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

Isabella Czedik-Eysenberg[1], Christoph Reuter[1]
[1]SInES, Department of Musicology, University of Vienna, Austria

DAS | DAGA 2025 COPENHAGEN

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***; Schubert & Wolfe, 2006), 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. Kazazis et al., 2017).

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** (Saitis & Siedenburg, 2023; Flexer, 2024), yielding mixed results. **Multimodal (audio-text) models**, which integrate both sound and textual descriptions, hold promise in representing and extracting timbre-related sound concepts.

Aims

Based on listener data, the current experiment aims to:

- … compare **timbre-related audio descriptors** across different audio feature extraction libraries in regards to their fitness to **human perceptual ratings** of ***"brightness"***, ***"roughness"*** and ***"percussiveness"*** – using stimuli on the same pitch vs. different pitches.
- … evaluate the potential of **multimodal embedding models** in capturing these timbre-related perceptual concepts.

Methods

Listening Experiment

Ratings were collected for **20 instruments** from the Vienna Symphonic Library (*violin*, *cello*, *guitar*, *harp*, *piano*, *harpsichord*, *celesta*, *harmonium*, *bassoon*, *clarinet*, *flute*, *saxophone*, *oboe*, *trumpet*, *horn*, *trombone*, *cornet*, *marimba*, *tubular bells*, *vibraphone*). The stimuli set comprised recordings of the instruments played:

1. … on the **same pitch** (E4) as well as
2. … on **different pitches** typical for the range of the instrument (⇒ 40 stimuli in total).

Each stimulus was rated on the scales ***"brightness"***, ***"roughness"*** and ***"percussiveness"*** on sliders by **31 subjects**.

Audio Feature Extraction

Subject ratings were compared to relevant **audio features** extracted via `Librosa` (McFee et al., 2015), `Essentia` (Bogdanov et al., 2013), `Praat`, `MIRtoolbox` (Lartillot et al., 2008), `Matlab Audio Toolbox`, `AudioCommons Timbral Models`, `PyTimbre` (Peeters et al., 2011) and Meyda (Rawlinson et al., 2015). In most cases, default parameters suggested by the library were used. Additionally, in some cases, custom implementations or multiple different parametric combinations were tested.

Multimodal Embedding Similarity Calculation

Further, the ratings of the timbre dimensions were compared with the **cosine similarity** of the stimuli to the verbal descriptions (***"bright"***, ***"rough"*** and ***"percussive"***) in **multimodal embedding spaces** based on `LAION-CLAP` (Wu et al., 2023). 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*

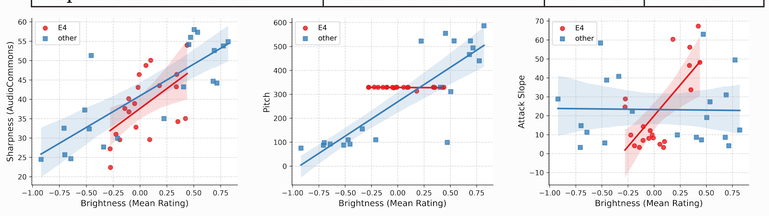
Results

Audio Feature Correlations on the Same Pitch vs. Different Pitches

Brightness

	Audio Feature	Library	r	p
All Pitches	F0 Peak	Custom	0.795	<0.001
	Pitch	Praat	0.785	<0.001
	Sharpness	AudioCommons	0.731	<0.001
	Sharpness DIN45692	Audio Toolbox	0.715	<0.001
	Spectral Centroid	Meyda	0.640	<0.001

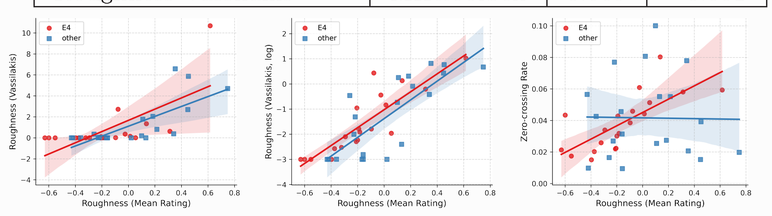
	Audio Feature	Library	r	p
Only E4	Attack Slope	MIRtoolbox	0.702	0.001
	Spectral Flatness	MIRtoolbox	0.654	0.002
	Sharpness	AudioCommons	0.567	0.009
	Spectral Centroid	MIRtoolbox	0.551	0.012



Roughness

	Audio Feature	Library	r	p
All Pitches	Spectral Crest	PyTimbre	-0.730	<0.001
	Roughness Vassilakis	MIRtoolbox	0.694	<0.001
	Roughness	AudioCommons	0.671	<0.001
	Dissonance	Essentia	0.669	<0.001
	Spectral Entropy	Essentia	0.649	<0.001

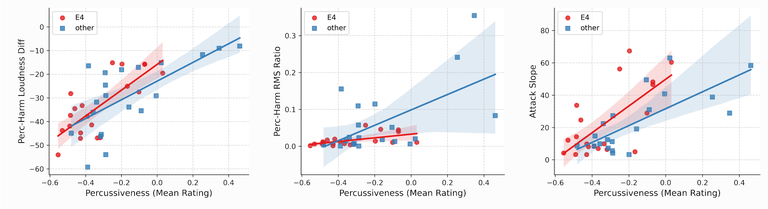
	Audio Feature	Library	r	p
Only E4	Zero-crossing Rate	MIRtoolbox	0.826	<0.001
	Spectral Flux (> 1000 Hz)	MIRtoolbox	0.818	<0.001
	Roughness Sethares	MIRtoolbox	0.774	<0.001
	Roughness Vassilakis	MIRtoolbox	0.669	0.001



Percussiveness

	Audio Feature	Library	r	p
All Pitches	Perc-Harm Loudness Δ	LibROSA	0.705	<0.001
	Percussive Loudness	LibROSA	0.679	<0.001
	Attack Slope	MIRtoolbox	0.640	<0.001
	Harmonicity	Praat	-0.611	<0.001
	Perc-Harm RMS Ratio	LibROSA	0.586	<0.001

	Audio Feature	Library	r	p
Only E4	Percussive Loudness	LibROSA	0.821	<0.001
	Perc-Harm Loudness Δ	LibROSA	0.761	<0.001
	Attack Slope	MIRtoolbox	0.674	0.001
	Shimmer	Praat	0.623	0.003



Interrater Reliability

Overall, subjects **agreed most** on the evaluation of ***brightness***, significantly less on ***roughness***, and least on ***percussiveness*** (see table).

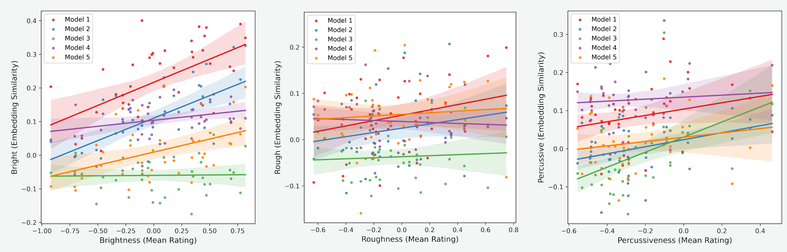
Rating Category	Mean Pairwise r	Cronbach's α
Brightness	0.648	0.980
Roughness	0.333	0.944
Percussiveness	0.212	0.881

Comparison with Multimodal Embeddings

Embedding similarity only partially aligned with human ratings, with the closest results for ***brightness***.

Model	Brightness r	Roughness r	Percussiveness r
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

*: p<0.05; **: p<0.01



Discussion

While brightness showed the strongest **pitch dependency** in this experiment, some features appeared relatively robust across conditions, such as **sharpness** models and **spectral centroid** in case of ***brightness***, **roughness (Vassilakis)** for ***roughness***, and the loudness difference between percussive and harmonic signal components (after **harmonic-percussive separation** via median filtering) for ***percussiveness***. **Multimodal embedding models** such as LAION-CLAP appear as a promising approach to extract predictions of **semantic timbre descriptions**, however **only partial correlations** to human ratings were observed in the current experiment.

References

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. | **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. In *14th Conference of the International Society for Music Information Retrieval (ISMIR)*, pp. 493–8. | **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. | **Kazazis, S., Esterer, N., Depalle, P., & McAdams, S. (2017).** A performance evaluation of the timbre toolbox and the mirtoolbox on calibrated test sounds. In *Proceedings of the 2017 International Symposium on Musical Acoustics* (pp. 144-147). | **Lartillot, O., Toiviainen, P. & Eerola, T. (2008).** A Matlab Toolbox for Music Information Retrieval. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_31 | **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). | **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. | **Rawlinson, H., Segal, N., & Fiala, J. (2015).** Meyda: an audio feature extraction library for the web audio API. In *The 1st web audio conference*. Paris. | **Saitis, C., & Siedenburg, K. (2023).** When ChatGPT talks timbre. *Proceedings of Timbre 2023: 3rd International Conference on Timbre*, Thessaloniki, Greece. | **Schubert, E., & Wolfe, J. (2006).** Does timbral brightness scale with frequency and spectral centroid?. *Acta acustica united with acustica, 92(5)*, 820-825. | **Steinmetz, C. J., & Reiss, J. (2021).** pyloudnorm: A simple yet flexible loudness meter in python. In *Audio Engineering Society Convention 150*. Audio Engineering Society.