Musical and emotional features individually and interactively predict perceived similarity of popular songs

Riesa Cassano-Coleman (rcassan2@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester

Kelly Jakubowski
Department of Music, Durham University

Elise A. Piazza
Department of Brain and Cognitive Sciences and Department of Neuroscience, University of Rochester

Abstract

Judging similarity between pieces of music is critical for interacting with it in everyday life. But how do musical and emotional features drive our subjective judgments of similarity? Much of the previous work has focused on low-level features and has largely ignored the impact of lyrics and emotion on perceived similarity. Here, we tested the influence of a comprehensive set of musical and emotional features on similarity, using original popular songs and cover versions to match clips on lyrics and melody. We found that tempo most strongly predicts lower similarity ratings, but key, voice type, and timbre differences predict similarity in an interactive manner. While emotional arousal did not predict similarity above and beyond tempo, emotional valence did. Together, these results suggest that both musical and emotional factors influence judgments of similarity, shedding light on the fine-grained explanatory mechanisms of listeners' everyday impressions of popular music.

Keywords: music cognition; similarity; timbre; emotional valence and arousal

Introduction

Similarity judgments are critical for our everyday cognition, for example allowing us to group novel items into previously learned categories based on shared features (Goldstone & Son, 2005; Nosofsky, 1991; Tversky, 1977). In music, judging the similarity between songs is a key aspect of how we search for, describe, and enjoy music in daily life. We may recommend a new artist to a friend, knowing that she likes an artist who makes similar-sounding music. Assessing similarity has been a core issue in the field of music information retrieval (MIR), underlying both the organization of huge digital music libraries since the early 2000s and the construction of effective music recommender systems employed by modern streaming services. Much of this work has relied heavily on spectral features, such as timbre (the "tone color of music", Terasawa, Slaney, & Berger, 2005), as measured by mel-frequency cepstral coefficients (MFCCs, Tzanetakis & Cook, 2002). In this same vein, much of the previous work on human perception of musical similarity has also focused primarily on the contributions of low-level features like MFCCs (Aucouturier & Pachet, 2002; Herre, Allamanche, & Ertel, 2003), rather than comprehensively testing the influence of a range of features.

While a few studies have expanded the features considered, this work is still typically limited by a lack of consideration of lyrics and emotional features. Novello et al. (2011) analyzed genre, tempo (fast vs slow), and a coarse measure of timbre (primary instrument: vocal, piano, guitar). They used a triad task, where participants heard three stimuli and were asked which two stimuli were most similar and which were least similar. They found a hierarchical relationship between features (i.e., participants were more likely to choose clips from the same genre as the most similar pair, then clips with the same tempo (fast vs slow), and then clips with the same primary instrument. However, this paradigm pits these features against each other, which reveals the priority (rank) of each feature but leaves unclear how they would interact or work together if pairs of clips were judged on a continuous similarity scale.

Bogdanov et al. (2011) aimed to explicitly link multiple levels of musical features in predicting similarity and found that a hybrid model, combining both lower-level (e.g., tempo, MFCCs) and higher-level features (e.g., genre, mood), outperformed models with just the low-level features. However, like Novello et al. (2011), this study also used a mix of vocal and non-vocal music and did not model potential impacts of the semantic content of the lyrics. While they also considered emotion, this aspect was limited to binary classifications (e.g., happy/non-happy) and did not consider the full spectrum of emotion (i.e. using the circumplex model of valence and arousal).

Many people engage with music because of the powerful emotions it elicits. Surprisingly, although much is known about the mapping between musical and emotional features (e.g., the minor mode sounds sad), it is still a fairly open question how distinct emotional features influence our *overall* perception of music (and comparisons between songs). Across many studies, it has been shown that certain features (spectral features, mode) are associated with emotional valence, whereas others (tempo and other rhythmic features) are more strongly associated with emotional arousal (summarized by Gabrielsson & Lindström, 2010; see also Grekow, 2017). However, the link between emotional

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

features and judgments of overall similarity has not been shown, and it is unknown whether emotional features add any information beyond the musical features.

In this project, we tested the influence of a comprehensive range of musical and emotional features on perception of musical similarity, while removing the potential influence of lyrics. To do this, we collected human similarity judgments between pairs of songs—originals and covers—which have identical lyrics, melody, and harmonic progressions, but vary widely along several other perceptual dimensions: timbre, tempo, key, and voice type. Cover versions of popular songs offer a rich naturalistic stimulus set, and our corpus spans a broad range of genres and years (1955-2022). Further, we collected emotional valence and arousal ratings for each clip to quantify the emotional difference between original and cover songs. Our novel assessment of how *perceived emotion* influences perceived similarity provides a new understanding of how perceptual *and* emotional features contribute to our high-level impressions of music.

First, we test the sets of musical and emotional features separately, because of known relationships between features at these different levels. This allows us to potentially capture subtler interactive relationships that might not be revealed by a full complex model. Based on previous work (Bogdanov et al., 2011; Novello et al., 2011), we expect tempo and timbre to predict similarity, but the influence of the other musical and emotional features is unknown. We then combine the musical and emotional features to understand how much of the variance in similarity ratings is uniquely accounted for by different features. Finally, we test whether the emotional features contribute any information above the musical features. Since we know that emotional valence and arousal can be at least partially explained by musical features (Gabrielsson & Lindström, 2010), we might expect that the effect of emotional features on similarity is either partially or fully mediated by musical features.

Materials and Methods

Stimuli

Our corpus has 70 songs from 1955-2022, with approximately 10 songs per decade. Songs were drawn primarily from the Billboard 100 Greatest Songs of All Time chart, but we used Year-End Hot 100 (Pop) charts as needed to ensure equal distribution of songs across years.

We searched YouTube to find covers. If fewer than five covers were available, we excluded the song. To be considered, covers needed to have suitable audio quality (minimal background noise) and at least 1000 views. We categorized each clip into one of the following genres: pop, acoustic, metal/hard rock, jazz, Latin, country (including bluegrass), a cappella, electronic dance music (EDM), classic rock, ballad, reggae, R&B, disco, hip hop/rap, and alternative/indie. It should be noted that the originals tended to be clustered in just a few genres (especially pop, classic rock, and R&B) but covers spanned a wider range of genres.

To maintain genre diversity in our corpus, we used a multi-step procedure to choose one cover per song. First, we pseudo-randomly chose five covers: we randomly chose one cover that was similar to the original, one cover that was either acoustic or ballad style (both "similar" and "acoustic" were by far the most common types of covers we found), and three additional covers in other genres, ensuring no genre was chosen more than once unless there was no other option. One cover was randomly chosen from the set of five.

For each original and cover, we extracted 5-second clips from first chorus (or the most recognizable section). Clips were matched on lyrical content, so some cover clips were slightly shorter or longer (range of 4-7s) than their corresponding original clips. We equated all clips in loudness.

Features

Acoustic features We extracted the following features from librosa (McFee et al., 2015): root-mean-square (RMS), spectral bandwidth, spectral centroid, spectral contrast, spectral flatness, spectral roll-off, zero-crossing rate, and mel-frequency cepstral coefficients (MFCCs, offset plus 12 coefficients). We averaged each feature across the 5-s clip.

Musical features For each clip, we manually coded tempo (using the tapping feature of a metronome), key (referencing online chord charts, using a keyboard to match chords), and voice type (treble versus tenor/baritone/bass, corresponding to the range of the lead vocalist's voice).

Emotion ratings

50 participants were recruited through Prolific (36 female, 13 male, 1 preferred not to say; ages 18-69).

For each trial, participants heard the clip and were asked to rate it by clicking on a two-dimensional compass representing valence on the x-axis and arousal (intensity) on the y-axis (-3 to +3 for both axes; see Fig. 1, adapted from McClay et al., 2023).

At the beginning of the experiment, participants heard several examples of longer clips from each quadrant to familiarize them with the compass before completing a practice trial.

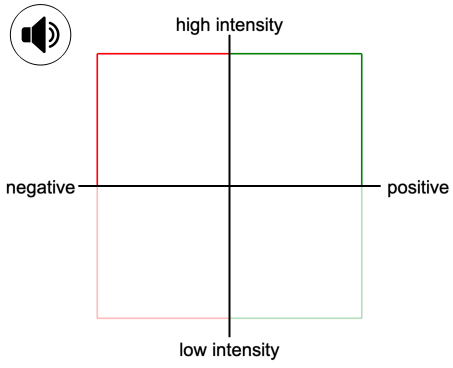

Figure 1: Emotion compass.

They had 5 seconds to respond to each trial. Participants were excluded if they did not respond in more than 10% of trials (n = 0). Missing data was ignored when averaging across ratings of all participants for each clip (total of 23 trials across all participants).

Stimuli were presented in 4 blocks of 35 trials each, with short breaks offered in between each block. The original and cover versions of each song were presented in separate blocks. Otherwise, the order of blocks was randomized and the order of the stimuli within the blocks was randomized.

At the end, participants completed a short questionnaire about their music listening habits and preferences.

Similarity ratings

50 additional participants were recruited through Prolific (28 female, 21 male, 1 preferred not to say; ages 20-61).

For each trial, participants heard the original and the cover clips (5s each, 1.5s of silence in between) and were asked to rate them on a continuous scale from extremely dissimilar (1) to extremely similar (7). Importantly, they were not given specific instructions for any particular features to pay attention to. The order of the stimuli was randomized, and the order of the original versus the cover within each trial was also randomized. Stimuli were presented in 2 blocks of 35 trials each, and participants were offered a short break halfway through.

Participants were excluded if more than 10% of their responses were over 10 seconds (n = 4). All trials over 10 seconds were excluded (53 additional trials). Participants completed the same post-survey questionnaire.

Analysis

Difference metrics For each feature, we calculated the differences between the original and selected cover and used that measure to predict similarity. For acoustic features and tempo, we simply calculated the difference between the original and cover. For emotional valence and arousal, we averaged across ratings for all participants and then calculated differences in the mean ratings.

For MFCCs, we computed the Euclidean distance between the original and cover. (This is equivalent to using simple difference for a single feature.)

Key difference is in terms of actual difference in semitones, not distance around the circle of fifths. We treat octaves as equivalent, so the possible range of absolute values is 0 to 6.

Voice type difference is coded as follows: 0 if lead vocalists for the original and cover share the same voice type ("same", e.g. both female singers); 1 if one clip had one voice type and the other clip had both voice types ("add opposite", n = 4, i.e. there were two lead singers—one male, one female—or sung by a choir); 2 if lead vocalists for the original and cover have opposite voice types ("switch").

For genre difference, we use whether or not the original and cover are in the same genre for sake of simplicity. (Coded as same genre = 1, different genre = 0).

Acoustic features and timbre Of all of the acoustic features, only MFCCs significantly predicted similarity ($\beta = -.30$, $p = .012$; $p > .05$ other features), so we used MFCCs difference as a succinct representation of timbre in further analyses.

Feature normalization For each feature, we took the absolute value of the difference, and then centered and scaled the differences so the overall distribution of each feature difference is centered at 0, with a standard deviation of 1. Similarity ratings are also centered and scaled. (Same/different genre is not scaled.)

Unscaled feature differences are used in visualizations, so the plot axes are meaningful. Scaled feature differences are used for models, so all of the features are on the same scale and coefficients are interpretable.

Packages All analysis was performed in R. We used `vegan` for variance partitioning and `eulerr` to generate area-proportional Venn diagrams. We used `mediation` for simple mediation analysis and `mma` for multiple mediation analysis.

Analysis approach All reported R^2 values are adjusted.

1. *Musical features*. First, we use each feature (tempo, key, voice type, timbre) to predict similarity in separate models. Then, we combine all four musical features into one model, and include interactions between pitch-related features (i.e. not tempo) to fit the best possible model (without overfitting the data).
2. *Genre*. Genre may be considered a higher-order, multidimensional musical feature, and so we explore it separately from the others. Genre is not included in the variance partitioning or mediation analyses. This is just an exploratory first pass: we plan to further quantify fine-grained genre distances (i.e., via embedding approaches) in future analyses.
3. *Emotional features*. We use emotional valence and arousal separately to predict similarity.
4. *Variance partitioning analysis*. We combine the different musical and emotional features in this analysis to understand the unique variance explained by the different features. Based on the best model of musical features (step 1), we include the pitch-related features and their interactions as one explanatory set, and include tempo, valence, and arousal individually.
5. *Mediation analysis*. Finally, to understand how the musical features might explain the relationships between valence and arousal and similarity, we use a mediation analysis to directly test what (if anything) valence and arousal contribute above and beyond the musical features. This analysis is performed for valence and arousal separately. First, we test the relationships between the emotional feature and individual musical features. If the relationship is significant, then the musical feature is considered a potential mediator. We then include the potential mediators in the model with the emotional feature. If the coefficient of the emotional feature is no longer

significant then there is full mediation, but if it is reduced (relative to the individual models in step 3) but still significant, then there is partial mediation.

Results

Musical features

Individual models Each of the musical features individually predicted similarity (tempo difference: $\beta = -.58$, $p < .001$, $R^2 = .33$; key difference: $\beta = -.34$, $p = .0040$, $R^2 = .10$; voice type difference: $\beta = -.31$, $p = .0099$, $R^2 = .081$; timbre [MFCCs distance]: $\beta = -.30$, $p = .012$, $R^2 = .076$). For each feature, if the original and the cover were more different for the given feature, then they were rated as less similar.

Model with all musical features In a model with all four musical features, tempo and timbre significantly predicted similarity ratings ($\beta = -.51$, $p < .001$; $\beta = -.23$, $p = .020$ respectively), but key and voice type did not ($\beta = -.089$, $p = .41$; $\beta = .15$, $p = .17$ respectively). This model explained 40.27% of the variance of similarity ratings.

Key, voice type, and timbre all relate to global spectral information in music, driven largely by pitch and instrumentation. These factors are independent from the global temporal information of music, as represented by tempo[1]. In a model including tempo and the possible interactions between key, voice type, and timbre, the model explained 47.89% of the variance[2]. Compared to the model with no interactions, this model was sufficiently more explanatory to merit the additional degrees of freedom (ANOVA on nested models: $F(4) = 3.37$, $p = .015$). In this model, tempo significantly predicted similarity, such that original-cover pairs that were more different in tempo were rated as less similar ($\beta = -.58$, $p < .001$, Fig. 2A). The interaction between key and voice type was significant ($\beta = -.26$, $p = .020$, Fig. 2B) and the interaction between timbre and voice type was significant ($\beta = -.25$, $p = .038$, Fig. 2C). This means that for songs where the voice type was the same between the original and cover (both female or both male voices), larger differences in key or timbre corresponded with pairs being rated as less similar (Figs. 2B/C, dots). Songs where the voice type switched between the original and cover were rated as less similar overall, regardless of how different they were on key or timbre (Figs. 2B/C, x's). No other coefficients in the model were significant, including main effects of key, voice type, or timbre.

Genre

Representations of genre are built up by a complex combination of instruments, tempo, rhythms, and cultural associations. Because of this complex relationship between genre and the other musical features, genre is analyzed here separately. It is not included in the variance partitioning or mediation analyses to allow us to focus on the emotional features.

There was a significant relationship between timbre and genre ($\beta = -.57$, $p = .045$), where pairs in the same genre had more similar timbre. This is expected, as different genres use instruments and textures in different ways, producing unique sound colors. There was a marginally significant relationship between tempo and genre ($\beta = -.53$, $p = .064$), suggesting that covers that were in a different genre from their original were

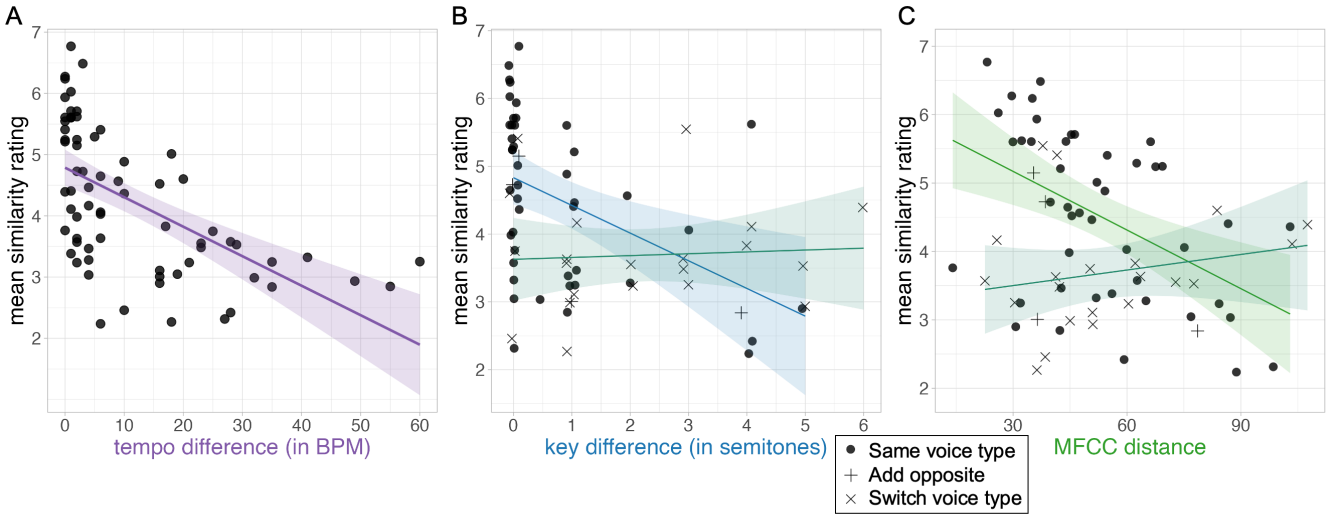

Figure 2. Musical features individually and interactively predict similarity. **(A)** Tempo. **(B)** Key and voice type: blue line is regression line fit to pairs with same voice type (solid dots), teal line is regression line fit to pairs with opposite voice type (crosses and x's). **(C)** Timbre and voice type: green line is regression line fit to pairs with same voice type (solid dots), teal line is regression line fit to pairs with opposite voice type (same as in B). (Legend only applies to B and C.)

[1] Although timbre does have a temporal aspect (i.e. attack, decay), our measure of timbre here averages in time, removing most of this temporal information.

[2] The model with all possible interactions explained 46.52% of the variance. However, this benefit was not worth the additional degrees of freedom ($F(11) = 1.69$, $p = 0.10$).

more likely to change tempo. There was no association between key difference and genre (β = -.077, p = .79) or voice type difference and genre (β = -.22, p = .44).

Genre difference itself predicted similarity (β = .69, p = .014, R^2 = .072), such that original-cover pairs that were in different genres were rated as less similar. We are currently exploring more fine-grained representations of genre to be included in the full analysis with emotional features.

Emotional features

Both valence and arousal predicted similarity (valence difference: β = -.55, p < .001, R^2 = .29; arousal difference: β = -.45, p < .001, R^2 = .19; Fig. 3), such that a greater difference in the emotional ratings for the original and cover on both valence and arousal was associated with lower similarity ratings. The interaction was not significant (β = .016, p = 0.87). The model with both valence and arousal explained 37.63% of the variance of similarity.

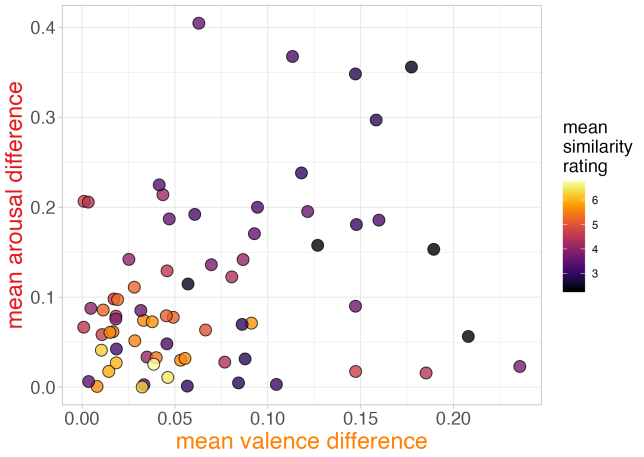

Figure 3: Emotional features individually predict similarity.

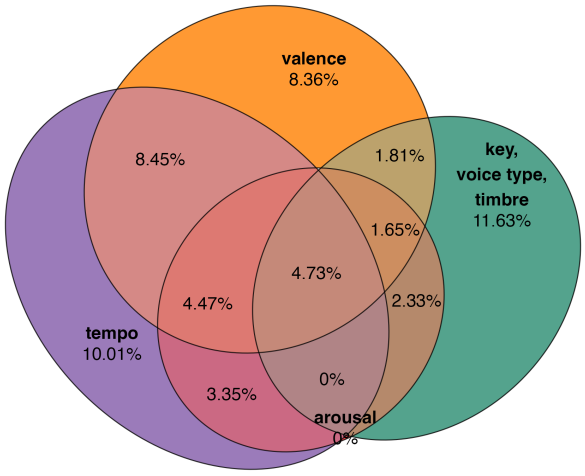

Figure 4: Venn diagram of variance partitioning results.

Variance partitioning analysis

Next, we performed variance partitioning analysis to understand how much of the variance in similarity ratings could be uniquely explained by each feature (or feature set, in the case of key, voice type, and timbre). The model with all musical and both emotional features accounted for 58.21% of the variance in similarity ratings. It should be noted that the musical features explained about 48% of the variance, so already we see that emotional features add some but not much explanatory power on top of the musical features.

Of the total of 58% variance explained, tempo uniquely accounted for 10.01% of the variance; key, voice type, and timbre accounted for 11.63%; valence accounted for 8.36%, and arousal uniquely accounted for 0% (Fig. 4). Together, tempo and valence accounted for an additional 8.45%.

Mediation analysis

Finally, we used mediation analysis to understand how the musical features might explain the relationships between the emotional features and similarity. Based on prior work (Gabrielsson & Lindström, 2010), musical features influence the perception of emotion in music. This hierarchical relationship is not explicitly captured by the variance partitioning analysis, so we test it directly here.

Valence Of the four musical features, tempo was the only one with a significant relationship with valence (β = .39, p < .001), so it was the only potential mediator considered.

When both tempo and valence were included in the model for similarity, both were significant (tempo: β = -.43, p < .001; valence: β = -.39, p < .001). The estimated direct effect of valence on similarity was -.38 (lower bound = -.55, upper bound = -.21) and the estimated indirect effect of tempo was -.17 (lower bound = -.29, upper bound = -.08). This means that valence remained significant, but its coefficient was reduced, suggesting that tempo was a partial mediator between valence and similarity (Fig. 5).

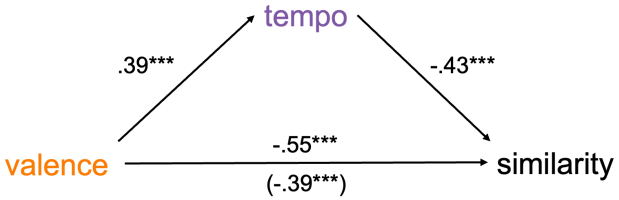

Figure 5: Summary of mediation results for valence.

Arousal Of the four musical features, tempo and timbre had significant relationships with arousal (tempo: β = .45, p < .001; timbre: β = .36, p = .0025), so both were considered as potential mediators.

When tempo, timbre, and arousal were included in the model for similarity, tempo and timbre were significant (tempo: β = -.50, p < .001; timbre: β = -.21, p = .046) but arousal was not (β = -.15, p = .18). The estimated direct effect

of arousal on similarity was -.17 (lower bound = -.36, upper bound = .060). The estimated indirect effect of tempo was -.22 (lower bound = -.37, upper bound = -.080), and the estimated indirect effect of timbre was -.081 (lower bound = -.16, upper bound = .012). This means that tempo fully mediated the relationship between arousal and similarity (Fig. 6). This is consistent with the variance partitioning analysis, which showed that arousal on its own doesn't account for any unique variance in the similarity ratings.

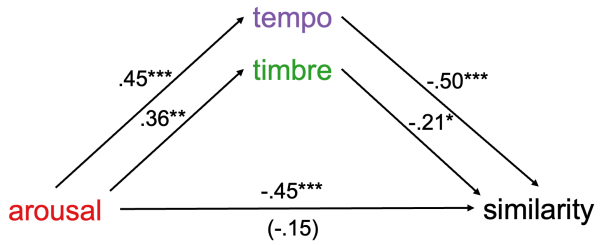

Figure 6: Summary of mediation results for arousal.

Discussion

In this project, we test the influence of features of music at multiple levels on overall perceived similarity. Importantly, we removed the potential impact of lyrics by gathering similarity ratings of pairs of original popular songs and cover versions. While we focused on the relationship between musical and emotional features in this analysis, we plan to include a continuous measure of genre, as a higher-order musical feature, to complement this work in future analysis.

Each of the musical features we looked at (tempo, key, voice type, timbre) predicted similarity. Tempo was the strongest predictor (Fig. 2A), but listeners were also sensitive to changes in other musical features. Key, voice type, and timbre interactively predicted similarity, such that a change in voice type (e.g. a switch from a male to female singer) reduced the impact of key or timbre changes on similarity ratings (Figs. 2B/C). This may be because a significant change in the pitch or timbre of the main vocal line draws listeners' attention away from other spectral features. While the influence of tempo and timbre on similarity ratings had been previously studied (Bogdanov et al., 2011, Novello et al., 2011), to our knowledge, no previous work had looked at key or voice type.

Both tempo and the other musical features (key, voice type, and timbre) uniquely accounted for a portion of variance, but those two categories didn't share any variance with each other unless that variance was also shared with at least one of the emotional features. This suggests that temporal information and global spectral information are processed somewhat independently in similarity perception.

We found that both emotional valence and arousal predicted similarity. Arousal by itself did not account for any unique variance in similarity ratings, and the relationship between arousal and similarity was fully mediated by tempo. This suggests that arousal doesn't contribute anything to similarity above and beyond the effect of tempo, which adds to our understanding of the relationship between arousal, low-level features, and overall impressions of music.

Valence, on the other hand, seemed to have a more complex relationship with the musical features and similarity. It did uniquely account for a portion of the shared variance in similarity ratings (i.e., above and beyond all other features we measured). And tempo only partially explained the relationship between valence and similarity. It is somewhat surprising that valence even varies enough to be meaningful: the traditional drivers of valence—semantic content of lyrics and mode of the music (major/minor)—are identical across originals and covers. This all suggests that valence contributes to similarity in unexpected ways, above and beyond musical features. There is more work to be done to fully understand this relationship. We suspect that if we had varied mode between originals and covers, we may have seen a smaller effect of tempo—this should be tested in the future.

One limitation of our acoustic feature measures is that they are simplified by averaging features across time. While previous work has used MFCCs averaged in time, this does remove potentially important time-varying information. One method for maintaining this information would be to perform time-warping to match each cover to its original and then compare shorter segments of the original and cover clips.

Additional analyses could also look at individual differences in listeners' musical expertise, listening habits, and musical preferences (Rentfrow & Gosling, 2003), which we predict could have some impact on similarity judgments. Further, we could potentially explore the impact of familiarity, which has been shown to affect listeners' neural responses to music (Pereira et al., 2011). However, we chose not to collect familiarity ratings of individual songs in the current study to avoid biasing their similarity judgments.

Future work could also collect similarity ratings across all possible original-original, cover-cover, and original-cover pairs. Having the full "similarity matrix" (rather than just the original-cover diagonal that we include here) would give us power to replicate the current findings as well as explicitly model the effect of semantic content of lyrics by including semantic distance between clips (as quantified by sentence embeddings from a model like BERT, Kenton & Toutanova, 2019). We could also use this expanded set of ratings to map and quantify the distances between genres represented in the corpus, allowing us to more precisely model how genre predicts similarity.

This work links multiple levels of musical features to overall perceptions of musical similarity. We find that similarity judgments between clips of music are driven largely by low-level musical features, with some additional influence of emotional valence, providing insight into the interplay between low-level perceptual features and higher-level emotional processing in naturalistic music perception.

References

Aucouturier, J.J., Pachet, F. (2002). Music similarity measures: What's the use? *ISMIR*.

Bogdanov, D., Serrà, J., Wack, N., Herrera, P., Serra, X. (2011). Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, *13*(4).

Gabrielsson, A., Lindström, E. (2010). The role of structure in the musical expression of emotions. In J. Slodoba & P. Juslin (eds.), *The Handbook of Music and Emotion* (pp. 367-400). New York: Oxford University Press.

Grekow, J. (2017) Audio features dedicated to the detection of arousal and valence in music recordings. *IEEE International Conference on Innovations in Intelligent SysTems and Applications (INISTA)* (pp. 40-44).

Goldstone, R., & Son, J. Y. (2005). Similarity. In K. Holyoak & R. Morrison (Eds.), *Cambridge Handbook of Thinking and Reasoning* (pp. 13–36). Cambridge University Press.

Herre, J., Allamache, E., Ertel, C. (2003). How similar do songs sound? Towards modeling human perception of musical similarity. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 83-86).

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (Vol. 1, No. 2).

McClay, M., Sachs, M. E., & Clewett, D. (2023). Dynamic emotional states shape the episodic structure of memory. *Nature Communications*, *14*(1), Article 1.

McFee, B., Raffel, C., Liang, D., Ellis, D.P.W., McVicar, M., Battenberg, E., Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference* (pp. 18-25).

Moore, B. C. J. (2003). *An introduction to the psychology of hearing* (5th ed.). Academic Press.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23*, 94-140.

Novello, A., McKinney, M.M.F., Kohlrausch, A. (2011) Perceptual evaluation of inter-song similarity in Western popular music. *Journal of New Music Research, 40*(1) (pp. 1-26).

Pereira, C.S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S.L., Brattico, E. (2011) Music and emotions in the brain: Familiarity matters. *PLoS ONE, 6*(11): e27241.

Rentfrow, P. J. & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology, 84,* 1236-1256.

Terasawa, H., Slaney, M., & Berger, J. (2005). The thirteen colors of timbre. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, 323–326. https://doi.org/10.1109/ASPAA.2005.1540234

Tversky, A. (1977). Features of Similarity. *Psychological Review*, *84*(4).

Tzanetakis, G., Cook, P. (2002) Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5) (pp. 293-302).