Feature-driven recognition of music styles

Pedro J. Ponce de León and José M. Iñesta

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante,
Ap. 99, E-03080 Alicante, Spain
{pierre,inesta}@dlsi.ua.es

Abstract. In this paper the capability of using self-organising neural maps (SOM) as music style classifiers of musical fragments is studied. From MIDI files, the monophonic melody track is extracted and cut into fragments of equal length. From these sequences, melodic, harmonic, and rhythmic numerical descriptors are computed and presented to the SOM. Their performance is analysed in terms of separability in different music classes from the activations of the map, obtaining different degrees of success for classical and jazz music. This scheme has a number of applications like indexing and selecting musical databases or the evaluation of style-specific automatic composition systems.
Keywords: Multimedia applications, computer music, self-organising maps, feature selection, content-based information retrieval.

1 Introduction

The automatic machine learning and pattern recognition techniques, successfully employed in other fields, can be also applied in music analysis. One of the tasks that can be posed is the modelization of the music style. Immediate applications are the classification, indexation and content-based search in digital music libraries, where digitised (MP3), sequenced (MIDI) or structurally represented (XML) music can be found. The computer could be trained in the user musical taste in order to look for that kind of music over large musical databases. Such a model could also be used in cooperation with automatic composition algorithms to guide this process according to a stylistic profile provided by the user.

Our aim is to develop a system able to distinguish musical styles from a symbolic representation of a melody using musicological features: melodic, harmonic and rhytmic ones. Our working hypothesis is that melodies from a same musical genre may share some common features that permits to assign a musical style to them. For testing our approach, we have initially chosen two music styles, jazz and classical, for our experiments. We will also investigate whether such a representation by itself has enough information to achieve this goal or, on the contrary, also timbric information has to be included for that purpose.

The key point of this work is to test the ability of self-organising maps (SOM) [1], to automatically perform this task. SOM are neural methods able to obtain approximate projections of high-dimensional data distributions in low-dimensional spaces, usually bidimensional. With the map, different clusters in the input data can be located. These clusters can be usually semantically labelled to characterise the training data and also hopefully future new inputs.

1.1 Related work

A number of recent papers explore the capabilities of SOM to analyse and classify music data. Rauber and Frühwirth [2] pose the problem of organising music digital libraries according to sound features of musical themes, in such a way that similar themes are clustered, performing a content-based classification of the sounds. Whitman and Flake [3] present a system based on neural nets and support vector machines, able to classify an audio fragment into a given list of sources or artists. Also in [4], the authors describe a neural system to recognise music types from sound inputs. In [5] the authors present a hierarchical SOM able to analyse time series of musical events and then discriminate those events in a different musical context. In the work by Thom [6] pitch histograms (measured in semitones relative to the tonal pitch and independent of the octave) are used to describe blues fragments of the saxophonist Charlie Parker. The pitch frequencies are used to train a SOM. Also pitch histograms and SOM are used in [7] for musicological analysis of folk songs.

These works pose the problem of music analysis and recognition using either digital sound files or symbolic representations as input. The approach we propose here is to use the symbolic representation of music that will be analysed to provide melodic, harmonic and rhythmic descriptors as input to the SOM (see Fig. 1) for classification of musical fragments into a, initially reduced, set of styles. We use standard MIDI files as the source of monophonic melodies.

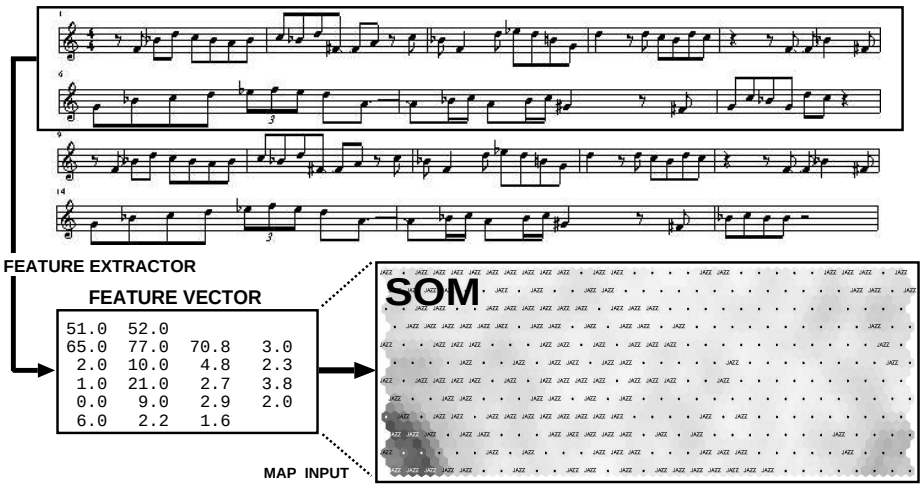

Fig. 1. Structure of the system: musical descriptors are computed from a window 8-bar wide and provided to the SOM for training and classification. Once trained, a style label is assigned to the units. During classification, the label of the winning unit provides the style to which the music fragment belongs to. This example is based on the Charlie Parker's jazz piece "Dexterity".

2 Methodology

The monophonic melodies are isolated from the rest of the musical content in the MIDI files. This way we get a sequence of musical events that can be either notes or silences. Other kind of MIDI events are filtered out. Each note can take a value from 0 to 127 (the pitch) and the duration is the distance in pulses from the event that onsets the sound of a note to the finishing event.

Here we will deal only with melodies written in 4/4. In order to have more restricted data, fragments of 8 bars are taken (enough to get a good sense of the melodic phrase in the context of a 4/4 signature). For this, each melody sequence has been cut into fragments of such duration.

We have chosen a vector of musical descriptors of the melodies as the input for the SOM, instead of the explicit representation of the melodies. Thus, a description model is needed. Firstly, three groups of features are extracted: melodic, harmonic and rhythmic properties. Then, from this initial set of features a selection procedure will be performed based on their values for the weight vectors of the trained SOM. This way, some reduced models have been constructed and their classification ability tested.

The features are computed using a time resolution of $Q = 48$ pulses per bar[1]. The initial set of 22 musical descriptors is:

- Overall descriptors:
  - Number of notes and number of silences in the fragment.
- Pitch descriptors:
  - Lowest, highest (provide information about the pitch range of the melody), average, and standard deviation (provide information about how the notes are distributed in the score).
- Note duration descriptors (these descriptors are measured in pulses):
  - Minimum, maximum, average, and standard deviation.
- Silence duration descriptors (in pulses):
  - Minimum, maximum, average, and standard deviation.
- Interval descriptors (distance in pitch between two consecutive notes):
  - Minimum, maximum, average, and standard deviation.
- Harmonic descriptors:
  - *Number of non diatonic notes.* An indication of frequent excursions outside tonality (extracted from the MIDI file) or modulations.
  - *Average degree*[2] *of non diatonic notes.* Describes the kind of excursions.
  - *Standard deviation of degrees of non diatonic notes.* Indicates a higher variety in the modulations.
- Rhytmic descriptor: *number of syncopations*: notes not beginning at the rhythm beats but in some places between them (usually in the middle) and that extend across beats.

[1] This is call quantisation. $Q = 48$ means that if a bar is composed of 4 times, each time can be divided, at most, into 12 pulses.

[2] Measured in distance in pitch from the key note of the diatonic scale.

2.1 SOM implementation

For SOM implementation and graphic representations the SOM_PAK software [8] has been used. For the experiments, a hexagonal geometry for unit connections and a bubble neighbourhood for training have been selected. The value for this neighbourhood is equal for all the units in it and decreases as a function of time.

The maps are displayed using the U-map representation, where the units are represented by hexagons with a dot or label in their centre. The grey level of unlabelled hexagons represents the distance between neighbour units (the clearer the closer they are). For the labelled units is an average of the neighbour distances. This way, clear zones are clusters of units, sharing similar weight vectors. The labels are a result of calibrating the map with a series of test samples and indicate the class of samples that activates more times each unit.

2.2 Feature selection procedure

The utilized features have been designed according to those used in musicological studies but there is no theoretical support for them. We have devised a selection procedure in order to keep those descriptors that actually contribute to make the classification. The procedure is based on the values for the features in the weight vectors of the trained SOMs. The maps are trained and labelled (calibrated) in an unsupervised manner (see Fig 2-a for an example. We try to find which descriptors provide more useful information for the classification. Some descriptor values for the weight vectors correlate better than others with the label distribution in the map. It is reasonable to consider that these descriptors contribute more to achieve a good separation between classes. See Fig. 2-b and 2-c for descriptor planes that correlate and that do not with the class labels.

Consider that the N descriptors are random variables $\{x_i\}_{i=1}^{N}$ that corresponds to the weight vector components for each of the M units in the map. We drop the subindex i for clarity, because all the discussion is related to each descriptor. We will divide the set of M values for each descriptor into two subsets: $\{x_j^C\}_{j=1}^{M_C}$ are the descriptor values for the units labelled with the classical style and $\{x_j^J\}_{j=1}^{M_J}$ are those for the jazz units, being M_C and M_J the number of units labelled with classical and jazz labels, respectively. We want to know whether these two set of values follow the same distribution or not. If false, it is an indication that there is a clear separation between the values of this descriptor for the two classes, so it is a good feature for classification and should be kept in the model and otherwise it does not seem to provide separability to the classes.

We have considered that both sets of values hold normality conditions and the following statistical for sample separation has been applied:

$$z = \frac{|\bar{x}_C - \bar{x}_J|}{\sqrt{\frac{s_C^2}{M_C} + \frac{s_J^2}{M_J}}} \quad , \tag{1}$$

where $\bar{x}_C$ and $\bar{x}_J$ are the means, and s_C^2 and s_J^2 the variances for the descriptor values for both classes. The larger the z value is, the higher the separation between both sets of values is for that descriptor. This value permits to order the

descriptors according to their separation ability and a threshold can be established to determine which descriptors are suitable for the model. This threshold, computed from a t-student distribution with infinite degrees of freedom and a 99.5% confidence interval, is $z = 2.81$.

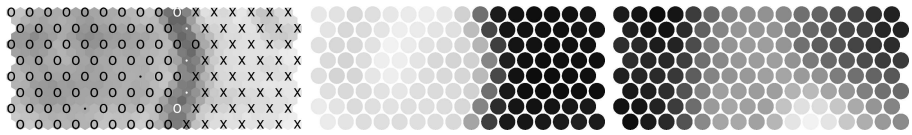

Fig. 2. Contribution to classification: (a:left) callibrated map ('X' and 'O' are the labels for both styles); (b:center) weight space plane for a feature that correlates with the areas; (c:right) plane for a feature that does not correlate.

3 Experiments and results

As stated above, we have chosen two given music styles: jazz and classical for testing our approach. The jazz samples were taken from jazz standards from different jazz styles like be-bop, hard-bop, big-band swing, etc., and the melodies were sequenced in real time. Classical tunes were collected from a number of styles like baroque, romantic, renaissance, impressionism, etc.

From the MIDI files, 430 jazz and 522 classical melodic samples have been extracted, all of them made up of 8 bars. From them, the 22 descriptors were computed. Two different SOM sizes have been used. Their parameters are displayed in the table below. Those maps have been trained with different subsets of descriptors.

map	coarse training			fine training		
size	iterations	neighb.rad.	learn.rate	iterations	neighb.rad.	learn.rate
16×8	3,000	12	0.1	30,000	4	0.05
30×12	10,000	20	0.1	100,000	6	0.05

After training and labelling, maps like that in figure 3 have been obtained. It is observed how the labelling process has located the jazz labels mainly on the left zone, and those corresponding to classical melodies on the right. Some units can be labelled for both music styles if they are activated by fragments from both styles. In these cases there is always a winner label (the one displayed) according to the number of activations. The proportion of units with both labels is the overlapping degree, that for the presented map was very low (8.0 %), allowing a clear distinction between styles.

In the Sammon projection of the map in figure 3 a knot separates two zones in the map. The zone at the left of the knot has a majority presence of units labelled with the jazz label and the zone at the right is mainly classical.

3.1 Feature selection results

Firstly we have trained the maps with the whole set of 22 features. This way a reference performance for the system is obtained. In addition, we have trained other maps using just melodic descriptors and also melodic and harmonic ones.

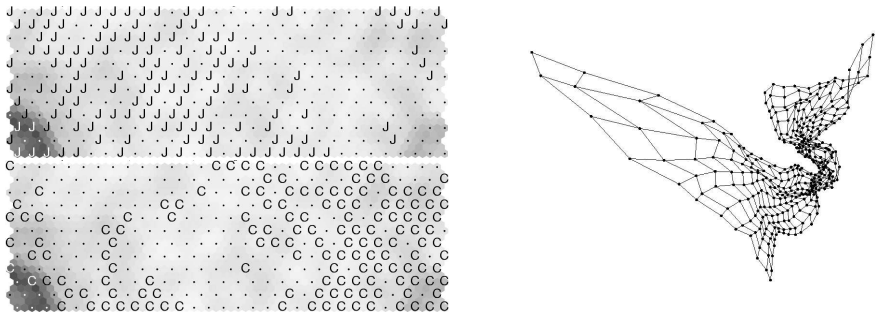

Fig. 3. Left: SOM map after being labeled with jazz (top) and classical (down) melodies. Note how both classes are clearly separated. Right: Sammon projection of the SOM, a way to display in 3D the organisation of the weight vector space.

We get a set of five trained maps in order to study the values of the weight space planes, using the method described in 2.2. This number of experiments has been considered enough due to the repetitivity of the obtained results. For each experiment we have ordered the descriptors according to their value for z_i (see eq. 1). In table 1 the feature selection results are displayed, including what descriptors have been considered for each model according to those results. Each model number denotes the number of descriptors included in that model. We have chosen four reduced model sizes: 6, 7, 10 and 13 descriptors. Descriptors with no entry in the order column are those having a z_i value under the threshold. Entries marked with a 'x' are not considered in that experiment.

Table 1. Feature selection results. For each descriptor the ordered position according to the statistical z_i for all the experiments and the average position are displayed. In the rightmost column, the models in which each descriptor is included are also presented.

descriptor	order in experiments					avg. order	models
Syncopation	x	x	1	x	1	1.0	7+10+13
Highest pitch	2	1	2	1	2	1.6	6+7+10+13
Max. interval	1	3	5	4	3	3.2	6+7+10+13
Dev. note duration	6	4	4	3	5	4.4	6+7+10+13
Max. note duration	7	5	3	5	4	4.8	6+7+10+13
Dev. pitch	3	2	6	7	7	5.0	6+7+10+13
Avg. note duration	4	7	7	2	6	5.2	6+7+10+13
Avg. pitch	9	8	8	8	8	8.2	10+13
Dev. interval	5	6	10	11	9	8.2	10+13
number of notes	8	9	9	6	10	8.4	10+13
number of silences	10	10	11	10	12	10.6	13
Min. note duration	11	11	12	9	–	10.8	
Min. silence duration	–	–	–	–	11	11.0	
Min. interval	12	12	13	12	13	12.4	13
Avg. interval	–	13	–	–	–	13.0	
Dev. non-diatonic degrees	–	–	14	x	–	14.0	
Num. non-diatonic notes	13	14	16	x	14	14.3	13
Lowest pitch	–	–	15	13	15	14.3	
Max. silence duration	–	–	–	–	–	–	
Avg. silence duration	–	–	–	–	–	–	
Dev. silence duration	–	–	–	–	–	–	
Avg. non-diatonic degrees	–	–	–	x	–	–	

3.2 Classification

For obtaining reliable results a scheme based on *leave-k-out* has been carried out. In our case $k = 10\%$ of the size of the whole database. This way, 10 sub-experiments were performed for each experiment and the results have been averaged. In each experiment the training set was made of a different 90% of the total database and the other 10% was kept for testing. The results are presented in table 2. The results in the table are those obtained in the next experiments:

- All descriptors: all the 22 melodic, harmonic and rhythmic features.
- 6 descriptors: max.pitch, max.interval, note number std.deviation, note number max., pitch std.deviation, and note number mean.
- 7 descriptors: all above plus syncopation.
- 10 descriptors: all above plus pitch mean, interval std.dev. and note number.
- 13 descriptors: all above plus silence number, min.interval, non-diatonic num.

The data presented in the table are successful classification rates for jazz and classical. Each model has been evaluated with the two different size SOM, and in each case the best partition and the average results for the 10 partitions of the leave-k-out experiment are displayed.

Table 2. Classification results (success rates are in percentages). "Best" results are not neccesarely with the same map. "Best" results for "Both" styles are averaged for jazz and classical styles with a particular map.

	JAZZ				CLAS				BOTH			
	16x8		30x12		16x8		30x12		16x8		30x12	
Descr.	BEST	AVG.	BEST	AVG.	BEST	AVG.	BEST	AVG.	BEST	AVG.	BEST	AVG.
All	89.8	72.7	87.8	61.2	93.2	79.6	85.1	70.8	90.8	76.1	80.7	66.0
6	98.0	79.4	81.6	68.4	95.2	82.1	90.5	**89.3**	92.5	80.8	78.3	73.3
7	96.0	**81.8**	83.7	74.1	97.3	86.5	97.3	76.6	96.0	**84.2**	85.1	75.4
10	98.0	78.8	87.8	63.3	96.0	82.7	90.5	74.6	88.8	80.7	89.2	68.9
13	87.8	72.0	89.8	67.1	97.3	82.6	85.1	68.8	84.4	77.3	78.0	68.0

The best average performances were consistently obtained with the smaller map, with a success classification rate around 80 %. The best average results were obtained for that map when using the 7-descriptor model (84.2 %). It is observed that 6-descriptor model performance are systematically improved when syncopation is included in the 7-descriptor model. In some experiments even a 98.0 % of success (96.0 % for both styles) has been achieved. The inclusion of more descriptors in the model worsens the results and the worst case is when all of them are used (76.1 % and 66.0 %).

4 Conclusions and future works

We have shown the ability of SOM to map symbolic representations of melodies into a set of musical styles using melodic, harmonic and rhythmic descriptions.

The best recognition rate has been found with a 7-descriptor model where syncopation, note duration, and pitch have an important role. The overlapping degree does not seem to be a key point when assessing the quality of a map.

Some of the misclassifications can be caused by the lack of a smart method for melody segmentation. The music samples have been arbitrarily restricted to 8 bars, getting just fragments with no relation to musical motives. This fact can introduce artifacts in the descriptors leading to less quality mappings. The main goal was to test the feasibility of the approach, and average recognition rates above 80% have been achieved, that is very encouraging keeping in mind these limitations and others like the lack of valuable information for this task like timbre.

A number of possibilities are yet to be explored, like the development and study of new descriptors. It is very likely that the descriptor subset models are highly dependent on the styles to be discriminated. To achieve this goal a large music database has to be compiled and tested using our system for multiple different style recognition in order to draw significant conclusions.

Acknowledgements

Thanks to the Spanish CICyT project TAR, code: TIC2000–1703–CO3–02, and to the CICyT TIC2001-5057-E pattern recognition thematic network.

References

1. T. Kohonen. Self-organizing map. *Proceedings IEEE*, 78(9):1464–1480, 1990.
2. A. Rauber and M. Frühwirth. *Automatically analyzing and organizing music archives*, pages 4–8. 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001). Springer, Darmstadt, Sep 2001.
3. Brian Whitman, Gary Flake, and Steve Lawrence. Artist detection in music with minnowmatch. In *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568. Falmouth, Massachusetts, September 10–12 2001.
4. Hagen Soltau, Tanja Schultz, Martin Westphal, and Alex Waibel. Recognition of music types. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1998)*. Seattle, Washington, May 1998.
5. O. A. S. Carpinteiro. A self-organizing map model for analysis of musical time series. In A. de Padua Braga and T. B. Ludermir, editors, *Proceedings 5th Brazilian Symposium on Neural Networks*, pages 140–5. IEEE Comput. Soc, 1998.
6. Belinda Thom. Unsupervised learning and interactive jazz/blues improvisation. In *Proceedings of the AAAI2000*, pages 652–657, 2000.
7. Petri Toiviainen and Tuomas Eerola. Method for comparative analysis of folk music based on musical feature extraction and neural networks. In *III International Conference on Cognitive Musicology*, pages 41–45, Jyvskyl, Finland, 2001.
8. T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. Som_pak, the self-organizing map program package, v:3.1. Lab. of Computer and Information Science, Helsinki University of Technology, Finland, April, 1995. `http://www.cis.hut.fi/research/som_pak`.