STRING QUARTET CLASSIFICATION WITH MONOPHONIC MODELS

Ruben Hillewaere and **Bernard Manderick**
Computational Modeling Lab
Department of Computing
Vrije Universiteit Brussel
Brussels, Belgium
{rhillewa,bmanderi}@vub.ac.be

Darrell Conklin
Department of Computer Science and AI
Universidad del País Vasco
San Sebastián, Spain
IKERBASQUE, Basque Foundation or Science
Bilbao, Spain
darrell_conklin@ehu.es

ABSTRACT

Polyphonic music classification remains a very challenging area in the field of music information retrieval. In this study, we explore the performance of monophonic models on single parts that are extracted from the polyphony. The presented method is specifically designed for the case of voiced polyphony, but can be extended to any type of music with multiple parts. On a dataset of 207 Haydn and Mozart string quartet movements, global feature models with standard machine learning classifiers are compared with a monophonic n-gram model for the task of composer recognition. Global features emerging from feature selection are presented, and future guidelines for the research of polyphonic music are outlined.

1. INTRODUCTION

In the field of music information retrieval, much research has been done in symbolic music genre classification, where a model has to assign an unseen score to a certain class, for example style, period, composer or region of origin. There are two main categories of models that have been widely investigated: *global feature models* and *n-gram models*. Global feature models express every piece as a feature vector and use standard machine learning classifiers, whereas n-gram models rely on sequential event features.

In a recent paper [10] the results of a thorough comparison of these types of models are reported for the task of classifying folk songs based on their region of origin on a large monophonic data set. That study demonstrates that the n-gram models are always outperforming the global feature models for this classification task. It is an interesting question whether this result still holds in a polyphonic setting.

In the literature, it appears that most research has been investigating classification or characterization of melodies (monophonic) [5, 14, 16], but only few efforts have been made to develop polyphonic models. In [15], orchestrated polyphonic MIDI files are classified using global features, including some features about musical texture and chords. A set of polyphonic features based on counterpoint properties was developed by [19], and applied to the task of composer classification. They find that the distinction between Haydn and Mozart string quartets, which is very interesting from a musicological point of view, is a hard classification task.

When considering polyphonic music, it is essential to qualify the form of input. Two formats can be considered:

voiced: a fixed and persistent number of parts; and,

unvoiced: free polyphony that is not available in, or cannot be easily divided into parts.

A typical example of voiced polyphony is a string quartet, consisting of 4 well-defined voices (Violin 1, Violin 2, Viola and Cello). Unvoiced polyphony is common, for example, in piano music.

Another way to view this dichotomy of polyphonic music is in terms of a MIDI file type: voiced (type 1), or unvoiced (type 0), realizing of course the grey area where tracks within a type 1 file may contain internal polyphony, and where type 0 files identify voices by use of channel numbers.

This paper investigates how monophonic global feature and n-gram models perform on the classification of Haydn and Mozart string quartets in their original voiced format. The voiced structure is exploited since these monophonic models are applied to separate voices. The initial database used in [19] containing 107 string quartet movements was extended to a total of 207 movements in order to measure statistically more relevant differences.

Two tentative hypotheses from previous work [11] are being verified in this paper:

1. n-gram models also perform better than global feature models on monophonic parts extracted from the polyphonic texture.

2. the first violin is the most distinctive voice of the string quartets.

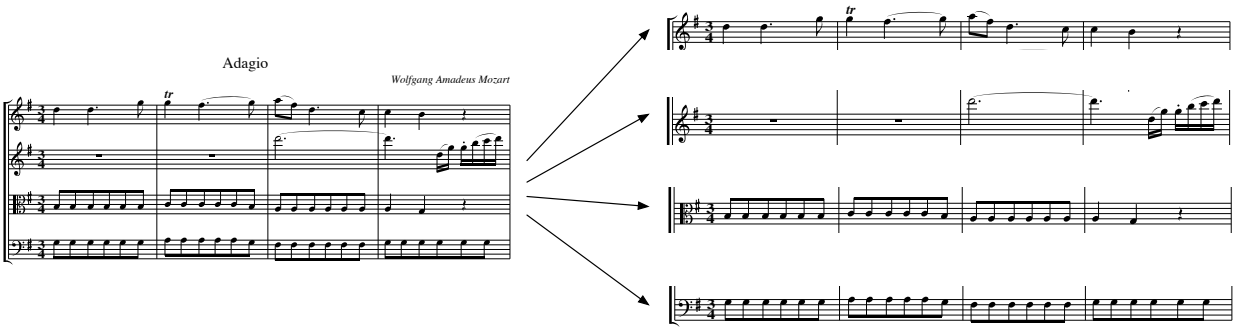


Figure 1. The voice separation of the string quartets into Violin 1, Violin 2, Viola and Cello.

For the global feature models, special care has been taken to apply feature selection within the inner loop of the cross validation scheme, in order to avoid overoptimistic evaluation estimates [8, 17]. A similar procedure has been set up to tune the parameters of the learning algorithms during the training phase. Features that emerge recurrently in the feature selection process are highlighted.

The remainder of this paper is structured as follows. We start by introducing the dataset and the music representations, global feature and n-gram models, and the classification methodology in the next section. Then, we give the results of the described models on the Haydn/Mozart dataset. We end with a discussion and some directions for future work.

2. METHODS

In this section we describe the dataset used for our experiments and we will give a short overview of both global feature and n-gram models. Furthermore, we introduce our classification methodology outlining the cross validation setup combined with supervised feature selection and SVM parameter tuning.

2.1 Dataset and music representation

The Haydn/Mozart dataset is composed of 112 string quartet movements from Haydn and 95 string quartet movements from Mozart, including most of the pieces from the dataset used in [19], but extending it as much as possible to nearly double its size. We chose to focus on the period 1770-1790 in which both composers were active, discarding early or late Haydn quartets which might be easy to classify. In order to maximize the number of quartet movements from Mozart, we included 8 movements from two flute quartets (K.285, K.298) and one oboe quartet (K.370), which are written for flute/oboe, violin, viola and cello, and thereby very similar to the string quartets. The original scores in **kern format were found on the website of the Center for Computer Assisted Research in the Humanities at Stanford University [1]. We transformed these to clean MIDI files, ensuring that the four voices appear on separate tracks and that all barlines are correctly synchronized in all voices by correcting several errors in note durations. We retained only the highest note of double stops, thereby reducing each voice to a purely monophonic sequence of notes. To enable the use of monophonic classification techniques, we created four monophonic data sets called Violin 1, Violin 2, Viola and Cello by isolating each voice of every string quartet movement, as illustrated in Figure 1.

2.2 Global feature models

In this section we introduce global features and the corresponding global feature models. A global feature summarizes information about a whole piece into a single attribute, which can be a real, nominal or boolean value, for example “average note duration”, “meter” or “major/minor”. With a set of global features, pieces can be simply re-expressed as feature vectors, and a wide range of standard machine learning algorithms can then be applied to evaluate the feature set.

Voiced polyphony presents the advantage of having a fixed number of monophonic parts, which enables us to isolate these parts and apply monophonic models. In this paper three global feature sets are used to represent the monophonic parts. These features describe melody characteristics, mainly derived from pitch or duration, whereby we mean that at least one pitch or duration value is inspected for the feature computation.

The global feature sets chosen are the following :

- The *Alicante* set of 28 global features, designed by P.J. Ponce de Léon and J.M. Iñesta in [16]. This set was applied to classification of MIDI tunes in jazz, classical, and pop genres. From the full set, we implemented the top 12 features that they selected for their experiments.

- The *Jesser* set, containing 39 statistics proposed by B. Jesser [13]. Most of these are derived from pitch, since they are basic relative interval counts, such as “amajsecond”, measuring the fraction of melodic intervals that are ascending major seconds. Similar features are constructed for all ascending and descending intervals in the range of the octave.

- The *McKay* set of 101 global features [15], which were used in the winning 2005 MIREX symbolic genre classification experiment and computed with McKay's software package jSymbolic [2]. These features were developed for the classification of orchestrated polyphonic MIDI files, therefore many features, for example those based on dynamics, instrumentation, or glissando, were superfluous for this analysis of monophonic single voices and we were able to reduce the set down to 61 features.

These global feature sets do not show many overlapping features, only some very basic ones occur in maximum two of the sets, such as the "pitch range". Therefore it is interesting to join the three feature sets to form the *Joined* set, which means every piece is represented as a data point in a 112-dimensional feature space. We are interested in finding out which features are relevant for this specific task of composer classification, therefore we will apply feature selection on this Joined set.

2.3 n-gram models

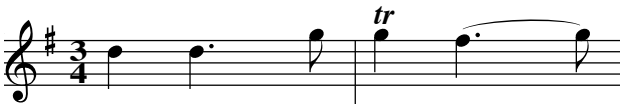
In this section we introduce n-gram models and how they can be used for classification of music pieces using event features. n-gram models capture the statistical regularities of a class by modeling the probability of an event given its preceding context and computing the probability of a piece as a product of event probabilities. This technique is particularly well-known for language modeling, a word in language being roughly analogous to an event in music. The context $\overline{e_{i-1}} = [e_1, e_2, \ldots, e_{i-1}]$ of an event e_i is usually limited to a short suffix $[e_{i-n+1}, \ldots, e_{i-1}]$, meaning the probability of the current event only depends on the $n-1$ previous events. The n-gram counts of the training data are used to estimate the conditional event probabilities $p(e_i \,|\, \overline{e_{i-1}})$, and the probability of a new piece $\overline{e_\ell}$ is obtained by computing the joint probability of the individual events in the piece:

$$p(\overline{e_\ell}) = \prod_{i=1}^{\ell} p(e_i \,|\, \overline{e_{i-1}}) \qquad (1)$$

To use an n-gram model for music classification, for each class a separate model is built, and a new piece is then simply assigned to the class with the highest piece probability.

For monophonic music, n-gram models and more powerful extensions are naturally applicable [6, 10], but polyphonic music needs first to be converted into a sequential form. One way to do this is to simply extract a voice (e.g., Violin 1) from the polyphonic texture.

To reduce the sparseness of the n-gram counts, we do not model the pitch or duration directly, but we first transform the music events by means of event features. An event feature assigns a feature-value to every event, in our case to every note in the music piece. The chosen event feature determines the level of abstraction of the data representation. The event feature we will use is the melodic interval. Models are constructed for a class by extracting the same voice (e.g., Violin 1) for every piece in a corpus, and viewing that piece as a sequence of melodic intervals.



pitch	74	74	79	79	78	79
melodic interval	⊥	0	+5	0	-1	+1
melodic contour	⊥	r	u	r	d	u
duration ratio	⊥	3/2	1/3	2	3/2	1/3
average pitch	77.1667					
rel. freq. m2	0.4					
Huron contour	ascending					

Figure 2. First measures of the first violon of the Adagio K.080 of W.A. Mozart, illustrating the contrast between global features (lower three) and event features (upper four).

Figure 2 illustrates the difference between global features and event features on an excerpt of the first violon of the Adagio K.080 of W.A. Mozart. A global feature describes a constant property of the whole piece, whereas an event feature is associated to one particular note. A global feature summarizes the data much more, but one uses a whole collection of global features to build a global feature model, whereas n-gram models are constructed using one single event feature.

2.4 Classification methodology

In this paper, two fundamentally different types of models are applied to the task of composer classification. In order to present any comparative results, we have to find a common way of evaluating the performance of these models. It is common practice to set up a cross validation scheme to obtain classification accuracies that generalize reasonably well.

Our data set is very small from a general machine learning point of view, only 207 samples, it is therefore preferable to do *leave-one-out cross validation*, where one uses as much training data as possible to train the model, discarding only one instance for testing purposes. For both global feature and n-gram models, a leave-one-out cross validation scheme was implemented.

Since global features represent every instance as a multidimensional feature vector, any standard machine learning classifier can be applied to get a performance accuracy. Simple classifiers such as Naive Bayes and k-nearest neighbours can give us a good indication, but in this work we opt for the more sophisticated Support Vector Machine, shortly SVM, which has been proven to be a state-of-the-art classifier [7]. An SVM makes use of a so-called kernel function to determine non-linear decision boundaries between classes, and a well-known kernel function, the

RBF-kernel, is used in this paper [4]. In this setting, an SVM has two parameters that need to be trained. The first is related to the softness of the decision margin, expressing the tradeoff between generality and classification accuracy, commonly denoted as C. The second is a parameter σ specific to the RBF kernel function. In practice, these parameters can simply be tuned by doing a "grid-search" over a large search-space of pairs (C, σ) as described in [12].

Another common machine learning technique is *feature selection*, which is often used to reduce the dimensionality of the data or to discover which features are highly correlated with the target class. In principle, feature selection is decreasing the size of the hypothesis space, which leads to a faster and more effective search for the learning algorithms and tends to avoid overfitting. Therefore, it has led to improved classification accuracies in some cases, or to a compact feature set that describes the data in a more interpretable, summarized way.

However, there is a subtlety in both feature selection and SVM parameter optimization, a pitfall to avoid when one uses *supervised* learning methods in combination with a cross validation scheme [8, 17]. In the simple case where a separate training and test set are given, one has to apply supervised preprocessing methods followed by the learning algorithm on the training set only, before testing the resulting model on the test set. Expanding this to a cross validation scheme, this means one must take care to apply these methods within the inner cross validation loop. As pointed out by [17], it is a common mistake to use both training and test set for supervised feature selection, which leads to overoptimistic and exaggerated performance results.

In this paper, SVM* denotes the model in which parameter tuning with a grid search has been done within the inner loop of the cross validation scheme. Feature selection is also implemented taking this special consideration.

3. RESULTS

In this section we describe the experimental results for the classification of the Haydn and Mozart string quartet movements. As a baseline, we keep in mind that the classes are quite equally distributed (112 Haydn, 95 Mozart), which means that 54.1% classification accuracy can be achieved by always predicting Haydn.

To evaluate the global feature approach, the SVM* classifier method is applied to the Joined set. As described above, this includes an SVM parameter tuning by doing a grid search within each loop of the leave-one-out cross validation. Furthermore, a supervised feature selection method called Correlation-based Feature Selection (CFS) is also applied. CFS is a filter method aiming to find a subset of features that are highly correlated with the class but have few intercorrelation among them [9]. The implementation of SVM* and the CFS make use of the Weka machine learning toolbox [3, 20].

For the n-gram model, we use a simple trigram model of the melodic intervals. For each Haydn and Mozart a separate model is built on the training set and a test piece is assigned to the class of which the model generates it with the highest probability according to Equation 1. A global classification accuracy is also computed with leave-one-out cross validation. The results for both the global feature models and the trigram models on the separate voices are reported in Table 1.

Voices	SVM*	SVM*+feat.sel.	3-grams
Violin 1	74.4	73.4	63.8
Violin 2	66.2	66.2	61.4
Viola	62.8	57.0	61.4
Cello	65.7	59.4	**75.4**

Table 1. The l.o.o. classification accuracies of the Joined global feature set and the trigram model on the separate voices extracted from the voiced string quartet movements.

It appears immediately that the results of previous work done on a smaller database of 107 pieces do not hold up [11]. Previously, we noticed a consistent tendency for n-gram models to perform better than global feature models regardless of the voice. Now we observe that the n-gram models perform well on the Cello dataset with an accuracy of 75.4%, but poorly on the other voices, whereas the global feature models achieve an almost equally high accuracy of 74.4% on the Violin 1. Our second hypothesis, about the first violin being the most predictive one for a composer, is also weakened because of this surprising result with n-gram models on the Cello. However, the global feature result on Violin 1 is still an indication of its predictive value. Additional computation of global feature models on the separate Alicante, Jesser and McKay sets confirm this indication, and show that we can order the voices according to there predictiveness with global feature models as follows: Violin 1, Cello, Violin 2 and Viola.

The second column of Table 1 is showing the results of the SVM* with CFS feature selection. These classification accuracies are slightly lower than without applying feature selection, which confirms that supervised feature selection does not necessarily lead to an improvement when it is applied in the inner loop of the cross validation scheme. Nevertheless, it is interesting for musicological reasons to see which features emerge in the selected feature subsets for each voice. Below we give three short examples of features that are selected in one or more voices.

- "dmajsec", i.e. the fraction of melodic intervals that are descending major seconds, is selected for both Violin 1 and Violin 2. Looking at the relative frequencies of this feature, it appears that Mozart uses more descending major seconds than Haydn for the two violins.

- "shortestlength" emerges in both the Violin 2 and the Viola. This is the shortest duration such that all durations are a multiple of this shortest duration (except for triplets). Again by looking at the relative distributions, one notices that Mozart tends to use smaller shortest lengths in these voices.

- "ImportanceMiddleRegister" is one of the features selected for the Cello. This denotes simply the fraction of notes with MIDI pitches between 55 and 72, which is roughly the upper range of the instrument. It seems that Haydn uses the cello more often in this range than Mozart in these string quartets.

4. CONCLUSIONS AND FUTURE WORK

This paper has applied monophonic classification models to the task of composer recognition in voiced polyphonic music, specifically Haydn and Mozart string quartets written in the period 1770-1790. An earlier dataset of string quartet movements is extended to a total of 207 pieces to obtain more statistical significance. The voiced structure is exploited by extracting the separate voices to enable the use of monophonic models. Several conclusions emerge: that a simple trigram model of melodic interval performs very well on the cello, achieving the best classification accuracy of 75.4%, but is outperformed by the global feature models on the other voices. Therefore, we are also unable to conclude that the first violin is indeed the most predictive voice for a composer, even though the results on Violin 1 are consistently best with the global feature approaches.

At first sight, these observations are rather disappointing, but they confirm the necessity of having a sufficiently large dataset before making any claims. Learning algorithms in symbolic music have to cope with this kind of challenge, what shows there is still room for improvement on a machine learning level.

Currently, we are investigating what causes this remarkable result with the trigram model on the cello and the low accuracy on the first violin, by looking carefully which pieces are correctly classified by one method and not by another, or correctly by both. Perhaps there is a core part of this dataset that is 'easy' to classify, or else we might consider using an ensemble model where one combines the global feature models and the n-gram models in order to improve the overall accuracies. One could also wonder how the so-called Haydn Quartets, six quartets written by Mozart but famously inspired by and dedicated to Haydn, influence these results. So far we have only found an indication that these particular movements are slightly harder to recognize, this topic will be part of further research.

Further future work will address the issue of polyphonic music in different ways. Figure 3 illustrates the global structure of these future directions. As we detailed earlier in this paper, polyphonic music can be voiced, like the string quartets used for this study, or unvoiced, for example piano sonatas. Each of these types of polyphony can be modelled by monophonic or polyphonic models. The models from this work were monophonic models, which are situated in the outer left branch of the tree. Polyphonic models for voiced polyphony can for example be based on polyphonic global features taking into account voice information or harmonic global features, such as those used in [15, 19]. To apply monophonic models to unvoiced polyphonic music, one has to apply some voice extraction algorithm first, for example the *skyline* method [18], which simply slices polyphony at each distinct onset and takes the highest pitch of every slice. The outer right branch of the tree is dealing with unvoiced polyphonic music by means of polyphonic models. One can easily imagine global features representing this kind of data, for example by computing relative frequencies of vertical intervals, i.e. intervals between simultaneous notes. However, building a truly polyphonic n-gram model remains a challenge, as one has to deal with segmentation and representation issues to cope with sparsity.

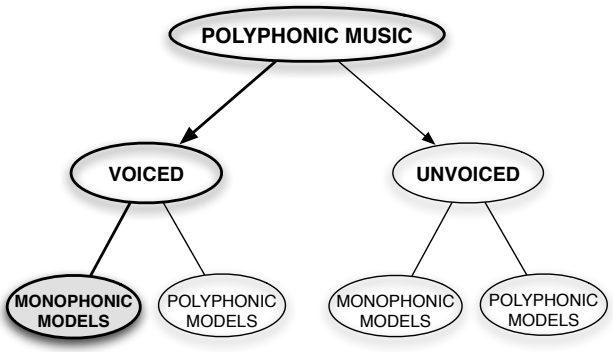


Figure 3. Tree structure outlining the possible ways to approach the classification of polyphonic music.

5. ACKNOWLEDGEMENTS

Darrell Conklin is supported by IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. Ruben Hillewaere is supported by the project Messiaen Weerspiegeld in collaboration with the Royal Conservatory of Brussels. Special thanks go to Stijn Meganck and Jonatan Taminau for their useful comments and support during this research.

6. REFERENCES

[1] `http://www.ccarh.org`.

[2] `http://jmir.sourceforge.net/jSymbolic.html`.

[3] `http://www.cs.waikato.ac.nz/ml/weka/`.

[4] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, USA, 1995.

[5] D. Conklin. Melodic analysis with segment classes. *Machine Learning*, 65(2):349–360, 2006.

[6] D. Conklin and I. H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

[7] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[8] R. Fiebrink and I. Fujinaga. Feature selection pitfalls and music classification. In *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 340–341, Victoria, Canada, 2006.

[9] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 359–366, Stanford, USA, 2000.

[10] R. Hillewaere, B. Manderick, and D. Conklin. Global feature versus event models for folk song classification. In *ISMIR 2009: 10th International Society for Music Information Retrieval Conference*, pages 729–733, Kobe, Japan, 2009.

[11] R. Hillewaere, B. Manderick, and D. Conklin. Melodic models for polyphonic music classification. In *Second International Workshop on Machine Learning and Music*, Bled, Slovenia, 2009.

[12] C.W. Hsu, C.C. Chang, C.J. Lin, et al. A practical guide to support vector classification. Technical report, 2003.

[13] B. Jesser. *Interaktive Melodieanalyse*. Peter Lang, Bern, 1991.

[14] M. Li and R. Sleep. Melody classification using a similarity metric based on Kolmogorov complexity. In *Sound and Music Computing*, Paris, France, 2004.

[15] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proceedings of the International Conference on Music Information Retrieval*, pages 525–530, Barcelona, Spain, 2004.

[16] P. J. Ponce de Léon and José M. Iñesta. Statistical description models for melody analysis and characterization. In *Proceedings of the 2004 International Computer Music Conference*, pages 149–156, Miami, USA, 2004.

[17] P. Smialowski, D. Frishman, and S. Kramer. Pitfalls of supervised feature selection. *Bioinformatics*, 26(3):440, 2010.

[18] A.L. Uitdenbogerd and J. Zobel. Matching techniques for large music databases. In *Proc. ACM Multimedia Conference*, pages 57–66, Orlando, Florida, 1999.

[19] P. van Kranenburg and E. Backer. Musical style recognition - a quantitative approach. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*, pages 106–107, Graz, Austria, 2004.

[20] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques. 2nd edition*. Morgan Kaufmann, San Francisco, 2005.