DMRN+16: Digital Music Research Network
One-day Workshop 2021

Queen Mary University of London
Tue 21 December 2021

Acoustic Representations for Perceptual Timbre Similarity

Cyrus Vahidi, Ben Hayes, Charalampos Saitis, George Fazekas[*]

Centre for Digital Music, Queen Mary University of London, United Kingdom, c.vahidi@qmul.ac.uk

Abstract— **In this work, we outline initial steps towards modelling perceptual timbre dissimilarity. We use stimuli from 17 distinct subjective timbre studies and compute pairwise distances in the spaces of MFCCs, joint time-frequency scattering coefficients and Open-L3 embeddings. We analyze agreement of distances in these spaces with human dissimilarity ratings and highlight challenges of this task.**

***Index Terms*—** timbre, acoustic representations, psychoacoustics

I. Method

We used 17 timbre dissimilarity datasets that were compiled in a previous meta-analysis publication [1]. We share an open-source repository containing 17 dissimilarity matrices and corresponding audio sampled at 44.1 kHz[1].

We extracted temporally averaged mel-frequency cepstral coefficients (MFCCs), joint time-frequency scattering coefficients (jTFS) [2] and OpenL3 embeddings [3] for 1000ms of audio of each stimulus. We consider jTFS as it characterises *spectrotemporal modulations*, analogously to the model used in [1]. We used a window length of 25ms for MFCCs with 40 coefficients. jTFS coefficients were computed using Kymatio[2] with maximum scale $J = 8$, $Q = 12$ filters per octave, temporal averaging of $T = 1000ms$ and frequential averaging of $F = 1$ octave, yielding 869 coefficients. 512-dimensional OpenL3 embeddings were extracted using an open-source Python package[3].

Pairwise euclidean distances of the form in Eqn. (1) were computed between all embeddings within each dataset.

$$D_e(x_i, x_j) = \sqrt{(Ix_i - Ix_j)^T(Ix_i - Ix_j)} \qquad (1)$$

II. Results

We collected all triplets (a, i, j) from a dissimilarity matrix, where i and j belong to the k-nearest neighborhood of an anchor a and satisfy the triplet inequality $D(a, i) < D(a, j)$.

Table 1: Mean triplet agreement using a $k = 5$ nearest neighborhood

Dataset	***MFCC***	***OpenL3***	***jTFS***
Barthet2010	0.71	0.77	0.88
Grey1977	0.57	0.64	0.61
Grey1978	0.41	0.48	0.45
Iverson1993_Onset	0.59	0.59	0.56
Iverson1993_Remainder	0.57	0.54	0.54
Iverson1993_Whole	0.59	0.66	0.64
Lakatos2000_Comb	0.55	0.53	0.55
Lakatos2000_Harm	0.64	0.73	0.61
Lakatos2000_Perc	0.53	0.55	0.48
McAdams1995	0.62	0.63	0.58
Patil2012_A3	0.65	0.65	0.65
Patil2012_DX4	0.48	0.6	0.54
Patil2012_GD4	0.58	0.64	0.45
Siedenburg2015_e2set1	0.73	0.71	0.65
Siedenburg2015_e2set2	0.68	0.69	0.61
Siedenburg2015_e3	0.58	0.56	0.5

Triplet agreement is the average number of triplets that satisfy $D_e(a, i) < D_e(a, j)$, i.e the distance ranking is respected in acoustic feature space e. Table 1 shows the mean triplet agreements per dataset using 5 nearest neighbors.

III. Conclusion

Initial experiments indicate that acoustic features alone are not sufficient to match perceptual distances. We highlight that the only dataset containing a homogeneous category for all stimuli, *Barthet2010*, produces a considerably higher figure than other datasets, which may suggest that its timbre space only encodes acoustical cues. Otherwise, we observe no clear differences between the representations. Further experiments will aim to learn a unified metric to approximate timbre space distances across datasets, considering specificity and categorical cues. This may give a clearer indication of the suitability of the proposed representations.

IV. References

[1] E. Thoret, B. Caramiaux, P. Depalle, and S. Mcadams, "Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre," *Nature Human Behaviour*, vol. 5, no. 3, pp. 369–377, 2021.

[2] J. Andén, V. Lostanlen, and S. Mallat, "Joint time–frequency scattering," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3704–3718, 2019.

[3] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[*] This work was supported by UK Research and Innovation [grant number EP/S022694/1]. Cyrus Vahidi is supported jointly by the UKRI and Music Tribe.

[1] https://github.com/ben-hayes/timbre-dissimilarity-metrics/

[2] https://github.com/kymatio/kymatio

[3] https://openl3.readthedocs.io/en/latest/index.html