

Bridging the Gap between Lyrical and Melodic Similarity: an Empirical Analysis of Song Pairs*

Christopher Lohse
Trinity College Dublin
lohsec@tcd.ie

Catalin Gheorghiu
Trinity College Dublin
cgheorgh@tcd.ie

Juan Montenegro
Trinity College Dublin
montenej@tcd.ie

Tharun Ajith
Trinity College Dublin
ajitht@tcd.ie

Adedolapo Adedokun
Trinity College Dublin
adedokua@tcd.ie

Abstract

A common approach for machine-readable music processing and music analysis, the ABC notation provides a straightforward language for transcribing music that opens the way to determining harmonic similarity through string-based methods. While prior studies have concentrated on either lyrical or musical similarity analysis, we inspect how the two methods are correlated, controlling for factors like genre and artist. After scraping a dataset of matching song lyrics and chords with embedded duration, we use the cosine similarity between TF-IDF vectors to model lyrical similarity and a re-scaled string edit distance to model harmonic (chord-based) similarity. We find a positive correlation between lyrical and harmonic similarity that is highly significant in spite of its low magnitude, mostly driven by the rock songs in the sample. We also find that the within-repertoire similarities of pop and rock artists relate differently to similarities with songs of other artists in the respective genres.

Keywords— lyrics, harmony, pairwise similarity, ABC notation, chords, pitch, duration, genre, term frequency, natural language processing, string edit distance

1 Introduction

1.1 Motivation

Music has been part of our human culture for many centuries. People who are aspiring musicians are often influenced by a variety of factors when it comes to their musical development such as family and friends, life experiences, education, and cultural background. However, the influence musicians exert upon each other can be a complex and multifaceted phenomenon, and it can be difficult to determine who influenced who in every case. A huge interest in the similarities between song lyrics and musical melodies has recently grown thanks to new technology. Nowadays, pairwise comparisons are the primary method used in studies of lyrical and musical similarity analysis, in which researchers assess how similar two songs are based on either their lyrics or their melodies (Sheikh Fathollahi and Razzazi, 2021; Post and Toussaint, 2011; Logan and Salomon, 2001; de Haas et al., 2008; Adam et al., 2010).

The problem at hand is the lack of a comprehensive framework for analyzing both lyrics and musical similarity in a single model, which limits our understanding of the relationship between these two factors in creating the emotional impact of music on listeners. While current studies on lyrical and musical similarity analysis rely on pairwise comparisons, these methodologies are limited in their capacity to make broader conclusions. As a result, there is an urgent need to bridge the gap between pairwise lyrical and musical similarity analysis for songs in order to acquire a more comprehensive understanding of the emotional effect that music has and create more sophisticated music recommendation systems which take both lyrics and harmony into account.

1.2 Research Questions

The starting point of our investigation should be determining to what extent lyrical and musical similarity co-occur in general, to then narrow down the field in search of specific factors that make this co-occurrence more (or less)

*Link to Github: <https://github.com/ChristopherLohse/TextAnalyticsPaper>

likely to present itself. To narrow down our study we model musical similarity as harmonic similarity based on the chords of a song.

We thus formulate three research questions to guide our analysis, each step increasing the chance for a link between similarities to become more apparent:

1. Are two songs similar in lyrics also similar in harmony?
2. Is lyrical similarity more closely related to harmonic similarity for a given genre than it is for another?
3. Is lyrical similarity more closely related to harmonic similarity within a given artist’s repertoire than it is between the repertoire and the rest of the genre?

We chose to focus on pop and rock as the genres to be contrasted – the data section gives clear reasons why. With this in mind, while we do not particularly expect lyrical similarity to be very strongly correlated with harmonic similarity for the average song, we find it likely that pop songs will present relatively more significant correlations thanks to the lower range of both feelings and chords employed. On the other hand, we expect rock artists to have much stronger correlations between songs in their own repertoire compared to the rest of the genre thanks to the opposite trend of experimentation and complexity that is inherent to rock music in our experience.

The next section contains a literature review meant to funnel previous findings in the field into a research framework, grounding our methods and giving some perspective on similarity analysis. Section 3 gives an overview of our data sources and pre-processing steps, ending with a description of the final-form dataset that we hope may enable future research in this direction beyond our paper. Section 4 outlines the methods used to answer the research questions, giving detailed explanations of the quantities we derive and the significance testing process. Section 5 relates our results and our interpretation of these findings. Finally, section 6 summarises our work, clarifies its limitations, and gives some specific directions for future research.

2 Related Work

This section reviews the literature on musical and textual similarity analysis, focusing on methods for determining textual similarity in song lyrics, musical similarity metrics, chord similarity measurements, and tree-based musical similarity approaches. We also discuss related work in calculating similarity measurements between two songs, addressing both lyrical and melodic/harmonical similarities.

2.1 Lyrical Similarity

We define lyrical similarity as the similarity of two songs on the basis of the words contained in the lyrics of the songs. TF-IDF (term frequency-inverse document frequency) is a numerical statistic that reflects the importance of a word in a document or a corpus. It is commonly used in information retrieval and text mining. Thus its application to lyrical similarity to encode the text is obvious (Van Zaanen and Kanters, 2010; Wang et al., 2011; Gossi and Gunes, 2016). These approaches are mood classification approaches of songs Van Zaanen and Kanters (2010); Wang et al. (2011), or music recommender systems based on lyrical similarity (Gossi and Gunes, 2016). Some approaches combine lyrical and medical information to improve the capabilities of music recommender systems (Laurier et al., 2008; Hu and Downie, 2010), making research into the correlation between lyrical and musical similarity an obvious next step.

2.2 Musical Similarity

We define musical similarity as the similarity between two songs based on melodic or harmonic features. To encode musical information in a computer-readable format the ABC notation (Walshaw, 1993, 2014) converts notes or chords into their alphabetical counterpart. After this conversion step the processing of musical information can be done with standard text analytics methods, and again the use of TF-IDF is a good way to encode this information (Su et al., 2014; Li et al., 2019).

After encoding two songs in a music vector or embedding standard distance measurements like cosine similarity or euclidian distance can be used to determine the similarity of two songs (Sheikh Fathollahi and Razzazi, 2021). Other more sophisticated approaches compare songs based on tree-based similarity metrics (McFee and Lanckriet, 2011; Rizo and Inesta, 2010). McFee and Lanckriet (2011) investigates using the classical KD-tree algorithm to efficiently index high-dimensional data by spatial partitioning for music information retrieval demonstrating the effectiveness of this approach (McFee and Lanckriet, 2011). Rizo and Inesta (2010) take advantage of the fact that music pieces can be represented by symbolic structures such as strings or trees containing the sequence of notes in the melody. They find that tree-based summaries of input musical data are both fast and accurate in similarity evaluation (Rizo and Inesta, 2010).

Analysing chord similarity is a well-known and frequently used process in music information retrieval and analysis tasks (Logan et al., 2004; de Haas et al., 2008; Adam et al., 2010). Determining the similarity of two songs by using chords can be considered as measuring the harmonical similarity between two songs because the melody of a song is not contained in its chords.

Some methods apply similarity measurements on the raw audio files and perform signal similarity analyses (Logan et al., 2004; de Haas et al., 2008). While other approaches work with the annotated chord progression in a textual format, after generating this annotation from the audio files (Adam et al., 2010). Adam et al. (2010) generate chord annotation by using a hidden Markov model approach.

de Haas et al. (2008) propose a method applying the k-means clustering algorithm to the spectral analysis of musical pieces. While some methods use standard similarity metrics such as cosine similarity or euclidean distance (Adam et al., 2010). Other methods use other similarity measurements such as the earth mover distance by Rubner et al. (2000) (Logan et al., 2004).

Logan et al. (2004) calculate the similarity of music by analysing the audio signals of musical pieces (Logan et al., 2004). Their method applies the k-means clustering algorithm to the spectral analysis of musical pieces. By using the earth mover distance (Rubner et al., 2000) pairwise similarity can be calculated (Logan et al., 2004). Another method used to calculate the similarity between two texts is the Levenshtein distance (Levenshtein et al., 1966), which is further outlined in 4.2. The Levenshtein distance is often used for harmonic (De Haas et al., 2009) or melodic (Huang et al., 2013) similarity measurements. And is considered computationally cheap due to the use of dynamic programming paradigms. Because of that, we decide to compare the harmonic similarity of the two songs based on the Levenshtein distance.

Considerable research has been done on encoding the melody or harmony of songs and determining their similarity, as mentioned earlier, some approaches combine lyrical and musical encoding to improve performance for downstream tasks based on this information, such as mood classification (Van Zaanen and Kanters, 2010; Wang et al., 2011) or song similarity for music recommendation (Gossi and Gunes, 2016). This is a step in the right direction, as (Liem et al., 2011) points out that there is a need to combine these two steps to further advance the field of music information retrieval (Liem et al., 2011). However, before doing so, it could be considered useful to explore the similarity between the lyrical and musical information contained in a song to justify the use of both information for encoding songs.

2.3 Correlation Measurements between two scores

In the natural language domain comparing two different scores occurs e.g. when a score calculated by a model based on the data is compared to a gold standard score based on human evaluation (Lin, 2004; Zhao et al., 2019; Gao et al., 2020). In these cases often three correlation score methods are used in order to obtain a measurement of the correlation between two scores: Pearson, Kendall and Spearman correlation (Lin, 2004; Zhao et al., 2019; Gao et al., 2020), which are explained in 4.3.

The Kendall and Spearman correlations are generally better at capturing a monotonic nonlinear relation between two variables, whereas the Pearson correlation is better at capturing a linear association between two variables (van den Heuvel and Zhan, 2022).

3 Data

The lyrics, as well as the pitches and durations of the chords, are extracted using a Python script that utilises the *Selenium WebDriver*¹ to gather the data. The chord's duration is calculated by the number of whitespaces until the next chord or the end of the line. The *pandas*² library is used to store the data for further processing. The chords of a song are used as a representation of the song's harmony. The lyrics and the corresponding chords for a song are taken from the website *ultimate-guitar*³.

We removed song duplicates, keeping only the version with the highest popularity score as we assume this popularity to be a proxy for fidelity to the original. To retain songs with only English words, we used the *langdetect*⁴ library to identify the overall language of the lyrics. We also re-balanced the dataset to contain a roughly equal amount of pop and rock songs, while removing all entries from artists with less than 3 unique songs in the corpus to definitively remove any non-English songs *langdetect* missed and increase the chances of significant results when the artist's role in similarity correlation is explored.

¹<https://github.com/SeleniumHQ/selenium/>

²<https://pandas.pydata.org/>

³<https://www.ultimate-guitar.com/>

⁴<https://pypi.org/project/langdetect/>

We performed a small amount of lyric cleaning at this stage by removing instructions in square brackets; further steps, such as removing punctuation or converting all characters to lowercase, will be done automatically as the lyrics are tokenised before TF-IDF is applied; a more thorough description of this process follows in the methodology. Table 1 should help the reader visualise the processed dataset we will be working with, transposed here for a lack of better formatting options.

Table 1: Table header after of dataset after pre-processing, to be used in the upcoming analysis.

Song ID	0	1	814
Artist ID	1	1	168
Genre	Pop	Pop	Rock
Pitch Series	[F, F, E, ...]	[Em, Cadd9, Em, ...]	[C, Em/B, Em/B, ...]
Pitch-Duration Series	[(F, 1), (F, 12), (E, 10), ...]	[(Em, 2), (Cadd9, 2), (Em, 2), ...]	[(C, 1), (Em/B, 4), (Em/B, 4), ...]
Enhanced Pitch Series	[F, F, F, ...]	[Em, Em, Cadd9, ...]	[C, Em/B, Em/B, ...]
Lyrics	NSYNC That girl (will never be...	Yeahyeah... oh... Hmm...	Lethargy got a hold of me...

After the preprocessing of the data, we have lyrics and chords of 815 unique songs. Out of these, 398 are pop and 417 are rock songs. The dataset comprises a total of 97 unique artists, with an average of just over 8 songs per artist; the average song has about 6 unique chords.

It should be noted that there are three series of pitches in the dataset. 'Pitch-Duration Series' is what we originally extracted, where chords are represented by tuples of pitch and a duration value re-scaled to be an integer multiple of the shortest duration in the corpus; this way, our durations start from 1 and the maximum duration is reduced substantially without compromising the proportions between values. The aforementioned re-scaling of the durations is very helpful for constructing the 'Enhanced Pitch Series' – dubbed 'Duration-Enhanced' from this point onwards – which transforms each chord tuple into a list that repeats the pitch as many times as the duration specifies. This series will be used to incorporate duration in the measurement of harmonic similarity. Finally, as it may be worth investigating the impact of ignoring duration in our experiment, the 'Pitch Series' is derived from the 'Pitch-Duration Series' by simply removing the durations in each chord's tuple.

4 Methodology

At the end of the data retrieval process, we are working with a corpus of song lyrics documents C_L and a separate corpus of chord string vector documents C_M , one for each song in C_L . For convenience, let there be a unified corpus $C = C_L \cup C_M$. The following sub-sections are meant to describe the basic technique suitable for analysing textual similarity, and a more advanced method specifically built for quantifying differences between strings – in our case, ABC-notation musical scores.

4.1 Lyrical Similarity

The Term Frequency / Inverse Document Frequency (TF-IDF) approach is a widely used heuristic in text analytics due to its robustness and independence of training dictionaries. In order to employ this framework for investigating lyrical similarity, we tokenise each document in C_L using the *nltk*⁵ package and lemmatise the result using the *WordNetLemmatizer* in order to incorporate knowledge of language structure for the effective merging of similar tokens, such as different forms of the same word. As lyrical expression does not necessarily follow the conventions of ordinary communication, we opt to keep stopwords in the corpus; this decision should also make the dataset more relevant to future research, as applications such as sentiment analysis are quite sensitive to the removal of high-frequency terms.

The *TfidfVectorizer* contained in the *Scikit-Learn*⁶ package would then map the sequence of tokens in the entire corpus to a single vector in bag-of-words style and convert the number of word occurrences to relative frequencies, giving the most weight to words that appear often in a given song but rarely overall. For term i and document d , the

⁵<https://github.com/nltk/nltk>

⁶<https://github.com/scikit-learn/scikit-learn>

first element is the term frequency TF , which is a measurement of how frequent a term i is in a single text d and can be expressed as.

$$TF_{i,d} = \frac{x_{i,d}}{\sum_i x_{i,d}}, \quad (1)$$

where $x_{i,d}$ is the number of times a term i occurs in a single text d . The second element, inverse document frequency IDF , can be considered as the uncommonness of a term i in the whole corpus of documents $d \in \{1, \dots, N\}$. The *scikit-learn* package computes this uncommonness as

$$IDF_{i,d} = \log \left(\frac{1 + N}{1 + \sum_d I(i,d)} \right) + 1, \quad (2)$$

$$\text{where } I(i,d) = \begin{cases} 1 & \text{if term } i \text{ in } d \\ 0 & \text{otherwise.} \end{cases}$$

The TF-IDF as a combination of TF and IDF and is calculated for a term i in a document d with

$$TFIDF_{i,d} = TF_{i,d} IDF_{i,d}. \quad (3)$$

The resulting TF-IDF vector is filled in for every song, allowing us to compute the pairwise similarity between lyrics using the standard cosine similarity formula from equation 4. Document vectors \mathbf{a} and \mathbf{b} have the same number of elements, with $\|\cdot\|$ denoting the vector's module and $|\cdot|$ denoting absolute value. The further away from 0, the higher the similarity between documents $S_{a,b}^{(L)}$, allowing us to quantify the (dis)similarity of different song pairs.

$$S_{a,b}^{(L)} = \frac{|\mathbf{a} \cdot \mathbf{b}|}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (4)$$

4.2 Harmonic Similarity

Levenshtein distance – or string edit distance – is defined as the minimum number of edit operations needed to change one string to match another (Levenshtein et al., 1966). Canonically, three operations are allowed: substituting one character for another, deleting one character, or inserting one character. We can use this metric of distance on duration-enhanced pitch sequences to assess how similar different melodies are. We implement the efficient dynamic programming "Algorithm X" of Wagner and Fischer (1974) to compute the edit distance $lev(a, b)$ between two musical scores; the pseudocode for this algorithm is given in the appendix (A.1). The higher the Levenshtein distance between scores, the less similar they are. with the unique advantage of preserving the order of the chords over alternatives such as the Jaccard similarity.

As identical scores should have a similarity of 1 and we wish to use the same 0–1 scale that cosine similarity is on, we obtain harmonic similarity by subtracting the min-max feature scaled Levenshtein distance from 1, as per equation 5. Note that minima and maxima are taken over the entire corpus, even in later settings where the paired songs (a, b) may be clustered by genre or artist – this preserves the validity of our results across settings, especially when harmonic similarity interacts with lyrical similarity.

$$S_{a,b}^{(M)} = 1 - \frac{lev(a,b) - \min_{(a,b)} lev(a,b)}{\max_{(a,b)} lev(a,b) - \min_{(a,b)} lev(a,b)} \quad (5)$$

4.3 Correlation Measurements

As we assume a linear relationship between the pairwise lyrical similarity $S^{(L)}$ and harmonic similarity $S^{(M)}$, the Pearson correlation r as a measurement of correlation is a sensible choice. The scope of our first research question is the entire corpus C , with no restrictions on artist or genre, thus covering every unique pair $(a, b) \in C \times C$. The Pearson correlation used to answer this first research question is given in equation 6, where $b > a$ means that song b appears later in the corpus than song a and \bar{S} is the mean of all similarities, lyrical or harmonic.

$$r_C = \frac{\sum_{a \in C} \sum_{b \in C, b > a} (S_{a,b}^{(L)} - \bar{S}^{(L)})(S_{a,b}^{(M)} - \bar{S}^{(M)})}{\sqrt{\sum_{a \in C} \sum_{b \in C, b > a} (S_{a,b}^{(L)} - \bar{S}^{(L)})^2 \sum_{a \in C} \sum_{b \in C, b > a} (S_{a,b}^{(M)} - \bar{S}^{(M)})^2}} \quad (6)$$

For the second research question, a separate correlation should be computed for each genre. Let then r_P and r_R the correlations between lyrical and harmonic similarity for pop and rock songs, respectively. If P contains all the

pop songs in the corpus, then the formula of r_P is trivially derived from equation 6, and the same applies to the corpus of rock songs R .

For the final research question, we compute four correlations for each artist in C . Let $A^{(P)}$ and $A^{(R)}$ denote the set of pop and rock artists respectively; $A^{(P)} \cap A^{(R)} = \emptyset$ must not necessarily hold. Furthermore, for an arbitrary genre $G \in \{P, R\}$, let G_i , $i \in A$ denote the songs of artist i in the given genre. We can then define the set $G_i^c = G \setminus G_i$ containing all of the songs within the given genre, but outside the repertoire of artist i . With all these definitions in place, the Pearson correlations between the lyrical and harmonic similarities of songs in the repertoire of artist i can be defined as r_{G_i} , $G \in \{P, R\}$. Finally, when pairing songs from the artist's repertoire to other songs in the genre for similarity and correlation results, we obtain $r_{G_i^c}$, defined explicitly in equation 7 due to its complexity.

$$r_{G_i^c} = \frac{\sum_{a \in G_i} \sum_{b \in G_i^c} (S_{a,b}^{(L)} - \bar{S}^{(L)})(S_{a,b}^{(M)} - \bar{S}^{(M)})}{\sqrt{\sum_{a \in G_i} \sum_{b \in G_i^c} (S_{a,b}^{(L)} - \bar{S}^{(L)})^2 \sum_{a \in G_i} \sum_{b \in G_i^c} (S_{a,b}^{(M)} - \bar{S}^{(M)})^2}} \quad (7)$$

Note that the formula above is only theoretical, and many of the correlations will not actually be computed because it is rare for a given artist to have songs in multiple genres. For most artists, only one of the pairs $(r_{P_i}, r_{P_i^c})$ and $(r_{R_i}, r_{R_i^c})$ will be computed. The next section details how the correlations' values are used to answer the research questions, and how we test the significance of our results.

4.4 Hypothesis Testing

The null hypothesis of the first research question boils down to $H_0 : r_C = 0$, with the two-way alternative $H_a : r_C \neq 0$. A student's T-test is the intuitive solution; equation 8 gives the test statistic with $N - 2$ degrees of freedom (d.o.f.), where $N = \text{card}[(a, b) : a \in C, b \in C, b > a]$ is the number of song pairs. The P-value of this test statistic informs the level of confidence we can reject the null hypothesis with.

$$t = \frac{r_C \sqrt{N-2}}{\sqrt{1-r_C^2}} \sim t(N-2) \quad (8)$$

The second research question requires a comparison of r_P and r_R . Assuming that the correlation between the lyrical and harmonic similarity of pop songs is virtually indistinguishable from that of rock songs leads to the null hypothesis $H_0 : r_P - r_R = 0$, with the two-way alternative $H_a : r_P - r_R \neq 0$. In order to test this difference of correlations, we perform Fisher Z-transformations (Fisher, 1915) on each term, yielding normally distributed variables for arbitrary genre G as per equation 9, where N_G denotes the number of song pairs within the genre.

$$z_G = \frac{1}{2} \ln \frac{1+r_G}{1-r_G} \sim N\left(\frac{1}{2} \ln \frac{1+r_G}{1-r_G}, \frac{1}{\sqrt{N_G-3}}\right) \quad (9)$$

By the properties of the normal distribution, the difference of the transformed correlations Z has a standard normal distribution under the null hypothesis, making it the perfect candidate for our test statistic. Its exact expression is

$$Z = \frac{z_P - z_R}{\sqrt{\frac{1}{N_P-3} + \frac{1}{N_R-3}}} \sim N(0, 1) \quad (10)$$

and the corresponding P-value can be compared against a Z table to identify the level of confidence we can reject the null hypothesis with.

Finally, the third research question requires an aggregation of comparisons between r_{G_i} and $r_{G_i^c}$ across artists and genres. For any given artist-genre combination, a test statistic $Z_{G,i}$ can be derived through the same Fisher Z-transformation employed for the second research question, taking care to modify the number of similarities N for every combination. Within each genre, we can count the number of significant P-value at a chosen confidence level (i.e. 5%), and compare how likely it is for the artist's lyrics and melody to be more strongly correlated within their repertoire. Proportions of significance above 50% would prompt us to answer the final research question positively. As a bonus, in order to see if different genres have different proportions of "original" artists, we conduct a two-sample Z-test on the proportions of significant results \hat{p}_P and \hat{p}_R for pop and rock, respectively.

The test statistic is given in equation 11, having a standard normal distribution under the null hypothesis $H_0 : \hat{p}_P = \hat{p}_R$.

$$Z = \frac{\hat{p}_P - \hat{p}_R}{\sqrt{\hat{p}_P(1-\hat{p}_P) \left(\frac{1}{N_P} + \frac{1}{N_R}\right)}} \sim N(0, 1) \quad (11)$$

Comparing the ensuing P-value against a Z table should deepen our insight into the differences (or lack thereof) between genres.

5 Results and Discussion

We derived similarities, correlations, and significance levels for two harmonic representations of each song. The first is dubbed 'Pitches' in the upcoming results tables – it is obtained by stripping the durations from the set of scraped chords described in section 3. The second relevant harmonic representation, dubbed 'Pitch Duration', is based on the 'Enhanced Pitch Series' also described in the Data section. Comparing the results of the two representations may give some insight into the relevance of duration for pairwise similarity studies at large.

Table 2 reveals that lyrical similarity is positively correlated with musical similarity across the entire corpus. Even if the correlations themselves are rather small, the large sample size of pairs makes their existence nevertheless highly significant. It is noteworthy that using the enhanced pitch series results in a larger observed r_C – almost double in magnitude – with a higher significance as well, corroborating the intuition that incorporating durations paints a clearer picture of the musical structure of the song.

Table 2: Results for question 1

	Correlation	p-Value
Pitches	0.00784	$1.57e^{-5}$
Duration-Enhanced	0.01288	$1.34e^{-12}$

The genre-based analysis of the second research question is captured in table 3. From the 'Pop' multi-column, it is clear that the correlation between lyrical and musical similarity is not significantly different from 0 for the average pop song. Including durations in the analysis does not change this result. This is not the case in the 'Rock' multi-column, however, where a positive correlation is identified and incorporating durations amplifies its magnitude, almost doubling it. The difference in correlations is significant at least at the 5% level for both settings, suggesting a clear difference between genres: while we cannot tell whether it is the music that determines the lyrics or the other way around, certain lyrics appear to be associated with specific musical structures in rock music.

Table 3: Correlation and p-value for pop and rock pitches and durations

	Pop		Rock	
	Pitches	Duration-Enh.	Pitches	Duration-Enh.
Correlation	0.0019	-0.0019	0.0139	0.0267
p-value	0.6160	0.6005	0.0001	0.0000
p-value of correlation diff.				
	Pitches	Duration-Enh.	Duration-Enh.	
	0.0194		0.0000	

Examples of this may include major chords overlapping with powerful statements, or low pitches overlapping with serious messages. Investigating such heuristics may be a worthy path for future research in the field to pursue. The same cannot be said for pop music, where there does not seem to be a pattern of what sort of music best fits specific lyrical segments; this conforms to our initial hypothesis, following the intuition that a bias towards using the most popular musical themes of the moment erodes the relationship between music and lyrics.

It is now understood that the comparatively high correlation between the lyrical and musical similarities of rock songs was driving the lower correlation found in the wider corpus, and that the corpus-wide correlation cannot be interpreted as necessarily applying to either genre; it should be seen as an average instead.

The artist-based analysis of the third research question is captured in table 4. The first and third column capture the proportion of artists whose within-repertoire correlation between lyrical and musical similarity is significantly different from the correlation between similarities taken over songs in their repertoire and songs outside of it, but still within the genre. Judging by the positive average difference in the second column, at least about 24% of pop artists present stronger correlations between the lyrical and musical similarities in their own repertoire. Meanwhile, the negative average difference in the fourth column would suggest that at most 17% of rock artists have a stronger

correlation between lyrical and musical similarity when their songs are paired with rock songs of other artists. The difference between proportions is insignificant judging by the P-values in column 5, regardless of whether the durations are incorporated in the analysis or not, suggesting there is no difference between genres in this regard.

Table 4: Proportion and average difference value for pop and rock pitches and durations

	Pop		Rock		p proportion diff.
	proport. signif. different pop artists	avg. diff.	proport. signif. different rock artists	avg. diff	
Pitches	0.2667	0.3028	0.1190	-0.2971	0.1557
Duration-Enh.	0.24444	0.3212	0.1667	-0.2400	0.4219

The average difference in column 4 is negative, meaning that the correlation between similarities is not determined by the artist when it comes to rock. It is interesting to note that incorporating duration in the analysis markedly increased the proportion of rock artists for whom the within-repertoire correlation is significantly different, suggesting that the overall increase in correlations observed in table 3 was mostly driven by higher within-repertoire correlations. In other words, with a deeper understanding of the musical structure comes a clearer pattern corresponding to the artist. Nevertheless, our hypothesis that rock artists would have stronger correlations within their own repertoire is disproven. The positive values in the second column would suggest the opposite of our prior belief is true: pop artists seem to generally have more of a personal style that permeates their repertoire than rock artists do.

6 Conclusion

We began our investigation of harmonic and lyrical similarity with three research questions: One, are two songs similar in lyrics and also similar in harmony? Two, is lyrical similarity more closely related to harmonic similarity for a given genre than for another? And three, is lyrical similarity more closely related to harmonic similarity within a given artist’s repertoire than in the rest of the genre?

In our results, we have shown that across the corpus, lyrical similarity is positively correlated with harmonic similarity. For the second question in our corpus, we noted that in rock there is a significant correlation between lyrical and harmonic similarity, whereas in pop there is little correlation. Finally, for our third question, we inferred that artists in both genres have within-repertoire correlations, consistent with our intuitions that artists tend to have a personal style within their repertoire. We also learned that pop artists tend to have a more personal style that saturates their repertoire than rock artists.

Further, our study also has some limitations. Namely, in the size of our dataset and the selection of artists and genres for analysis. Our study examined only two genres with less than 1000 artists, far fewer than what exists in the documented musical world. Future research may consider a larger musical corpus and greater musical variety in the analysis, as pop and rock are just two of many contemporary and relevant genres today.

An improved metric of chord duration could be implemented to accurately provide an overview of the melody and have the span of the chord under a specific unit. In the future one could also incorporate functional music theory into our analysis to reduce the chords to their function: for example, an Em and Em7 in pop tend to have the same function and can be thought of as musical synonyms of different flavours. This may increase similarity and correlation results by aggregating equivalences. Lastly, another exciting dimension to consider in future research would be to investigate how the musical and lyrical similarity might change over time, as music and genre are dancing landscapes; there are constantly growing, shifting, and changing over time.

Future work may also extend our analysis by using state-of-the-art text processing techniques. A potential further analysis could be looking at sentiment and emotion; investigating whether the emotional meaning conveyed in the song is a connecting factor in its harmonic similarity with other songs of the same genre. The distinctive variations among the genres hold great potential for exciting future endeavours. As discussed earlier in the results section, we found that certain lyrics appear to be associated with specific musical structures in rock music, such as major chords overlapping with powerful statements, or low pitches overlapping with serious messages. These musical devices and themes give solid grounds for future research in the field to pursue.

References

- Adam, E., Nouné, E., and Yared, Y. (2010). A system for music similarity search based on harmonic content. *Beirut, Lebanon*.
- de Haas, W., Veltkamp, R., Wiering, F., et al. (2008). Tonal pitch step distance: a similarity measure for chord progressions. In *Proceedings of the Eighth International Conference on Music Information Retrieval (ISMIR)*.
- De Haas, W. B., Rohrmeier, M., Veltkamp, R. C., Wiering, F., et al. (2009). Modeling harmonic similarity using a generative grammar of tonal harmony. In *Proceedings of the Tenth International Conference on Music Information Retrieval (ISMIR)*.
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.
- Gao, Y., Zhao, W., and Eger, S. (2020). SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.
- Gossi, D. and Gunes, M. H. (2016). Lyric-based music recommendation. In *Complex networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016*, pages 301–310. Springer.
- Hu, X. and Downie, J. S. (2010). Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 159–168.
- Huang, T., Xia, G., Ma, Y., Dannenberg, R., and Faloutsos, C. (2013). Midifind: fast and effective similarity searching in large midi databases. In *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research*, pages 209–224. sn.
- Laurier, C., Grivolla, J., and Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *2008 seventh international conference on machine learning and applications*, pages 688–693. IEEE.
- Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Li, S., Jang, S., and Sung, Y. (2019). Automatic melody composition using enhanced gan. *Mathematics*, 7(10):883.
- Liem, C. C., Müller, M., Eck, D., Tzanetakis, G., and Hanjalic, A. (2011). The need for music information retrieval with user-centered and multimodal strategies. In *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, MIRUM '11, page 1–6, New York, NY, USA. Association for Computing Machinery.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Logan, B., Kositsky, A., and Moreno, P. (2004). Semantic analysis of song lyrics. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, volume 2, pages 827–830 Vol.2.
- Logan, B. and Salomon, A. (2001). A music similarity function based on signal analysis. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 745–748.
- McFee, B. and Lanckriet, G. R. (2011). Large-scale music similarity search with spatial trees. In *ISMIR*, pages 55–60. Citeseer.
- Post, O. and Toussaint, G. (2011). The edit distance as a measure of perceived rhythmic similarity. *Empirical Musicology Review*.
- Rizo, D. and Inesta, J. M. (2010). Trees and combined methods for monophonic music similarity evaluation. *Proceedings of the Annual Music Information Retrieval Evaluation exchange*.
- Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover’s distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99.

- Sheikh Fathollahi, M. and Razzazi, F. (2021). Music similarity measurement and recommendation system using convolutional neural networks. *International Journal of Multimedia Information Retrieval*, 10:43–53.
- Su, L., Yeh, C.-C. M., Liu, J.-Y., Wang, J.-C., and Yang, Y.-H. (2014). A systematic evaluation of the bag-of-frames representation for music information retrieval. *IEEE Transactions on Multimedia*, 16(5):1188–1200.
- van den Heuvel, E. and Zhan, Z. (2022). Myths about linear and monotonic associations: pearson’s r , spearman’s ρ , and kendall’s τ . *The American Statistician*, 76(1):44–52.
- Van Zaanen, M. and Kanthers, P. (2010). Automatic mood classification using TF*IDF based on lyrics. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, pages 75–80.
- Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.
- Walshaw, C. (1993). Abc2mtex: An easy way of transcribing folk and traditional music, version 1.0. *University of Greenwich, London*.
- Walshaw, C. (2014). A statistical analysis of the abc music notation corpus: Exploring duplication. In *Workshop on Folk Music Analysis (FMA2014)*, page 2.
- Wang, X., Chen, X., Yang, D., and Wu, Y. (2011). Music emotion classification of chinese songs based on lyrics using tf* idf and rhyme. In *ISMIR*, pages 765–770.
- Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. (2019). MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Wagner-Fischer Algorithm

Algorithm 1 Wagner-Fischer Algorithm

```
1: function LEV( $a, b$ )
2:    $m \leftarrow \text{len}(a)$ 
3:    $n \leftarrow \text{len}(b)$ 
4:    $dist \leftarrow \text{zeros}(m, n)$ 
5:   for  $i \leftarrow 0$  to  $m$  do
6:      $dist[i, 0] \leftarrow i$ 
7:   end for
8:   for  $j \leftarrow 0$  to  $n$  do
9:      $dist[0, j] \leftarrow j$ 
10:  end for
11:  for  $j \leftarrow 1$  to  $n$  do
12:    for  $i \leftarrow 1$  to  $m$  do
13:       $dist[i, j] \leftarrow \min(dist[i - 1, j] + 1, dist[i, j - 1] + 1, dist[i - 1, j - 1] + (a[i] \neq b[j]))$ 
14:    end for
15:  end for
16:  return  $dist(m, n)$ 
17: end function
```

A.2 Authors' Contributions

Christopher Lohse:

Conducted related work on chord similarity
Developed problem statement and research questions
Served as a verifier for the team
Fixed references and edited subsections
Clarified tf-idf description
Coordinated gap in literature from motivation to what's in the literature review
Optimized performance of some parts of the code
Rewrote literature review according to the reviews received
Proofread the whole paper and edited accordingly


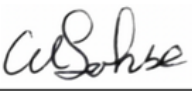

Catalin Gheorghiu, Christopher Lohse, Tharun Ajith, Juan Montenegro, Adedolapo Adedokun

, , , , 

Catalin Gheorghiu:

Served as the chair for the team
Contributed to identifying directions for future research
Abstract rewrites
Synthesised research questions
Hypothesised answers
Contributed to dataset description
Data preprocessing
Written methodology
Written results and discussion
Code underlying methodology and results

Catalin Gheorghiu, Christopher Lohse, Tharun Ajith, Juan Montenegro, Adedolapo Adedokun

, , , , 

Juan Montenegro:

Developed the abstract and identified the motivation for the study
Conducted related work on similarity measurements
Served as a recorder for the team
Fixed contributions and edited subsections
Conducted tf-idf and Levenshtein distance analysis on papers in literature review

Catalin Gheorghiu, Christopher Lohse, Tharun Ajith, Juan Montenegro, Adedolapo Adedokun

, , , , 

Adedolapo Adedokun:

Conducted related work on tree-based approaches and text similarity
Served as an ambassador for the team
Identified how the quantities found in the research help answer the research question
Identified the gap in knowledge that leads to the research question in the introduction.
Contributed to harmonic similarity in literature review as well as writing conclusion in final paper
Identified potential for future research using functional music theory

Catalin Gheorghiu, Christopher Lohse, Tharun Ajith, Juan Montenegro, Adedolapo Adedokun

, , , , 

Tharun Ajith:

Wrote the web crawler to gather data
Contributed to data and preprocessing descriptions
Served as an accountant for the team
Identified the Python packages for data
Contributed dataset specifications
Identified future work for the dataset, including adding more features and conducting genre analysis

Catalin Gheorghiu, Christopher Lohse, Tharun Ajith, Juan Montenegro, Adedolapo Adedokun

, , , , 