A Machine Learning Approach to Musically Meaningful Homogeneous Style Classification

William Herlands
Electrical Engineering
Princeton University
Princeton, NJ 08544
herlands@princeton.edu

Ricky Der
Dept. of Mathematics
University of Pennsylvania
Philadelphia, PA 19104
rickyder@sas.upenn.edu

Yoel Greenberg
Dept. of Music
Bar-Ilan University
Ramat Gan, Israel
yoel.greenberg@biu.ac.il

Simon Levin
Ecology & Evln. Biology
Princeton University
Princeton, NJ 08544
slevin@princeton.edu

Abstract

Recent literature has demonstrated the difficulty of classifying between composers who write in extremely similar styles (homogeneous style). Additionally, machine learning studies in this field have been exclusively of technical import with little musicological interpretability or significance. We present a supervised machine learning system which addresses the difficulty of differentiating between stylistically homogeneous composers using foundational elements of music, their complexity and interaction. Our work expands on previous style classification studies by developing more complex features as well as introducing a new class of musical features which focus on local irregularities within musical scores. We demonstrate the discriminative power of the system as applied to Haydn and Mozart's string quartets. Our results yield interpretable musicological conclusions about Haydn's and Mozart's stylistic differences while distinguishing between the composers with higher accuracy than previous studies in this domain.

1 Introduction

Differentiating stylistically homogeneous musical works is exceedingly difficult. For certain classes of problems this classification is vexing for musical professionals, let alone machines. However, automatically classifying high-dimensional, stylistically similar compositions is of interest to both the machine learning community and industry. Classification can aid content-based music retrieval and recommendation systems for commercial vendors, as well as aid in topic modeling for researchers. Traditional musicological research would also benefit from the ability to quantitatively validate musical theories or demonstrate previously undiscovered musical trends.

Early machine-aided music style classification studies addressed classification of music into clearly distinct styles. For example, (Leon and Inesta 2004) explored extracting features for music categorization, focusing exclusively on jazz and Western classical music. Similarly, (McKay and Fujinaga 2004), (Li and Sleep 2004), and (Shan, Kuo, and Chen 2002) addressed music classification between Western classical, jazz, rock, New Age, or Chinese music. (Cuthbert, Ariza, and Friedland 2011) reported on the technical aspects of quantitative musicological research and on distinguishing between Chinese and Central European music.

Recent literature has addressed more nuanced stylistic questions focusing on composer classification as well as emotion and performance difficulty categorization (Sturm 2012; Chiu and Chen 2012). Using melodic patterns, (Conklin 2009) distinguished between geographic sub-classes within folk music. (Van Kranenburg and Backer 2004) demonstrated the ability to classify between Western classical composers using relatively simple melodic and rhythmic features. Their study analyzed a number of nuanced classification problems including Haydn's and Mozart's string quartets. Expanding on that initial research, (Hillewaere, Manderick, and Conklin 2010) used Haydn's and Mozart's string quartets to test the efficiency of two types of features sets on multiple-voice music. This research compared global features, which are calculated over the entire length of a movement, with *n*-gram features, such as those described in (Chai and Vercoe 2001). Finally, (Dor and Reich 2011) used sets of time-ordered pitch patterns to classify between classical composers. Their results indicate that discriminating between Haydn's and Mozart's string quartets was the most difficult two-composer comparison problem.

While our study does not concentrate on audio music classification (discussed extensively in (Cunningham, Bainbridge, and Downie 2012)) it is significant to note that (Widmer 2003) explored the possibility of utilizing machine learning to understand how the differences between classical music performers reveal fundamental principles of expressive musical performance.

Homogeneous Classification

The research presented in this paper builds upon the literature in a number of ways, with three goals:

1. Introduce a repertoire of more complex and musicologically meaningful features. We expanded the global feature set generally used in the literature by introducing higher-order features which provide a greater depth of melodic, rhythmic, and multi-voice analysis. Additionally, we introduce a new set of features which measure local extrema at particular moments within a score. Since all our features are composed of basic musical components such as pitch, rhythm, and harmony, they are generally applicable to discriminate between any sub genre of music which can be represented as sheet music.
2. Yield musically meaningful insight into the compositional style of the composers in question. Much of the style classification literature has focused on musical styles remote from one another, their objective being directed primarily at gaining insights into the technological, rather than musicological, aspects of machine learning research. We wanted to be able to draw musical conclusions from the system beyond a mathematical description of the classification decision boundary.
3. Classify stylistically homogeneous composers. The present research is applied to Haydn's and Mozart's string quartets, which are musical scores that employ four instruments: a 1st violin, 2nd violin, viola, and cello. To increase stylistic homogeneity and maximize the meaningful variance between composers we exclusively considered the first movement of the string quartet. Additionally, only quartets in so-called "sonata form" were analyzed. In this manner we demonstrate the power of our system on a difficult classification problem.

Franz Joseph Haydn and Wolfgang Amadeus Mozart present an ideal case of stylistically homogeneous composers who lived contemporaneously, wrote in the same Western classical style, and admitted each other's influence on their work. Indeed, in an informal Stanford test, users self-rate their knowledge of the composers' music and then classify random selections from either Mozart or Haydn's string quartets. The graph in Figure 1 shows the results of this study on 20,822 participants (Sapp and Liu 2013).

Even self-rated experts only achieved a 65% classification accuracy. While not a scientific test of discrimination difficulty, this quiz corroborates previous machine learning studies which demonstrated that distinguishing between Haydn's and Mozart's quartets is significantly harder than classifying between distinct styles, such as rock and classical music, a distinction where most laymen can easily succeed (Dor and Reich 2011; Hillewaere, Manderick, and Conklin 2010).

The remainder of the paper is organized as follows. Section 2 briefly describes our data collection. The musical analysis and feature extraction are detailed in Section 3, and the machine learning methodology is delineated in Section 4. Classification results are presented in Section 5 with an expanded musicological analysis of the findings. Finally, Section 6 concludes with some observations about the research and directions for future study.

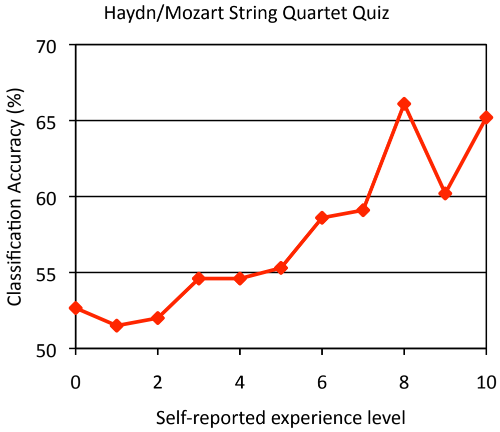

Figure 1: Human classification of segments of Haydn's and Mozart's string quartets. Data points estimated based on published charts.

2 Data

Since our system focused on differentiating between compositional styles, data for musical scores were extracted from sheet music in MIDI format. Recordings were not analyzed because variations between performers would distort the compositional data.

Data were provided by Music21 (Cuthbert and Ariza 2010). Due to the limited number of Haydn's and Mozart's scores as well as limitations in the Music21 database, we used 49 of Haydn's and all 23 of Mozart's string quartets. Additionally, we included 2 Mozart flute quartets to increase our sample size. Our study was restricted to first movements of the quartets in sonata form, the most common musical form of the Classical period which has a trajectory roughly described as leading from stability, through instability, back to stability. We extracted notes and rests from each composition with their associated duration from each voice (a voice being the musical line of a single instrumental part). A note is defined as a unit at some time offset from the start of a musical piece with a duration and pitch. A rest is defined as a unit at some time offset from the start of a musical piece with a duration and no pitch. Rests indicate silence of a particular instrument, whereas notes indicate sound.

The MIDI format discretizes the spectrum of continuous frequencies so that each octave (frequency doubling) is divided into twelve notes, spaced logarithmically. While the absolute intervals of the MIDI scale facilitate mathematical analysis it is limited insofar as it does not distinguish between enharmonic equivalents (e.g. C sharp vs. D flat).

To retain consistency only the highest pitch was used when the score indicated two simultaneous notes or a chord in a single voice. Based on the overall key of the sonata form, each composition was transposed into C major if in a major key, and into A minor, if in a minor key, to obtain a basic normalization.

3 Feature Definition and Extraction

Four sets of features were developed which we term global first-order, global higher-order, local first-order, and local higher-order features. Table 1 lists the types of first- and higher-order features, and the end of this section discusses how those features can be applied globally, over the entire musical score, or locally over specific regions of the score. These four categories of features were constructed to probe the effects of increasingly complex aspects of the music. Each of these categories contain elements of traditional musicological single-voice and multiple-voice features.

Single-voice features consider stylistic elements of individual voices and primarily consist of properties of a melody and its rhythm. Multiple-voice features analyze the interactions between two or more voices and relate to questions of harmony and simultaneity.

First-Order Features

Our first-order features measure the empirical distributions of simple musicological elements defined at every time-step in a musical score. Table 1 gives a listing of the types of first-order features used. Being macroscopic, these features do not contain information about the stylistic unfolding of a musical composition. However, these first-order features do point to certain biases in the underlying distribution of musical elements for each composer. It is reasonable to think that such biases could exist either as secondary consequences of stylistic choices or as an unconscious proclivity towards certain musical ranges.

Each first-order feature is evaluated over an individual voice. The distribution of pitches was analyzed from a number of different angles. The empirical histogram of all MIDI pitches, both absolute pitch and modulo 12 (to group them as scale degrees, irrespective of register), were employed as features. From these histograms, further features were extracted. Melodic range was calculated by taking both the absolute pitch range and the standard deviation of the pitch range. Similarly, histograms and functions of the histograms were calculated for the intervals between successive pitches.

The final melodic feature measured was the percentage of "chromatic" notes employed versus "diatonic" notes, the diatonic notes being the ones included in the standard 7-note major or minor scale, and the chromatic notes their complement[1]. Generally speaking, chromatic notes are statistically less common than diatonic ones, and are generally considered to play a more expressive role.

Aspects of rhythm were measured by quantifying the distribution of note and rest durations. Histograms as well as the absolute range, mean, and standard deviation of note durations measured the prevalence of basic rhythmic elements. Prevalence of rests were measured by the percentage of significant rests equal to or longer than 1 bar since a listener tends to hear smaller rests as "filled" in by the preceding note. Additionally the percentage of cumulative silence measured the duration of rests relative to the duration of a score.

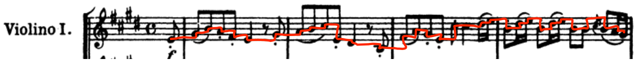

Figure 2: Haydn Opus 17 no 1 sheet music of 1st violin with potential melodic curve.

Higher-Order Features

While the first-order features measure the statistics of musical elements defined at every time-step within a piece, the higher-order features were designed to measure more complex stylistic elements ranging across larger segments of a score. This larger perspective provides some understanding of overarching musical characteristics of a composition.

Single-Voice Features Higher-order single-voice features pertain to melodic and rhythmic stylistic elements of individual voices. These features were designed on the premise that each instrument has a different melodic and rhythmic role within a string quartet. Analyzing the behavior of each voice can demonstrate how Haydn and Mozart use individual voices. This analysis should indicate whether these composers are consistent with their use of instruments between scores and if there are categorical differences between how Haydn and Mozart used individual instruments.

In order to analyze the content and progression of melody and rhythm within a voice, we constructed representations of a voices melodic and rhythmic curves

Melodic curves, as illustrated in Figure 2, are linear representations of the notes assigned to a particular instrument. We used a zero-order hold interpolation between successive notes to create a melodic curve of the discrete time and pitch information from sheet music (see Figure 3 for an example). This interpolation is true to a listener's experience of music where musical notes take on constant values that do not change over their duration. Rests were handled as with first order features, by representing significant rests equal to or great than 1 bar as discontinuities in the melodic curve.

Rhythmic curves were similarly constructed by interpolation between the duration of notes and rests in a score.

For both melodic and rhythmic curves we developed features that could measure their complexity and variation (see Table 1). Through these metrics we hoped to capture some stylistic differences between the overall structure of Haydn's and Mozart's melodies and rhythms. Complexity of the curves for each voice was measured by the fractal dimension, which is a measurement of the total length of a curve.

In addition, the discrete derivative of each curve (i.e. the difference time series for a discrete time series) was calculated in order to quantify the speed of melodic and rhythmic change. Complexity of the derivative was again measured by fractal dimension. A number of different measures of variation in the derivative were employed, including standard deviation (i.e. L^2 norm), and the size of the zero set of the derivative. Both these measures give a quantitative sense of the "consistency" of a voice. For example, a zero set with larger cardinality indicates a more consistent voice.

[1] In the case of the minor scale, which has variants at the sixth and seventh degrees, we define 9 diatonic notes instead of 7.

	First-Order	Higher-Order, Single-Voice	Higher-Order, Multiple-Voice
Melodic	Pitch range Absolute range Standard deviation Pitch intervals Histogram Standard deviation Histogram over all pitches Percentage of chromatic notes	Melodic curve Fractal dimension $\frac{d}{dt}$ of melodic curve Fractal dimension Standard deviation Zero set	Melodic curve difference Average Fractal dimension Standard deviation $\frac{d}{dt}$ of melodic curve difference Fractal dimension Standard deviation Zero set Dissonance
Rhythmic	Note duration Histogram Absolute range Standard deviation Rests % of significant rests % of cumulative silence	Rhythmic curve Fractal dimension $\frac{d}{dt}$ of rhythmic curve Fractal dimension Standard deviation Zero set	Rhythmic curve difference Average Fractal dimension Standard deviation $\frac{d}{dt}$ of rhythmic curve difference Fractal dimension Standard deviation Zero set
Simultaneity			% of simultaneous onset of notes % of simultaneous onset of rests

Table 1: List of features.

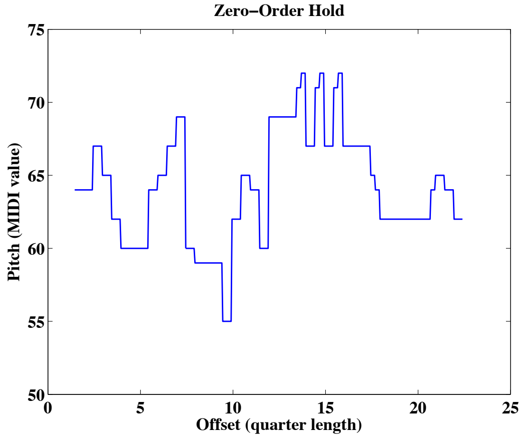

Figure 3: Haydn Opus 17 no 1 melodic curve of 1st violin constructed by zero-order hold.

Multiple-Voice Features Multiple-voice features (see Table 1) were created under the assumption that stylistically significant interdependency between voices exists in string quartets, and that composers use these styles with some regularity among their musical works.

Some multiple-voice features measure the difference between two voices' melodic or rhythmic curves. Such features attempt to quantify the complexity and variation of the resulting curve in order to understand how a composer combines instruments within a score. For example, we computed the difference between the melodic curves of the first and the second violin. Since the difference between these two curves is itself a curve, its properties can be measured.

In this manner the mean, standard deviation, and fractal dimension are calculated for the melodic and rhythmic pairwise differences between voices. These statistics were also calculated on the derivative of the melodic and rhythmic difference in order to give more nuanced stylistic interactions. Additionally, dissonance metrics were calculated between concurrent notes of pairs of voices[2]. All of these features were measured over connected regions of the curves, not interrupted by discontinuities due to significant rests.

The synchronization of various instruments is another significant multiple-voice feature which captures the pairing of voices and the independence of certain instruments. We quantified this element by calculating the rate of simultaneous onset of notes and rests between two or more voices. Both the frequency of notes as well as the frequency of rests that begin at the same onset between voices, as a function of length of a score, were measured. For example, if $T = \{0, 1, \ldots, D\}$ where D is a terminal time, and $v_i(t)$ is a boolean function which is 1 if voice i begins a new note at time t, and 0 otherwise, then we may define a simultaneity measure S as

$$S = \sum_{t \in T} \prod_{i \in I} v_i(t) \tag{1}$$

which gives the number of simultaneous notes throughout a set of instruments, I, in the duration $t \in T$.

Local Features

Both first- and higher-order global features are fundamentally limited insofar as they process a musical score as a single unit of information, thereby reducing a musical section's varied elements to a single, macroscopic value. Yet neither composers nor listeners think in terms of global statistical distributions. Rather, they hear and understand music mostly

[2] Our quantitative measures of dissonance followed standard notions in musicology: unisons and fifths were rates most consonant, thirds slightly less, and semitones and tritones least consonant.

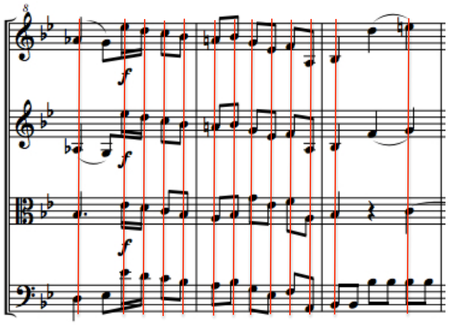

Figure 4: Measures 8-10 of Mozart's K. 172 with strong local simultaneity of notes between all voices.

in a segmented, local manner, with particular sections having distinct stylistic elements. Thus it is important to think of the global underlying distribution of those elements as a baseline, which can then be used to identify deviations from this baseline in specific subsections of a score.

For example, consider the higher-order feature which measures the simultaneous onset of notes between all voices, defined in Equation 1. While over the entirety of Mozart's K. 172 first movement, 47% of the notes in the 1st violin are simultaneously played with the other three voices, in the section in Figure 4, 87% percent of the notes in the 1st violin are simultaneously played with the other instruments. Additionally, the voices are in melodic unison. Though this section is quite distinctive to the human ear and an essential part of the score, the global higher-order feature does not reflect such information.

Local implementations of the global first-order and higher-order features in Table 1 were created which evaluate those features over subsets of the score. Our procedure for localization was as follows. We conducted a linear search using a fixed sliding window over the length of a score. A first-order or higher-order feature was then evaluated within the sliding window, giving a time-series for the evolution of that feature over the score.

Functions of these time-series then define local features. For example, we measured the local and global maxima of the feature time-series. We also measured the location of the global maxima relative to the length of the score. The extrema provide a measurement of the magnitude of the local variation while the location of the extrema are designed to detect whether a composer has a bias for creating regions of extreme variation at particular locations in the score, as might be expected in a sonata form movement.

4 Classification Methodology

We used standard machine learning techniques to train and test classifiers, comparing their performance by separately using global first-order, global higher-order, local first-order, and local higher-order features as inputs. Experiments used the scikit-learn package (Pedregosa et al. 2011).

The dataset was balanced by under-sampling Haydn to choose 25 quartets for each run. Training and testing sets were formed by 80/20 cross validation. Averaging the results for the cross validation over all runs provided our results.

Feature selection was performed by correlating feature values (generally real-valued) with the class label over the training data. Features with a Pearson correlation coefficient less than 0.05 were eliminated from the final training phase.

Using the training data, an inner cross-validation loop further subdivided the data into smaller sets. On these smaller validation sets, five standard classifiers were iteratively trained and tested on the reduced feature set, in order to select an optimal classifier. The library of classifiers employed consisted of (1) a linear support vector machine (SVM) with regularization parameter chosen via grid search; (2) a naive Bayes classifier, (3) decision tree and random forest classifiers using Gini impurity measure, (4) an AdaBoost ensemble classifier with decision tree stump base learners.

The best classifier resulting from this inner cross-validation loop on the training data was then selected, retrained on the entire set of training data, and its generalization capability computed on the hitherto unseen test data.

This entire procedure was run over multiple splits of the data between training and test sets in order to obtain an estimate of the average classification accuracy. Throughout the experimentation we took care to not perform feature selection or optimize the system over any testing data either explicitly or inadvertently by manually shifting parameters and hyper-parameters over the course of several tests.

In a second set of experiments, we used the above system to study the discrimination potential of individual features. These experiments measured the classification power of each feature employed singly, and, therefore, did not employ a feature filtering stage.

While the system's feature space was large relative to the sample size, we chose not to incorporate dimensionality reduction or decorrelation methods such as principal component analysis. Although such methods could potentially increase the system's classification accuracy, mapping the feature space to a lower-order space would reduce the musicological significance of the results, as individual stylistic features would be less interpretable in the reduced space.

Similarly, we only report results for classifiers trained on the four musical categories in isolation. Though not investigated here, it seems possible that a blending of the classifiers across our four feature sets could lead to significant gains in classification accuracy.

5 Results and Discussion

Our learning system chose linear SVM classifiers in the overwhelming number of cross-validations trials, though Bayes classifiers were also chosen with some frequency. The choice of an SVM seems reasonable given the high dimensionality of our feature space. Additionally, the Bayes classifier have demonstrated significant classification power even when features are not strictly independent and identically distributed, conditional on the class label (Zhang 2004).

The average classification accuracy for each group of features (first-order, higher-order, and local) is shown in Table 2. In another experiment, we subdivided our feature

Feature category	Classification accuracy
Global first-order	80%
Global higher-order	76%
Local first-order	57%
Local higher-order	77%
Melody (single-voice)	80%
Rhythm (single-voice)	72%
Multiple-voice	73%

Table 2: Classification results for feature categories and musical feature type.

set along musicologically relevant lines and then ran our learning system with those groupings (see Table 2). Our results indicate that the system is able to discriminate between Mozart and Haydn string quartets at an accuracy that exceeds previous literature in this domain. The results also surpass the accuracy rate of self-defined human experts, as measured by the Stanford online quiz.

The excellent performance of the global first-order features suggests that overall statistical variations in the basic musical elements of Haydn and Mozart's quartets can be used to successfully discriminate between the composers. The success of our expanded set of first-order features features is most probably due to secondary effects of stylistic variations. By contrast, the poor performance of the local first-order features emphasizes the point that first-order features are more indicative perhaps of a similar overall vocabulary shared by the composers, and are therefore not as useful in distinguishing between them.

The relatively strong classification power of both the global and local higher-order features suggest that there are meaningful quantitative stylistic differences between Haydn and Mozart's quartets. These differences are manifest both in the general use of voices and interactions between voices over the length of a score as well as in their local nuances.

Individual Feature Results

While the previous section addressed the question of using feature categories to discriminate between Haydn and Mozart, it is of considerable musical interest to ask which individual features were most informative in distinguishing between the composers. Table 3 identifies the most significant features, and below we discuss the musicological significance of a few selected features.

The global first-order feature of a pitch interval of size 2 was discriminating in our tests. Figure 5 shows the pitch interval histogram for our data, grouped into stepwise motion (1-2 intervals), skips of thirds (3-4) and larger skips. As one can see, Mozart has significantly more seconds (MIDI pitch interval of size 1 and 2) than Haydn.

The significance of seconds could indicate a predilection for more scalar and hence less virtuosic passages in Mozart. Haydn's greater virtuosity in the first violin probably emanates from personal and social circumstances. For many of his works Haydn had at his disposal excellent first violinists. In contrast, Mozart wrote largely for the amateur market.

Global First-Order:
1st violin
Pitch intervals of 2 MIDI notes
Note duration
Global Higher-Order:
Derivative of 1st violin melodic curve
Fractal dimension
Standard deviation
Zero set
Std of the melodic difference
Between 1st violin and 2nd violin
Between 1st violin and viola
Simultaneous onset of rests in 2nd violin, viola, cello
Local Higher-Order:
Derivative of 1st violin melodic curve
Maxima of fractal dimension
Derivative of cello melodic curve
Maxima of fractal dimension
Maxima of standard deviation

Table 3: List of most discriminating features for each feature category. Note that no local first-order features were significantly discriminating.

Three higher-order features relating to the melodic curve of the first violin demonstrate that Haydn's use of that voice is more complex (as judged by our mathematical definitions of complexity), has greater variation, and changes more often than in Mozart's quartets. Perhaps the most telling feature measured the complexity of the derivative of the melodic curve which is a measure of the complexity of changes in that voice's melody. Overall, Haydn's compositions had a fractal dimension complexity 39% greater than Mozart. Additionally, the standard deviation of that curve was 22% greater in Haydn. This indicates that Haydn employs changes in the first violin which are of greater magnitude and more frequent than Mozart, again, a reflection of more virtuosic writing for the first violin.

The value of the first violin as a distinguishing element is also seen in the level of coordination with the other instruments. Haydn has a 44% higher rate of simultaneous onsets of rests between the 2nd violin, viola, and cello. This indicates that Mozart employs varying groupings between the four voices more readily, while Haydn more frequently using the three lower voices as a homogeneous accompanying unit. Haydn's employment of the bottom voices as an accompanying section again allows for writing which is relatively more concentrated in the first violin.

6 Conclusion

We presented a machine classification system that can distinguish between stylistically homogeneous music with high accuracy. The system used more complex global features than previous research and introduced a novel class of local features to detect nuanced stylistic elements. Meaningful musicological results were derived from the system's output which extend beyond traditional machine learning studies. Additionally, since the feature sets exploit foundational

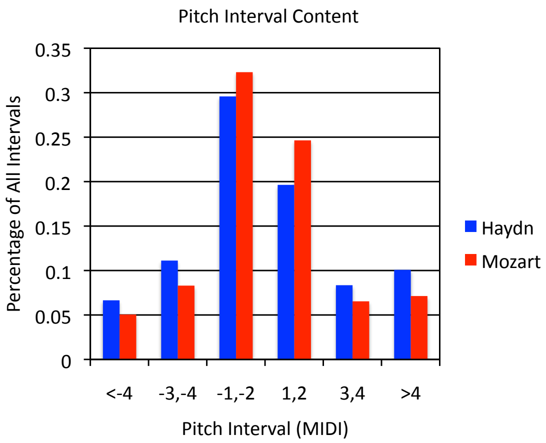

Figure 5: Comparison of successive pitch intervals.

elements of music present in any tonal work, the features are not restricted to any particular composers. Instead, the same feature classes could discriminate between any subgenres of music which can be represented as sheet music. Applying the system to Haydn's and Mozart's string quartets demonstrated the features' utility by yielding a number of significant stylistic differences between the composers. Future studies could expand the power of local features or develop features to capture long-range dependencies in musical scores. Through such inter-disciplinary research musicologists may be able to increasingly turn to statistical tools when considering questions of stylistic similitude.

References

Chai, W., and Vercoe, B. 2001. Folk music classification using hidden markov models. *Proceedings of International Conference on Artificial Intelligence*.

Chiu, S.-C., and Chen, M.-S. 2012. A study on difficulty level recognition of piano sheet music. In *Multimedia (ISM), 2012 IEEE International Symposium on*, 17–23. IEEE.

Conklin, D. 2009. Melody classification using patterns. In *Second International Workshop on Machine Learning and Music*, 37–41.

Cunningham, S. J.; Bainbridge, D.; and Downie, J. S. 2012. The imapct of mirex on scholarly research.

Cuthbert, M. S., and Ariza, C. 2010. Music21: A toolkit for computer-aided musicology and symbolic music data. In *ISMIR*, 637–642.

Cuthbert, M. S.; Ariza, C.; and Friedland, L. 2011. Feature extraction and machine learning on symbolic music using the music21 toolkit. *12th International Society for Music Information Retrieval Conference*.

Dor, O., and Reich, Y. 2011. An evaluation of musical score characteristics for automatic classification of composers. *Computer Music Journal* 35(3):86–97.

Hillewaere, R.; Manderick, B.; and Conklin, D. 2010. String quartet classification with monophonic models. In *ISMIR*, 537–542.

Leon, P. J. P. d., and Inesta, J. M. 2004. Statistical description models for melody analysis and characterization. *Proceedings of the International Computer Music Conference*.

Li, M., and Sleep, R. 2004. Improving melody classification by discriminant feature extraction and fusion. *Universitat Pompeu Fabra*.

McKay, C., and Fujinaga, I. 2004. Automatic genre classification using large high-level musical feature sets. In *ISMIR*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Sapp, C., and Liu, Y.-W. 2013. The haydn/mozart string quartet quiz.

Shan, M.-K.; Kuo, F.-F.; and Chen, M.-F. 2002. Music style mining and classification by melody. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, volume 1, 97–100 vol.1. ID: 1.

Sturm, B. L. 2012. A survey of evaluation in music genre recognition. *Proc. Adaptive Multimedia Retrieval, Copenhagen, Denmark*.

Van Kranenburg, P., and Backer, E. 2004. Musical style recognition-a quantitative approach. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*, 106–107.

Widmer, G. 2003. Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries. *Artificial Intelligence* 146(2):129–148.

Zhang, H. 2004. The optimality of naive bayes. *Florida Artificial Intelligence Research Society Conference*.