Detecting Music Similarity - A Novel Framework Combining Metadata and Content-Based Similarity Measurements Using Deep Learning

Lavonnia Newman[1], Dhyan Shah[1], Chandler Vaughn[1], Faizan Javed[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Adjunct Lecturer, Data Science, Southern Methodist University,
Dallas, TX 75275 USA
{lavonnian, dhyans, cvaughn, fjaved}@smu.edu

Abstract. Music is incorporated into our daily lives whether intentional or unintentional. It evokes responses and behavior so much so there is an entire study dedicated to the psychology of music. Music creates the mood for dancing, exercising, creative thought or even relaxation. It is a powerful tool that can be used in various venues and through advertisements to influence and guide human reactions. Population segmentation can be achieved through the musical genres. Identification with genres such as classical, pop, rock, country, gospel can be used as avenues to connect with the targeted audience to influence purchases, associate with candidates running for office, or even trigger long ago memories. Classification of music into one genre or across genres increases the targeted audience pool, hence detecting similarities and classifying music into those various groups is a useful tool for companies and organizations. With the age of the Internet, it is now easier than ever to access, distribute, modify, and create derivative musical works. Yet programmatically assigning genre classification, and determining music similarity is still a nacent field of study. Through the course of this paper, we explore various approaches to detecting music similarity between songs. We then propose a unique, data-driven, approach to detecting music similarity across a diverse feature space.

1 Introduction

Almost everyone you speak with loves Music. It evokes emotion and enriches our lives. In many ways it is a core component to our human existence, and it is embedded in our daily lives so much that it can often go unnoticed. Unintentional music exposure occurs through commercial advertisements, elevators, bars, restaurants, stores, trains, planes, movies, sporting events and many more venues too numerous to list. It is used to set the mood for the occasion such as dancing, aerobic exercise, praise and worship, or even relaxation or work [1]. When combining intentional music listening with background music, the average American is exposed to music more than 30% of their waking hours per day [1].

Individual musical tastes vary considerably. Musical genres such as rock, classical, pop, country, folk, easy listening, and gospel have been used to segment society, to varying degrees of success. But, over the past few decades, we have seen an explosion of musical categories. This has largely been driven by massive technological shifts in computing, networking, data processing, and music production equipment. Anyone with a few thousand dollars can produce an album in their garage and distribute it via the Internet. This, coupled with considerable business model changes for the music industry towards *streaming*, have created a rapidly advancing need for automation and intelligence in categorization and analysis of musical works.

Motivation: With the age of digital technology, it is now easier than ever to explore inherent musical genre and its characteristics. Musical genres are categorical labels created by artists, listeners, and producers to characterize music. These labels are carefully crafted descriptions of audio instrumentation, rhythmic structure, and harmonic content which forms the overall musical genre category. Music services today attempt to streamline music choices to targeted audiences, largely to the benefit of the service provider. There are opportunities to identify *music crossover* where a musical work can be classified across two or more genres, thus providing a potentially larger target audience for an artist or genre. Genre annotation is critical to this process and is unfortunately largely performed manually in the industry. Given the state of computing and data science, we believe that we can build on the current state of the art music similarity research, as well as existing machine learning models to improve music similarity search. [17]

Commented [VC1]: Remove before final draft print

Machine learning classification of musical genre can provide a semantic meta-data framework on feature extraction for a content-based analysis of any music. Most music forensics analysis has its roots in signal processing, but many of the practices are still nascent and evolving. Have you ever heard of a “forensic musicologist”? The term describes someone that analyzes music not as a singer, composer or writer, but as someone that is focused on analyzing and detecting structural similarity in music. Our research seeks to describe a framework of feature analysis via machine learning and apply it to music genre classification and music similarity. [18]

The breadth of research yet devoted to music similarity measurement is relatively finite. This is especially true when considering the proliferation and acceleration of adoption of digital music, and the increasing practice of sampling during the creative process for new works. Zhang, et al. classified research areas in music similarity as generally belonging to the categories of “Metadata-based similarity, content-based similarity, or semantic description-based similarity” [7]. Given the nascent field, we chose to restrict our research to content-based methods. This rich area provides a rich feature playground, with low-level, but highly dimensional, features to work from.

Tangential research exists that we can learn from as well. This includes studies on how chroma features and fingerprinting can be used to identify cover songs, arguably a related problem to sampling. However, few approaches specifically target building algorithmic approaches to detecting music similarity, or sampling as “a song within [another] song” [7, 8].

Problem Statement: How to evaluate, model, and analyze music similarity remains one of the more challenging and diverse research areas in audio analysis today. Through the course of this research, we consider various approaches to detecting similarity in target songs, and then utilize the rich feature space of raw

Commented [VC2]: Remove before final draft print audio, along with machine learning, to propose a unique, data-driven, approach to detecting music similarity.

The remainder of this paper is organized as follows:

- Section 2 provides an overview of how music is currently analyzed, and the associated properties that are measured,
- Section 3 covers ethics and ethical implications for this work,
- Section 4 provides a survey of techniques that are important to understand
- Section 5 covers our proposed framework and approach for music plagiarism detection,
- Section 6 describes the empirical results to our testing and experimentation,
- Section 7 suggests future avenues for study,
- Section 8 outlines lessons learned from this research, and;
- Section 9 concludes our findings.

2 Music Analysis & Properties

To understand how we intend to analyze audio for similarity, its important for us to take a moment to educate the reader to the feature-rich space that audio provides. While the terms *unstructured data* typically elicts ideas of numerous data points or text attributes, on several topics, with no obvious way to statistically analyze them, in this case we use it to definitively describe raw audio. Both the process of feature extraction, and the features themselves, deserve some review. While we will not attempt to cover the full sphere of audio analysis, we provide the below tutorial on audio analysis as a baseline starting point for the reader to further understand our work, and to appreciate the complexities for audio analysis.

2.1 Tutorial – Background Definitions

Most readers will relate to common audio concepts such as frequency and amplitude. *Frequency* represents the speed of a vibration, which thus determines *pitch*. *Amplitude*, in contrast, is the size of that vibration. So if frequency determines pitch, amplitude can be thought of deterministic to how loud that pitch is. The spectrum of human-detectable hearing ranges to 12 distinct pitches, through various octaves. This, at its core, makes up the basis of a sort of DNA for music and coupled with temporal, rhythmic and meolodic features, it is the basis of modern western music that we know today. This also provides a basic framework for understanding some of the audio feature extraction techniques we describe below.

One principle area of audio feature extraction centers on a term described as *chroma*. Chroma is representative of tonal and pitch content. In literal terms, it is the “color” of a musical pitch, decomposed such that it is octave-invariant into 12 pitch classes. Chroma features help capture musical characteristics in a condensed, potentially visual, form while “being robust to changes in timbre and instrumentation”

[14]. Chroma features are typically extracted from raw audio content utilizing Short Time Fourier Transforms, Constant Q Transforms, and normalized Chroma Energy methods, as shown in **Fig. 1. Chroma Feature Extraction Process.**

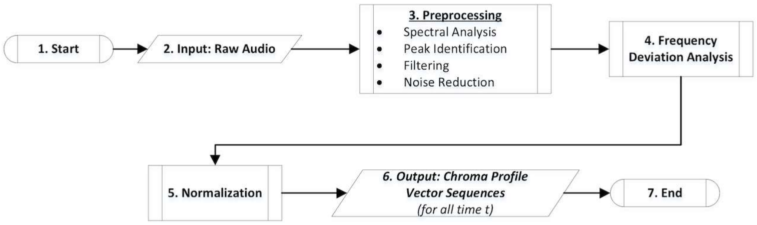

Fig. 1. Chroma Feature Extraction Process

It is logical to think that for any time *t* in a local time window of a song, it would have distinct chroma features. If we think of building these chroma feature windows across all windows for the song, we can build a representative pitch profile of the song over all 12 pitch chroma bands. When combined, this creates what is known as a *chromagram* for the song (see Fig. 2. Chromagram Example). Chromagrams are sometimes displayed as the squared magnitude of the Fourier coefficients at each section. In these cases, it is known as a *spectagram*. This spectagram is one way to densely represent a songs pitch structure.

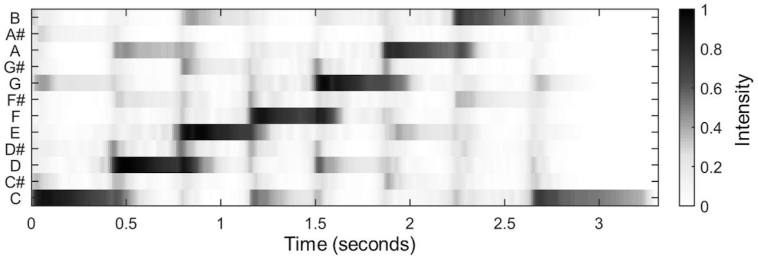

Fig. 2. Chromagram Example

Chromagrams can be thought of as one of the many building block features to audio processing. We cover it here only to instruct the reader on the type of processing that is done in order to extract features from audio. Several different pre-processing and post-processing techniques can be applied to raw audio to yield varying spectral, rhythmic, temporal, and melodic features for raw audio. Some of these subsequent features are tabulated below. While an exhaustive review of how these features are calculated is beyond the scope of this paper, we encourage the reader to review references for these feature types to learn more. For those features we utilize directly, we will explain in line with their use.

Table 1. Example Acoustic Features

Category	Category Description	Feature Space
Timbral	Tonal texture	Harmonic Pitch-Class Profile Spectral Centroid Spectral Contrast Rolloff Low-Energy Mel-frequency Cepstral Coefficient
Temporal	Time domain signals	Zero Crossing Rate Autocorrelation Waveform Moments Amplitude Modulation (loudness)
Spectral	Musical characteristics by spectra	Auto-regressive features Spectral asymmetry Kurtosis Flatness Crest factors Slope Decrease Variation Frequency derivative of Constant-Q Octave-band signal intensities
Rhythmic	Musical timing	Beat histogram Rhythm strength Regularity Average tempo
Melodic	Melodic content	Pitch histogram

Some of the more common tools used today for music and speech recognition include mel-frequency cepstral coefficients (MFCC), and spectrogram analysis. Unfortunately, both of these methods are sensitive to additive noise and pitch dynamics. As a result, it is common to see log scaling as a post-processing step. (see Fig. 3. Spectogram Log Scale Example and Fig. 4. Mel-frequency Cepstral Coefficient (MFCC) Example)

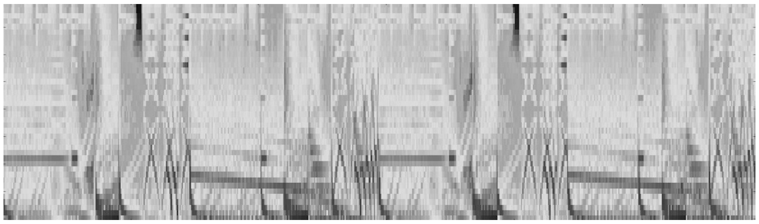

Fig. 3. Spectogram Log Scale Example

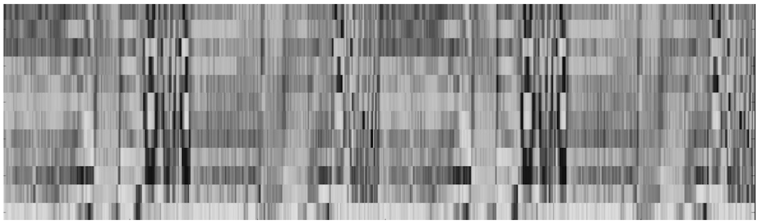

Fig. 4. Mel-frequency Cepstral Coefficient (MFCC) Example

2.2 Feature Extraction

Chroma Pitch

Chroma Log-Pitch

Chroma Energy Normalized Statistics

Spectogram

Short-Time Fourier Transform (STFT)

Constant-Q Transform

Log-Frequency Spectogram

Mel-Frequency Cepstral Coefficients (MFCC)

Log-Mel-Frequency Cepstral Coefficients (Log-Mel)

3 Ethics

There are several ethical considerations in this analysis. Obviously in music there are the general concerns over plagiarism and copyright infringement. A use or close imitation of a substantial part of another musician's work claiming one's own music without a proper credit to original artist constitutes music plagiarism. Further, the music industry attempts to restrict classifications of "sound-a-likes," making it difficult for anyone to make a claim that their song sounds like another artist. These issues complicate technological evolution since finding music similarity is, by its definition, the process of finding songs that sound like other songs.

Further, from a researcher perspective, there are ethical considerations around bias. Researchers may tend to choose known or readily available algorithms and constructs to leverage in their research, which may inherently bias music similarity to the potential end detriment to certain artists music. Plus, music is still very much a human experience. Emotion, mood, and feeling are continuums and not necessary ordinal or categorical measurements. This has direct implications for ethical considerations [14].

3. Music Industry Ethical Considerations

Music similarity is only considered plagiarism if the original work has copyright protection. However, plagiarism does not necessarily result in a copyright infringement. Worldwide, the music industry is suffering from what might be considered as plagiarism, as end-users have instantaneous, easy access to digital media and contents over the internet, and can easily sample, re-sample, and modify existing artistic works by incorporating them into new derivative works.

While there are no defined standards to evaluate music similarity or plagiarism, the intellectual property laws in the U.S. fall short of clearly and legally defining the issue. The most recent Digital Millennium Copyright Act [14] defines the statute of limitation to 70 years after the death of the creator of a musical work, which has blurred the lines between the original song and an allegedly infringed song.

Venturing into music forensics and applying machine learning techniques, one must be cautious in evaluating music similarity. Ethically speaking, there can be three possible categories for musical similarity [16]:

1. Inspiration,
2. Coincidence, and;
3. Plagiarism/Copying/Wrongful appropriations.

Inspiration is considered a legitimate element of similarity in any form of artistic production. For centuries musicians have genuinely inspired from other variants of music (e.g. folk music) as long as inspiration is limited to only fragments of artistic bytes and synthesized into a musician's own perspective to create original music.

Additionally, there are only 12 notes in Western music and therefore coincidence of musical similarity between two songs have a non-zero probability. Thankfully, there are various other dimensions to a song. In legal terms, if more dimensions within a song appear to be similar, then it should not be considered coincidence. However, these are subjective measures. Measures that may be routinized with the tests of time, but in no time soon.

When taken together in the context of music similarity, the conditions of the first two categories help prove the third. If there is the potential for inspiration, and there the probability for coincidence is low, then there is a high likelihood that plagiarism has occurred. This is when a song potentially violates copyright law. The prevalence of sampling in pop music today, makes this a particularly difficult and widespread issue in the Music Industry today. Where does inspiration start and stop? How different should a rythem or melody be to be original? These are questions that the entire industry is grappling with.

3.2 Research Ethical Considerations

In the area bias as it relates to research, the questions are different. There are certain considerations we can be safe to consider moot. For example, the idea that music similarity research results could cause physical harm to listeners of the music is not plausible under normal circumstance. However, could this research be used to economically harm certain groups or individuals? Could it disadvantage an artist or group of artists as a result? Unfortunately, without proper controls and forethought the answer would be 'yes' there is a probability where benefits to some happen at the expense of others.

Unexpected or undue harm can occur as a result of "unintentional power and bias" dynamics in the research process [14]. Due to the complexities of algorithms, and the embedded nature of this intelligence into products, it creates a *black-box* problem we must be aware of. Holzapfel, et. al claim bias come in three forms:

- **Pre-existing** – relating to existing socio-cultural norms.
- **Technical** – relating to the data available, methods, or evaluation techniques
- **Emergent** – relating to how algorithms behave when faced with new *emergent* types of data that they have never encountered before.

As researchers, we would also posit that we have our own form of *research bias* that we must consider. We obviously want to help solve a problem, get good marks from our advisors, and be well known for the quality of our work. We have potential bias to create a favorable looking body of work, and are thus motivated to make choices about datasets, about problem scope, and about our technical approach that may inherently be biased. Each of these forms of bias have the propensity to potentially impact artists and other market participants.

To address these concerns, we carefully considered our research design, validating our choices along the way with our advisors. As guidelines we evaluated these decisions through the following lense:

Table 2. Bias Containment Challenge Questions

Bias Category	Challenge Questions
Pre-existing - Cultural	Are we considering the proper breadth of data to represent cultural minority classes, under-represented groups, and emergent categories?
Technical	Does our data choice contain a representative sample for the problem scope? Is the quality of the data such that we can be reasonably assured that it will not materially impact the results?

Emergent	Have we designed tests to validate results against new types of data properly?
Researcher Judgement	Do we have proper controls to document data or technical related issues to combat error propagation? Have we properly documented "ground truth" measures?

4 Survey of Techniques

The survey of techniques content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

5 Solution Approach

The solution approach content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

5.1 Data Acquisition

The data acquisition content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

5.2 Exploratory Data Analysis

The data handling content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

5.3 Technical Approach

The technical approach content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

6 Analysis and Results

The analysis and results content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

7 Future Work

The future work content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

8 Lessons Learned

The lesson learned content will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

9 Conclusion

The conclusion will be added here. The remaining lines have been added as filler to maintain formatting rules. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum

References

1. Rentfrow, Peter J.; Goldberg, Lewis R.; Levitin, Daniel J. (1 January 2011). "The Structure of Musical Preferences: A Five-Factor Model" *Journal of Personality and Social Psychology*. 100 (6): 1139–1157.
2. Hardjono, Thomas, et al. "Towards an Open and Scalable Music Metadata Layer." arXiv preprint arXiv:1911.08278 (2019).
3. Müllensiefen, Daniel, and Marc Pendzich. "Court decisions on music plagiarism and the predictive value of similarity algorithms." Musicae Scientiae 13.1_suppl (2009): 257-295.
4. Dittmar, Christian, et al. "Audio forensics meets music information retrieval—a toolbox for inspection of music plagiarism." 2012 Proceedings of the 20th European signal processing conference (EUSIPCO). IEEE, 2012.
5. De Prisco, Roberto, et al. "Music plagiarism at a glance: metrics of similarity and visualizations." 2017 21st International Conference Information Visualisation (IV). IEEE, 2017.
6. Dredge, Stuart. "Music Industry Enters 2020 on a Wave of Growth – and Optimism." Music Industry Enters 2020 on a Wave of Growth – and Optimism, Music Ally Ltd., Dec. 2019, musically.com/2020/01/03/analysis-music-industry-2020-growth/.
7. International Federation of the Phonographic Industry. IFPI Global Music Report 2019, 2 Apr. 2019, ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2019.

8. Zhang, Bingjun, et al. “CompositeMap.” Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 09, 2009, doi:10.1145/1571941.1572011.
9. Purwins, Hendrik, et al. "Deep learning for audio signal processing." IEEE Journal of Selected Topics in Signal Processing 13.2 (2019): 206-219.
10. Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).
11. “Music.” Wikipedia, Wikimedia Foundation, 9 Feb. 2020, en.wikipedia.org/wiki/Music.
12. Stav, Iyar. "Musical plagiarism: a true challenge for the copyright law." DePaul J. Art Tech. & Intell. Prop. L 25 (2014): 1.
13. “How to Tell If a Song's Been Copied - from a Trained Musicologist - BBC Newsbeat.” BBC News, BBC, 23 Sept. 2015, www.bbc.co.uk/newsbeat/article/34282895/how-to-tell-if-a-songs-been-copied---from-a-trained-musicologist.
14. Office, U.S. Copyright. “Legislative Developments.” Copyright, www.copyright.gov/legislation/dmca.pdf.
15. Shah, Ayush & Kattel, Manasi & Nepal, Araju & Shrestha, D.. “Chroma Feature Extraction.” (2019).
16. Holzapfel, Andre, Bob Sturm, and Mark Coeckelbergh. "Ethical dimensions of music information retrieval technology." Transactions of the International Society for Music Information Retrieval 1.1 (2018): 44-55.
17. G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, July 2002.
18. Tao Feng. “Deep Learning for music genre classification”