturbation strengths α beyond the range of consistent monotonic improvement, defined as the first point where the metric decreases for two consecutive values of α, even if the metric later increases.

In the case of AA, there is a continuous rise in metrics as perturbation strength increases up to a certain threshold, beyond which performance either stagnates or declines gradually. For the timbral quality "bright", precision, recall and MAP rise sharply from all

the way up to $\alpha = 1.75$ or 2, while MRR reaches its peak value of 1 early on at $\alpha = 0.25$ (Table 1) (Figure 1). These trends are also evident in the UMAP plots in Figure 2 that depict the clustering of audio embeddings around the text query as perturbation strength is increased. As α increases, a larger number of audio embeddings labeled as "bright" cluster closer to the query, and replace the incorrectly retrieved "dark" and "distorted" audio embeddings. Comparable trends are observed for dark and distorted, with slightly lower optimal α thresholds (1–1.5) and steeper early rises in MAP. These observations collectively reinforce the two key implications:

I] Applying the identified audio perturbation vector to audio embeddings labeled as a given timbral adjective steers them further towards the adjective query, i.e., the embeddings move monotonically in this direction up to a certain threshold. This suggests that there is a well defined direction for this timbral quality in CLAP's audio embedding space.

II] The identified audio perturbation vector successfully encodes this direction.

In the case of TT, the changes in metric values across perturbation strengths are less consistent as compared to those in AA— bright and distorted showed slight, uneven gains in precision and recall, while dark declined beyond $\alpha = 1$ MAP rose modestly overall, and MRR reached 1 for all adjectives across several α values. These observations imply that while there is a direction for each timbral adjective in CLAP's text embedding space, they may not be as well defined as their counterparts in the audio embedding space and do not move as monotonically. The magnitude of the directions may also vary from one adjective to another in the text embedding space.

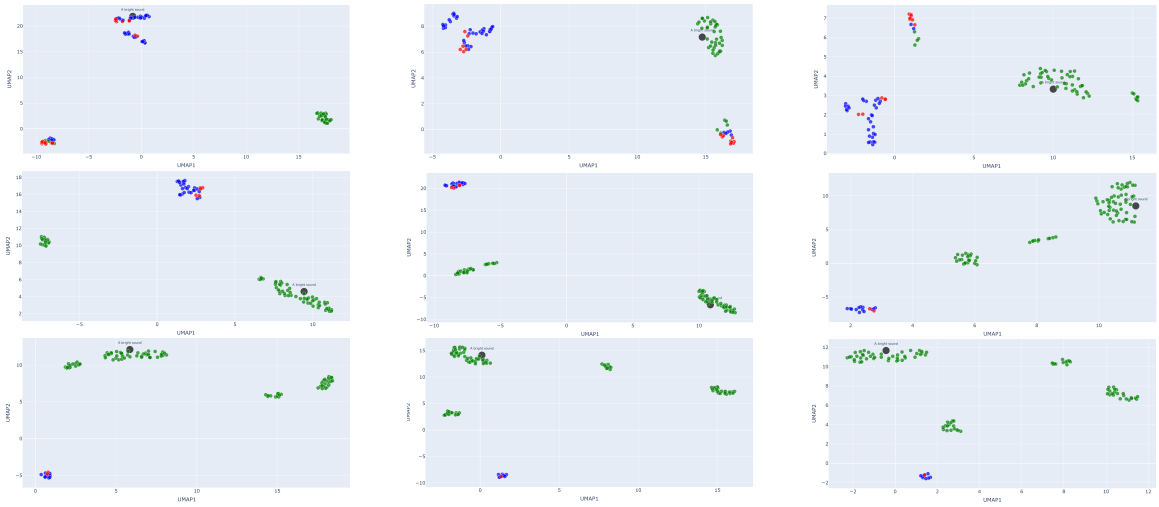


Figure 2: UMAP plots for AA perturbation of "bright" across increasing perturbation strength α. The black point marks the text-query embedding; green, blue, and red indicate audio embeddings for "bright", "dark", and "distorted" samples, respectively.

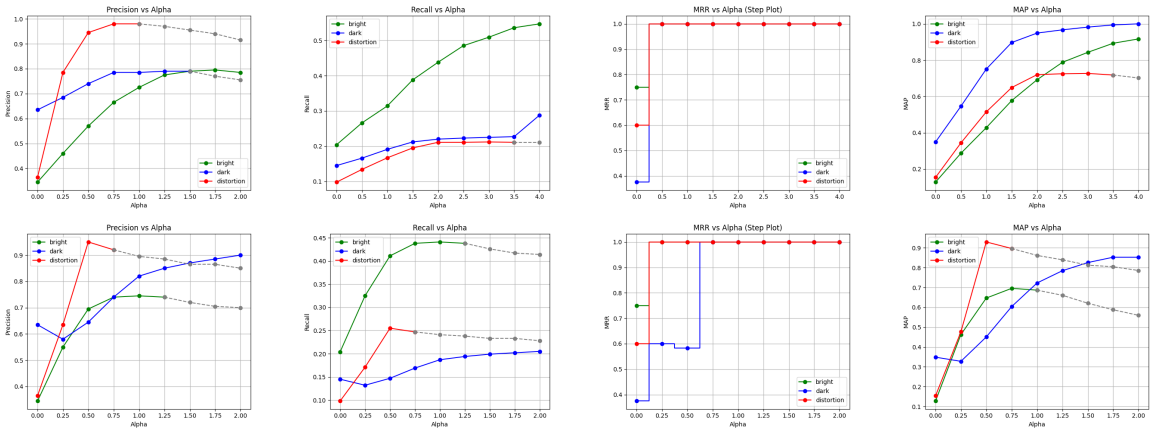


Figure 3: From left to right: Precision, Recall, MRR, and MAP vs Perturbation Strength (α) for AT (top) and TA (bottom). Gray segments mark perturbation strengths α beyond the range of consistent monotonic improvement, defined as the first point where the metric decreases for two consecutive values of α, even if the metric later increases.

The results for AT in Figure 3 and Table 2 show that text perturbation vectors have a much stronger effect on the audio embeddings than they did on the text query embeddings in the case of TT. When the same text perturbation vectors are applied to the corresponding audio embeddings, an increase in α results in a much sharper rise in both precision and MAP for all three adjectives as compared to those in TT. This is especially evident for the "dark" adjective, where both precision and MAP reach a peak of 1. This suggests that, for the same scalings, text perturbation vectors are more effective in steering the audio embeddings in the desired timbral direction, than they are in steering the text embeddings. The MRR values also remain at 1 for many more values of α , indicating that there is a higher threshold for over-steering in this case. In the case of TA, there is a rise in precision, recall and MAP for all three adjectives, but especially for "distorted", are extremely sharp. This trend is similar to what was seen in AA, however, the curves are slightly steeper in TA as compared to AA. This implies that for the same scalings, the audio perturbation vectors are equally, or even slightly more effective in steering text embeddings in the desired timbral direction.

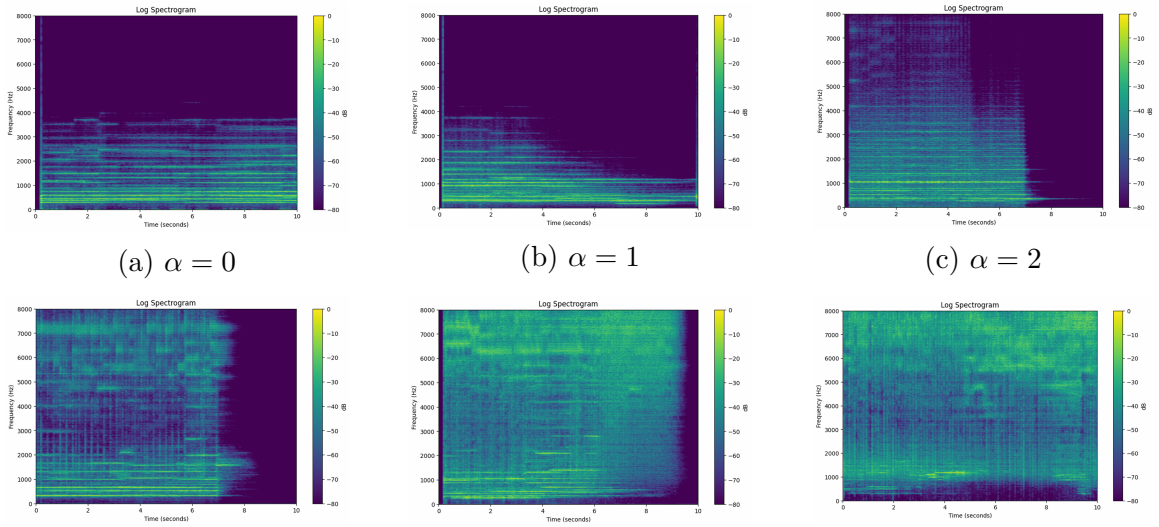
These observations have two further implications:

III] Cross-modal perturbations reveal a shared timbre subspace: Cross-modal perturbations reveal that the timbre directions identified in each modality are not just modality-specific artefacts, but constitute a geometrically aligned sub-space across CLAP's audio and text encoders.

IV] Steering the audio embeddings along the text direction by even a small amount, results in higher cosine similarities, as compared to when the text embeddings are steered

along the same direction. This implies that timbre variance may be more linearly spaced– and therefore, more exploitable, in CLAP's audio space as compared to the text space.

4.2. Text-to-Music Generation



(a) $\alpha = 0$ (b) $\alpha = 1$ (c) $\alpha = 2$

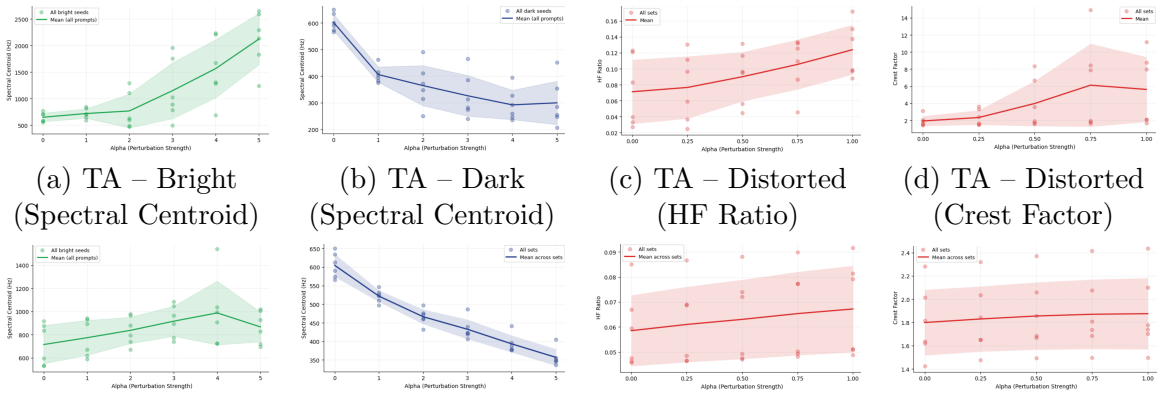
(d) $\alpha = 3$ (e) $\alpha = 4$ (f) $\alpha = 5$

Figure 4: Spectrograms for one set of audio clips generated from AA perturbations for "bright" across increasing Perturbation strengths (α values).

For the timbral adjective "bright", spectral centroid is measured for each generated audio clip, where higher values correlate with greater brightness. Figure 5 shows spectral centroid values (across two "bright" text prompts and all seed values) for TA and TT. In both cases, spectral centroid increases almost linearly with perturbation strengths. This is also reflected in Figure 4, where spectrograms for one set of audio generations are shown. High frequencies are visibly enhanced and low frequencies reduced as perturbation strength increases. The TA generations exhibit a slightly steeper slope and higher overall centroid range as compared to TT, suggesting that audio-side perturbations produce stronger perceptual changes in brightness.

Similarly, spectral centroid is also measured for "dark", where lower values indicate more darkness. In both cases, spectral centroid decreases almost linearly with increasing perturbation strength (Figure 5), with TT resulting in more gradual and subtle acoustic shifts again.

Distortion is less straightforward to quantify than "bright" or "dark," so we measure both the High-Frequency Energy ratio (HF ratio) and Crest factor. A higher HF ratio generally indicates more spectral saturation or distortion, while a lower Crest factor is expected under strong clipping. In the current results, the HF ratio increases almost linearly with perturbation strength for all conditions, suggesting a progressive rise in high-frequency har-



(a) TA – Bright (Spectral Centroid)
(b) TA – Dark (Spectral Centroid)
(c) TA – Distorted (HF Ratio)
(d) TA – Distorted (Crest Factor)
(e) TT – Bright (Spectral Centroid)
(f) TT – Dark (Spectral Centroid)
(g) TT – Distorted (HF Ratio)
(h) TT – Distorted (Crest Factor)

Figure 5: Psychoacoustic feature variation with perturbation strength (α). Each subplot shows mean (line) and individual sample values (points) for a given adjective ('bright', 'dark', or 'distorted') and perturbation type (TA or TT). Shaded regions represent standard deviation across samples.

monic content consistent with increased distortion. On the contrary, the Crest factor shows a gradual increase that plateaus at higher perturbations for AA—indicating that while the signal's high-frequency energy increases, its dynamic range does not compress as it would under hard clipping. For TT and TA, both HF ratio and Crest factor increase only very slightly, implying subtler spectral changes. Overall, movement along the "distorted" directions appears to capture the buildup of high-frequency content characteristic of distortion, but not the amplitude flattening typical of physically clipped signals, suggesting that the embeddings emphasize harmonic density more than waveform nonlinearity.

5. CONCLUSIONS AND FUTURE WORK

By following the SVD based pipeline presented in the methodology, we have been able to identify potential timbral directions in CLAP's audio embedding space. Potential directions for the same timbral adjectives were also discovered in the text embedding space. On applying these vectors as perturbations in a unimodal setting and using the perturbed embeddings in a downstream retrieval task, we are able to confirm that audio embeddings move monotonically in each timbral direction, and that the computed perturbation vectors successfully encode these directions. On performing cross-modal retrieval, we have also confirmed that CLAP has a shared timbal subspace– i.e., timbre directions in the audio encoder and text encoder constitute a geometrically aligned sub-space. Further, cross-modal retrieval reveals that timbre variance may be more linearly spaced– and therefore, more exploitable, in CLAP's audio space as compared to the text space.

The perturbation vectors are also used in a TTM generation task in order to numerically quantify the correlations between each timbral direction and perceptually relevant spectral

features that are associated with the timbral quality in consideration. We observe that CLAP's representations of brightness and darkness align well with perceptual cues, whereas its representation of distortion appears to rely more on high-frequency harmonic content than on amplitude clipping. One limitation of this analysis is that the "distorted" sounds in Nsynth consist of synthesized tones with added harmonics rather than clipped signals; as a result, feature changes such as crest factor may reflect spectral structure rather than clipping.

The generation task also reveals that by perturbing the text encoder with timbre directions, it is possible to have direct control over not just the presence of the timbral quality in the generated audio, but also the amount of the timbral quality. Essentially, the implemented method allows direct and fine-grained control of timbre via natural language. This method could therefore have powerful uses in AI tools for music production and sound design, such as text-based sample generation or text-based SFX generation. Finally, the methods for identifying timbre directions in this study can be applied to the latent spaces of any pre-trained audio and language models, respectively. In the future, these methods can be extended to more complex and realistic audio, including longer musical phrases and polyphonic or mixed-instrument recordings, to evaluate whether the identified timbre directions generalize beyond isolated single notes. They can also be applied to a wider range of timbral qualities and descriptors. The pipeline presented in this paper can thus be used as a framework for analyzing how multimodal models encode perceptual concepts like timbre, thereby advancing interpretability research in language-audio modeling.

6. Acknowledgements.

Generative AI (ChatGPT) was used as a writing aid for paraphrasing text in some sections of this paper (mainly Section 2.2 and Section 4) in order to enhance overall readability of the paper.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. Musiclm: Generating music from text. *ArXiv*, abs/2301.11325, 2023. URL https://api.semanticscholar.org/CorpusID:256274504.

Victor Deng, Changhong Wang, Ga¨el Richard, and Brian McFee. Investigating the sensitivity of pre-trained audio embeddings to common effects. *ArXiv*, abs/2501.15900, 2025. URL https://api.semanticscholar.org/CorpusID:275921744.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. *ArXiv*, abs/1704.01279, 2017. URL https://api.semanticscholar.org/CorpusID:3697399.

Haohe Liu, Zehua Chen, Yiitan Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion

models. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256390486.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bodganov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: a corpus of audio captions for music-and-language evaluation. *ArXiv*, abs/2311.10057, 2023. URL https://api.semanticscholar.org/CorpusID:265221180.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Charalampos Saitis and Stefan Weinzierl. The semantics of timbre. *Timbre: Acoustics, Perception, and Cognition*, 2019. URL https://api.semanticscholar.org/CorpusID:164817977.

Malcolm Slaney. Semantic-audio retrieval. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4108–IV–4111, 2002. doi: 10.1109/ICASSP.2002.5745561.

Sangeeta Srivastava, Ho-Hsiang Wu, Joao Rulff, Magdalena Fuentes, Mark Cartwright, Claudio Silva, Anish Arora, and Juan Pablo Bello. A study on robustness to perturbations for representations of environmental sound. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 125–129, 2022. doi: 10.23919/EUSIPCO55093.2022.9909557.

Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008. doi: 10.1109/TASL.2007.913750.

Yusong Wu, K. 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. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022. URL https://api.semanticscholar.org/CorpusID:253510826.

7. APPENDIX A.

The following tables show evaluation results for Text-to-Music Retrieval using baseline vs perturbed CLAP embeddings. Each row reports performance for a given timbral adjective ("bright", "dark", or "distorted") across four metrics — Precision@K, Recall@K, MRR@K, and MAP@K. The Baseline row corresponds to unmodified embeddings, while subsequent rows (AA, TT, AT, TA) show the best scores obtained using perturbations along the learned timbre directions. The associated perturbation strength (α) at which each metric achieved its maximum value is indicated in parentheses.

Table 1: Baseline vs. AA and TT.

	Prec@K	**Rec@K**	**MRR@K**	**MAP@K**
		"Bright"		
Baseline	0.345	0.204	0.750	0.128
AA	0.795 (α=1.75)	0.470 (α=1.75)	1.0 (α=0.25)	0.720 (α=1.5)
TT	0.400 (α=3)	0.237 (α=3)	1.0 (α=0.5)	0.303 (α=4)
		"Dark"		
Baseline	0.635	0.145	0.375	0.349
AA	0.790 (α=1.25)	0.180 (α=1.25)	0.750 (α=1.25)	0.666 (α=1.25)
TT	0.610 (α=0.5)	0.139 (α=0.5)	1.0 (α=1)	0.384 (α=1)
		"Distorted"		
Baseline	0.365	0.098	0.6	0.154
AA	0.980 (α=0.75)	0.263 (α=0.75)	1.0 (α=0.25)	0.977 (α=0.75)
TT	0.485 (α=3.5)	0.130 (α=3.5)	1.0 (α=1)	0.285 (α=2.5)

Table 2: Baseline vs. AT and TA.

	Prec@K	**Rec@K**	**MRR@K**	**MAP@K**
		"Bright"		
Baseline	0.345	0.204	0.750	0.128
AT	0.925 (α=4)	0.547 (α=4)	1.0 (α=0.5)	0.917 (α=4)
TA	0.745 (α=1)	0.441 (α=1)	1.0 (α=0.25)	0.696 (α=0.75)
		"Dark"		
Baseline	0.635	0.145	0.375	0.349
AT	1.0 (α=4)	0.288 (α=4)	1.0 (α=4)	1.0 (α=4)
TA	0.900 (α=2)	0.205 (α=2)	1.0 (α=2)	0.876 (α=2)
		"Distorted"		
Baseline	0.365	0.098	0.6	0.154
AT	0.790 (α=3)	0.212 (α=3)	1.0 (α=3)	0.727 (α=3)
TA	0.950 (α=0.5)	0.255 (α=0.5)	1.0 (α=0.5)	0.930 (α=0.5)