

QUANTIFYING MUSICAL STYLE: RANKING SYMBOLIC MUSIC BASED ON SIMILARITY TO A STYLE

Jeff Ens
Simon Fraser University
jeffe@sfu.ca

Philippe Pasquier
Simon Fraser University
pasquier@sfu.ca

ABSTRACT

Modelling human perception of musical similarity is critical for the evaluation of generative music systems, musicological research, and many Music Information Retrieval tasks. Although human similarity judgments are the gold standard, computational analysis is often preferable, since results are often easier to reproduce, and computational methods are much more scalable. Moreover, computation based approaches can be calculated quickly and on demand, which is a prerequisite for use with an online system. We propose StyleRank, a method to measure the similarity between a MIDI file and an arbitrary musical style delineated by a collection of MIDI files. MIDI files are encoded using a novel set of features and an embedding is learned using Random Forests. Experimental evidence demonstrates that StyleRank is highly correlated with human perception of stylistic similarity, and that it is precise enough to rank generated samples based on their similarity to the style of a corpus. In addition, similarity can be measured with respect to a single feature, allowing specific discrepancies between generated samples and a particular musical style to be identified.

1. INTRODUCTION

Measuring musical similarity is a fundamental challenge, related to many tasks in Music Information Retrieval (MIR). In this paper, we focus on measuring the similarity between a MIDI file and an arbitrary musical style. In a musical context, the term style can refer to historical periods, composers, performers, sonic texture, emotion, and genre [8]. Here, we use the term style to denote the musical characteristics exhibited by a corpus $\mathcal{C} = \{C_1, \dots, C_n\}$, as expressed by a feature set \mathcal{F} . Depending on the contents of \mathcal{C} , style may correspond to something as specific as a subset of a composer's work, as general as the entirety of Western Classical Music, or as personal as the musical preferences of an individual.


We propose StyleRank¹, a method for ranking MIDI files based on their similarity to a style delineated by \mathcal{C} . It can be used as a tool for musicological research, to evaluate Style Imitation (SI) systems, and to filter the output of an SI system. An SI system aims to generate music that exhibits the stylistic characteristics of \mathcal{C} [27]. The primary contributions are as follows: a collection of novel features for symbolic music representation; an efficient MIDI feature extraction tool written in C++ with bindings in Python; a measure of similarity with respect to an arbitrary style delineated by \mathcal{C} ; and two experiments demonstrating that this measure is robust, and highly correlated with human perception of stylistic similarity.

2. MOTIVATIONS

There are several motivating factors for this research. In general, modelling human perception of musical similarity is of particular interest within the areas of Musicology, Music Cognition, and Music Theory [42]. Moreover, robust measures of musical similarity are critical for many MIR tasks, including database querying, music recommendation, and genre recognition. Although human perception is the gold standard for measuring musical similarity, natural human limitations place restrictions on the quantity and speed at which judgments can be collected, directly motivating automated measures of musical similarity.

More specifically, there are inherent challenges in designing a robust and reproducible listening experiment to evaluate SI systems. There are many variables which directly effect the quality of an experimental result, such as the number of participants, the listening environment, the sound equipment, and the number of samples selected for comparison. Even controlling for those variables, there is significant variability in how music is perceived, based on one's level of training [5] and musical background [13, 16, 32], which can result in a limited inter-rater agreement [34]. This is a particular issue, as it may hamper reproducibility and comparison with previously published results.

In most cases, sampling from an SI system is a stochastic process, and as a result, generated samples vary in quality. Developing a filtering process for generated material is a high priority concern, as low quality samples are undesirable when using a generative model in a production setting. Although measuring the log-likelihood of a sample can be

 © Jeff Ens, Philippe Pasquier. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Jeff Ens, Philippe Pasquier. "Quantifying Musical Style: Ranking Symbolic Music based on Similarity to a Style", 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 2019.

¹ The code is available at https://github.com/jeffreyjohnens/style_rank

useful as a proxy for quality, there are cases where log-likelihood significantly diverges from human perception. Theis et al. provide examples of generated images with high log-likelihood and extremely low quality [37]. To the best of our knowledge, there are no pre-existing methods for ranking generated samples with respect to an arbitrary style.

3. RELATED WORK

A wide variety of similarity measures have been developed to measure melodic [41], harmonic [9,25,31] and rhythmic similarity [38]. Many of these algorithms measure similarity by comparing two symbolic sequences [43]. Stylistic similarity, however, is rarely exhibited through sequence similarity, but rather through the repeated use of particular musical devices (i.e. melodic phrases, voice leading, and chord voicing) interspersed throughout the material [43]. In order to address this concern, approaches based on compression or pattern extraction have been proposed to measure similarity [2,6,21]. Since we aim to measure similarity with respect to \mathcal{C} , a more suitable approach will leverage information about the discriminative aspects of the entire corpus \mathcal{C} , rather than only taking two MIDI files into consideration.

In the context of SI system evaluation, the Turing Test [40] and the Consensual Assessment Technique [3] have been used to measure the stylistic similarity between generated artifacts $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_m\}$ and a particular style \mathcal{C} [20,28]. Objective measures have also been used to evaluate SI systems. Dong et al. measure the ratio of empty bars, pitch class diversity, note duration, rhythmic consistency, and tonal distance [10]. Trieu and Keller propose a variety of metrics ranging from rhythmic variety to harmonic consistency [39]. Since these metrics produce a single scalar value, it is easy to compare \mathcal{C} and \mathcal{G} . However, these high-level metrics are likely only capable of measuring stylistic similarity in a very general sense. Sturm and Ben-Tal. plot distributions of meter, mode, number of tokens, pitch and pitch class for \mathcal{C} and \mathcal{G} , but do not provide an automated method for analyzing discrepancies [36].

More comprehensive methodologies have been proposed, which involve computing all pairwise inter-set distances between samples in \mathcal{C} and \mathcal{G} ($D_{CG} = [\text{dist}(c, g) : (c \in \mathcal{C}) \wedge (g \in \mathcal{G})]$), as well as all pairwise intra-set distances for samples within a set ($D_{GG} = [\text{dist}(g_i, g_j) : (g_i \in \mathcal{G}) \wedge (g_j \in \mathcal{G}) \wedge (g_i \neq g_j)]$).² CAEMSI [14], a domain independent framework for the analysis of SI systems, provides a statistical method to test the null hypothesis $H_0 : (D_{GG} \neq D_{CG}) \vee (D_{CC} \neq D_{CG}) \vee (D_{CC} \neq D_{GG})$ against the alternative hypothesis $H_1 : D_{GG} = D_{CC} = D_{CG}$. Yang and Lerch extract multi-dimensional features from each MIDI file [44]. For each feature, D_{CG} and D_{CC} are constructed using Euclidean distance and smoothed using kernel density estimation [26,33]. The distance between D_{CC} and D_{CG} is measured using (1) the area of over-

lap and (2) the Kullback–Leibler Divergence [18]. In contrast to both of these approaches, which involve evaluating the similarity between \mathcal{G} and \mathcal{C} , StyleRank is optimized to evaluate the similarity of a single sample $g \in \mathcal{G}$ to \mathcal{C} .

4. FEATURES

Although the features extracted by jSymbolic2 [23] are quite comprehensive, many features are high-level, and thus, ill-suited for the fine-grained distinctions that are necessary to rank stylistically similar MIDI files. For example, the Chord Type Histogram feature contains only 11 categories. In order to capture the complexity of the musical material being analyzed, we extract a variety of high-dimensional categorical distributions from a single MIDI file. A categorical distribution is a discrete probability distribution describing a random variable that has k possible distinct states. In what follows we adopt the following notation. Given a set x , $||x||$ denotes the number of elements in the set x , $\min(x)$ and $\max(x)$ denote the minimum and maximum element in x respectively, and x_i denotes the i^{th} element in x . $x \setminus y$ is the set difference between x and y , and $x \times y$ is the Cartesian product of x and y . \ll indicates a left bitwise shift and \gg indicates a right bitwise shift. $\&$, \vee , and $|$ refer to the bitwise AND, XOR, and OR operations, respectively.

4.1 Pitch Class Set Representations

In order to reduce the number of chords, we discard octave information and represent chords as pitch class sets, using a 12-bit integer to denote the presence or absence of a particular pitch class ($C = 0, C\# = 1, \dots, B = 11$). For example, the C-major chord $\{60, 64, 67\}$ corresponds to the pitch class set $x = \{0, 4, 7\}$, which corresponds to the integer $\sum_{i=1}^{||x||} (1 \ll x_i) = 2^0 + 2^4 + 2^7 = 145$. Since there are 12 pitch classes, there are $2^{12} = 4096$ pitch class sets, which greatly reduces the possible number of chords. However, it is possible to further reduce this space if we create an equivalence class for all transpositionally equivalent pitch class sets. For example, the pitch class sets $\{0, 4, 7\}$ and $\{2, 5, 10\}$ are transpositionally equivalent, as both are major chords, the only difference being their root. This results in 352 distinct pitch class sets (PCD). Using Eq. (1c) a PCD can be calculated, where x is an 12-bit integer. Notably, pitch class sets are considered equivalent under the reversal operation when calculating the Forte number of a pitch class set [15]. Consequently, the pitch class sets $\{0, 4, 7\}$ and $\{0, 3, 7\}$ have the same Forte number, but correspond to different PCD's.

$$\text{rot}(x, n, i) = (x \ll i) | (x \gg (n-i)) \& (2^n - 1) \quad (1a)$$

$$\text{reduce}(x, n) = \min(\{\text{rot}(x, n, i) : 0 \leq i < n\}) \quad (1b)$$

$$\text{pcd}(x) = \text{reduce}(x, 12) \quad (1c)$$

Alternatively, a pitch class set x can be represented as the set of scales which are supersets of x . Given a scale \mathbb{S} , let $\mathbb{S}_i = \{(s + i) \bmod 12 : s \in \mathbb{S}\}$. The scale representation can be calculated with Eq. (2), where

² Note that we adapt the set-builder notation to construct a list (e.g., $[i/2 : 0 \leq i < 4] = [0, 0, 1, 1]$), which unlike a set, may contain duplicate values.

$\mathbb{S}^M = \{0, 2, 4, 5, 7, 9, 11\}$ and $\mathbb{S}^H = \{0, 2, 3, 5, 7, 8, 11\}$ denote the major and harmonic minor scales respectively. $\phi(\cdot)$ returns 1 if the predicate \cdot is true and 0 otherwise.

$$sc(x) = \left(\sum_{i=1}^{12} \phi(x \subseteq \mathbb{S}_i^M) \ll i \right) + \left(\sum_{i=1}^{12} \phi(x \subseteq \mathbb{S}_i^H) \ll (12+i) \right) \quad (2)$$

4.2 Feature Definitions

Given a MIDI file M , for each note $n \in M$, $ons(n)$ returns the onset time of n in ticks, $dur(n)$ returns the duration of n in ticks, and $pitch(n)$ returns the pitch. An ordered set containing the unique onsets $O = \{ons(n) : n \in M\}$ is constructed, and the i^{th} chord is the set of notes $\mathbb{C}^i = \{n : (ons(n) \leq O_i) \wedge (ons(n) + dur(n) > O_i)\}$. $isOns(\mathbb{C}, n)$ and $isTie(\mathbb{C}, n)$ are functions that return 1 if n is an onset or a tie respectively, and 0 otherwise. The function $pc_i(\mathbb{C}, n)$ returns 1 if n corresponds to the pitch class i and 0 otherwise. In order to simplify the feature definitions, we use Eq. (3d), which accepts a chord \mathbb{C} and a set of functions F , and only returns 1 if there is an element in X for which each $f \in F$ evaluates to 1. As a result, $\mathbf{I}(\mathbb{C}, \{isOns, pc_i\})$ is 1 if there is a note $n \in \mathbb{C}$ that is an onset and is equivalent to the pitch class i .

$$pc_i(\mathbb{C}, n) = \begin{cases} 0, & \text{if } pitch(n) \bmod 12 \equiv i \\ 1, & \text{otherwise} \end{cases} \quad (3a)$$

$$isOns(\mathbb{C}, n) = \begin{cases} 0, & \text{if } \max(\{ons(n) : n \in \mathbb{C}\}) > ons(n) \\ 1, & \text{otherwise} \end{cases} \quad (3b)$$

$$isTie(\mathbb{C}, n) = 1 - isOns(\mathbb{C}, n) \quad (3c)$$

$$\mathbf{I}(\mathbb{C}, F) = \begin{cases} 0, & \text{if } \max(\{\prod_{i=1}^{|F|} F_i(\mathbb{C}, n) : n \in \mathbb{C}\}) < 1 \\ 1, & \text{otherwise} \end{cases} \quad (3d)$$

Table 1 provides formal definitions of all the features, where \mathbb{C}^t denotes the t^{th} chord, \mathbb{M}^t denotes the t^{th} melody pitch, $\mathbb{P}^t = \{pitch(n) : n \in \mathbb{C}^t\}$, $\mathbb{O}^t = \{ons(n) : n \in \mathbb{C}^t\}$, and $\mathbb{K}^t = \{pitch(n) : (n \in \mathbb{C}^t) \wedge isOns(n)\}$. $popcount(\cdot)$ is a function that counts the number of set bits in an integer, $pc(x) = x \bmod 12$ and $popcount(x) = |(x \bmod 12) - 6|$. Dissonance is calculated using Stolzenburg’s periodicity function [35], which we refer to as $stol(\cdot)$. Let $diss(\mathbb{P}, \mathbb{T}) = \frac{1}{|\mathbb{T}|} \sum_{x \in \mathbb{T}} stol(\mathbb{P}^x)$, where \mathbb{P} and \mathbb{T} are pitch sets, and $\mathbb{P}^x = \{\mathbb{P}_i - x : \mathbb{P}_i \in \mathbb{P}\}$. $voiceMotion(\cdot)$ is a function that accepts two successive pitch sets ($\mathbb{P}^t, \mathbb{P}^{t+1}$) and returns an integer corresponding to the type of voice motion. $tonnetzLength(\cdot)$ is a function that accepts a pitch class set and returns the length of the shortest path through Tonnetz [24] vertices containing each pitch class.

Each function is calculated for all valid values of t , resulting in a categorical distribution with unsigned 64-bit integers as the categories. For example, given a standard 4-voice Bach chorale containing m chords, the function $ChordSize$ is calculated for $0 \leq t < m - 2$, producing

a categorical distribution with the categories $\{0, 1, 2, 3, 4\}$. In some cases, we weight values by chord duration, denoted by a \star in the table. In the case that a function returns a set of values (`IntervalDist`), we combine the returned sets to form the categorical distribution. Since the number of categories k grows exponentially large for some features (e.g., `ChordShape`), we restrict $k \leq 1000$ by ranking categories according to the number of samples they appear in, removing infrequently occurring categories.

4.3 Implementation

We implement the feature extraction tool in C++, using `pybind11` [17] to create Python bindings. The `Midifile` library³ is used to parse MIDI files.

5. SIMILARITY COMPUTATION

In the most general sense, we are interested in measuring the similarity between a single MIDI file \mathcal{X} and a corpus $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_n\}$. We represent each MIDI file by applying a non-empty set of feature transformations $\mathcal{F} = \{f_1, \dots, f_k\}$, producing a set of categorical distributions for each MIDI file. For each $f_i \in \mathcal{F}$, we aim to measure the similarity between a single categorical distribution $f_i(\mathcal{X})$ and a set of categorical distributions $f_i(\mathcal{C}) = \{f_i(\mathcal{C}_1), \dots, f_i(\mathcal{C}_n)\}$. Using a distance metric \mathcal{D} , the average similarity could be calculated $\frac{1}{n} \sum_{i=1}^n 1 - \mathcal{D}(f_i(\mathcal{C}_i), f_i(\mathcal{X}))$. However, this approach does not leverage information about the discriminative aspects of the entire corpus. The results in Experiment 1 demonstrate the deficiencies of this approach. Instead, we use Random Forests [7] to construct an embedding space before measuring the average similarity. Although neural networks are often ideal for learning embeddings, the time required to train k neural networks is prohibitive for an online system.

Decision trees are commonly used to model complex data. When used to classify data, each terminal node represents a discrete class label, and an arbitrary input is classified based on the terminal node it reaches. Using a trained Random Forest, an input can be represented based on the terminal node it reaches in each decision tree. Given a Random Forest containing N decision trees each with L terminal nodes, an input can be represented as a vector $v \in \{0, 1\}^{N \times L}$. To learn an embedding for a single feature transformation $f_i \in \mathcal{F}$, we train a Random Forest to discriminate between a collection of items $f_i(\mathcal{G}) = \{f_i(\mathcal{G}_1), \dots, f_i(\mathcal{G}_m)\}$ and a corpus $f_i(\mathcal{C}) = \{f_i(\mathcal{C}_1), \dots, f_i(\mathcal{C}_n)\}$. Concretely, each $f_i(\mathcal{G}_i) \in f_i(\mathcal{G})$ is given the label 0, and each $f_i(\mathcal{C}_i) \in f_i(\mathcal{C})$ is given the label 1. We refer to the vector produced for a sample \mathcal{X} as $\mathbf{R}_{\mathcal{X}}^{\mathcal{G}, \mathcal{C}, f_i}$. Breiman measures the similarity of two vectors using the dot product [7]. In order to weight each feature transformation ($f_i \in \mathcal{F}$) equally, we use cosine similarity (Eq. (4a)), which is simply the normalized dot product. The similarity between \mathcal{X} and \mathcal{C} with respect to a set of

³ <https://midifile.sapp.org/>

Feature Name	Function	Description
ChordDissonance *	$[\text{diss}(\mathbb{K}^t, \mathbb{K}^t)]$	the dissonance of onsets based on periodicity [35]
ChordDistinctDurationRatio	$(1 \ll \ \{\text{dur}(n) : n \in \mathbb{C}^t\}\) \mid 2^{\ \mathbb{C}^t\ }$	the ratio of distinct note durations to chord size
ChordDuration	$\max(\mathbb{O}^{t+1}) - \max(\mathbb{O}^t)$	the duration of a chord
ChordLowestInterval	$\min(\mathbb{P}^t \setminus \{\min(\mathbb{P}^t)\}) - \min(\mathbb{P}^t)$	the difference between the lowest two notes
ChordOnset	$(\sum_{i=1}^{\ \mathbb{C}^t\ } (\text{isOns}(\mathbb{C}_i^t) \ll (i-1))) \mid 2^{\ \mathbb{C}^t\ }$	an integer representing which notes are onsets
ChordOnsetPCD *	$\text{pcd}(\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{isOns}, \text{pc}_i\}) \ll i))$	distinct pitch class set excluding ties
ChordOnsetRatio	$(1 \ll \sum_{n \in \mathbb{C}^t} \text{isOns}(n)) \mid 2^{\ \mathbb{C}^t\ }$	the ratio of onsets to chord size
ChordOnsetShape *	$\sum_{i=1}^{\ \mathbb{C}^t\ } (\text{isOns}(\mathbb{C}^t, \mathbb{C}_i^t) \ll (\mathbb{P}^t - \min(\mathbb{P}^t)))$	piano roll type representation of onset pitches
ChordOnsetTiePCD *	$\text{pcd}(\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{isOns}, \text{pc}_i\}) \ll i)) + \text{pcd}(\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{isTie}, \text{pc}_i\}) \ll i)) \ll 12$	concatenated distinct pitch class set of onsets and distinct pitch class set of ties
ChordOnsetTieReduced *	$\text{reduce}((\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{isOns}, \text{pc}_i\}) \ll i)) + (\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{isTie}, \text{pc}_i\}) \ll (12+i))))$	concatenated pitch class set of onsets and pitch class set of ties reduced using Eq. (1b)
ChordPCD *	$\text{pcd}(\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{pc}_i\}) \ll i))$	distinct pitch class set
ChordPCDWBass *	$\text{pcd}(\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{pc}_i\}) \ll i)) + 2^{12+\text{pc}(\min(\mathbb{P}^t))}$	distinct pitch class set with bass pitch class
ChordPCSizeRatio	$(1 \ll \ \{\text{pc}(p) : p \in \mathbb{P}^t\}\) \mid 2^{\ \mathbb{P}^t\ }$	the ratio of distinct pitch classes to chord size
ChordRange (ϕ_1)	$\max(\mathbb{P}^t) - \min(\mathbb{P}^t)$	the range of pitches in a chord
ChordShape *	$\sum_{p \in \mathbb{P}^t} (1 \ll (p - \min(\mathbb{P}^t)))$	piano roll type representation of chord pitches
ChordSize	$\ \mathbb{C}^t\ $	the number of notes in a chord
ChordTonnetz *	$\text{tonnetzLength}(\{\text{pc}(x) : x \in \mathbb{P}^t\})$	length of shortest path through Tonnetz [24] vertices
ChordSizeNgram	$\ \mathbb{C}^t\ + (\ \mathbb{C}^{t+1}\ \ll 8) + (\ \mathbb{C}^{t+2}\ \ll 16)$	an n -gram of chord sizes ($n = 3$)
ChordTranBassInterval	$\text{pc}(\min(\mathbb{P}^{t+1}) - \min(\mathbb{P}^t))$	pitch class interval between two lowest notes
ChordTranDissonance	$[\text{diss}(\mathbb{P}^t, \mathbb{P}^{t+1})]$	the dissonance of intervals based on periodicity [35]
ChordTranDistance	$ \min(\mathbb{P}^{t+1}) - \min(\mathbb{P}^t) + \max(\mathbb{P}^{t+1}) - \max(\mathbb{P}^t) $	approximated voice leading distance
ChordTranOuter	$\text{pc}(\phi_1(\mathbb{P}^t)) + (\text{pc}(\phi_1(\mathbb{P}^{t+1})) \ll 8) + (\text{pc}(\min(\mathbb{P}^t) - \min(\mathbb{P}^{t+1})) \ll 16)$	pitch class transition using only the outer notes
ChordTranPCD	$\text{reduce}((\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^t, \{\text{pc}_i\}) \ll i)) + (\sum_{i=0}^{11} (\mathbf{I}(\mathbb{C}^{t+1}, \{\text{pc}_i\}) \ll (12+i))), 24)$	transition between distinct pitch class sets
ChordTranRepeat	$(\prod_{n \in \mathbb{C}^t} \text{isOns}(n)) (\mathbb{P}^t = \mathbb{P}^{t+1})$	chord repetition with onsets
ChordTranScaleDistance	$\text{popcount}(\text{sc}(\mathbb{P}^t) \vee \text{sc}(\mathbb{P}^{t+1}))$	hamming distance between scale representations
ChordTranScaleUnion	$\text{popcount}(\text{sc}(\mathbb{P}^t) \mid \text{sc}(\mathbb{P}^{t+1}))$	the union between scale representations
ChordTranVoiceMotion	$\text{voiceMotion}(\mathbb{P}^t, \mathbb{P}^{t+1})$	type of voice motion (contrary, oblique, etc.)
Mel.		
MelodyNgram	$\sum_{i=0}^3 (\mathbb{M}_{t+i+1} - \mathbb{M}_{t+i} \bmod 12) \ll 8i$	n -gram of melodic intervals ($n = 3$)
MelodyPCD	$\text{pcd}(\sum_{i=0}^{11} (\mathbf{I}(\{\mathbb{M}^{t+i} : 0 \leq i < 5\}, \{\text{pc}_i\}) \ll i))$	distinct pitch class of successive melody notes
Inter.		
IntervalClassDist	$\{\text{pcc}(p_i - p_j) : (p_j < p_i) \wedge (p_i, p_j \in \mathbb{P}^t \times \mathbb{P}^t)\}$	interval class for each combination of chord pitches
IntervalDist	$\{\text{pc}(p_i - p_j) : (p_j < p_i) \wedge (p_i, p_j \in \mathbb{P}^t \times \mathbb{P}^t)\}$	interval for each combination of chord pitches

Table 1. Definitions for Chord features, Chord Transition features, Melody features (Mel.), and Interval features (Inter.). The * symbol indicates that a categorical distribution is weighted by chord duration.

features \mathcal{F} is computed using Eq. (4b), which produces a scalar value on the range $[0, 1]$.

$$\cos(X, Y) = \frac{X \cdot Y}{\sqrt{\sum_{i=1}^N X_i^2} \sqrt{\sum_{i=1}^N Y_i^2}} \quad (4a)$$

$$S_{\mathcal{X}}^{\mathcal{G}, \mathcal{C}, \mathcal{F}} = \frac{1}{\|\mathcal{C}\| \|\mathcal{F}\|} \sum_{c \in \mathcal{C}} \sum_{f \in \mathcal{F}} \cos(\mathbf{R}_{\mathcal{X}}^{\mathcal{G}, \mathcal{C}, f}, \mathbf{R}_{\mathcal{C}}^{\mathcal{G}, \mathcal{C}, f}) \quad (4b)$$

6. EXPERIMENTS

In the following experiments, we train a Random Forest [7] using the scikit-learn python module [30]. We set the maximum tree depth at 5, the number of trees to 500, and measure the quality of the split using entropy. The class weight is balanced to be robust against size discrepancies between \mathcal{C} and \mathcal{G} .

6.1 Experiment 1 : Analytic Testing

We test StyleRank with styles delineated by a single composer, and by an entire genre, using the Classical Archives MIDI dataset⁴. In total there are 75 composers, and 6 musical genres. More details on the composition of the dataset can be found in the Appendix⁵. We keep only one MIDI file per composition. Each MIDI file is represented as a list of pitches, sorted lexicographically according to onset and pitch. To compare two pieces, the Levenshtein distance [19] is measured twice, once for the first 100 pitches in each piece, and once for the last 100 pitches. We eliminate pieces which have a Levenshtein distance less than 0.75, after normalizing the distance on the range $[0, 1]$. We choose this conservative value to ensure all duplicates are removed.

Given two styles $A = \{a_1, \dots, a_m\}$ and $B = \{b_1, \dots, b_n\}$, where $m = 2n$, let $\mathcal{C} = \{a_i : 1 \leq i \leq n\}$,

⁴ <https://www.classicalarchives.com/midi.html>

⁵ https://github.com/jeffreyjohnens/style_rank/tree/master/appendix

$\mathcal{G}_A = \{a_i : n < i \leq 2n\}$, $\mathcal{G}_B = B$, and $\mathcal{G} = \mathcal{G}_A \cup \mathcal{G}_B$. By construction $\mathcal{G} \cap \mathcal{C} = \emptyset$. We train a Random Forest and compare two distributions $x = [S_g^{\mathcal{G}, \mathcal{C}, \mathcal{F}} : g \in \mathcal{G}_A]$ and $y = [S_g^{\mathcal{G}, \mathcal{C}, \mathcal{F}} : g \in \mathcal{G}_B]$, where \mathcal{F} denotes the set of features described in Table 1. Ideally, each value in x should be larger than all values in y , since elements in \mathcal{G}_A and \mathcal{C} belong to the same style (A). However, depending on the specificity of the style, there may be some degree of overlap between A and B . In order to determine if there is a measurable difference between x and y we directly compare the means ($\bar{x} > \bar{y}$), and we calculate the p -value ($p^{\bar{x} > \bar{y}}$) for a One-Sided Mann-Whitney test [22] with the alternative hypothesis that $\bar{x} > \bar{y}$.

In cases where multiple statistical comparisons are performed, it is common practice to apply a correction to the raw p -values. The Bonferroni correction [11] is calculated by dividing the desired level of significance ($\alpha = 0.05$) by the number of comparisons. The Benjamini–Yekutieli procedure [4] controls the false discovery rate under arbitrary dependence assumptions, and is less conservative than the Bonferroni correction. Given m null hypotheses and their corresponding p -values P_1, \dots, P_m , the p -values are sorted in ascending order. For a given level of significance, in our case $\alpha = 0.05$, reject the null hypothesis for the first k values that satisfy $P_k \leq k\alpha / (m * c(m))$ where $c(m) = \sum_{i=1}^m 1/i$.

Table 2 shows the results of 1000 trials, reporting the percentage of trials where $\bar{x} > \bar{y}$, and the percentage of trials where $p^{\bar{x} > \bar{y}}$ is significant, applying no correction ($\alpha = 0.05$), the Benjamini–Yekutieli procedure (FDR), and the Bonferroni correction (Bon). We compare StyleRank against three distance measures, Cosine, Manhattan and Euclidean, replacing $S_g^{\mathcal{G}, \mathcal{C}, \mathcal{F}}$ with $\frac{1}{\|\mathcal{C}\| \|\mathcal{F}\|} \sum_{c \in \mathcal{C}} \sum_{f \in \mathcal{F}} 1 - \mathcal{D}(f(c), f(g))$.

6.2 Experiment 2: Congruity with Human Perception

In order to evaluate how well StyleRank correlates with human perception, we use data from the BachBot [20] experiment. In total, there were 5,967 participants, including 1329 novices, 2786 intermediate, 1341 advanced and 511 experts. Liang et al. generated 36 samples (\mathcal{G}) from a neural network trained on a collection of Bach Chorales (\mathcal{C}). Participants were asked to discriminate between a generated musical excerpt and an actual Bach chorale. They were each asked to complete 5 comparisons.

For each $g \in \mathcal{G}$, we count the number of times it was mistakenly classified as a Bach chorale N_g^{miss} , and the number of times it was correctly identified as computer generated N_g^{corr} . The raw count data can be found in the Appendix. We take the relative frequency of misclassifications $T_g = N_g^{\text{miss}} / (N_g^{\text{miss}} + N_g^{\text{corr}})$ as an indication of how similar g is to the style of Bach’s Chorales (\mathcal{C}). This results in $\binom{36}{2} = 630$ pairwise comparisons for which we have a ground truth ranking. Using a chi-square contingency test [29] we can measure the degree to which we are certain that there is a difference between two samples. We measure accuracy using Eq. (5b), where p_{ij} is the p -value for the chi-square contingency test comparing the counts

for the i^{th} and j^{th} examples, $\phi(\cdot)$ is a function returning 1 if the predicate \cdot is true and 0 otherwise, and α denotes the threshold for significance.

$$f(x, y) = \begin{cases} 1, & \text{if } \phi(S_x^{\mathcal{G}, \mathcal{C}, \mathcal{F}} < S_y^{\mathcal{G}, \mathcal{C}, \mathcal{F}}) = \phi(T_x < T_y) \\ 0, & \text{otherwise} \end{cases} \quad (5a)$$

$$\text{acc}(\mathcal{G}, \mathcal{C}, \alpha) = \frac{\sum_{i=1}^{|\mathcal{G}|} \sum_{j=i+1}^{|\mathcal{G}|} f(\mathcal{G}_i, \mathcal{G}_j) \phi(p_{ij} < \alpha)}{\sum_{i=1}^{|\mathcal{G}|} \sum_{j=i+1}^{|\mathcal{G}|} \phi(p_{ij} < \alpha)} \quad (5b)$$

The results for Experiment 2 are presented in Table 3. We report the accuracy, calculated using Eq. (5b), for a random ranking (Random), StyleRank with the jSymbolic [23] features (jSymbolic), Log-likelihood (Loglik), and StyleRank. All the default features are extracted using jSymbolic, and features with zero standard deviation are removed. This results in a single feature vector with dimension of 453, for which we train a single Random Forest. Using the Performance RNN [12], which was trained with the same representation and data as the original BachBot, we evaluate the negative log-likelihood \mathcal{L}_g of each of the generated examples (loglik). To calculate the accuracy we simply replace the term $S_X^{\mathcal{G}, \mathcal{C}, \mathcal{F}} < S_Y^{\mathcal{G}, \mathcal{C}, \mathcal{F}}$ with $\mathcal{L}_X < \mathcal{L}_Y$ in Eq. (5a).

7. DISCUSSION

Collectively, the results of both experiments demonstrate that StyleRank is robust to corpora of varying sizes, and highly correlated with human perception of stylistic similarity. In the Appendix, we expand Experiment 1 to demonstrate that StyleRank’s performance is robust, even when the number of distinct styles in \mathcal{G} is increased. In Experiment 1, there is a large difference between raw distance measures and StyleRank. This highlights the limitations of the approach described by Yang and Lerch, which uses euclidean distance to measure the distance between feature vectors [44]. Although euclidean distance works well in low-dimensional settings, it does not scale well to high dimensions. In fact, it has been shown that Manhattan distance performs better than Euclidean distance in high dimensional settings [1], which we also see in our own experimental results. Understandably, there is a decrease in performance when analyzing styles delineated by genre, as these styles have more variance, and are less consistent than the work of a single composer. Overall, these results demonstrate that StyleRank can proficiently rank MIDI files with different styles.

The results for Experiment 2 demonstrate that StyleRank is capable of making fine-grained distinctions between MIDI files that correspond with human perception of stylistic similarity. It is worth noting that participants found it difficult to discriminate between generated and human-composed samples in the BachBot experiment, evidenced by the average classification accuracy of novice (0.57), intermediate (0.64), advanced (0.68), and expert

size	StyleRank				Cosine				Manhattan				Euclidean				
	μ	Sig	FDR	Bon	μ	Sig	FDR	Bon	μ	Sig	FDR	Bon	μ	Sig	FDR	Bon	
Composer	10	0.963	0.86	0.725	0.0	0.837	0.624	0.381	0.0	0.879	0.662	0.413	0.0	0.827	0.565	0.28	0.0
	25	0.951	0.888	0.807	0.609	0.808	0.583	0.422	0.24	0.793	0.578	0.415	0.244	0.729	0.532	0.363	0.226
	50	0.926	0.905	0.873	0.78	0.705	0.559	0.454	0.333	0.751	0.599	0.468	0.34	0.717	0.565	0.428	0.3
	100	1.0	0.986	0.973	0.951	0.713	0.636	0.59	0.515	0.723	0.633	0.568	0.486	0.715	0.626	0.571	0.504
Genre	10	0.81	0.379	0.0	0.0	0.68	0.193	0.0	0.0	0.686	0.2	0.0	0.0	0.645	0.176	0.0	0.0
	25	0.867	0.578	0.376	0.198	0.729	0.348	0.084	0.038	0.74	0.374	0.053	0.021	0.691	0.298	0.06	0.022
	50	0.88	0.715	0.59	0.432	0.776	0.484	0.266	0.126	0.747	0.489	0.253	0.088	0.714	0.344	0.158	0.082
	100	0.927	0.847	0.774	0.671	0.766	0.555	0.406	0.265	0.755	0.566	0.44	0.284	0.785	0.462	0.269	0.178

Table 2. The normalized frequency over 1000 trials where $\bar{x} > \bar{y}$ (μ), $p^{\bar{x} > \bar{y}} < 0.05$ (Sig), $p^{\bar{x} > \bar{y}}$ is significant after applying the FDR correction (FDR), and $p^{\bar{x} > \bar{y}}$ is significant after applying the Bonferonni correction (Bon). Size denotes the size of the corpus $||\mathcal{C}|| = ||\mathcal{G}_A|| = ||\mathcal{G}_B||$.

	Novice				Intermediate			
	$\alpha = 5.0$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.005$	$\alpha = 5.0$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.005$
Random	.482 ± .025	.479 ± .031	.466 ± .044	.440 ± .062	.500 ± .023	.500 ± .026	.502 ± .033	.499 ± .037
jSymbolic	.471 ± .006	.463 ± .008	.472 ± .012	.491 ± .015	.478 ± .011	.474 ± .013	.467 ± .014	.456 ± .017
Loglik	.629 ± .000	.669 ± .000	.764 ± .000	.817 ± .000	.654 ± .000	.668 ± .000	.690 ± .000	.732 ± .000
StyleRank	.716 ± .001	.774 ± .002	.855 ± .004	.899 ± .005	.702 ± .002	.715 ± .002	.758 ± .002	.808 ± .002
	Advanced				Expert			
Random	.511 ± .010	.514 ± .013	.512 ± .017	.515 ± .019	.493 ± .019	.492 ± .025	.492 ± .032	.485 ± .038
jSymbolic	.481 ± .011	.480 ± .011	.470 ± .014	.474 ± .013	.452 ± .008	.449 ± .009	.482 ± .012	.464 ± .013
Loglik	.673 ± .000	.694 ± .000	.730 ± .000	.724 ± .000	.657 ± .000	.692 ± .000	.741 ± .000	.800 ± .000
StyleRank	.718 ± .001	.756 ± .001	.806 ± .002	.808 ± .002	.692 ± .002	.745 ± .003	.821 ± .004	.881 ± .005

Table 3. The accuracy of each model, calculated using Eq. (5b), with standard error calculated over 10 trials.

(0.71) participants [20]. Based on our experimental results, the jSymbolic [23] feature set is no better at predicting rankings than a random model. This is likely due to the fact that high level features are not sufficiently discriminative for this task. In contrast to the jSymbolic feature set, our method involves full categorical distributions, which we believe are critical in measuring fine-grained differences. Importantly, there is a substantial difference between the accuracy of rankings based on log-likelihood and StyleRank. Interestingly, both log-likelihood and StyleRank best model high certainty ($\alpha = 0.005$) comparisons made by self identified novices. This may be an artifact of increased variance as the number of ground truth comparisons decreases as α increases.

It should be noted that participants in the BachBot experiment were not directly asked to rank samples according to their similarity to the style of Bach’s chorales. We extrapolated a ranking from the number of times a sample was miss-classified, which is an indirect way of measuring stylistic similarity. However, since these rankings were based on a large sample size, we are confident that they are reflective of human perception.

8. APPLICATION

StyleRank can be used in a variety of settings. Importantly, we must note that there are no limitations on the composition of \mathcal{G} . For example, one could compare k different sets with $\mathcal{G} = \{\mathcal{G}_i^1, \dots, \mathcal{G}_{n_1}^1, \mathcal{G}_1^2, \dots, \mathcal{G}_{n_2}^2, \dots, \mathcal{G}_1^k, \dots, \mathcal{G}_{n_k}^k\}$. First of all, the method can be used to rank samples generated by an SI system, based on their similarity to \mathcal{C} . StyleR-

ank can be used to filter highly dissimilar samples automatically. Filtering is as simple as taking the samples $g \in \mathcal{G}$ with a similarity $S_g^{\mathcal{G}, \mathcal{C}, \mathcal{F}}$ above some threshold, and discarding the rest. Secondly, StyleRank can be used to rank models. Given k models, let $\mathcal{G} = \{\mathcal{G}^1, \dots, \mathcal{G}^k\} = \{\mathcal{G}_i^1, \dots, \mathcal{G}_{n_1}^1, \dots, \mathcal{G}_1^k, \dots, \mathcal{G}_{n_k}^k\}$, where \mathcal{G}^i denotes the set of samples generated by the i^{th} model. Then the distributions $x_i = [S_g^{\mathcal{G}, \mathcal{C}, \mathcal{F}} : g \in \mathcal{G}^i]$ can be compared using an appropriate statistical test. Third, the method can be used to isolate the specific features f that deviate from the style delineated by \mathcal{C} by comparing the distributions $x_f = [S_g^{\mathcal{G}, \mathcal{C}, f} : g \in \mathcal{G}]$ for each f in a set of features \mathcal{F} . In addition, StyleRank can be used as a tool for musicologists to explore variations in style.

9. CONCLUSION

Quantifying musical stylistic similarity is a difficult task. We propose StyleRank, a method to rank individual MIDI files based on their similarity to an arbitrary style. Experimental evidence supports our approach, demonstrating that our method is robust, and is highly correlated with human perception of stylistic similarity. Future work involves applying this approach to other domains where SI systems are being developed. Additional features can be added to the current collection, in particular rhythm-based features, as the current collection is pitch-centric. Although we believe our experiments to be fairly comprehensive, continued validation of the proposed method on additional data is always beneficial.

10. ACKNOWLEDGMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Helmut Hugo Eppich Family Graduate Scholarship.

11. REFERENCES

- [1] C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434, 2001.
- [2] T.E. Ahonen, K. Lemström, and S. Linkola. Compression-based similarity measures in symbolic, polyphonic music. In *ISMIR*, pages 91–96, 2011.
- [3] T. M. Amabile. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013, 1982.
- [4] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- [5] M. Besson, D. Schön, S. Moreno, A. Santos, and C. Magne. Influence of musical expertise and musical training on pitch processing in music and language. *Restorative Neurology and Neuroscience*, 25(3-4):399–410, 2007.
- [6] P. Boot, A. Volk, and W.B. de Haas. Evaluating the role of repeated patterns in folk song classification and compression. *Journal of New Music Research*, 45(3):223–238, 2016.
- [7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [8] R.B. Dannenberg. *Style in Music*, pages 45–57. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
- [9] W.B. de Haas, F. Wiering, and R.C. Veltkamp. A geometrical distance measure for determining the similarity of musical harmony. *International Journal of Multimedia Information Retrieval*, 2(3):189–202, Sep 2013.
- [10] H.W. Dong, W.Y. Hsiao, L.C. Yang, and Y.H. Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *32nd AAAI Conference on Artificial Intelligence*, pages 34–41, 2018.
- [11] O.J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.
- [12] D. Eck, A. Roberts, J. Engel, C. Hawthorne, and I. Simon. Magenta, 2019. <https://github.com/tensorflow/magenta>.
- [13] T. Eerola, T. Himberg, P. Toiviainen, and J. Louhivuori. Perceived complexity of western and african folk melodies by western and african listeners. *Psychology of Music*, 34(3):337–371, 2006.
- [14] J. Ens and P. Pasquier. Caemsi : A cross-domain analytic evaluation methodology for style imitation. In *International Conference on Computational Creativity*, pages 64–71, 2018.
- [15] A. Forte. *The Structure of Atonal Music*. Yale [paperbacks]. Yale University Press, 1973.
- [16] E.E. Hannon and S.E. Trehub. Metrical categories in infancy and adulthood. *Psychological Science*, 16(1):48–55, 2005.
- [17] W. Jakob, J. Rhineland, and D. Moldovan. pybind11 – seamless operability between c++11 and python, 2017. <https://github.com/pybind/pybind11>.
- [18] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [19] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, 10(8):707–710, 1966.
- [20] F. T. Liang, M. Gotham, M. Johnson, and J. Shotton. Automatic stylistic composition of bach chorales with deep lstm. In *Proc. of the International Symp. on Music Information Retrieval*, pages 449–456, 2017.
- [21] N.H. Liu, Y.H. Wu, and A.L. Chen. Efficient knn search in polyphonic music databases using a lower bounding mechanism. *Multimedia systems*, 10(6):513–528, 2005.
- [22] H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.
- [23] C. McKay, J. Cumming, and I. Fujinaga. jsymbolic 2.2: Extracting features from symbolic music for use in musicological and mir research. In *Proc. of the International Symp. on Music Information Retrieval*, 2018.
- [24] A.V. Oettingen. *Harmoniesystem in dualer Entwicklung*. W. Glaser, Dorpat, 1866.
- [25] J.F. Paiement, D. Eck, and S. Bengio. A probabilistic model for chord progressions. In *Proc. of the International Symp. on Music Information Retrieval*, pages 11–15, 2005.
- [26] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.
- [27] P. Pasquier, A. Eigenfeldt, O. Bown, and S. Dubnov. An introduction to musical metacreation. *Computer Entertainment*, 14(2):3–17, 2017.

- [28] M. T. Pearce and G. A. Wiggins. Evaluating cognitive models of musical composition. In *Proceedings of the 4th international joint workshop on computational creativity*, pages 73–80. Goldsmiths, University of London, 2007.
- [29] K. Pearson. *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling*, pages 11–28. Springer New York, New York, NY, 1992.
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [31] J. Pickens and T. Crawford. Harmonic models for polyphonic music retrieval. In *Proc. of the International Conference on Information and Knowledge Management*, pages 430–437, 2002.
- [32] J.B. Prince, M.A. Schmuckler, and W.F. Thompson. The effect of task and pitch structure on pitch-time interactions in music. *Memory & Cognition*, 37(3):368–381, 2009.
- [33] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.
- [34] M. Schedl, A. Flexer, and J. Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.
- [35] F. Stolzenburg. Harmony perception by periodicity detection. *Journal of Mathematics and Music*, 9(3):215–238, 2015.
- [36] B. L. Sturm and O. Ben-Tal. Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *Journal of Creative Music Systems*, 2(1):1–29, 2017.
- [37] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016. arXiv:1511.01844.
- [38] G.T. Toussaint. A comparison of rhythmic similarity measures. In *Proc. of the International Symp. on Music Information Retrieval*, pages 242–245, 2004.
- [39] N. Trieu and R. Keller. Jazzgan: Improvising with generative adversarial networks. In *6th International Workshop on Musical Metacreation*, 2018.
- [40] A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.
- [41] V. Velardo, M. Vallati, and S. Jan. Symbolic melodic similarity: State of the art and future challenges. *Computer Music Journal*, 40(2):70–83, 2016.
- [42] A. Volk, E. Chew, Elizabeth Hellmuth M., and C. Anagnostopoulou. Music similarity: Concepts, cognition and computation, 2016.
- [43] A. Volk, W.B. de Haas, and P. Kranenburg. Towards modelling variation in music as foundation for similarity. In *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*. School of Music Studies, Aristotle University of Thessaloniki, 2012.
- [44] L.C. Yang and A. Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 1:1–12, Nov 2018.