

# Statistical description models for melody analysis and characterization\*

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

*The analysis of melodies symbolically represented as digital scores (standard MIDI files) is studied. A number of melodic, harmonic, and rhythmic statistical descriptors are computed. Their representation ability is analyzed using the principal components technique (PCA). Also their ability to separate two particular musical styles, like jazz and classical, using a per-feature separability test. A qualitative discussion on such a description model based on these results is included. **Keywords:** Melody description, principal component analysis, music information retrieval.*

## 1 Introduction

For a number of musical information retrieval (MIR) tasks, like style classification or author recognition, it is sometimes desirable to rely on a description of a melody, rather than its explicit representation. When dealing with symbolic representations of music, often one has to use only on melodic properties, like pitch and duration. This is the situation when trying to characterize a melody for which we do not know any information about timbre, and we want to explore the intrinsic properties of the melody to obtain an answer about its quality.

Works based on score-like material (i.e., symbolically represented music) include a work by Dannenberg, Thom, and Watson (1997), where some statistical descriptors, similar to the ones presented here, are used to build interpretation style classifiers for interactive performance systems, reporting more than 98% accuracy with four styles.

Pitch histograms and self-organising maps are used by Toivainen and Eerola (2001) for musicological analysis of folk songs. In a recent work, Cruz-Alcázar et al. (2003) show the ability of grammatical inference methods for modelling musical style. The authors also discuss about the encoding schemes that can be used to achieve the best recognition result.

## 2 Objectives

In this paper a description model based on statistics is presented. Previous work on the application of similar models to music style recognition have been encouraging (see de León and Iñesta (2003)). In the present work we present a more in-deep analysis of the behaviour and information provided by the different studied models.

First, the proposed methodology will be presented, describing the musical data, the description model, the analysis techniques employed, and some music style recognition experiments used to test the model. Second, the analysis and experiment results will be addressed, and finally, conclusions and possible lines of further work will be discussed.

## 3 Methodology

First in this section the music sources from which the datasets are obtained are presented. Second, the details of the statistical feature extraction from the musical data are described. Then, the analysis tools are presented. Finally, the application of the description model to the musical style discrimination task is discussed.

### 3.1 Musical data

MIDI files from jazz and classical music, were used. These styles were chosen due to the general agreement in the musicology community about their definition and limits (Reimann 2003). Classical melody samples were taken from works by Bach, Beethoven, Brahms, Chopin, Dvorak, Grieg, Haendel, Mendehlson, Mozart, Paganini, Schubert, Schumann and Vivaldi. Jazz music samples were standard tunes from well known jazz authors including Miles Davis, Duke Ellington, Bill Evans, Charlie Parker, etc. The corpus is made up of a total of 110 MIDI files, 45 of them being classical music and 65 being jazz music. The length of the corpus is around 10,000 bars (40,000 beats). It is an heterogeneous corpus, not specifically created to fit our purposes but collected from different web sites without any processing before using them.

\* This work was supported by the Spanish CICYT project code TIC2003-08496-C04.

The MIDI files are composed of several tracks, one of them being the melody track from which the input data are extracted<sup>1</sup>. These melody tracks sound like monophonic sequences, but they may not be strictly monophonic (i.e. some overlapping between consecutive notes may exist in the real-time sequenced MIDI tracks). Selecting at every moment the note with the highest pitch, these quasi-polyphonic sequences are converted into actual monophonic sequences. Similar approaches preserving the highest pitch for the melody have been used in Uitdenbogerd and Zobel (1998) and judged to be more effective for melody extraction than other more sophisticated techniques.

Thus, the monophonic melodies consist of a sequence of musical events that can be either notes or silences. Each note can take a value from 0 to 127 (the pitch), encoded together with the MIDI note onset event. Each of these events at time  $t$  has a corresponding note off event at time  $t + d$ , being  $d$  the note duration measured in pulses<sup>2</sup>. Time gaps between a note off event and the next note onset event are silence events.

### 3.2 Description model

A description scheme has been designed based on descriptive statistics that summarise the content of the melody in terms of pitches, intervals, durations, silences, harmonicity, rhythm, etc. This kind of statistical description of musical content is sometimes referred to as *shallow structure description* (Pickens 2001).

Given a melody track, the statistical descriptors are computed from equal length segments, defining a window size of  $\omega$  measures. Once the descriptors of a segment have been extracted, the window is shifted  $\delta$  measures forward to obtain the next segment to be described. Given a melody with  $m > 0$  measures, the number of segments  $s$  of size  $\omega > 0$  obtained from that melody is

$$s = \begin{cases} 1 & \text{if } \omega \geq m \\ 1 + \lceil \frac{m-\omega}{\delta} \rceil & \text{otherwise} \end{cases} \quad (1)$$

Note that, at least, one segment is extracted in any case ( $\omega$  and  $s$  are positive integers;  $m$  and  $\delta$  may be positive fractional numbers). A combination of window size and shifting is denoted as  $\langle \omega, \delta \rangle$ .

Each window size defines a different corpus, since descriptors obtained from segments with different lengths are considered qualitatively different. The  $\delta$  shifting affects the number of samples obtained, but not the intrinsic quality of the descriptors. In this work, the values used to investigate the description model are  $\delta = 1$  and  $\omega = 1, 2, 4, 8, 16, 32, 64$  and  $\infty$  (whole melody).

A corpus is thus made up of vectors of musical descriptors computed from each melody segment available. Each vector

is labelled with the style of the melody that the segment belongs to. An initial set of descriptors has been defined based on three groups of features that assess the melodic, harmonic and rhythmic properties of a musical segment, respectively. From the statistical point of view, descriptors are grouped as counter, range, average, deviation and normality descriptors.

Melodic features have to do with note pitches, durations and silences durations. Silence related information is often not considered when describing melodic lines, relying exclusively on note related information or, if considered at all, silences work merely as stopwords between melodic phrases. Being an important part of the melody concept, silences are treated as notes with no sound, i.e. only their duration is used, in the description model presented here. Their actual importance relative to other description groups would be discussed in section 4.

Harmonic features are modeled analyzing non-diatonic notes present in the melody. Rhythmic features include Inter Onset Interval (IOI) descriptors and qualitative rhythm descriptors like a syncopation counter.

Thus, the initial description model is made up of the following 28 descriptors (descriptor abbreviations appear between parenthesis):

- General descriptors:
  - *Number of notes* (notNUM), *number of significant silences* (silsigNUM), and *number of not significant silences* (silintNUM). The adjective *significant* stands for silences explicitly written in the underlying score of the melody. In MIDI files, short gaps between consecutive notes may appear due to interpretation nuances like *stacatto*. These gaps (interpretation silences) are not considered significant silences since they should not appear in the score. To make a distinction between kinds of silences is not possible from the MIDI file and it has been made defining a silence duration threshold. This value has been empirically set to a duration of a sixteenth note. All silences with duration greater or equal than this threshold are considered significant.
- Pitch descriptors:
  - *Pitch range* (pchRNG, the difference in semitones between the highest and the lowest note in the melody segment), *average pitch* (pchAVG) relative to the lowest pitch, and *standard deviation of pitches* (pchDEV, provide information about how the notes are distributed in the score).
- Note duration descriptors (measured in pulses and computed using a time resolution of  $Q = 48$  pulses per

<sup>1</sup> Without loosing generality, all the melodies are written in the 4/4 meter.

<sup>2</sup> A *pulse* is the basic unit of time in a MIDI file and is defined by the resolution of the file, measured in pulses per beat.

bar<sup>3</sup>):

- *Range* (durRNG), *average* (durAVG, relative to the minimum duration), and *standard deviation* (durDEV) of note durations.

- Significant silence duration descriptors (in pulses):

- *Range* (dslRNG), *average* (dslAVG, relative to the minimum), and *standard deviation* (dslDEV).

- Inter Onset Interval (IOI) descriptors (an IOI is the distance, in pulses, between the onsets of two consecutive notes<sup>4</sup>):

- *Range* (ioiRNG), *average* (ioiAVG, relative to the minimum), and *standard deviation* (ioiDEV).

- Interval descriptors (distance in pitch between two consecutive notes):

- *Range* (itvRNG), *average* (itvAVG, relative to the minimum), and *standard deviation* (itvDEV).

- Harmonic descriptors:

- *Number of non diatonic notes.* (ndNUM) An indication of frequent excursions outside the song key (extracted from the MIDI file) or modulations.
- *Average degree of non diatonic notes.* (ndAVG) Describes the kind of excursions. This degree is a number between 0 and 4 that indexes the non diatonic notes of the diatonic scale of the tune key, that can be major or minor key<sup>5</sup>.
- *Standard deviation of degrees of non diatonic notes.* (ndDEV) Indicates a higher variety in the non diatonic notes.

- Rhythmic descriptor:

- *Number of syncopations* (syncop): notes that do not begin at the rhythm beats but in some places between them (usually in the middle) and that extend across beats.

- Normality descriptors. They are computed using the D’Agostino statistic for assessing the distribution normality of the  $n$  values  $v_i$  in the segment for pitches, durations, intervals, etc. The test is performed using this equation:

$$D = \frac{\sum(i - \frac{n+1}{2})v_i}{\sqrt{n^3(\sum v_i^2 - \frac{1}{n}(\sum v_i)^2)}} \quad (2)$$

The descriptors of this kind computed for the analysed segment are the normality values of:

- *pitch distribution.* (pchNORM)
- *note duration distribution.* (durNORM)
- *IOI distribution.* (ioiNORM)
- *silence duration distribution.* (dslNORM)
- *interval distribution.* (itvNORM)
- *non-diatonic notes distribution.* (ndNORM)

The development of this descriptor set aimed to have values independent from pitch transposition and duration scaling (only the syncopation counter would be affected), while capturing essential distribution of features in the melody.

From the symbolic point of view, pitch is naturally measured as halftones rather than frequencies. In order to compare pitches, their difference in halftones is computed. This way pitch becomes an arithmetic quantity. Thus, for pitch and interval properties to be independent to transposition, the range descriptors are computed as maximum minus minimum values, and the average-relative descriptors are computed as the average value minus the minimum value.

For durations (note duration, silence duration) or time intervals (IOI), in order to compare them musically, the ratio between their duration values is computed, i.e., a half note is twice a quarter note, a dotted half note is three times a quarter note, and so. Thus, to make note duration, silence duration and IOI descriptors independent from scaling, the range descriptors are computed as the ratio between the maximum and minimum values, and the average-relative descriptors are computed as the ratio between the average value and the minimum value.

This descriptive statistics is similar to histogram-based descriptions used by other authors (Thom 2000; Toiviainen and Eerola 2001) that also try to model the distribution of musical events in a music fragment. Computing the range, mean, and standard deviation from the distribution of musical items like pitches, durations, intervals, IOIs, and non-diatonic notes we reduce the number of features needed (each histogram may be made up of tens of features).

### 3.3 Feature selection

The features described above have been designed according to those used in musicological studies, but there is no theoretical support for their melody characterization capability. To analyze the features for their separability capabilities, samples are grouped by style, giving two different sample classes, jazz and classical. A selection procedure has been applied in order to keep those descriptors that better contribute to the separability between styles. The method assumes feature independence, that is not true in general, but it tests the

<sup>3</sup> This is called quantisation.  $Q = 48$  means that if a bar is composed of 4 beats, each beat can be divided, at most, into 12 pulses.

<sup>4</sup> Two notes are considered consecutive even in the presence of a silence between them.

<sup>5</sup> Non diatonic degrees are: 0: bII, 1: bIII (♯III for minor key), 2: bV, 3: bVI, 4: bVII. The key is encoded at the beginning of the melody track.

separability provided by each descriptor independently, and uses this separability to obtain a descriptor ranking. This per-feature tests can help to identify and discard easily recognizable 'bad' features, keeping more elaborated techniques from unnecessary burden.

Consider the  $M$  descriptors as random variables  $\{x_j\}_{j=1}^M$  whose  $N$  sample values are those of a dataset corresponding to a given  $\omega$  window size. We drop the subindex  $j$  for clarity, because all the discussion applies to each descriptor. We will divide the set of  $N$  values for each descriptor into two subsets:  $\{x_{C,i}\}_{i=1}^{N_C}$  are the descriptor values for classical samples and  $\{x_{J,i}\}_{i=1}^{N_J}$  are those for the jazz samples, being  $N_C$  and  $N_J$  the number of classical and jazz samples, respectively.  $x_C$  and  $x_J$  are assumed to be independent random variables, since both sets of values are computed from different sets of melodies. We want to know whether these random variables belong to the same distribution or not. We have considered that both sets of values hold normality conditions, and assuming that the variances for  $x_C$  and  $x_J$  are different in general, the test contrasts the null hypothesis  $H_0 \equiv \bar{x}_C = \bar{x}_J$  against  $H_1 \equiv \bar{x}_C \neq \bar{x}_J$ . If  $H_1$  is concluded, 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 style classification. Otherwise, it does not seem to provide separability between the classes.

The following statistical for sample separation has been applied:

$$z = \frac{|\bar{x}_C - \bar{x}_J|}{\sqrt{\frac{s_C^2}{N_C} + \frac{s_J^2}{N_J}}}, \quad (3)$$

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. A threshold to decide when  $H_0$  is more likely than  $H_1$ , that is, the descriptor passes the test for the given dataset, must be established. This threshold, computed from a t-student distribution with infinite degrees of freedom and a 99.7% confidence interval, is  $z = 2.97$ . Furthermore, the  $z$  value permits to arrange the descriptors according to their separation ability.

When this test is performed on a number of datasets with different window sizes, a threshold on the number of passed tests can be set as a criterion to select descriptors. This threshold is expressed as a minimum percentage of tests passed. Once the descriptors are selected, a second criterion for grouping them permits to build several descriptor models incrementally. First, selected descriptors are ranked according to their  $z$  value averaged over all tests. Second, descriptors with similar  $z$  values in the ranking are grouped together. This way, several descriptor groups are formed, and new descriptor models can be formed by incrementally combining these groups. See the section 4.2 for the models that have been obtained.

### 3.4 Feature space transformation

The feature analysis discussed in the previous section does not take into account dependencies between descriptors. One technique that deals with the correlation between features to obtain new, uncorrelated features is Principal Component Analysis (PCA). This is achieved transforming the original feature space of  $n$  dimensions into a new space where features are uncorrelated and calculated as lineal combinations of the original ones. PCA finds the orthogonal axes of the new space. The first axis is the one where the variability of the original data is maximum, that is, its direction is that of a line in the original feature space where the projection of sample points gives the maximum dispersion. The second axis is one orthogonal to the first where dispersion is maximum, and so on. The transformation can be expressed as

$$Y = W^T X \quad (4)$$

where  $Y$  are the transformed samples,  $W$  is a  $n \times n$  orthogonal transformation matrix, and  $X$  are the original samples. It can be proven that

$$\begin{aligned} \mu_Y &= W^T \mu_X \\ \Sigma_Y &= W^T \Sigma_X W \end{aligned} \quad (5)$$

where  $\mu_X, \mu_Y$  are the sample means, and  $\Sigma_X, \Sigma_Y$  the covariance matrices in the original and transformed space, respectively.  $\Sigma_Y$  is a  $n \times n$  diagonal matrix, showing that new space features are uncorrelated. The values  $\lambda_i$  in the diagonal of this matrix are the eigenvalues of  $\Sigma_X$ , and the columns of  $W$  are its eigenvectors.  $\Sigma_X$  is a symmetric positive semidefinite matrix, with real eigenvalues. Furthermore, these values are ordered:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \quad (7)$$

corresponding to the variance of the transformed samples in each dimension. Thus, the dimension associated with the first eigenvalue has the greater variance, and the first column of  $W$ —the eigenvector associated with  $\lambda_1$ —defines the direction for this dimension. This is the first principal component—the first axis in the transformed space—and it explains the maximum variance direction of the original samples. The same applies to the other transformed dimensions. The axes in the transformed space are ordered by eigenvalue, in such a way that selecting only the first  $m$  axes that provide almost 100% of the total original variance. Thus, the dimensionality of the original space can be reduced at a low cost. But the interesting point here is that the coefficients of  $W$  can be interpreted to understand how original features contribute to the total variance of the samples. Features corresponding to coefficients with larger values are considered to contribute more to the variance explained by the principal component under study. Section 4.3 discuss the exploration of these coefficients.

### 3.5 Style classification

The description models presented here have been tested on the musical style discrimination problem with a variety of classification techniques, including supervised and unsupervised, parametric and non-parametric methods, like Bayesian,  $k$ -nearest-neighbours ( $k$ NN), linear discriminant analysis (Duda and Hart 1973), and self-organising maps (Kohonen 1990). Some of the results obtained with the Bayesian and  $k$ NN classifiers are presented in this paper, to illustrate the performance of the shallow description approach. See de León and Iñesta (2003) for a detailed discussion about style classification results.

The Bayesian classifier is parametric and, when applied to a two-class problem, computes a discriminant function:

$$g(X) = \log \frac{P(X | \omega_1)}{P(X | \omega_2)} + \log \frac{\pi_1}{\pi_2} \quad (8)$$

for a test sample  $X$  where  $P(X | \omega_i)$  is the conditional probability density function for class  $i$  and  $\pi_i$  are the priors of each class. Gaussian probability density functions for each style are assumed for each descriptor, with means and variances estimated from the training data. The classifier assigns a sample to  $\omega_1$  if  $g(X) > 0$ , and to  $\omega_2$  otherwise.

The  $k$ -NN classifier is a classical non-parametric approach to classification. It uses an Euclidean metrics to compute the distance between the test sample and those in the training set. The style label is assigned to the test sample by a majority decision among the nearest  $k$  training samples (the  $k$ -neighbourhood).

## 4 Experiments and results

### 4.1 Descriptor space

A quick look at the descriptor distributions in the chosen datasets helps to understand what can be expected when assuming normality for them at the feature selection or classification test experiments. Figure 1 shows the notNUM descriptor distribution. It is an example of the kind of distribution encountered among the remaining descriptors. Almost all of them follow Gaussian-like or Poisson-like distributions, except for the normality descriptors that, due to their statistics test nature, have a chi square distribution shape. The normality assumption is therefore a low risk simplification, because the Poisson distribution approximates the gaussian distribution in many cases, especially for large  $\omega$  values.

The other descriptors with a gaussian distribution shape are: pitch and pitch interval descriptors, IOI deviation, and non-diatonic average and deviation. Descriptors with a Poisson shaped distribution include the significant silence counter, note duration and silence duration descriptors, IOI range and average, non-diatonic note counter and syncopation descriptors.

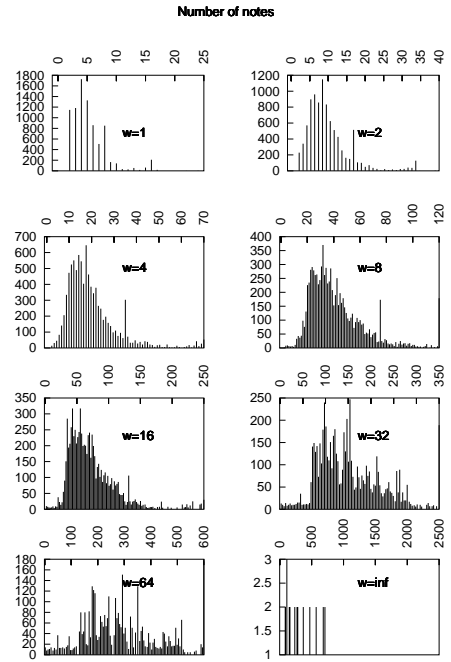


Figure 1: Number of notes for several  $\omega$  values. It has a gaussian-like distribution.

### 4.2 Feature selection results

The feature selection test presented in section 3.3 has been applied to datasets corresponding to 100 randomly selected  $\langle \omega, \delta \rangle$  pairs. This is motivated by the fact that the descriptor computation is different for each  $\omega$  and the set of values is different for each  $\delta$ , so the best descriptors may be different for different  $\langle \omega, \delta \rangle$  values. Thus, the sensitivity of the classification to the feature selection procedure can be analysed, minimising the risk of biasing this analysis in favour of particular  $\langle \omega, \delta \rangle$  values. The range for  $\omega$  is limited from 1 to 100, and the range for  $\delta$  varies from 1 to its corresponding  $\omega$  value. Datasets with less than a hundred sample values have not been taken into account. These datasets are in the range of  $\omega > 50$  and  $\delta > 20$ .

The descriptors were sorted according to the average  $z$  value ( $\bar{z}$ ) computed for the descriptors in the tests. The descriptors are shown sorted in table 1. The  $\bar{z}$  values for all the tests and the percentage of passed tests for each descriptor are displayed. In order to select descriptors, a threshold on the number of passed tests has been set to 95%. This way, those descriptors that failed the separability hypothesis in more than a 5% of the experiments were discarded from the reduced models. Only 12 descriptors out of 28 were selected. In the rightmost column, the reduced models in which the descriptors were included are presented. Each model is denoted with the number of descriptors included in it. Three reduced

descriptor	$\bar{z}$	passed tests	models
Number of notes	22.5	100%	6,10,12
Pitch average	22.3	100%	6,10,12
Pitch range	22.2	100%	6,10,12
Interval range	20.3	100%	6,10,12
Syncopation	19.6	100%	6,10,12
Pitch deviation	18.7	100%	6,10,12
Number of significant silences	14.2	100%	10,12
Interval distrib. normality	14.2	100%	10,12
Interval deviation	14.0	100%	10,12
IOI deviation	13.2	97%	10,12
Note duration deviaton	9.3	95%	12
Non-diatonic degrees dev.	9.1	100%	12
Silence duration deviation	6.3	94%	—
Silence duration range	6.1	87%	—
Note duration distrib. normality	6.0	89%	—
Note duration average	5.6	71%	—
Silence duration average	5.1	85%	—
Non-diatonic degrees avg.	4.9	66%	—
IOI range	4.7	53%	—
Number of non-significant silences	4.5	76%	—
Silence duration distrib. normality	4.3	45%	—
IOI average	4.2	53%	—
Non-diatonic degree distrib. normality	3.5	39%	—
Note duration range	3.3	34%	—
Pitch distrib. normality	2.6	25%	—
Num. non-diatonic notes	2.5	32%	—
IOI distrib. normality	2.2	20%	—
Interval average	1.7	14%	—

Table 1: Separability ranking between styles

	Counter	Avg.	Dev.	Norm.	Range
Pitch	6	6	6	(none)	6
Note dur.	—	(none)	12	(none)	(none)
IOI	—	(none)	10	(none)	(none)
Silence dur.	10	(none)	(none)	(none)	(none)
Pitch interval	—	(none)	10	10	6
Harmonicity	(none)	(none)	12	(none)	—
Rhythm	6	—	—	—	—

Table 2: Feature selection by descriptor category and statistics type. Numbers correspond to reduced description models where each descriptor appears. A dash means that the descriptor was not computed.

size models have been chosen, with 6, 10, and 12 descriptors. The biggest gaps in the  $\bar{z}$  values for the sorted descriptors led us to group the descriptors in these reduced models.

It is interesting to note that at least one descriptor from each category of those defined in section 3.2 were selected for a reduced model. Table 2 summarizes these results. The numbers in the table stand for the reduced model number for which the descriptor was assigned. The best represented categories were pitches and intervals, suggesting that the pitches of the notes and the relation among them are the most influent features for the style recognition problem. From the statistical point of view, standard deviations were the most important features, since five of them from six possible ones were selected. Counter and range descriptor types seem to be also of relevance.

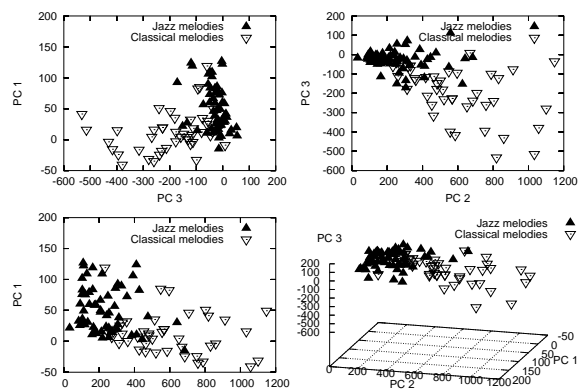


Figure 2: Melodies in the first three principal components space (whole melody dataset).

### 4.3 Feature space transformation results

In order to investigate the coefficients of the PCA transformation matrix, several datasets corresponding to

$$\omega = 1, 2, 4, 8, 16, 32, 64, \infty$$

and  $\delta = 1$  were used. These are representative datasets for small, medium, large  $\omega$ , and whole melody ( $\omega = \infty$ ). A PCA for each dataset was performed independently, and the transformation coefficients for the first three principal components have been studied for each analysis. Table 3 summarizes these results for the first principal component coefficients. The first column is the dataset segment size, and the second column is a list of contributing descriptors with a coefficient absolute value greater than 0.1. For small segment sizes ( $\omega = 1, 2, 4$ ), descriptors related to note duration have a dominant role. For average segment sizes ( $\omega = 4, 8$ ) counters for notes and silences appear among the most contributing descriptors. For large segment sizes ( $\omega = 16, 32, 64$ ), silence descriptors become important factors, while note duration descriptors become less relevant to this first principal component. Finally, with whole melody segment samples ( $\omega = \infty$ ), silence descriptors become clearly important.

From the statistical point of view, range descriptors have the largest coefficients (not present in the table for clarity) among duration related features for small to medium segment sizes. For large  $\omega$  values, the silence duration deviation has the biggest coefficient among silence descriptors.

The first principal component alone explains more than 50% of total variance for small segment sizes, and more than 35% for larger sizes. More than 75% of total variance is accumulated between the three first principal components for all  $\omega$  values tested.

The remarkable absence here are pitch related descriptors. Neither pitch, nor pitch interval nor non-diatonic note information appears among the first three components for any  $\omega$ .

$\omega$	contributing descriptors (listed by their order of component weight)
1	durRNG, ioiRNG, durAVG, durDEV, ioiDEV, ioiAVG
2	durRNG, ioiRNG, durAVG, durDEV, ioiDEV, ioiAVG
4	ioiRNG, durRNG, ioiDEV, durDEV, ioiAVG, durAVG, notNUM
8	ioiRNG, durRNG, notNUM, silintNUM, ioiDEV, durDEV, ioiAVG, durAVG
16	notNUM, ioiRNG, durRNG, silintNUM, silsigNUM, ioiDEV, durDEV
32	notNUM, dsIDEV, silintNUM, dsIRNG, silsigNUM, ndNUM, durRNG
64	dsIDEV, dsIRNG, notNUM, ioiRNG, silintNUM, ioiDEV, silsigNUM
$\infty$	ioiRNG, dsIRNG, dsIDEV

Table 3: First PCA component composition for every  $\omega$ .

It seems that, compared to duration descriptors –we include silence descriptors here–, they are of little importance for the statistically description of melodies. In other words, information about rhythm seems to be more important than information about pitch from this statistical description approach, at least for this pair of styles.

Inspecting the last six principal components, whose contribution to the total variance is virtually 0% ( $< 0.00001$ ) for every  $\omega$ , we find the normality descriptors being the ones having larger coefficient values. This is clearly an indication that they have no characterization ability for our datasets.

For the fourth to twenty-second principal components, a mixing of a variety of descriptors can be found contributing hardly a 25% of the so far unexplained total variance. Furthermore, more than 97% of the total variances can be explained by the first ten components for every  $\omega$ . This fact can be used to construct a description model in the transformed space with only 10 from a total 28 components with very little loss of information. There is even a slight tendency of total variance to concentrate on the first principal components as  $\omega$  increases.

The whole melody segment size case is of especial interest here, because only with its two first principal components, a 94% of the total dataset variance is explained (97% if the third component is included). Table 4 shows the most contributing descriptors for these first three principal components. As said before, the silence duration descriptors are very important for the first component—ioiRNG has the largest coefficient and also captures information about silences. The most contributing descriptors for the second component are all the counters present in the model. The third component adds the syncopation descriptor to the second component descriptor list. A picture of the melody distribution in the space formed by the three first components can be seen in figure 2.

#### 4.4 Style classification results

A leaving-10%-out crossvalidation scheme has been used for all the classification experiments, and averaged success rates are presented here. The experiments have been performed for  $\omega$  values between 1 and 100, and the range for  $\delta$  varies from 1 to its corresponding  $\omega$  value. Datasets with less than a hundred sample values have not been take into

PC 1	PC 2	PC 3
ioiRNG (.93)	notNUM (.87)	silsigNUM (−.69)
dsIRNG (.32)	silintNUM (.43)	silintNUM (.60)
dsIDEV (.14)	ndNUM (.17)	ndNUM (.28)
–	silsigNUM (.15)	notNUM (−.23)
–	–	syncop (.11)

Table 4: Most contributing descriptors for first three principal components from whole melody dataset. Component weights appear in parenthesis.

model	Bayes	NN
6	93.2 <sub>(100,2)</sub>	94.0 <sub>(91,16)</sub>
10	95.5 <sub>(98,1)</sub>	92.6 <sub>(99,19)</sub>
12	93.2 <sub>(58,1)</sub>	92.6 <sub>(98,19)</sub>
28	89.5 <sub>(41,33)</sub>	96.4 <sub>(95,13)</sub>

Table 5: Best style recognition percentages

account. This datasets are in the range of  $\omega > 50$  and  $\delta > 20$ .

**Bayes classifier** All the parameters needed to train the Bayesian classifier are estimated from the training set, except for the priors of each style, estimated as the percentage of samples from this style found in the test set.

All three reduced models outperformed the 28-descriptor model. The overall best result (95.5% of average success) for the Bayesian classifier have been obtained with the 10-descriptor model in the point (98, 1). See Table 5 for a summary on best results – indices represent the  $\langle \omega, \delta \rangle$  values for which the best success rates were obtained –.

**k-NN classifier** Before performing the main experiments for this classifier, a study of the evolution of the classification as a function of  $k$  has been designed, resulting in no significant improvements for  $k$  values larger than 1. Thus, the simplest classifier was selected:  $k = 1$ , to avoid unnecessary time consumption due to the very large number of experiments performed.

All models performed comparatively for  $\omega \leq 35$ . For  $\omega > 35$ , the 28-descriptor model begins to perform better than the reduced models. The best results (96.4%) were obtained for the point (95, 13) with the 28-descriptor model. The best results for all the models have been consistently obtained with very large segment lengths (see Table 5). The NN classifier obtained an 89.2% in average with the 28-descriptor model, while the other models yielded similar rates around 87%. Table 6 shows average style recognition percentages for all  $\langle \omega, \delta \rangle$  combinations tested.

model	Bayes	NN
6	84.2 ± 2.0	87.4 ± 2.9
10	88.5 ± 3.2	86.9 ± 2.5
12	89.5 ± 1.7	87.1 ± 2.5
28	71.1 ± 6.3	89.2 ± 4.5

Table 6: Averages and standard deviations of style recognition percentages

## 5 Conclusions and future work

A description model for monophonic melodies based on statistical features has been presented. Its representation capability for them and the separation ability for a musical style identification task has been also analysed.

From the feature selection stage, where descriptors have been analyzed as independent random variables, a number of interesting conclusions can be drawn. Pitches and intervals have shown to be the most discriminant features. Other very important features have been the number of notes and the rhythm syncopation. Although the former set of descriptors may be probably important in other style classification problems, probably these latter two have shown their importance in this particular problem of classical versus jazz. From the statistical point of view, standard deviations were the most important features, since five of them from six possible ones were selected.

From the feature space transformation results drawn from the analysis of the first principal components of the transformed feature space, it can be concluded, based on the  $\omega$  values tested, that short melodies are better described using note duration related features, for example IOI statistics, or note duration range descriptors. For large melody segments or whole melodies it seems that silence information is useful to describe them, as it provides greater total variance than descriptors based on other musical properties.

It is important to recall that range and average descriptors are relative to minimum values. When to select statistics is needed to measure a melody property, relative descriptors seems to be a good choice for any melody length, while simply counting the number of occurrences may work if quite large melodies are studied.

The general behaviour in the style classification task for all the models and classifiers has been to improve the results for larger  $\omega$  values. This general trend supports the importance of using large melody segments to obtain good classification results with this kind of models. The preferred values for  $\delta$  were small, because they provide a higher number of training data.

In the future, we plan to make use of all this methodology to test other kind of classifiers, like feed-forward neural nets or support vector machines, and to explore the performance of

this statistic description approach with a number of different styles.

An extension to this methodology is under development, where a voting scheme for segments is used to collaborate in the classification of the whole melody. Our experimental framework permits the training of a large number of classifiers that, combined in a multiclassifier system, could produce even better results.

## Acknowledgment

The authors would like to thank Carlos Pérez-Sancho, Francisco Moreno-Seco and Jorge Calera for their help, advise, and programming. Without their help this paper would have been much more difficult to finish.

## References

- Cruz-Alcázar, P. P., E. Vidal, and J. C. Pérez-Cortes (2003). Musical style identification using grammatical inference: The encoding problem. In *Proc. of CIARP 2003*, La Habana, Cuba, pp. 375–382.
- Dannenberg, R., B. Thom, and D. Watson (1997). A machine learning approach to musical style recognition. In *Proc. of the 1997 Int. Computer Music Conf.*, pp. 344–347.
- de León, P. J. P. and J. M. Iñesta (2003). Feature-driven recognition of music styles. In *Ist Iberian Conference on Pattern Recognition and Image Analysis. Lecture Notes in Computer Science*, 2652, Majorca, Spain, pp. 773–781.
- Duda, R. O. and P. E. Hart (1973). *Pattern classification and scene analysis*. John Wiley and Sons.
- Kohonen, T. (1990). Self-organizing map. *Proceedings IEEE* 78(9), 1464–1480.
- Pickens, J. (2001). A survey of feature selection techniques for music information retrieval. Technical report, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts.
- Reimann, H. (2003). Jazz versus classical music: their objects and criteria for aesthetical evaluation. In *Proc. of the Hawaii Int. Conf. on Arts and Humanities*, Honolulu, USA.
- Thom, B. (2000). Unsupervised learning and interactive jazz/blues improvisation. In *Proc. of the AAAI2000*, pp. 652–657.
- Toiviainen, P. and T. Eerola (2001). Method for comparative analysis of folk music based on musical feature extraction and neural networks. In *III Int. Conf. on Cognitive Musicology*, Jyväskylä, Finland, pp. 41–45.
- Uitdenbogerd, A. and J. Zobel (1998). Manipulation of music for melody matching. In *Proc. of ACM Int. Multimedia Conf.*, Bristol, UK, pp. 235–240.