

# Evaluating Timbre-Related Audio Descriptors Across Different Libraries and Multimodal Embeddings

Isabella Czedik-Eysenberg<sup>1</sup>, Christoph Reuter<sup>1</sup>

<sup>1</sup> *SiNES, Musicological Department, University of Vienna, 1090 Vienna, E-Mail: isabella.czedik-eysenberg@univie.ac.at*

## Background

Timbre-related audio features are a useful tool in music psychology and related fields to investigate the relationship between perceptual phenomena and their physical correlates in the signal. Although the relation between certain timbre dimensions and specific audio signal features is well-documented (e.g. *brightness* perception and *spectral centroid* [14]), often multiple similar descriptors exist and the effectiveness of different descriptors for predicting perceptual ratings of timbre dimensions can depend on implementation details and parameter choices (see e.g. [7]).

Beyond low-level audio features and acoustic models, recent studies have investigated the potential of Large Language Models to interpret timbre semantics and music similarity [13, 5], yielding mixed results. Multimodal (audio-text) models, which integrate both sound and textual descriptions [17], hold particular promise in representing and extracting timbre-related sound concepts.

## Aims

Based on listener data, the current experiment aims to: (1) compare timbre-related audio descriptors across different feature extraction libraries in terms of how well they align with human perceptual ratings of *“brightness”* [14], *“roughness”* [16] and *“percussiveness”* [4]. To this end, we include both stimuli with a uniform pitch (E4) as well as varied pitches for higher ecological validity; and (2) evaluate the potential of multimodal embedding models to capture these perceptual timbre concepts.

## Methods

After collecting listener ratings for a set of instrumental tones, the average ratings were tested for correlations with audio features extracted from the stimuli as well as with a proximity measure in multimodal embedding spaces (see upcoming sections).

### Listening Experiment and Audio Stimuli

Ratings were collected for 20 instruments from the Vienna Symphonic Library, including *violin, cello, guitar, harp, piano, harpsichord, celesta, harmonium, bassoon, clarinet, flute, saxophone, oboe, trumpet, horn, trombone, cornet, marimba, tubular bells, and vibraphone*. The stimulus set consisted of 40 examples, with two recordings of each instrument:

1. One at the same pitch (E4), referred to as 'E4' condition.
2. One at a (different) pitch typical for the instrument's register, referred to as the 'other' condition.

Each stimulus was rated by 31 subjects on the perceptual scales of *“brightness”*, *“roughness”* and *“percussiveness”* using sliders.

### Audio Feature Extraction

Subject ratings were compared to relevant audio features extracted via Librosa [9], Essentia [3], Praat/Parselmouth [2,6], MIRtoolbox [8], Matlab Audio Toolbox, AudioCommons Timbral Models [10], PyTimbre [11] and Meyda [12].

In most cases, default parameters suggested by the library were used. Additionally, in some cases, custom calculations (i.e. for percussiveness measures) or multiple different parametric combinations were tested (e.g. 'Sethares' and 'Vassilakis' for *roughness* calculation via MIRtoolbox).

### Multimodal Embedding Similarity Calculation

Multimodal embedding models generate latent representations of elements from different modalities (e.g., text and audio) within a shared embedding space. To evaluate their potential for capturing timbre semantics, we compare the perceptual ratings with the cosine similarity of the stimuli to the verbal descriptions (*“bright”*, *“rough”* and *“percussive”*) in multimodal embedding spaces based on LAION-CLAP [17]. The following models were used:

**Model 1:** *630k-audioset-best*

**Model 2:** *630k-audioset-fusion-best*

**Model 3:** *music\_audioset\_epoch\_15\_esc\_90.14*

**Model 4:** *music\_speech\_epoch\_15\_esc\_89.25*

**Model 5:** *music\_speech\_audioset\_epoch\_15\_esc\_89.98*

## Results

### Interrater Reliability

Overall, subjects showed the highest agreement in the evaluation of *brightness*, significantly less agreement on *roughness*, and the least agreement on *percussiveness* (see Table 1).

**Table 1:** Interrater-agreement measures for the different timbre dimensions.

Rating category	Mean pairwise r	Cronbach's $\alpha$
Brightness	0.648	0.980
Roughness	0.333	0.944
Percussiveness	0.212	0.881

### Audio Feature Correlations and Pitch-Dependency

The following sections present the results of the correlation analysis for each tested timbre dimension.

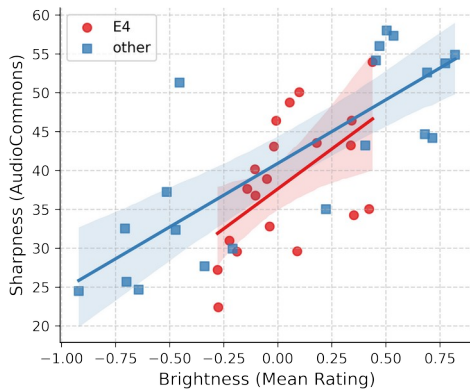
### Brightness

In the case of brightness, a strong pitch dependency is evident when comparing the results for all stimuli vs. those including only the 'E4' stimuli (same pitch condition) (see Table 2).

**Table 2:** Pearson correlations of average *brightness* ratings vs. extracted audio features; \*:  $p < 0.05$ ; \*\*:  $p < 0.01$ .

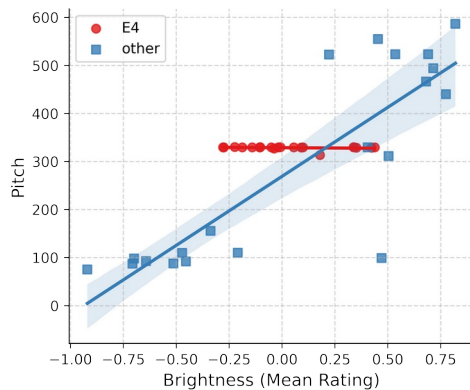
Audio Feature	Library	all	E4
Pitch (F0)	Praat	0.785**	-0.160
Sharpness	AudioCommons	0.731**	0.567**
Sharpness DIN45692	MATLAB Audio Toolbox	0.715**	0.451*
Spectral Centroid	MIRtoolbox	0.628**	0.551*
Attack Slope	MIRtoolbox	0.180	0.702**
Spectral Flatness	MIRtoolbox	0.356*	0.654**

Models for *sharpness*, as implemented in AudioCommons Timbral models or MATLAB audio toolbox, consistently showed correlations with human *brightness* ratings across both pitch conditions (see Figure 1).



**Figure 1:** Average *brightness* ratings vs. *timbral sharpness* as calculated via AudioCommons Timbral Models.

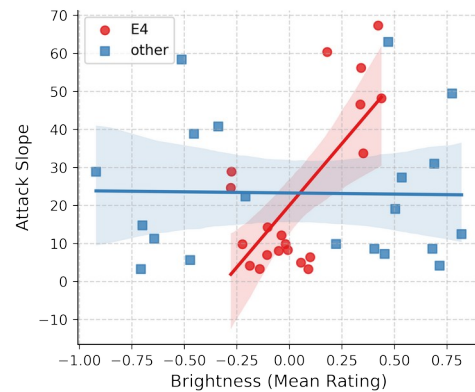
Although in fact *pitch* (F0) showed the highest fit to *brightness* ratings when considering all pitches, it is understandably not a suitable feature when comparing tones at the same pitch (see Figure 2).



**Figure 2:** Average *brightness* ratings vs. *pitch* (F0) as calculated via Praat.

Conversely, while the tones' *attack slope* was related to *brightness* perception when considering only the same pitch,

it failed to show a significant correlation as soon as different pitches were compared (see Figure 3).



**Figure 3:** Average *brightness* ratings vs. *attack slope* as calculated via MIRtoolbox.

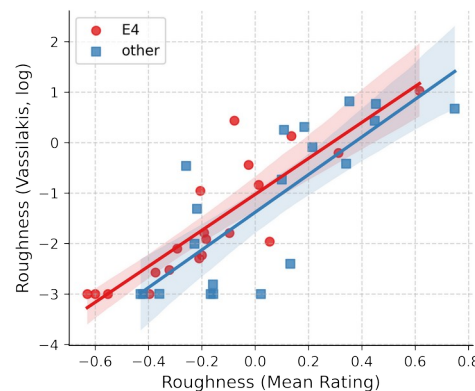
### Roughness

In contrast to *brightness*, correlations between audio features and *roughness* were less dependent on pitch (see Table 3).

**Table 3:** Pearson correlations of average *roughness* ratings vs. extracted audio features; \*:  $p < 0.05$ ; \*\*:  $p < 0.01$ .

Audio Feature	Library	all	E4
Roughness (Vassilakis, log)	MIRtoolbox	0.815**	0.867**
Roughness (Vassilakis)	MIRtoolbox	0.694**	0.669**
Roughness (Sethares)	MIRtoolbox	0.564**	0.772**
Roughness	AudioCommons	0.671**	0.711**
Dissonance	Essentia	0.669**	0.736**
Spectral Crest	PyTimbre	-0.730**	-0.745**
Zero-crossing Rate	MIRtoolbox	0.369*	0.826**

While most of the included *roughness* models showed relatively strong correlations, the *roughness* values calculated using MIRtoolbox with Vassilakis weighting [16] were skewed in distribution, but exhibited a very strong linear correlation with perceptual ratings after log-transformation (see Figure 4).



**Figure 4:** Average *roughness* ratings vs. *roughness* (log-transformed) as calculated via MIRtoolbox using the 'Vassilakis' parameter.

While the *zero-crossing rate* might appear as a suitable descriptor for *roughness* when considering stimuli at the same pitch, it was sensitive to pitch differences (Figure 4).

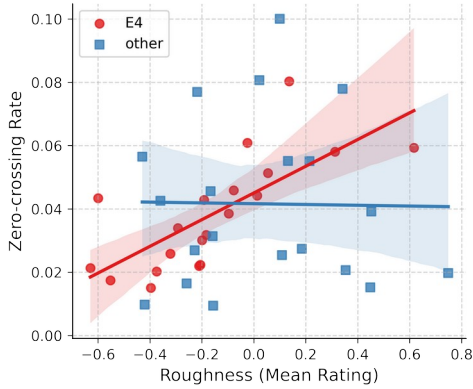


Figure 5: Average *roughness* ratings vs. *zero-crossing rate* as calculated via MIRtoolbox.

**Percussiveness**

For *percussiveness*, we considered various measures derived from *harmonic-percussive source separation* [4], with the *loudness* difference between the percussive and harmonic components showing the strongest correlation with perceptual *percussiveness* ratings (see Figure 6). Loudness was calculated using pyloudnorm [15]. Next to that, also other measures such as a steeper attack (*attack slope*) and higher *shimmer* were associated with greater perceived *percussiveness* (see Table 4).

Table 4: Pearson correlations of *percussiveness* ratings vs. extracted audio features; \*:  $p < 0.05$ ; \*\*:  $p < 0.01$ .

Audio Feature	Library	all	E4
Percussive-Harmonic Loudness $\Delta$	LibROSA/pyloudnorm	0.705**	0.761**
Percussive Loudness	LibROSA/pyloudnorm	0.679**	0.821**
Attack Slope	MIRtoolbox	0.640**	0.674**
Shimmer (Local)	Praat	0.629**	0.623**
Harmonicity	Praat	-0.611**	-0.504*
Perc/Harm RMS Ratio	LibROSA	0.586**	0.530*

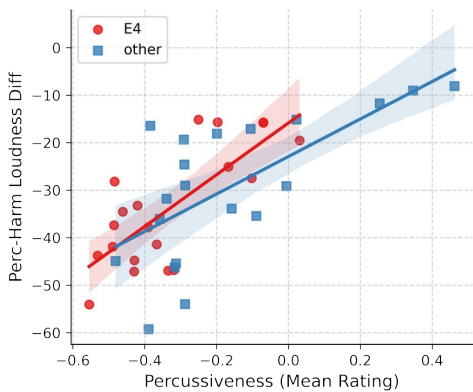


Figure 6: Average *percussiveness* ratings vs. loudness difference of the percussive and harmonic signal component as calculated via LibROSA and pyloudnorm.

**Comparison with Multimodal Embeddings**

Table 5 and Figures 7–9 display the results of the comparison between the perceptual timbre ratings and the proximity of the stimuli to the corresponding verbal descriptions (e.g. “*bright*”) in the embedding space for each of the models. Embedding similarity only partially aligned with human ratings, with the closest results for *brightness*.

Table 5: Pearson correlations of timbre description ratings vs. embedding similarity; \*:  $p < 0.05$ ; \*\*:  $p < 0.01$ .

Model	Brightness	Roughness	Percussiveness
Model 1	0.570**	0.234	0.240
Model 2	0.645**	0.210	0.328*
Model 3	0.021	0.071	0.501**
Model 4	0.419**	0.079	0.178
Model 5	0.339*	-0.066	0.097

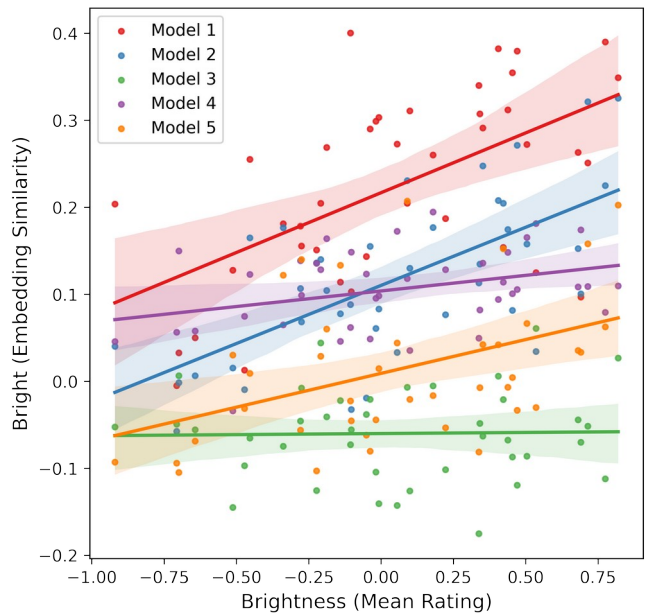


Figure 7: Average *brightness* ratings vs. embedding similarity of the audio to the verbal description “*bright*”.

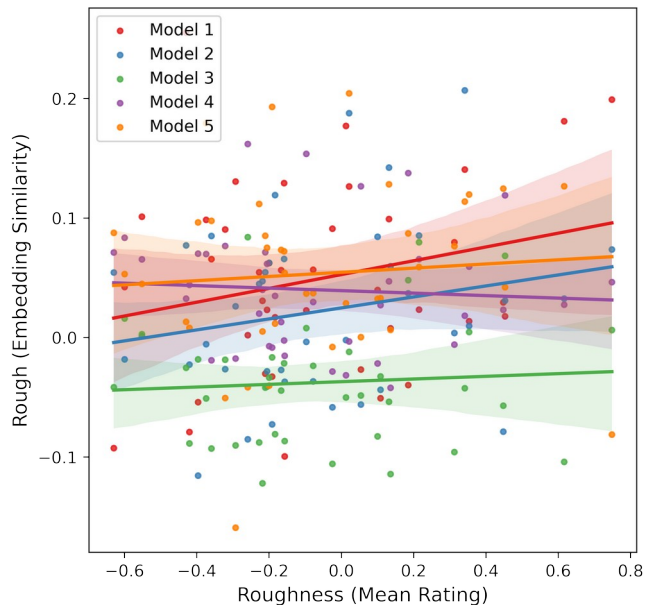
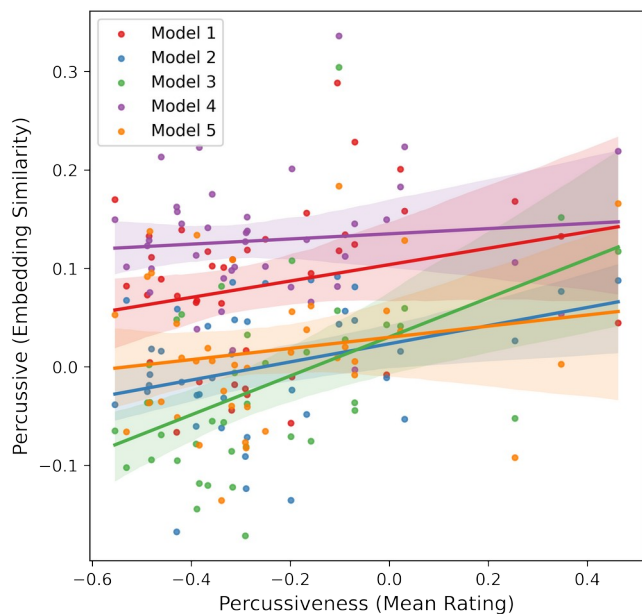


Figure 8: Average *roughness* ratings vs. embedding similarity of the audio to the verbal description “*rough*”.



**Figure 9:** Average *percussiveness* ratings vs. embedding similarity of the audio to the verbal description “*percussive*”.

## Discussion & Conclusion

While *brightness* showed the strongest pitch dependency in this experiment, some features appeared relatively robust across conditions, such as several *sharpness* models and *spectral centroid* in case of *brightness*, log-scaled *roughness* (Vassilakis [16]) for *roughness*, and the loudness difference between percussive and harmonic signal components (after harmonic-percussive separation via median filtering [4]) for *percussiveness*.

Multimodal embedding models such as LAION-CLAP [17] appear as a promising approach to extract predictions of semantic timbre descriptions, however only partial correlations to human ratings were observed in the present experiment.

## References

- [1] Alonso-Jiménez, P., Bogdanov, D., Pons, J., & Serra, X. (2020). Tensorflow audio models in essentia. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 266-270). IEEE.
- [2] Boersma, P., & Van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glott International*, 5(9/10), 341-347.
- [3] Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. *Proceedings of the 14th Conference of the International Society for Music Information Retrieval (ISMIR)*, pp. 493–8.
- [4] Fitzgerald, D. (2010). *Harmonic/percussive separation using median filtering*. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFX10)*, Graz, Austria.
- [5] Flexer, A. (2024). On the validity of employing ChatGPT for distant reading of music similarity. *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, San Francisco, United States.
- [6] Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, 1-15.
- [7] Kazazis, S., Esterer, N., Depalle, P., & McAdams, S. (2017). A performance evaluation of the timbre toolbox and the mirtoolbox on calibrated test sounds. *Proceedings of the 2017 International Symposium on Musical Acoustics* (pp. 144-147).
- [8] Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Bordeaux, France.
- [9] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference* (Vol. 8, 18–25).
- [10] Pearce, A., Safavi, S., Brookes, T., Mason, R., Wang, W., & Plumbley, M. (2019). *Deliverable D5.8 - Release of timbral characterisation tools for semantically annotating non-musical content*. Audio Commons Initiative. Technical Report.
- [11] Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. Timbre Toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, 130(5), 2902-2916.
- [12] Rawlinson, H., Segal, N., & Fiala, J. (2015). Meyda: an audio feature extraction library for the web audio API. *The 1st web audio conference*. Paris.
- [13] Saitis, C., & Siedenburg, K. (2023). When ChatGPT talks timbre. *Proceedings of Timbre 2023: 3rd International Conference on Timbre, Thessaloniki, Greece*.
- [14] Schubert, E., & Wolfe, J. (2006). Does timbral brightness scale with frequency and spectral centroid?. *Acta acustica united with acustica*, 92(5), 820-825.
- [15] Steinmetz, C. J., & Reiss, J. (2021). pyloudnorm: A simple yet flexible loudness meter in python. *Audio Engineering Society Convention 150*. Audio Engineering Society.
- [16] Vassilakis, P. N. (2001). *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance*. Doctoral Dissertation. Los Angeles: University of California, Los Angeles; Systematic Musicology.
- [17] Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.