

Melodic models for polyphonic music classification

Ruben Hillewaere¹, Bernard Manderick¹, and Darrell Conklin²

¹ Computational Modeling Lab, Department of Computing, Vrije Universiteit Brussel, Brussels, Belgium

{rhillewa,bmanderi}@vub.ac.be

² Department of Computing, City University London, United Kingdom
conklin@city.ac.uk

Abstract. The classification of polyphonic music still presents challenges for current music data mining methods. In this paper we explore the performance of classifiers specifically created for melody on the polyphonic classification task. On a small dataset of string quartet movements of Haydn and Mozart, the melodic n -gram model outperforms the melodic global feature model for composer recognition. Furthermore, a simple model that combines the predictions made from different instrumental parts outperforms models created from any single voice. The results indicate that models taking into account polyphonic information achieve higher classification accuracy.

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 [1] we report the results of a thorough comparison of these types of models for the task of folk song classification on a large monophonic dataset, and it appears that the n -gram models are always outperforming the global features in this case. 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 [2–4], but only few efforts have been made to develop polyphonic models. In [5], 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 [6], 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.

In this paper, we will investigate how melodic models perform on single voices extracted from this Haydn/Mozart dataset. Taking into account the small amount of data, 107 movements in total, we will not be able to have very conclusive results, unless the discrepancy between the model performances is large enough. However, we hope to be able to find a preliminary result suggesting that n -gram models are still performing best. A second interest of this research is to verify the latent assumption that the voice carrying the melody, in this case the first violin, is the most characteristic for a composer.

2 Methods

In this section we describe the dataset used for our experiments and we will give a short overview of the different types of models that we will compare.

2.1 Dataset and music representation

The Haydn/Mozart dataset contains 54 string quartet movements from Haydn and 53 string quartet movements from Mozart, as selected by [6]. The original scores in `**kern` format were found on the website of the Center for Computer Assisted Research in the Humanities at Stanford University [7]. 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. To enable the use of melodic 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. A fifth artificial dataset, Concat, was made by concatenating all the voices one after the other, in order to get richer statistics.

2.2 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. We chose three global feature sets that we discuss extensively in [1]. All 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 (see Table 1) :

- The *Alicante* set of 28 global features, proposed by P.J. Ponce de León and J.M. Iñesta in [2]. This set was applied to classification of MIDI tunes in jazz/classical/pop genres. From this set, we implemented the top 12 features that they selected for their experiments.

Global feature set	# features	pitch	duration
Alicante	12	7	2
Jesser	39	31	6
McKay	60	35	5
Joined	111	73	13

Table 1. Global feature sets used in our experiments. The last two columns show the number of features that are derived from pitch and duration.

- The *Jesser* set, containing 39 statistics designed by B. Jesser [8]. The pitch-based features are very basic relative interval counts, like “amajsecond”, measuring the fraction of ascending major seconds, for all ascending and descending intervals in the range of the octave.
- The *McKay* set of 101 global features [5], which were used in the winning 2005 MIREX symbolic genre classification experiment and computed with McKay’s software package jSymbolic [9]. Since many features were superfluous for this analysis of monophonic single voices, we could reduce the set down to 60 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 might improve the results, as was the case for the folk songs in [1].

2.3 n -gram models

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 $\overline{e_{i-1}} = [e_1, e_2, \dots, e_{i-1}]$ of an event e_i is usually limited to a short suffix $[e_{i-n+1}, \dots, e_{i-1}]$, 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:

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

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. To reduce the sparseness of the n -gram counts, we do not model the pitch or duration directly, but we first map the music events into melodic intervals.

	Global feature models				n -gram models
	Alicante	Jesser	McKay	Joined	$n = 3$, mel. interval
Violin 1	57.9	64.5	63.6	63.6	73.0
Violin 2	54.2	51.4	58.9	53.3	58.9
Viola	51.4	44.9	51.4	55.1	55.1
Cello	68.2	56.1	61.7	60.7	66.4
Concat	57.9	60.7	N/A	N/A	72.0

Table 2. Classification accuracies of the global feature sets on the *Europa-6* collection, obtained by leave-one-out cross validation. Global features were evaluated with a 3-nearest neighbours classifier.

2.4 Combining voice models

Confronted with (voiced) polyphonic data, a straightforward classification method using melodic models is to make the naive assumption that voices are independent of one another. Therefore, we simply combine the models $p_j(\overline{e}_\ell)$ for individual voices j by multiplication of probabilities:

$$p(\overline{e}_\ell) = \prod_{j=1}^k p_j(\overline{e}_\ell) \quad (2)$$

For each voice, a separate model is built, and a new piece is then simply assigned to the class with the highest piece probability according to (2).

3 Results

In this section we describe the experimental results for the classification of the Haydn and Mozart string quartet movements. Since this is a preliminary study, we restricted ourselves to the use of very simple models for both the global features and the n -gram models. To evaluate the global feature sets, we used the k -nearest neighbours classifier with $k = 3$ from the Weka machine learning toolbox [10, 11], obtaining classification accuracies with leave-one-out cross validation. For the n -gram model, we use a simple trigram model of the melodic intervals, which was also cross validated with a leave-one-out scheme. The results are reported in Table 2. As a baseline, we keep in mind that the classes are very equally distributed (54 Haydn, 53 Mozart), which means that 50.5% can be achieved by always predicting Haydn. We see that these results confirm that n -gram models tend to have a higher classification accuracy, except for the Cello dataset on which the Alicante features perform remarkably well. Looking at the global features only, we note that the McKay features yield the best results on average, which had also been found on the folk tune dataset in [1]. Unfortunately, computing the McKay features on the concatenated data was practically impossible with the jSymbolic tool due to computation time. The results on the

Voices	McKay	3-grams
$k = 2$ (Violin 1, Cello)	68.2	79.4
$k = 4$ (all voices)	69.2	75.4

Table 3. The results of the combined voice models for the McKay features and the 3-gram model. $k = 2$ is the combination of Violin 1 and *Cello*, whereas $k = 4$ is the combination of all four voices.

Joined feature set were rather disappointing, but we believe this is due to the large number of features (111) compared to the number of instances (107).

Looking at the average performance on every dataset, we see that best results were achieved on Violin 1, which indicates that the first violin in a Haydn or Mozart string quartet is indeed the most characteristic voice. The other voices seem to be ordered consistently as the cello, then the second violin and the viola at last. Concatenating the voices does not seem to be effective to improve the performance, since the results on the Concat data set are always slightly lower than on the first violin.

To combine the voice models, we have chosen to compare the 3-gram model with the best global feature set on average, the McKay set. We report two voice combinations in Table 3, one with $k = 2$ picking the best two voices Violin 1 and Cello, and one with $k = 4$ combining the four voices of the string quartets. There is a remarkable gain on the performance for both methods, and especially with the trigram models of the first violin and the cello we get an accuracy of 79.4%. This indicates that using polyphony can actually lead to an improvement.

4 Conclusions and Future Work

This paper has applied melodic classification models to the task of composer recognition in polyphonic music, specifically Haydn and Mozart string quartet movements. Several conclusions emerge: that models based on the first violin are generally most predictive of Haydn versus Mozart; that a melodic n -gram model outperforms global feature models on the first violin; and that a simple voice combination scheme leads to an improvement over individual voices. These results are preliminary, since the dataset is so small. They should also be validated by pushing both the global feature and the n -gram models, by performing feature selection methods and more powerful classifiers on the global features, and by using more powerful n -gram models [1]. The models used in this paper may be called *pseudo-polyphonic*, since they are still melodic models, however they have been adapted for application to voiced polyphonic music. We intend to develop a truly polyphonic event model, taking into account vertical relationships between overlapping notes, which might then be compared with polyphonic global features.

References

1. Hillewaere, R., Manderick, B., Conklin, D.: Global feature versus event models for folk song classification. In: ISMIR 2009: 10th International Society for Music Information Retrieval Conference, Kobe, Japan (2009)
2. Ponce de León, P.J., Iñesta, J.M.: Statistical description models for melody analysis and characterization. In: Proceedings of the 2004 International Computer Music Conference. (2004) 149–156
3. Conklin, D.: Melodic analysis with segment classes. *Machine Learning* **65**(2) (2006) 349–360
4. Li, M., Sleep, R.: Melody classification using a similarity metric based on Kolmogorov complexity. *Sound and Music Computing* (2004)
5. McKay, C., Fujinaga, I.: Automatic genre classification using large high-level musical feature sets. In: Proceedings of the International Conference on Music Information Retrieval, Barcelona, Spain (2004) 525–530
6. van Kranenburg, P., Backer, E.: Musical style recognition - a quantitative approach. In: Proceedings of the Conference on Interdisciplinary Musicology (CIM). (2004) 106–107
7. <http://www.ccarh.org>
8. Jesser, B.: *Interaktive Melodieanalyse*. Peter Lang, Bern (1991)
9. <http://jmir.sourceforge.net/jSymbolic.html>
10. <http://www.cs.waikato.ac.nz/ml/weka/>.
11. Witten, I., Frank, E.: *Data Mining: Practical machine learning tools and techniques*. 2nd edition. Morgan Kaufmann, San Francisco (2005)