Quantifying the Influence of Musical Features on Perceptual Similarity of Popular Songs

Riesa Cassano[1], Elise Cole[1], Kelly Jakubowski[2], Elise A. Piazza[1,3]

[1]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY; [2]Department of Music, Durham University, Durham, England
[3]Department of Neuroscience, University of Rochester Medical Center, Rochester, NY
Contact: rcassan2@ur.rochester.edu

Background & Motivation

- To understand the world, we constantly judge the similarity of the things around us, e.g. whether one cat is fluffier than another or one dish is spicier than another.
- When judging the overall similarity of two songs, what features do we use–low-level acoustic properties, emotional qualities, or something else?

- **Cover songs** offer a rich space to test this question, by varying low-level features but keeping the high-level song identity.
- Cover versions keep the same lyrics and chord structure as the original, but may vary on other features, such as genre, tempo, key, and gender of the vocalist.

- To what extent do acoustic or musical features and emotional features (valence and arousal) influence the perceived similarity between pairs of originals and covers?

Methods

Example: 2010 #1

original

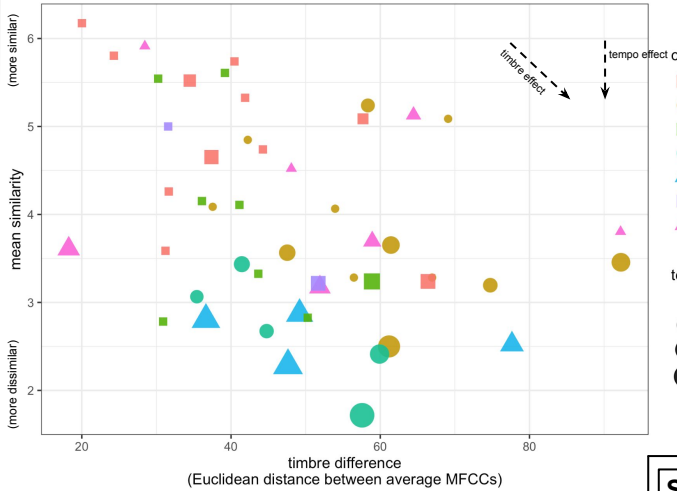
cover

- We found covers on YouTube for 50 songs from 2008-2019 from the Billboard Year-End charts and pseudo-randomly selected one cover per song.
- We manually coded key, gender, and cover category. We extracted tempo and mel-frequency cepstral coefficients (MFCCs: a concise description of the spectral envelope) using librosa (McFee et al., 2015).
- We averaged MFCCs over time for the original and cover and computed Euclidean distance (Tzanetakis & Cook, 2002).

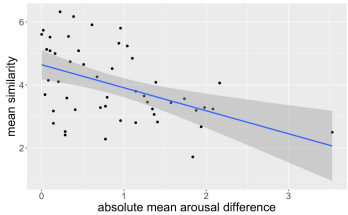
- Subjects (N=50) rated similarity between the original and cover on a scale from 1 (extremely dissimilar) to 7 (extremely similar).
- Other subjects (N=50) rated each clip on valence (negative to positive) and arousal (passive to active).

Results

Acoustic and musical features



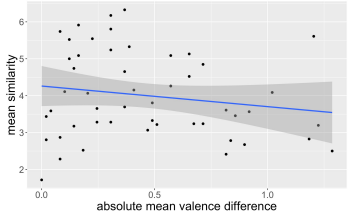
Emotional features



Originals and covers that are rated as more different in **arousal** are also rated as less similar (β = -0.73, p < 0.001***, R^2 = 0.20).

This relationship is likely mediated by tempo.

The relationship is still significant when the "outlier" (difference > 3) is removed (β = -0.79, p = 0.0017**, R^2 = 0.17).



Originals and covers that are rated as more different in **valence** are *not* rated as significantly less similar (β = -0.56, p = 0.24, R^2 = 0.0089).

This isn't surprising: each original/cover pair keeps the same lyrics and mode (major vs minor), so these differences in valence are relatively small and uninformative.

Best model of similarity: tempo, timbre, and cover category

	Tempo	Timbre	Rock/metal+ (vs)	Jazz+ (vs)
β	-0.56	-0.32	-0.92	-1.42
p	0.0074**	0.028*	0.017*	0.0062**

+compared to category 1 (pop or like original)
Not significant: key, gender of the vocalist, any interactions between features

Summary & Implications

- Previous work in music information retrieval (MIR) has established the relationship between timbre and similarity for music in general.
- This work takes a more controlled approach to specifically test similarity judgements when low-level features (e.g. timbre) vary, but the high-level identity doesn't.
- When comparing originals and covers, similarity judgements are driven by tempo and genre, as well as timbre.
- These "surface-level" features seem to better capture patterns in similarity judgments than valence and arousal ratings.
- The dissociation between low-level features and high-level abstract representations will be useful in future studies to investigate the relationships between music perception and cognitive processes like memory.
- Understanding the relationship between acoustic and musical features and perceived similarity is the first step in bridging that gap.

References

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and Music Signal Analysis in Python*. 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302. https://doi.org/10.1109/TSA.2002.800560