Music style recognition with language models – beyond statistical results

María Hontanilla, Carlos Pérez-Sancho, and Jose M. Iñesta

University of Alicante, Spain
mariahontanilla@yahoo.es, {cperez,inesta}@dlsi.ua.es
http://grfia.dlsi.ua.es

Abstract. In previous works we have shown that $n-$gram language models may be successfully applied to style recognition tasks. Now we present a different study, beyond the purely numerical results, examining more closely some of those results. We have found that some of the compositions might be considered as 'outliers' from a musical point of view, and newer experiments allow us to confirm those musical analyses, showing that our models' musical features are useful for this task.

Keywords: $n-$gram language models, style recognition, composer recognition

1 Introduction

Music style recognition has been an increasing research field in the last years. Up to now, most of the studies done aim to perform a better classification task, according to the success rate.

Sometimes it is difficult, even to expert musicians, to distinguish between composers when their styles are quite similar, often because the compositions belong to the same stylistic context or time, but also because sometimes composers wrote some of their pieces as an homage to other composers, writing in their style. Apparently, the classification should be easier between composers that are further in time than between those temporally closer. Particularly difficult is the known task of distinguishing between Mozart and Haydn: both composers not only shared the same style, but they even dedicated some of their string quartets to the other. A quiz with human listeners performed by the CCARH center at Stanford University [4] shows that non-experts where capable of properly identifying the composer of a 51% of the pieces, whereas this figure raises to only a 66% for self-reported expert listeners. In another recent work [5], the authors reached 80.4% success rate for this task, using a visual representation of musical scores and support vector machines.

When doing classification, maximizing the success rate is usually the target. In this paper we take a second view over the results, trying to find, if possible, a musical explanation to them. Starting from previous works [1], we take a closer look at some of the individual musical pieces, finding that, in some cases, the classifier grounds its decision on very small differences among several models. So the question is: should those results be rejected? On the other hand, we have observed that some of the results seem not to correspond with the general style of the composer, i.e., there are some compositions whose numerical results are far from the mean of the rest of pieces in the training corpus.

These observations have led us to another important question: are those pieces that seem to be quite different from the composer's style different from a musical point of view too? A simple musical comparison has shown that, in some cases, some of those pieces are atypical within the whole musical production of the composer. Then, we have performed new $n-$gram experiments in order to support this musical analysis. The results allow us to verify that, indeed, they agree with this musical analysis.

As an example, we will show two of the examples observed: the first movement from Mozart's string quartet KV 158 as an outlier, and the fourth movement from quartet KV 168 as a tie.

2 Corpora and methodology

2.1 Methodology

From MIDI files, simple musical features (relative pitch intervals and duration ratios) are extracted and converted to ASCII characters, following [2], so that every MIDI file becomes a character sequence. For details on the method, the reader is referred to [1]. Although the MIDI files considered are polyphonic, they are structured in different tracks per voice, so for this particular study, as explained below, only the soprano track (upper voice) will be considered and encoded. Then, an $n-$gram model is built with all the text sequences of every composer, following a leave-one-out scheme. The smoothing method applied is a simple linear interpolation with models of lower $n-$order:

$$\begin{aligned} p_{\mathrm{I}}(w_i|w_{i-n+1}\cdots w_{i-1}) = {} & \lambda_n p_{\mathrm{V}}(w_i|w_{i-n+1}\cdots w_{i-1}) \\ & + \lambda_{n-1} p_{\mathrm{V}}(w_i|w_{i-n+2}\cdots w_{i-1}) \\ & + \cdots \\ & + \lambda_1 p_{\mathrm{V}}(w_i) \\ & + \lambda_0 p_{\mathrm{U}}(w_i) \end{aligned} \tag{1}$$

where p_{I} stands for the interpolated probability of the $n-$gram $(w_{i-n+1}\cdots w_i)$, being w_i a word of the music word sequence $w = w_1\cdots w_k$, p_{V} is the maximum likelihood estimator, and p_{U} is the uniform probability distribution. The weights $\lambda_n\cdots\lambda_0$ are adjusted using a validation set. For that, the whole data set for a particular author is divided in 5 parts: 4 parts are used for training and 1 part for validation. This is done 5 folds, obtaining 5 different models.

For a given new target work w, its perplexity against each model c is computed as:

$$PP_c(w) = \sqrt[k]{\frac{1}{\prod_{i=1}^{k} p_I(w_i|w_{i-n+1}\cdots w_{i-1})}} \tag{2}$$

These results are averaged, obtaining a mean perplexity and a deviation for the work.

2.2 Corpora

In our previous work [1] five composers were studied and their works compared. The works used were: 52 from G. Ph. Telemann (1681–1767), 51 from G. F. Händel (1685–1759), and 75 from J. S. Bach (1685–1750) from the Baroque style; and 46 from W. A. Mozart (1756–1791) and 49 from F. J. Haydn (1732–1809) from the Classical style. This corpus, which is varied in instrumentation, musical form, and number of pieces, is the same as in [3]. For the present analysis we have focused in the results obtained for individual works, trying to identify those that deserve a closer study.

As shown in [1], the upper voice seems to be useful enough to perform a good classification task, though its results are lower than those achieved when using all the instrument's information. For simplicity in the music analysis, we have done the new experiments using information from the upper voices only.

3 Results

Figure 1 shows the results for every piece of one of the models for the same composer. As an example, we show the results of Mozart using a *decoupled* representation for pitch and duration (both properties for each note are represented independently) and $n-$gram length $n = 3$. In this graph, the horizontal lines represent the mean perplexity and the vertical ones represent the range of the standard deviation. The result labeled `MODEL` is the result of the whole model, followed by the individual results of every test piece in the model.

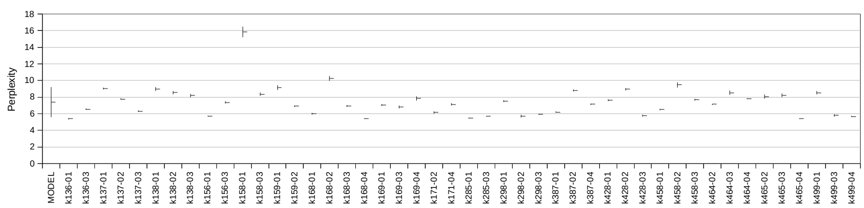

Fig. 1. Graphical comparison of the perplexity of every Mozart's quartet in the training set against a model built with the rest of Mozart's works in it, using independent representations for pitch and duration, and $n = 3$.

3.1 Mozart's string quartet KV 158, 1st movement

Here we can see that there is a piece (KV 158-01) whose result is significantly appart from the whole model (we consider to be significantly distant those pieces whose mean is more than $2 \times stdev$ from the mean of the model). Therefore, the question arises as to whether it is actually a musically atypical piece in Mozart's style or it has some particular features not present in the rest of this particular corpus. We have observed that this is indeed a non typical Mozart's piece, particularly in the rhythm of the main melodic motif (Figure 2), which was not an usual rhythm in that period, and that it would rather seem to be a sort of musical innovation or a kind of surprise tried by Mozart in this first movement of the string quartet.

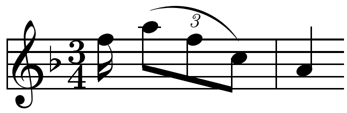

Fig. 2. Main melodic motif from the beginning of the first movement of Mozart's string quartet KV 158 and its rhythmic representation. The characters are the encoding [2] of the inter-onset ratios of every pair of consecutive notes (e.g. 'Z' encodes the ratio = 1 of a pair of equal duration notes).

In order to verify that the anomaly is found in the rhythm, we have performed two additional experiments using both the melodic and rhythmic information alone, and we have compared their results with the previous experiment using them combined with the decoupled representation. These results are shown in Table 1, where it can be observed that the rhythmic feature is the discordant one, as its perplexity is much further from the whole model than when using just pitch intervals.

Table 1. Mean and standard deviations of the perplexity obtained by the model and the first movement of the string quartet KV 158, using intervals only, durations only, and both combined.

	3−grams		
	Intervals	Duration ratios	Intervals and duration ratios
Model	10.6 ± 1.9	**3.8 ± 1.3**	7.4 ± 1.8
KV 158-01	14.2 ± 0.4	**10.0 ± 0.6**	15.9 ± 0.6

We have additionally done a ranking of the rhythmic 3-grams generated from this piece. We have found that, besides the expected most frequent 3-gram (`Z Z`

Z, representing three consecutive notes of the same duration), the two other most frequent 3-grams in this movement belong to this rhythmic motif, whereas they are extremely rare in the whole model. These frequencies are shown in Table 2 and allow us to state that the peculiarity of this piece lies in the rhythm, which is not a usual rhythm in Mozart's language.

Table 2. Frecuencies for the most seen rhythmic 3-grams in the 1st movement of Mozart's quartet KV 158 and for the same 3-grams in the whole Mozart corpus.

	Frequency (%)	
3-gram	KV 158-01	All
Z Z Z	9.9	37.1
Z Z Y	6.8	1.8
C Z Z	6.5	0.1

3.2 Mozart's string quartet KV 168, 4th movement

On the other hand, when comparing every piece against every model, we can see that some of the classification decisions are taken by a very small difference between the models. We wonder whether these decisions should actually be taken into account, or whether they should be addressed in a somehow different way. As an example, we show the results of the fourth movement of Mozart's string quartet KV 168 (Figure 3) against all the five composers' models.

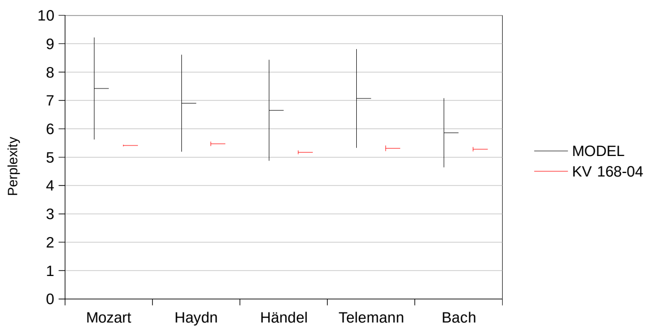

Fig. 3. Comparison of 4th movement from Mozart's quartet KV 168 against every model, using pitch and duration independent representations and $n = 3$. Model perplexities are the averages and deviations for all the Mozart pieces in the dataset.

This movement is actually a fugue, and this might be the reason why the system founds itself in trouble to distinguish one of the composers as the actual author. This is probably due to, although the fugue was a musical form specially preferred in the Baroque period, other composers from different periods have used it as well in their compositions. In this case, the system succeeds in recognizing a style which is closer to the Baroque period than to Mozart's and Haydn's Classical style, assigning a lower perplexity to these composers, but fails in selecting its actual author.

4 Conclusions and future work

From the results shown in this paper, we think that it is desirable going beyond the simple statistical analysis when trying to classify musical compositions in the style of the studied composers. We have realized that the classification of some of the pieces is done with a very little difference from one model to another, and also that there are some pieces that do not fit well in the model built from their author. This shows that it is quite difficult to build a model that is general enough to capture the style of a composer, while being able to identify the subtleties of each individual piece at the same time.

However, a musical analysis on a few pieces has shown that the numerical results for them agree with these analysis indeed, so we think that our system is still able to capture some of the musical features characterizing music styles. Nevertheless, deeper research with other musical compositions needs to be done in this way, in order to study whether these results might be generalized and, if so, to open a discussion about what to do with this kind of pieces when doing classification tasks. Another open question to be addressed is whether or not these models are able to capture higher level stylistic traits, such as musical form, in other to answer some issues that have arised during this study.

References

1. Hontanilla, M., Pérez-Sancho, C., and Iñesta, J. M. (2011) Composer recognition using language models. In: Proc. of Signal Processing, Pattern Recognition, and Applications (SPPRA 2011), pp. 76–83.
2. Doraisamy, S., Rüger, S. (2003) Robust polyphonic music retrieval with n-grams. Journal of Intelligent Information Systems, **21**(1): 53–70.
3. van Kranenburg, P., Backer, E. (2004) Musical style recognition - a quantitative approach. In: Proc. of the Conference on Interdisciplinary Musicology (CIM 2004), pp. 106–107.
4. The Haydn/Mozart String Quartet Quiz. [online] Available at: http://qq.themefinder.org/ [Accessed Sep. 2017].
5. Velarde, G., Weyde, T., Cancino Chacón, C., Meredith, D., Grachten, M. (2016) Composer recognition based on 2D–filtered piano–rolls.In. Proc. of the 17th International Sociecy for Music Information Retrieval Conference (ISMIR 2016), pp. 115–121.