Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction

EMILIA GÓMEZ
Music Technology Group, Universitat Pompeu Fabra
Sonology Department, Escola Superior de Música de Catalunya

PERFECTO HERRERA
Music Technology Group, Universitat Pompeu Fabra
Sonology Department, Escola Superior de Música de Catalunya

ABSTRACT: The automatic analysis of large musical corpora by means of computational models overcomes some limitations of manual analysis, and the unavailability of scores for most existing music makes necessary to work with audio recordings. Until now, research on this area has focused on music from the Western tradition. Nevertheless, we might ask if the available methods are suitable when analyzing music from other cultures. We present an empirical approach to the comparative analysis of audio recordings, focusing on tonal features and data mining techniques. Tonal features are related to the pitch class distribution, pitch range and employed scale, gamut and tuning system. We provide our initial but promising results obtained when trying to automatically distinguish music from Western and non-Western traditions; we analyze which descriptors are most relevant and study their distribution over 1500 pieces from different traditions and styles. As a result, some feature distributions differ for Western and non-Western music, and the obtained classification accuracy is higher than 80% for different classification algorithms and an independent test set. These results show that automatic description of audio signals together with data mining techniques provide means to characterize huge music collections from different traditions and complement musicological manual analyses.

Submitted 2008 March 5; Accepted 2008 April 2; Reviewed 2008 August 20.

KEYWORDS: *audio description, tonality, machine learning, comparative analysis*

INTRODUCTION

Goals and motivation

THERE is a wealth of literature on music research that focuses on the comparative study of different musical genres, styles and traditions. According to Toiviainen and Eerola (2006, p. 1), "typical research questions in this area of inquiry involve the evolution of a musical style, typical musical features in the works a composer, or similarities and differences across music traditions from various geographical regions". Traditional research methods have been mainly based on either manual analysis of notated scores or aural analysis of music recordings. Although these manual analyses provide very accurate and expert information, they might have two potential limitations, as pointed out in the work by Toiviainen and Eerola (2006). First, manual annotation is a time consuming task, and this makes these studies to be based on a relatively small music collection that might not be then representative, in a statistical sense, of the corpora in study. Second, manual annotations might be subjective or prone to errors, especially if they are generated by different people with slightly different criteria without a common methodology (Lesaffre et al., 2004).

One way to overcome these limitations is to introduce the use of computational methods, which allows automating (in different degrees) the analysis of large musical collections. Many recent studies have

been devoted to apply computational models to comparative music research. Toiviainen and Eerola (2006) provide a very good overview of computational approaches to comparative music research, including issues related to music representation, musical feature extraction and data mining techniques. They also provide some examples of visualization of large musical collections based on these techniques, focusing on the analysis of MIDI representations.

Some studies in the field of Music Information Retrieval (MIR) have been also devoted to apply these methods to the analysis of audio recordings, mainly in some applied contexts such as music genre classification or artist identification (e.g. Tzanetakis & Cook, 2002). Extracted features are related to different musical facets and a varied set of data mining techniques are afterward applied to this set of descriptors. Timbre and rhythmic features are the most commonly used ones. They provide a way to characterize differences in instrumentation and meter, which are usually enough to discriminate diverse musical styles. Timbre is usually characterized by a group of descriptors directly computed from the signal spectrum (Tzanetakis & Cook, 2002), and common rhythmic features are tempo and Inter-Onset-Intervals (IOIs) histograms, which represent the predominant pulses (Gouyon & Dixon, 2005).

These methods are also considered when trying to measure the similarity between two musical pieces based on different criteria (used instruments, rhythmic pattern or harmonic progression). The definition of a music similarity measure is a very complex and somehow subjective task.

Until now, MIR research has mainly focused on the analysis of music from the so-called "Western tradition", given that most of MIR systems are targeted toward this kind of music. Nevertheless, we might ask if the available descriptors and techniques are suitable when analyzing music from different traditions. The term *Western* is generally employed to denote most of the cultures of European origin and most of their descendants, and it is often used in contrast to other cultures including Asians, Africans, Native Americans, Aboriginals, Arabs, and prehistoric tribes. Tzanetakis et al. (2007) have recently introduced the concept of *Computational Ethnomusicology* to refer to the use of computer tools to assist in ethnomusicological research, providing some guidelines and specific example of this type of multidisciplinary research. In this context, we present here an example of the use of audio analysis tools for comparative analysis of music from different traditions and genres.

The goal of the present study is to provide an empirical approach to the comparative analysis of music audio recordings, focusing on tonal features and a music collection from different traditions and musical styles. These descriptors are related to the pitch class distribution of a piece, its pitch range or tessitura and the employed scale and tuning system, being the feature extraction process derived from mathematical models of Western musical scales and consonance. We provide our initial but promising results obtained when trying to automatically distinguish or classify music from Western and non-Western traditions by means of automatic audio feature extraction and data mining techniques. Having in mind this goal, we analyze which features might be relevant to this task and study their distribution over a large music collection. We then apply some data mining techniques intended to provide an automatic classification of a music recording into Western and non-Western categories. From an applied point of view, we investigate if it is possible to automatically classify music into Western and non-Western by just analyzing audio data.

Musical scales in Western and non-Western music

As mentioned above, we hypothesize that tonal descriptors related to the pitch class distribution of a piece, pitch range and employed scale and gamut may be useful to differentiate music from different traditions and styles.

A *scale* is a small set of notes in ascending or descending order (usually within an octave), being a sequence long enough to define a mode, tonality or another linear construction which starts and ends on its main note. These scales establish the basis for melodic construction. On the other side, the *gamut* is defined as the full range of pitches in a musical system, appearing when all possible variants of all possible scales are combined.

Scales in traditional Western music generally consist of seven notes (i.e. scale degrees), repeat at the octave, and are separated by whole and half step intervals of tones and semitones. Western music in the Medieval and Renaissance periods (1100-1600) tends to use the diatonic scale C-D-E-F-G-A-B, with rare and unsystematic presence of accidentals. Music of the common practice period (1600-1900) uses three types of scales: the diatonic scale and the melodic and harmonic minor scales, having 7 notes. In the 19th and 20th centuries, additional types of scale are explored, as the chromatic (12 notes), the whole tone (6

notes), the pentatonic (5 notes), the octatonic or diminished scales (Drabkin, 2008). As a general observation, in traditional Western music, scale notes are most often separated by equally-tempered tones or semitones, creating a gamut of 12 pitches per octave, so that the frequency ratio between consecutive semitones is equal to $st = \sqrt[12]{2}$, i.e. the interval value in the logarithm 'cent' metric is equal to 100 cents.

Many other musical traditions employ scales that include other intervals or a different number of pitches. According to Burns (1998, p. 217), the use of discrete pitch relationships is largely universal, pitch glides (as glissandos or portamentos) are used as embellishment and ornamentation in most musical cultures, and the concept of octave equivalence, although far from universal in early and structurally simpler music, seems to be common to more advanced musical systems.

For instance, gamelan music uses a small variety of scales including Pélog and Sléndro, not including equally tempered intervals (Carterette & Kendall, 1994). Ragas in Indian classical music often employ intervals smaller than a semitone, as both musical systems of India (Hindustani and Karnatic) are based on 22 possible intervals per octave and are not equal interval. Arabic music may use quarter tone intervals. According to Burns (1998), there are different theories as to the number of used intervals (ranging from 15 to 24) and some controversy as to where they are true quarter tones or merely microtonal variations of certain intervals.

Central and southern African music is characterized by the dominance of rhythmic and percussive devices, and the scales of musical instruments do not seem to be an approximation of the Western tempered scale (Merriam, 1959; VV.AA, 1973). Chinese and Japanese tuning also differ from equal-tempered scale (Piggott, 1891-1892).

In addition to the mentioned musical traditions, there are also other musical genres (apart from classical Western music) that may employ scale intervals smaller than a semitone. For instance, the blue note is an interval that is neither major nor minor, but in between, giving it a characteristic flavor. In blues, a pentatonic scale is often used, and in jazz many different modes and scales are found (being chromatic scales commonly used), often in the same piece.

MUSIC COLLECTION

A relevant step for a comparative study of music material is the definition of a proper audio collection. This collection should be representative of the different styles present in music from Western and non-Western tradition, which is an arduous task, given the variety of both categories.

We have tried to cope with the variety of both classes of music, gathering a music made of 500 audio recordings from non-Western music (distributed by region: Africa, Java, Arabic, Japan, China, India and Central Asia). These samples consist of recordings of traditional music from different areas, and we discarded those having some Western influence (equal-tempered instruments as the piano, for instance).

We also considered 1000 recordings from Western music, gathered from commercial CDs and distributed across the musical genres presented in Figure 1.

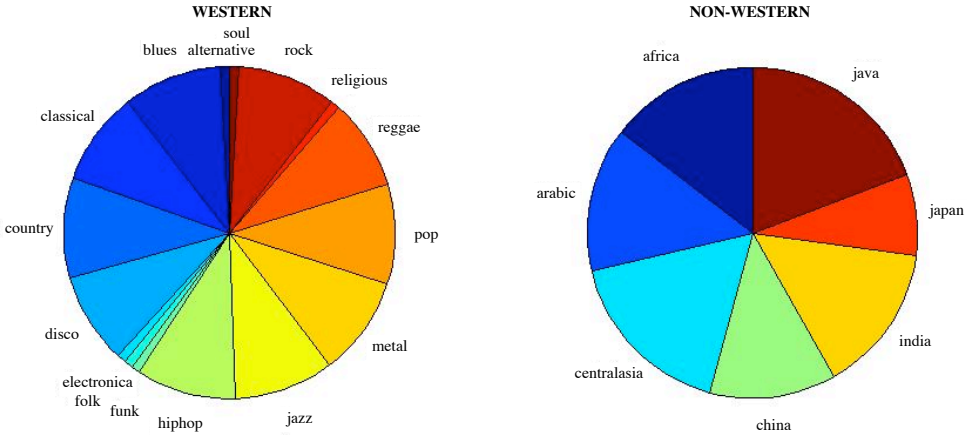

Fig. 1. Distribution of musical genres within the music collection

Non-Western music was chosen to be representative of the musical tradition of each geographical region, and Western music was chosen to cover a set of varied musical genres, which might be more representative of the different types of music than geographical ordering. The "Western" collection that was chosen has been widely used in automatic genre recognition (Tzanetakis, 2001; Holzapfel, 2007; Rentfrow, 2003).

AUDIO FEATURE EXTRACTION

Another relevant task is the definition and computation of a representative set of musical features for the studied problem. These features should be automatically extractable from audio recordings.

Based on previous studies, we hypothesize that features derived from a tonal analysis of the piece might be relevant for comparative analysis and music similarity in this particular context, as they represent the pitch class distribution of the piece and they are not influenced by instrumentation or tempo (Toiviainen & Eerola, 2006; Gómez, 2006, p. 153-183).

We compute these features over the first 30 seconds of each musical piece, for two different reasons. First, 30 seconds are being considered enough in the MIR community to recognize a certain musical style and a musical key. According to Gómez (2006, p. 134), the accuracy rate of a key estimation algorithm is similar if we only consider the beginning of a piece (from 15 seconds) than the whole piece. Second, and from a practical point of view, the computation speed is reduced. We then assume that analyzing the first 30 seconds of a musical recording should be enough to classify if the piece belongs or not to the Western music tradition. In order to check this fact, we listened to the starting segments of all the pieces and discarded few non representative parts (containing silences or ambiguous introductions). We present here the set of audio features that has been considered in this study.

Tuning frequency

One of the features that we consider relevant for this task is the frequency used to tune a musical piece, which is close to 440 Hz in the Western tradition. We estimate the tuning frequency (i.e. its difference with 440 Hz) as the value that minimizes the deviation of the main spectral peaks from an equal-tempered scale. These spectral peaks are obtained after frequency analysis of the audio signal through the Discrete Fourier Transform (DFT). An estimation of the tuning frequency is computed in a frame basis (frames have a 100 ms duration and are 50% overlapped), and a global estimate is derived from the frame values by means of a histogram. A more detailed description of the algorithm is presented in (Gómez, 2006, p. 71-76). An example is provided in Figure 2.

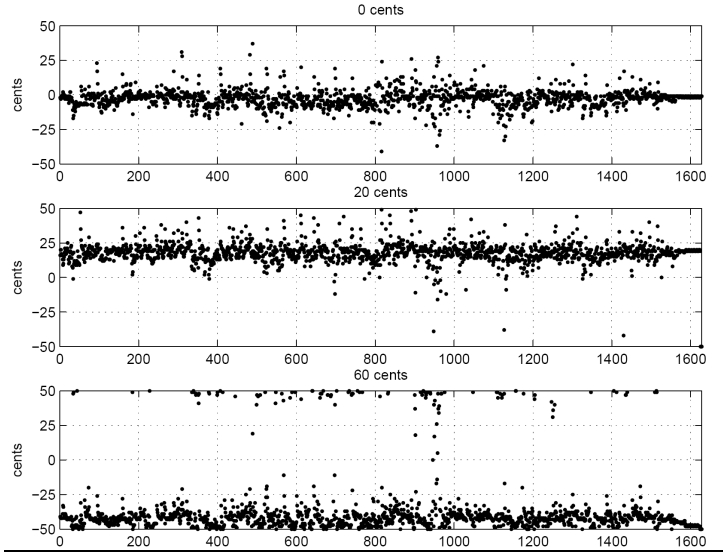

Fig. 2. Frequency deviation with respect to 440 Hz (in cents) vs. frame index, computed for a piece tuned to 440 Hz (up), with a deviation of 20 cents (middle) and 60 cents (down).

High-resolution pitch class distributions

Pitch-class distributions from symbolic data are already considered in Toiviainen and Eerola (2006) for the comparative analysis of symbolic data from different geographical regions. We extend this idea to the analysis of audio recordings by computing pitch class distributions from audio signals.

We call the obtained features the Harmonic Pitch Class Profile (HPCP), and the procedure for its computation is presented in detail in Gómez (2006) and illustrated in Figure 3. The HPCP is computed in a frame basis, considering 100 ms overlapped frames. We perform a spectral analysis of each audio frame followed by a peak estimation procedure. The peak frequencies are then mapped into pitch-class values according to the tuning frequency value previously estimated. The HPCP vector is computed using an interval resolution of 10 cents per semitone, so that the vector size is equal to 120 points (10 values per semitone). This resolution is chosen in order to achieve a more detailed representation of the pitch class distribution of the piece and to cope with the small pitch variations obtained through different tuning systems and scales. Finally, we consider the HPCP average of the frames belonging to the considered audio excerpt (30 first seconds).

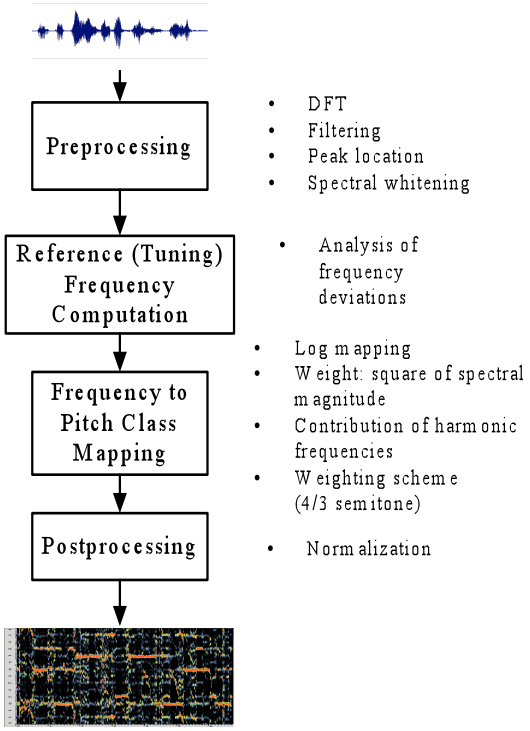

Fig. 3. Block diagram for HPCP computation.

We also compute a "transposed" version of the HPCP, called the THPCP, by ring-shifting the HPCP vector according to the position of the maximum value (*shift*):

$$THCP[n] = HPCP[\mathrm{mod}(n - shift, size)] \; n = 1, \; \ldots size.$$

This feature can be considered to be invariant to transposition, i.e. a piece which is transposed to a different key will have different HPCP but same THPCP values. In this sense, pieces are considered to be similar if they share the same tonal profile regardless of its absolute position.

Figure 4 shows an example of HPCP for an audio excerpt from Burkina Faso (labeled as *African*) including percussion and singing voice. We observed that the local maxima are not located on the exact positions of the tempered semitones. Figure 5 shows a comparison of high-resolution HPCP for a Western and non-Western musical excerpt. We also observe that the local maxima are not located on the exact positions of the tempered semitones for the Chinese piece, as it in fact happens for the Beethoven excerpt.

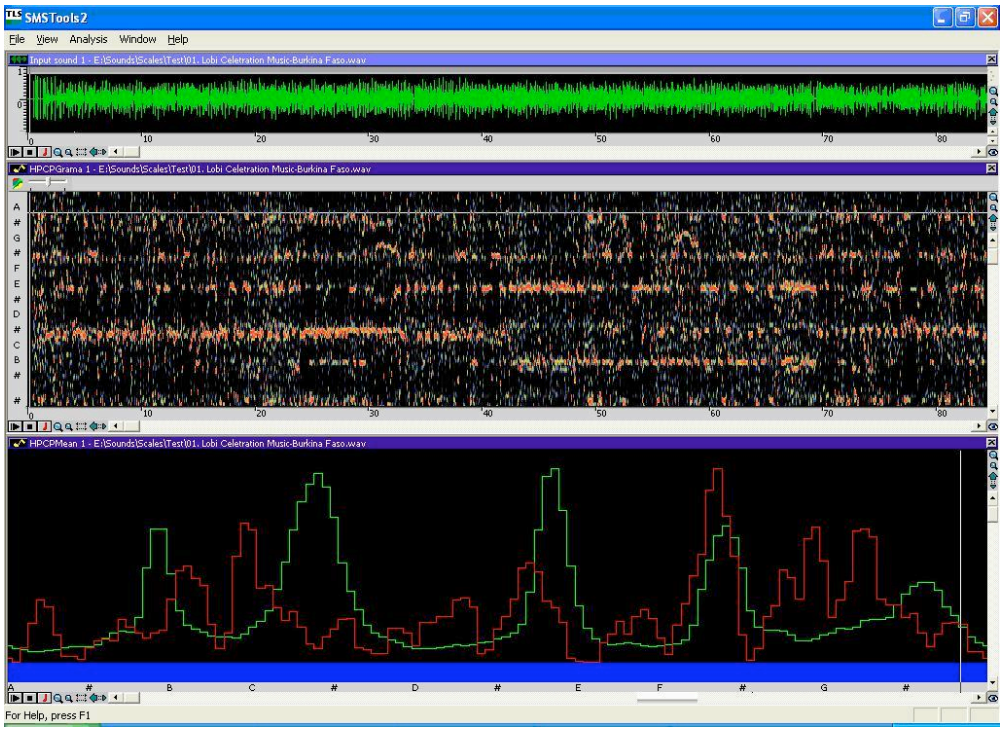

Fig. 4. Example of HPCP evolution for an African audio excerpt. Audio signal vs. time (top), high resolution HPCP vs. time (mid) and average HPCP (bottom, green line). The red line at the bottom shows a short-tem average of HPCP.

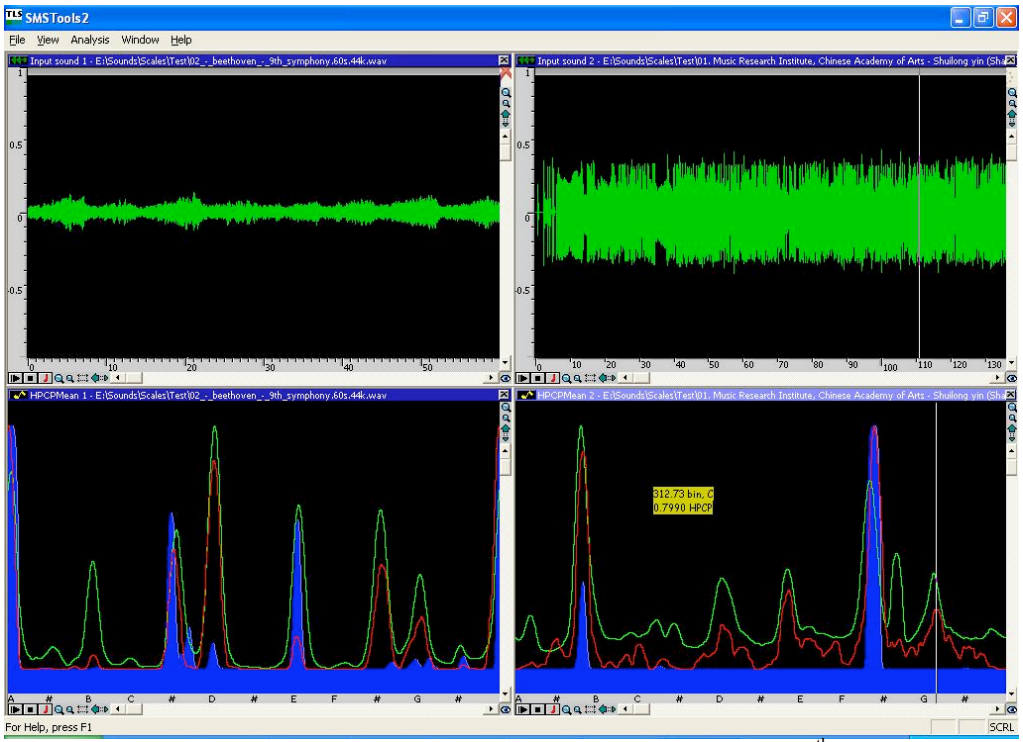

Fig. 5. HPCP for a Western and non-Western musical excerpt (Left: Beethoven 9th symphony. Right: Chinese folk instrumental piece). Audio signal vs. time (top), and global HPCP (bottom, green line). The red line at the bottom shows a short-tem average of HPCP, while the blue plot indicates an instantaneous value.

Features derived from pitch class distributions

We also compute two features from the ones described above. They are intended to distinguish between music composed using equal-tempered and non-equal tempered scales, as it is a relevant aspect to distinguish music from Western tradition.

EQUAL-TEMPERED DEVIATION

The equal-tempered deviation measures the deviation of the HPCP local maxima from equal-tempered bins. In order to compute this descriptor, we first extract a set of local maxima from the HPCP, *{k}, k=1…K*, and we then compute their deviations from closest equal-tempered bins, weighted by their magnitude and normalized by the sum of peak magnitudes:

$$Etd = \frac{\sum_{k=1}^{K} HPCP[k] \cdot abs(k - et_k)}{\sum_{k=1}^{K} HPCP[k]}$$

where et_k represents the closest equal-tempered bin from HPCP bin *k*.

NON-TEMPERED ENERGY RATIO

Non-tempered energy ratio represents the ratio between the amplitude of non-tempered HPCP bins and the total amplitude:

$$ER = 1 - \frac{\sum_{i=1}^{12} HPCP[pos_i]}{\sum_{i=1}^{size} HPCP[i]}$$

where *size = 120* (size of the HPCP vector) and pos_i are given by the HPCP positions related to the equal-tempered pitch classes.

DIATONIC STRENGTH

This descriptor represents the maximum correlation of the HPCP vector and a diatonic major profile ring-shifted in all possible positions. This diatonic major profile is represented in Figure 6.

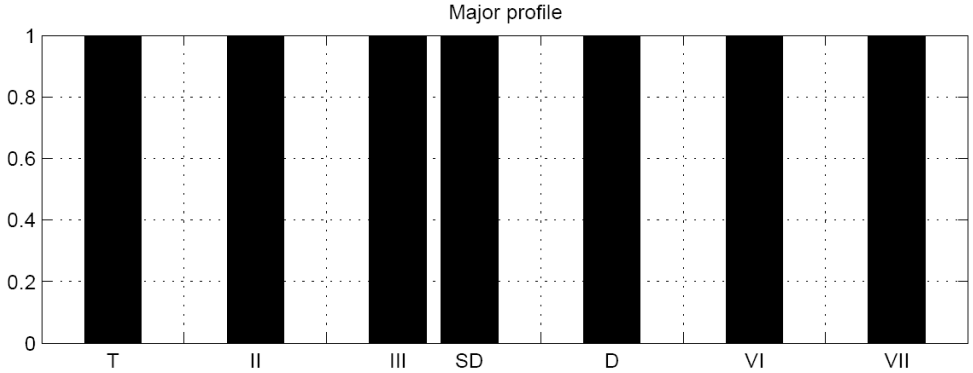

Fig. 6. Diatonic major profile represented as the presence (1 or 0) of each of the 12 semitones of the chromatic scale.

We hypothesize that this correlation should be higher for a piece using a diatonic major scale, which is characteristic of Western music.

Octave centroid

HPCP and related descriptors do not consider the absolute pitch height of the analyzed piece, as these descriptors map all the pitch class values to a single octave. This means that two pieces having the same pitch-class distribution but played in different octaves will have similar values for the extracted features. In order to take into account the octave location of the played piece, and to analyze its distribution over different musical pieces, we have defined a feature called *octave centroid*, which represents the geometry centre of the played pitches.

In order to estimate this descriptor, we first perform a spectral analysis and apply a multipitch estimation method based on Klapuri (2004). This method outputs an estimation of the predominant pitches found in the analyzed spectrum. A centroid feature is then computed from this representation on a frame basis. We finally consider different statistics of frame values (mean, median, standard deviation, inter-quartile range, kurtosis, skewness) as global descriptors for the analyzed piece.

Roughness

Complementary to pitch-class distribution and derived features, we also compute a roughness descriptor, as a measure of sensory dissonance. Roughness is a perceptual sensation arising from the presence of energy in very close frequencies, as it happens in the case that 2 instruments are less-than perfectly tuned. In order to experience roughness the frequency differences between partials has to be larger than 10 Hz (Zwicker and Fastl, 1999). Roughness can be typically experienced when listening to gamelan music, as the involved instruments are slightly mistuned on purpose (Tenzer, 1998).

We have used the roughness estimation model proposed in Vassilakis (2001, 2005). We compute a roughness value for each analysis frame, by summing the roughness of all pairs of components in the spectrum. We consider as the main frequency components of the spectrum those with spectral magnitudes higher than the 14% of the maximum spectral amplitude, as indicated in Vassilakis (2001, 2005). Then, a global roughness value is computed as the median of instantaneous values. We finally consider different statistics of frame values (mean, median, standard deviation, inter-quartile range, kurtosis, skewness) as global roughness descriptors for the analyzed piece.

DISTRIBUTION OF FEATURES

Once the set of features used to characterize the target collection has been defined and computed, we can study their distribution through the different types of music within the test collection, in order to have a preliminary idea of their usefulness for comparative analysis and classification. We present here some results on statistical analysis of the audio features for different group of pieces, in order to illustrate the differences between musical genres and origins.

Figure 7 shows the distribution of the tuning descriptor for Western and non-Western music. As expected, the distribution of tuning deviation with respect to 440 Hz is centered on 0 cents for Western music. On the other hand, it appears to be equal distributed between -50 and 50 cents for non-Western pieces. This is confirmed by a goodness-of-fit chi-square statistical test, where a p-value=0.061indicates that the distribution of tuning frequencies for the non-Western pieces roughly follows a uniform distribution. On the other hand, the distribution for the Western pieces does not follow such a distribution.

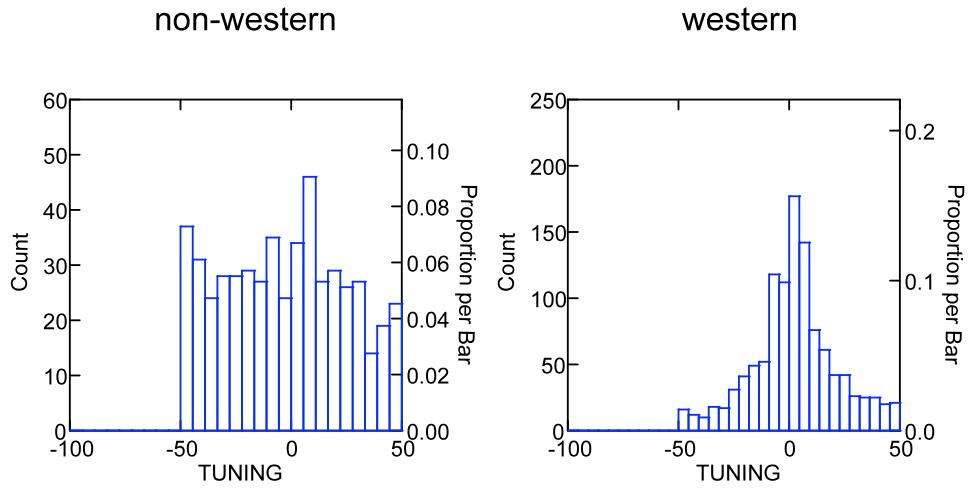

Fig. 7. Distribution of tuning deviation (in cents) for Western and non-Western music.

We also find some differences in the main transposed HPCP features, which represents the intensity of the different degrees of a diatonic major scale. For the main degrees, we find larger values for Western than for non-Western music. Figure 7 shows the distribution of THPCP3, which represents the intensity of the second degree of a diatonic scale. According to the distribution, it appears to be lower for non-Western music than for Western music, and differently distributed. In fact, a goodness-of-fit chi-square statistical test yields a p-value=0.132, indicating the correspondence between THPCP3 distribution of non-Western pieces and a Gompertz distribution (with parameters b = 3.2 and c = 2.62). The same statistical test yields a p-value=0 for the same distribution considering Western pieces.

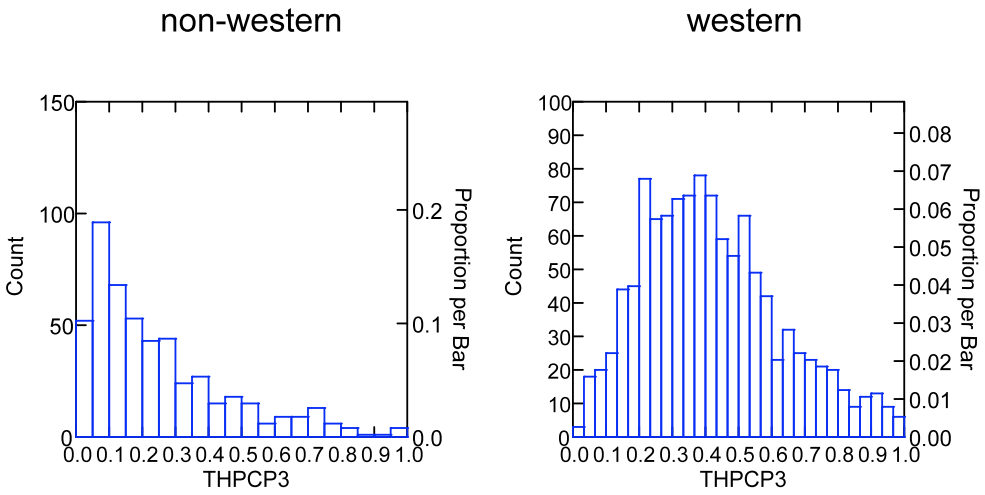

Fig. 8. Distribution of THPCP3 for Western and non-Western music.

The THPCP6 feature represents the intensity of the fourth (sub-dominant) degree of a diatonic scale. This feature also shows lower values for non-Western music than for Western music and they are differently distributed, as shown in figure 9.

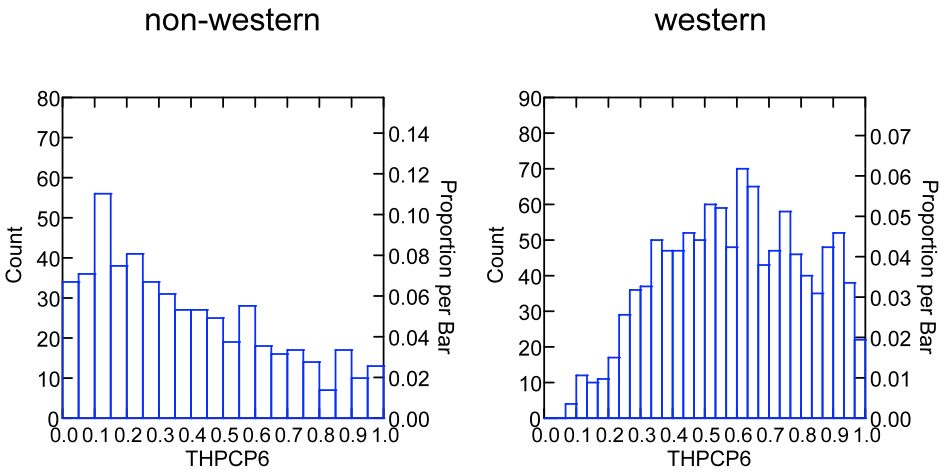

Fig. 9. Distribution of THPCP6 for Western and non-Western music.

THPCP8 represents the intensity of the fifth (dominant) degree of a diatonic scale. As for the previous descriptors, it seems to be lower for non-Western music than for Western music and differently distributed, as shown figure 10. In fact, a goodness-of-fit chi-square statistical test yields a p-value=0.652, indicating the correspondence between THPCP8 distribution from non-Western pieces and a Logit Normal distribution (with parameters μ=-0.206171 σ=1.293628). The same statistical test yields a p-value=0 for the same distribution considering Western pieces.

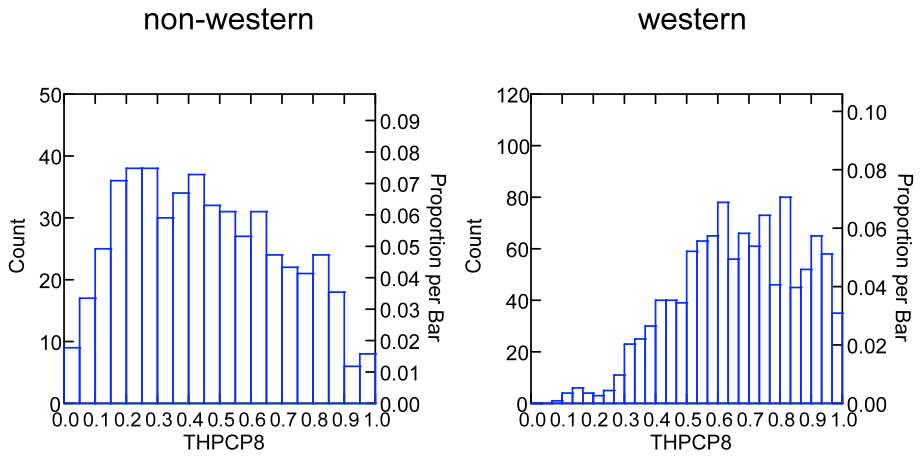

Fig. 10. Distribution of THPCP8 for Western and non-Western music.

Regarding descriptors derived from HPCP, the equal-tempered deviation, representing the deviation from an equal-tempered scale, also appears to be lower for Western than for non-Western music, as shown in figure 11. In fact, a goodness-of-fit chi-square statistical test yields a p-value=0.652, indicating the correspondence between the equal-tempered deviation distribution from non-Western pieces and a Weibull distribution (scale=0.185 and shape=2.156). The same statistical test yields a p-value=0 for the same distribution considering Western pieces.

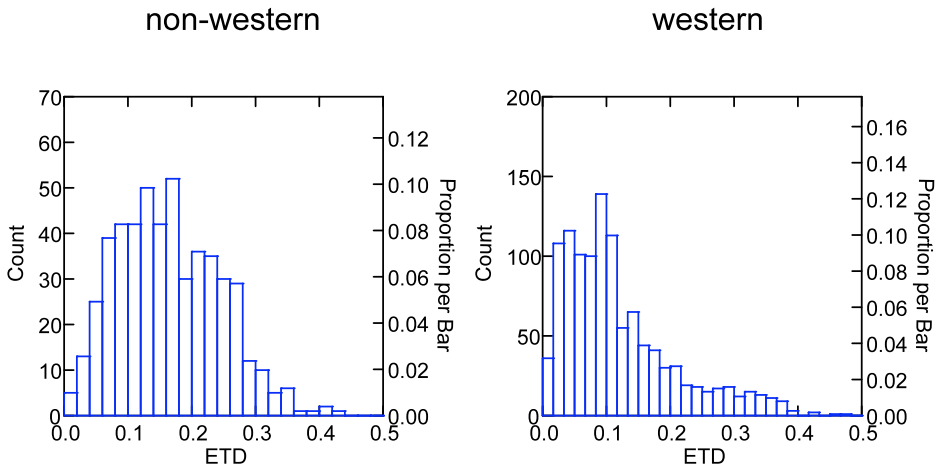

Fig. 11. Distribution of equal-tempered deviation for Western and non-Western music.

AUTOMATIC CLASSIFICATION

After this preliminary statistical analysis of the computed features, we present here some results of automatic classification based on the extracted descriptors. Our goal here is to have a classifier that automatically assigns the label "Western" or "non-Western" to any audio file that is input and analyzed according to the procedure explained in the previous sections.

We have used different machine learning techniques implemented in the WEKA machine learning software, a collection of classification algorithms for machine learning tasks (Witten & Frank, 2005b). Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. The algorithms are applied directly to our dataset.

Evaluation results using different machine learning techniques

We adopt an evaluation procedure based on 10-fold cross-validation over equally-distributed classes. In 10-fold cross-validation, the data is divided into 10 subsets of approximately equal number of examples, and the different machine learning algorithms are trained 10 times, each time leaving out one of the subsets from training, but using only the omitted subset for accuracy estimation, this way we test the learnt model using previously "unseen" examples. We do this train-test cycle ten times in order to provide a better estimate of the generalization error. In other words, we try to get a good estimation of the errors the system will yield when classifying examples that do not belong to our current collection.

We have approached several classification methods but, for the sake of clarity and conciseness we only present two of them, which are explained in detail in (Witten & Frank, 2005a). One of these approaches (decision trees) provides a clearly understandable output that has allowed identifying the most relevant features for classification. In addition, decision trees are easy to implement as a collection of "if-then" rules in any programming environment. The other approach we have explored (support vector machine) is considered as one of the best-performing learning algorithms currently available.

- **Decision trees** are massively used for different machine learning and classification tasks. One of the main reasons for their acceptance lies in the fact that their output is a model that can be interpreted as a series of {**if** a descriptor has a value bigger or smaller than x **then** classify the observation as C} clauses. Decision trees are constructed top-down, beginning with the feature that seems to be the most informative one, that is, the one that maximally reduces entropy. Branches are then created from each one of the different values of this feature. The training examples are sorted to the appropriate descendant node, and the entire process is then repeated recursively using the examples of one of the descendant nodes, then those of the other. An in-depth treatment of decision trees can be found in (Mitchell, 1997). As shown in Figure 12, which depicts the decision tree computed to model our data, the test at a node compares the descriptor with a constant value. Leaf nodes give a classification that

applies to all instances that reach the leaf. There are different algorithms to build decision trees. The one we have used, called *J4.8* in Weka is an implementation of the version 8 of the so-called C4.5 (Quinlan, 1993; Witten & Frank, 2005a, p. 189-200), which is probably the most often decision tree used in the scientific community. We have also tested different values for the parameter *minObj*, which specifies the minimum number of objects (or instances) allowed at a leaf. This allows getting very compact trees without sacrificing precision as figure 12 illustrates.

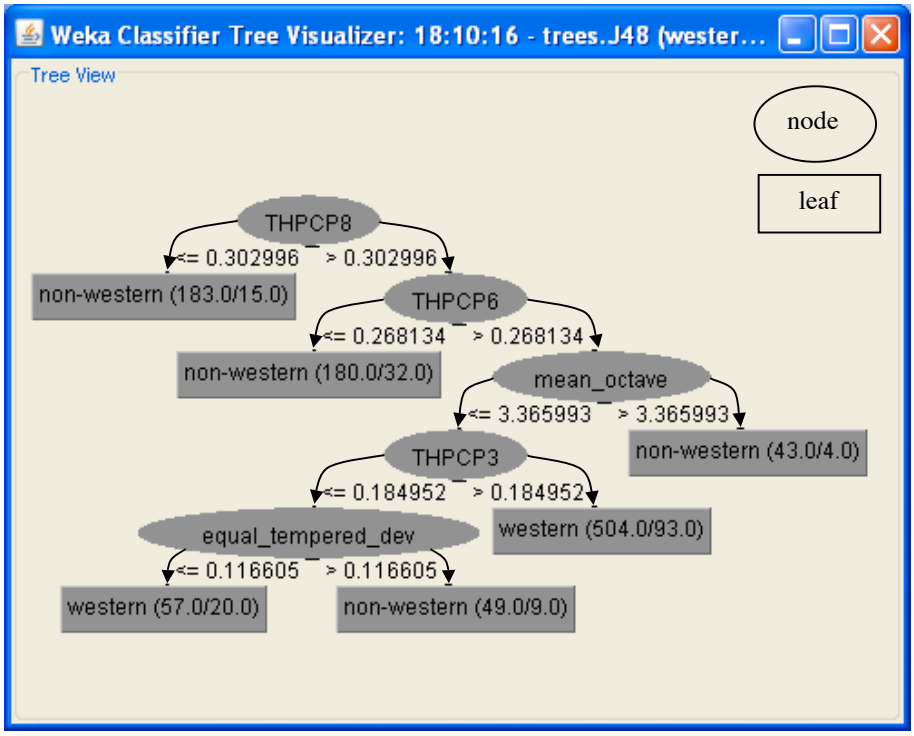

Fig. 12. Example of a Decision tree (J48 with a minimum of 40 examples per leaf) with 6 leaves. Parentheses indicate the number of correctly/incorrectly classified files for each branch. The overall classification accuracy for this tree is provided in Table 1.

- **Support Vector Machines** (SVM) are classifiers based on statistical learning theory (Vapnik, 1998). The basic training principle underlying SVMs is finding the optimal linear hyperplane that separates two classes, such that the expected classification error for unseen test samples is minimized (i.e., they look for good generalization performance). This hyperplane can be delimited by a subset of the available instances, which define the "support vectors" for it. Based on this principle, a SVM uses a systematic approach to find a linear function with the lowest complexity. For linearly non-separable data, SVMs can (non-linearly) map the input to a high dimensional feature space where a linear hyperplane can be found. This mapping is done by means of a so-called *kernel* function. Although there is no guarantee that a linear solution will always exist in the high dimensional space, in practice it is quite feasible to construct a working solution. In other words, it can be said that training a SVM is equivalent to solving a quadratic programming with linear constraints and as many variables as data points. Weka implements support vector machines using the SMO (Sequential minimal optimization) algorithm (Witten & Frank, 2005a, p. 214-235; Platt, 1998). It is advisable to tune certain parameters of this algorithm, as its complexity parameter and the exponent for the used polynomial kernel, in order to decrease its classification error.

For the Western versus non-Western categories, and using these different techniques, we achieve the classification results summarized in Table 1. F-measure is a common measure to evaluate the performance of information retrieval systems, and it is defined as the weighted harmonic mean of precision and recall:

$$F\text{-measure} = 2 \cdot (precision \cdot recall)/(precision + recall)$$

where precision is the fraction of the retrieved instances that belong to the correct category and recall is the fraction of the documents that belong to the correct category which are successfully retrieved.

Method		**Global accuracy**	**Accuracy per class**	
Classification algorithm	**WEKA Parameters**	**Correctly Classified Instances**	**F-Measure Western**	**F-Measure non-Western**
Decision Trees (J48)	30 minObj	80.41 %	0.808	0.8
	40 minObj	**82.38 %**	0.827	0.82
	50 minObj	79.63 %	0.795	0.797
Support Vector Machine (Binary SMO Classifier)	Complexity = 1, Exponent = 1.5	86.12 %	0.865	0.857
	Complexity = 1.5, Exponent = 1.5	**86.51 %**	0.869	0.861

Table 1. Classification results for different machine learning techniques provided by WEKA.

As a general conclusion, we observed that the classification accuracy is higher than 80% for all the machine learning techniques used. It could be argued that the Western-non Western distinction that we are aimed at is an artificial and biased one. If that was the case, an analysis that would not superimpose those two categories as an "a priori" could yield different groupings of the data. Cluster analysis allows for that kind of unsupervised, emergent type of data arrangement. K-means clustering is one of the most usual clustering techniques (Witten & Frank, 2005a), and it tries to group data into homogeneous clusters, and, at the same time, to separate heterogeneous data into different clusters. The homogeneity criterion is defined by means of a Euclidean distance, which the algorithm tries to minimize for the examples that are clustered together and maximized among different clusters.

In our experiment, we have clustered the data asking for the algorithm to find 2 clusters. We performed 10 runs of the algorithm and cross-tabulated the cluster assignment against the Western/non-Western distinction, observing that the error varied from 29.677% to 29.9817%, which indicates an extremely robust solution: the two clusters found in a non-supervised way coincide, to a large extent, with the two categories used in the supervised experiments. In addition, clustering solutions using more than 2 groups yielded larger errors than the 2-clusters solution. It could, then, be the case that the distinction is not so artificial, and that it emerges when we consider music according to the tonality-related descriptors that we have computed.

Feature selection

The classifiers used in the experiments presented above are designed to detect the most appropriate descriptors and even the most appropriate instances for optimizing their decisions. However, there are some automatic methods that specifically give some hints on the usefulness of the available features.

We have tested an attribute evaluation method for attribute selection, correlation-based feature selection (CFS) (Hall, 2000). This algorithm selects a near-optimal subset of features that have minimal correlation between them, and maximal correlation with the to-be-predicted classes. In the set of descriptors selected by this algorithm we observed that the most relevant descriptors are THPCP3, THPCP8, THPCP10, tuning, equal tempered deviation, roughness (mainly its median and standard deviation along the considered excerpt) and octave centroid (average along the audio excerpt).

The relevance of these features has already been noticed when performing an analysis of their value distributions and they also coincide with the descriptors found on the generated decision trees, as shown in the previous section.

Classification accuracy for different musical genres and traditions

It is very informative to study the distribution of classification errors, in order to have some insights about the limitations of the method. For instance, we observed in Table 1 that the accuracy (F-measure) for Western and non-Western categories has no significant differences.

In order to study the accuracy distribution over musical genres, we have built the decision tree shown in Figure 12, which provides a classification accuracy of 82.38% over the music collection under study. Figure 13 shows the distribution of correct classification for different Western musical genres and non-Western traditions (grouped by region).

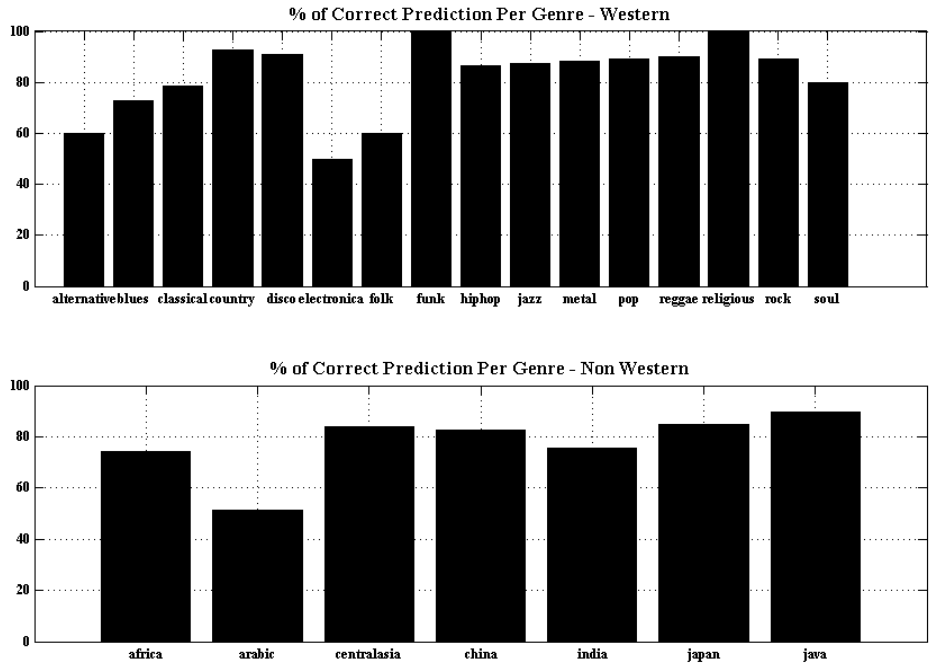

Fig. 13. Percentage of correctly classified instances distributed among genres using the Decision Tree from Figure 12. The classification accuracy for this tree is provided in Table 1.

We observed that *electronica* is the most misclassified Western style, and it is usually labeled as non-Western. We can justify this by the fact that the excerpts under electronic music usually include non-quantized pitched sounds, as well as non-pitched sounds that can increase the values of features such as equal tempered deviation or tuning, which have been found to be high for non-Western music. This is also the case for the few items under the *alternative* category. In the case of the 10 *folk* examples, 40% of them are misclassified due to a low value of THPCP6, related to the relative intensity of the fourth degree of the scale.

Regarding non-Western material, *Arabic* music is sometimes labeled as Western. This can be due to the tonal similarity between some of the audio excerpts under this category and the Western musical tradition, including some tempered instruments and scale degrees. An analysis of the misclassified instances revealed high values for the descriptors THPCP8 (related to the fifth), THPCP6 (related to the fourth) and THPCP3 (related to the second) in 81.82% of the misclassified instances, and high values for THPCP8 and THPCP6 together with small equal tempered deviation for the rest.

Classification accuracy for an independent test set

In order to test the generalization power of the classification algorithm, we have performed an evaluation using an independent test set from the collection used for training the models. The chosen independent set is the NASA Voyager Golden Record [1]. NASA placed aboard the Voyager spacecraft a time capsule intended to represent the history of our world to extraterrestrials. This capsule contains a record with sound and images of Earth, selected by a committee chaired by Carl Sagan of Cornell University and including 27

musical pieces from Eastern and Western classics and a variety of ethnic music (Sagan, 1984). These 27 pieces were manually classified by us into 12 Western and 17 non-Western pieces, and then analyzed in order to obtain the set of descriptors from the 30 first seconds of each piece. These data were then used as an independent test set for the selected classifiers obtaining the classification results shown in Table 2.

Method		**Global accuracy**	**Accuracy per class**	
Classification algorithm	**Parameters**	**Correctly Classified Instances**	**F-Measure Western**	**F-Measure non-Western**
Decision Trees (J48)	30 minObj	70.34 %	0.667	0.733
	40 minObj	**85.18 %**	0.818	0.875
	50 minObj	**85.18 %**	0.818	0.875
Binary SMO Classifier	Complexity = 1, Exponent = 1	81.48 %	0.737	0.857
	Complexity = 1, Exponent = 1.5	74.07 %	0.632	0.8
	Complexity = 1.5, Exponent = 1.5	74.07 %	0.632	0.8

Table 2. Classification results for different machine learning techniques provided by WEKA, using an independent test set from the Voyager Golden Record.

We observed that 85.18% of the pieces were correctly classified by the system, even if the obtained performance slightly varied for the two classification algorithms. The best results were obtained by using decision trees with different configuration parameters. The classifier reaches nearly the same performance with these "unseen" files than in the train-test cycle, which demonstrates the generalization ability of the proposed models.

CONCLUSIONS AND FUTURE WORK

In this study we provided an empirical approach to the comparative analysis of music audio recordings, focusing on tonal features and a music collection from different traditions and musical styles. We presented some encouraging results obtained when trying to automatically distinguish or classify music from Western and non-Western traditions by means of automatically audio feature extraction and data mining techniques. We obtained a high rate of classification accuracy of 80% for a music collection of 1500 pieces from different musical traditions using a restricted set of tonal features. From this, we can argue that it can be possible to automatically classify music into western and non-western by just analyzing audio data.

We are aware that there are larger issues involved in the determination of musical genres. We are also aware of the limitations of the concept of Western as opposed to non-Western music. Ideally we should be able to define and formalize stylistic features proper to different kinds of music or "stylistic areas" and approach genres not just geographically but as a set of traits and then refine our descriptors accordingly.

Future work will then be devoted to different issues. One important aspect that we would like to achieve is to contrast and complement this group of descriptions from an ethnomusicology perspective, analyzing in detail some of the used excerpts. We will also investigate the main variations inside Western and non-Western styles by comparing in details the different musical genres and musical traditions. We also plan to analyze how automatically-extracted audio features related to timbre and rhythmic aspects (which were out of our scope in this paper) can improve the classification and complement the current feature set.

The present work shows that automatic audio description tools, together with data mining techniques can help to characterize huge music collections and complement musicological manual analyses. It also confirms that tonal features extracted from audio data are representative of the pitch class

distribution, scale, gamut and tuning system of the analyzed piece, and that they provide means of characterizing different traditions and styles. We believe that audio description tools have a great potential to assist in ethnomusicological research and we hope that our work will contribute to the understanding of the world's musical heritage by means of computational modeling.

ACKOWLEDGMENTS

The authors would like to thank Dr. Ricardo Canzio, from the Graduate Institute of Musicology, National Taiwan University, for his expertise about ethnomusicology and his assistance and review of this project. Special thank also go to Professor Paul Poletti, from the *Escola Superior de Música de Catalunya* for his knowledge and recommendations about scales and tuning systems in different musical traditions, and to Dr. Ramon Pelinski for his insightful comments related to the ethnomusicological issues involved in our research.

NOTES

[1] http://voyager.jpl.nasa.gov/spacecraft/goldenrec.html

REFERENCES

Burns, E. M. (1998). *Intervals, scales and tuning*. In *The Psychology of Music*, 2nd Edition, edited by Diana Deutsch, pp. 215-264. New York: Academic Press. ISBN 0-12-213564-4.

Carterette, E. C., & Kendall, R. A. (1994). *On the tuning and stretched octave of Javanese gamelan*. Leonardo Music Journal, Vol 4. pp. 59-68.

Drabkin, W. (2008) *Scale*. Grove Music Online ed. L. Macy (Accessed [31 January 2008]), http://www.grovemusic.com

Gómez, E. (2006). *Tonal description of music audio signals*. PhD dissertation, Universitat Pompeu Fabra. http://mtg.upf.edu/~egomez/thesis

Gouyon, F., & Dixon, S. (2005). *A review of automatic rhythm description systems*. Computer Music Journal Vol. 29, No. 1, pp. 34-54.

Hall, M. A. (2000). *Correlation-based feature selection for discrete and numeric class machine learning*. In proceedings of the Seventeenth International Conference on Machine Learning,

Holzapfel, A., & Stylianou, Y. (2007). *A Statistical Approach To Musical Genre Classification Using Non-Negative Matrix Factorization*, Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 2(15-20), April 2007, pp. II-693 - II-696.

Klapuri, A. (2004). *Signal processing methods for the automatic transcription of music*, Ph.D. thesis, Tampere University of Technology, Finland.

Lesaffre, M., Leman, M., De Baets, B., & Martens, J.-P. (2004). *Methodological considerations concerning manual annotation of musical audio in function of algorithm development*, Proceedings of the International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, October 9-14.

Merriam, A. P. (1959). *Characteristics of African Music*. Journal of the International Folk Music Council, Vol. 11, pp. 13-19.

Mitchell, T. M. (1997). *Machine learning*. Boston, MA: McGraw-Hill.

Piggott, F. T. (1891-1892). *The Music of Japan*. Proceedings of the Musical Association, 18th Sess., pp. 103-120.

Platt, J. (1998). *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Rentfrow, P. J., & Godsling, S. D. (2003). *The Do Re Mi's of Everyday Life: The Structure and Personality Correlates of Music Preferences.* Journal of Pers. Soc. Psychology, Vol. 84, No. 6, pp. 1236-1254.

Sagan, C (1984). *Murmurs of Earth: The Voyager Interstellar Record*. Ballantine Books.

Tenzer, M. (1998). *Balinese Music.* Periplus Editions: Singapore.

Toiviainen, P., & Eerola, T. (2006). *Visualization in comparative music research*. In A. Rizzi & M Vichi (Eds.), COMPSTAT 2006 - Proceedings in Computational Statistics. Heidelberg: Physica-Verlag, pp. 209-221.

Tzanetakis G., Kapur, A., Andrew Schloss, W., & Wright, M. (2007). *Computational Ethnomusicology*, Journal of Interdisciplinary Music Studies, Vol. 1, No. 2, pp. 1-24.

Tzanetakis, G., Essl, G., & Cook, P.R. (2001). *Automatic Musical Genre Classification of Audio Signals*, Proceedings of International Symposium on Music Information Retrieval (ISMIR), Bloomington, Indiana.

Tzanetakis, G., & Cook, P. (2002). *Musical Genre Classification of Audio Signals*, IEEE Transactions on Speech and Audio Processing , Vol. 10, No. 5, July 2002.

Vapnik, V.N. (1998). *Statistical learning theory.* New York: Wiley

Vassilakis, P. N. (2001). *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance*. Doctoral Dissertation. Los Angeles: University of California, Los Angeles; Systematic Musicology.

Vassilakis, P. N. (2005). *Auditory roughness as a means of musical expression*, Selected Reports in Ethnomusicology 12 (Perspectives in Systematic Musicology), pp. 119-144.

VV, A A. (1973). *The scales of some African Instruments*. International Library of African Music, Sound of Africa Series – LP Records, pp. 91-107 http://anaphoria.com/depos.html

Witten, I. H., & Frank, E. (2005a). *Data Mining. Practical Machine Learning Tools and Techniques*, Second Edition, Elsevier, San Francisco, CA, USA.

Witten, I. H., & Frank, E. (2005b). *Weka 3: Data Mining Software in Java (Version 3.4)* [Computer Software]. Available from http://www.cs.waikato.ac.nz/ml/weka/

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and models*. Berlin: Springer.