Artificial Neural Network genre classification of musical signals

Fergani Lamya*, Amrane Houacine*

* Electronic & Informatic Faculty, University of Sciences & Technology, Algiers, Algeria

lamifer@msn.com
Ahouacine@lycos.com

Abstract: Searching and organizing growing musical collections for the Algerian radio requires classifying the music signals into a hierarchy of genres to structure them. Musical genres are defined as categorical labels that auditors use to characterize pieces of music. So, a musical genre can be characterized by a set of common perceptive parameters shared by its members. These perceptive parameters are closely related to the instrumentation, rhythmic structure and also harmonic content of the music. An automatic genre classification would actually be very helpful to replace or complete human genre annotation, which is actually used. In this paper, we explore the automatic genre classification of a musical Algerian database. More specifically, two feature sets composed of signal objective descriptors and which are closely related to perceptive ones (in this case timber) are proposed. The automatic classification of this database is then evaluated through an artificial neural network, more specifically a multiplayer perceptron (MLP). The parameters of this MLP are optimized to obtain the best scores in each case. We thus obtain scores of 60% to 80% for eight genres. Interesting comparative results are reported and commented.

Key words: audio classification, feature extraction, musical genre classification, music information retrieval, artificial neural network, multiplayer perceptron.

INTRODUCTION

No one can deny the incredible growing sizes of digital music libraries. A way of identifying music to listen to or find a particular song among a large collection of songs (as the collection of the Algerian radio) is to create labels, named musical genres. Musical genre is thus a way of describing what an item shares with other items as well as what differentiates it from others. There are no strict boundaries between genres for they depend highly on many factors: cultural, historical, public and marketing. But for sure, members of a particular genre share specifications relative to instrumentation, rhythmic structure and pitch content of the music itself.

The different approaches to genre classification that are actually in use are: the manual approach (needs musicologists and is time-consuming), the automatic approach based on the musical signal analysis. An emergent approach uses human-entered meta-data to group things together and therefore structure music pieces into genres. Most users, for music information retrieval; currently use genre hierarchies, typically created by human experts, as reference. Automatic genre classification can automate and accelerate this process. Beside this, it provides an efficient framework for developing and evaluating objective signal descriptors associated to subjective descriptors such as timber, pitch and rhythm.

Such objective descriptors or features may also form the basis of audio similarity, audio thumbnailing, and classification and segmentation analysis techniques.

In this paper, the problem of automatically classifying musical signals, extracted from Algerian radio database is addressed. Our goal is to verify if recent techniques available for occidental music can be adopted for the Algerian music and can lead to good musical genre classification rates. As mentioned above, we propose two sets of features to describe our musical signals. The lack of specific features for music timbral texture led us to derive most of them from earlier studies on speech and general sounds. Though the two sets attempt to represent the same subjective music descriptor, timber, they are constituted of different objective signal descriptors. The first set represents the timbral texture of the music signals through the well known Mel frequency Cepstral Coefficients (MFCC) associated with spectral temporal coefficients.

For the second set of features, we have associated timbral features to Linear prediction coefficients (LPC) combinated also with spectral and temporal
coefficients. The performances and importance of the proposed feature sets are evaluated by training a neural network, using music collections from different physical supports old disks, compact disks, radio and the web. Those pieces of music hierarchies are in fact defined by human auditors and accepted by most musicologists.

The paper is structured as follows. A short state of the art on genre classification is provided in section II. Section III deals with feature extraction and the two sets of parameters specific for music and describing timbral texture of the pieces. The evaluation of the performances for the proposed sets of features via automatic classification is described in section IV. Section IV deals with a comparative study of the two feature sets and combination of them through a specific neural network: the Multilayer Perceptron. Experiments and results are reported. The last section is devoted to conclusions and future directions.

1. Related work

The heart of any automatic musical classification or analysis system is the extraction of feature vectors. Though different classifiers have been compared [4], the choice of features has a large much effect to the recognition accuracy than the selected classifiers have. Even if artificial neural networks classifiers give satisfactory scores Many different sets of parameters have been proposed so far. A large number of them are mainly originating from speech recognition or analysis area. There are a wide variety of different features that can be used to characterize audio signals. They can be divided generally into time-domain and frequency-domain (spectral) features.

Mel-frequency cepstral coefficients (MFCC) are a set of perceptually motivated features that have been widely used in speech. MFCC analysis is a non-parametric method modeling the human auditory perception system. It provides a compact representation of the spectral envelope, such that most energy of the signal is concentrated in first coefficients (13 for speech and 5 for music). In order to model spectral variation of the data, the differentials of MFCC are also sometimes added to the feature set. MFCC have proven their efficiency for speech recognition, speaker recognition and musical instrument recognition [9]. Cepstrum coefficients, which are also a compact way to represent the coarse shape of the spectrum, are also commonly used [10]. Linear Prediction Coding (LPC), where the signal is approximated as a linear combination of past signal samples with an all-pole filter is a valid source-model for periodic signals produced by instruments or vocals: the poles of the filter correspond the peaks in the power spectrum. It provides an efficient way for calculating spectrum. In addition to the previously presented features, statistics over spectral distribution can represent pertinent characteristics of the musical spectrum. Spectral Centroid, Spectral Flux, Spectral Roll-off, Low Energy and Zero Crossing Rate were successfully used in [4] for musical genre classification. MFCC, LPC, Cepstrum and spectral statistics describe signal spectrum. Many authors use combinations of those coefficients to construct different timbral texture feature sets. Two different classification approaches have been used in the musical genre recognition systems. In short time audio analysis, the signal is broken into small and overlapping segments in time. These segments are called analysis windows and we assume that the signal for that short amount of time is stationary. Frames are then classified each separately; combining these classification results over a larger window gives the global classification result. The second approach takes account of the sound texture that arises from the temporal relationships between frames i.e. their temporal order. Therefore, the running means and variances of the extracted features described in the previous section are computed over a number of analysis windows. This larger window is called texture window. A single vector of features will represent each musical signal over the texture window and then genre classified.

2. Feature extraction

Feature extraction provides a compact numerical representation of the musical pieces. We have chosen to represent each signal by a unique feature vector over the texture window. This vector will only contain timbral descriptors. The Timber is closely related to the instrumentation of the music performance. It is a perceptual attribute of sound that enables us to distinguish it from other sound with same pitch, loudness and duration. Most objective signal descriptors associated to timbre describe the spectral distribution of sound. We will use the following features to characterize timbral texture. These latter can be divided into time-domain, spectral-domain, MFCC and LPC coefficients.

2.1. Mel Frequency Cepstral Coefficients (MFCC):

MFCC computation is based on STFT. The necessary steps for calculating MFCC are the following: a preprocessing step involves pre-emphasizing of the signal (filtering out high frequencies). The signal is then segmented into stationary frames of 20ms. Hamming window is used to weight the pre-emphasized frames. We then compute the Discrete Fourier Transform (DFT) for each frame. Since the human auditory system does not perceive pitch linearly, the frequencies are scaled nonlinearly to the Mel-frequency scale; This process imitate the behaviour of human hearing. After the Mel-scale filterbank, logarithm is applied to the amplitude spectrum .The Mel-spectral components are decorrelated with the Discrete Cosine transform (DCT). For each frame, 6 MFCC coefficients are obtained using this transform and the first coefficient is discarded, as it is function of channel gain. We have chosen 5 coefficients, for we are going to complete the timbre vector with some other spectral features.
2.2. Time-domain coefficients:

These coefficients are called time-domain because they are computed using the time envelop of the signal. We have chosen two time-domain coefficients that are widely used in audio signal processing.

2.2.1. Zero Crossings rate:

The zero crossings rate gives a measure of the noisiness of a signal. Zero crossings rate for musical signals is higher for musical signals than speech.

\[ Z_t = \frac{1}{2} \sum_{n=1}^{N} \text{sign}(x(n)) \# \text{sign}(x(n)) \]

Where the sign function is 1 for positive arguments and 0 for negative ones; x[n] is the time domain signal for frame \( t \).

2.2.2. Low Energy:

Low-Energy is defined as the percentage of analysis windows that have less RMS energy than the average RMS energy across the texture window.

2.3. Spectral-domain coefficients:

The spectral-domain coefficients are computed via the signal spectrum.

2.3.2. Spectral Centroid:

The spectral centroid is the center of gravity of the magnitude spectrum of the Short Time Fourier Transform (STFT). It measures the spectral brightness of a sound.

\[ C_t = \frac{\sum_{n=1}^{N} M_t(n) \# n}{\sum_{n=1}^{N} M_t(n)} \]

Where \( M_t(n) \) is the magnitude of the Fourier transform at frame \( t \) and frequency bin \( n \).

2.3.3. Spectral Flux:

The spectral flux is a measure of local spectral changes in the signal.

\[ F_t = \sum_{n=1}^{N} (N_t(n) \# N_{t-1}(n))^2 \]

where \( N_t(n) \) and \( N_{t-1}(n) \) are the normalized magnitude of the Fourier transform at current frame \( t \) and previous frame \( t-1 \) respectively.

2.3.4. Spectral Rolloff:

The Spectral rolloff is a measure of spectral shape. It’s defined as the frequency \( F \) below which 85% of the magnitude distribution is concentrated.

\[ \sum_{n=1}^{K} M_t(n) \# 0.85 \leq \sum_{n=1}^{N} M_t(n) \]

2.4. Linear Prediction Coefficients (LPC):

Linear prediction is one of the most powerful speech analysis techniques. The concept which is associated to it is very simple: predict the current value of the signal from stored past values of the signal. Each sample of the signal is a linear combination of previous samples expressed through a difference equation. This equation is called a linear predictor and is defined as follows:

\[ s(n) = \sum_{k=1}^{p} a_s(n-1) - GU(n) \]

\( p \) is the order of prediction. The coefficients of equation (5) , \( a_s \), must be estimated. The estimation is done by minimizing the mean-square error between the predicted signal and the actual signal. Several efficient methods (autocorrelation, covariance, recursive lattice formulation) exist that assure convergence to a unique solution. So, we can replace the signal by \( p \) prediction coefficients (LPC) which will completely describe its characteristics.

Before the features extraction, the time-domain musical signals are normalized to have zero mean and unity variance. After that, music signals are divided into 20ms frames (analysis windows); Using hamming windowing minimizes edge effects and successive analysis windows overlap each other 10ms. We then compute for each frame the descriptors cited in the previous section (MFCC, LPC, time-domain and spectral-domain).

The means and variances of spectral centroid, rolloff, flux; zero-crossings, MFCC and LPC over the texture window of 3s are computed (3s is shown to be the minimum time amount necessary to identify a particular piece of music by human listeners). The low energy descriptor is already computed over the texture window. The dimension of the resulting timbral feature vector is depending on which combination of the coefficients is used.

3. Classification

In order to classify our data we have chosen, among classical classifiers, artificial neural networks for this task. And among different artificial neural networks (ANN), we will optimize a Multilayer Perceptron (MLP) to evaluate the proposed feature sets. This MLP is trained using real Algerian music pieces.

3.1. Datasets

Training the classifiers needs a large collection of
Algerian musical signals. Various recording qualities were used to choose the musical excerpts: radio broadcasting, compact disks, and old disks. Two different audio format files were also used: the wav and the MP3 format. The Algerian musical genres were labeled following a study over 50 subjects of different ages and positions. The most cited labels have been adopted as Algerian genres dataset. These are: Andalou, Chaabi, Chaoui, Haouzi, Kabyle, Malouf, Rai, Staifi. 8 excerpts represent each of the 8 genres. It leads to a database of 64 excerpts. The files were stocked at 22050 Hz, 16 bits, mono audio files.

3.2. Artificial neural networks

The ANN very loosely simulates a biological neural system. We have implemented a specific ANN classifier: a Multi layer Perceptron, which is a feedforward network. The training algorithm used with this network is back propagation (RPG), which is the mostly used. The network is composed of input, hidden, and output layers. The inputs of the first layer are the feature vector described in the previous section. So the size of this vector gives the number of neurons of this layer. The number of neurons of the output layer is equal to number of genres (eight for our case). The output uses the code “one for N”: one neuron represents a genre and is activated if the genre is the good one. The values of the network parameters (weights and biases) are computed iteratively in the process of network training (learning). After training the network should be tested. The classification results are calculated using a ten-fold cross-validation where the dataset is randomly partitioned so that 10% is used for testing and 90% is used for training. The process is iterated with different partitions and the results are averaged. This ensures that the calculated accuracy will not be biased because of a particular partitioning of training and testing.

4. Results

These results are obtained using a single vector representing each audio file. This vector consists in timbral feature set computed over the texture window. Listening tests permitted the best choice for the texture window position in the whole file. The dimension of this vector varies for we have tested different combinations of MFCC, LPC, time-domain and spectral-domain coefficients. The number of MFCC and LPC coefficients varies also. We propose in this paper 9 combinations to be evaluated.

After training the MLP, we have found that the best number of neurons for the hidden layer is six. The results presented in this section will thus take account of this.

4.1. Classification accuracy percentage

Figure 1 shows the best classification accuracy percentage results for each of the six genres proposed. These results are obtained with the feature vector containing time-domain, spectral-domain and 5 MFCC’s. The global score for all the genres is of 86.11%. We thus obtain satisfactory results; We can also see that the genres: Haouzi, Chaoui and Chaabi obtain the best scores (100%) while the genre labeled Rai obtain the worst one (58.33%). This is in part due to the label Rai which may contain musical pieces that doesn’t share the same instrumentation.

4.2. Confusion Matrix:

Table 2 shows the details of genre classification performances in the form of confusion matrix. The columns of that matrix correspond to the actual genre and the rows to the predicted genre. The percentage of correct classification lies in the diagonal of the confusion matrix. For example, the cell of row 2, column 2 with value 87.5% means that 87.5% of the Haouzi was classified as Haouzi (good classification accuracy for Haouzi: 87.5% of Haouzi are classed as Haouzi). The cell of row 3, column 2 with value 12.5% means that 12.5% of Haouzi are wrongly classed as Kabyle.

We can see that the best scores lie really on the diagonal of the confusion matrix. Best scores are of 87.5% for Haouzi and Chaoui; Worse scores are of 50% for Staifi, Rai and Andalou. The confusion between genres is essentially due to the choice of the excerpts and to the broad limits between genres as Malouf and Andalou, which maybe classified as subgenres in a larger class Classical. The same conclusion could be done for genres as Staifi and Rai. They maybe sub genres for a more general class labeled Modern.

4.3. Importance of feature sets:

Figure 2 shows the importance of the different feature set proposed. The results are obtained for the MLP classifier. We give the classification accuracy averages for all genres. In all cases, feature sets containing MFCC coefficients gives the best scores; The best score 86.11% is obtained when combining time-domain, spectral-domain and 5 MFCC’s. It seems that a higher number of MFCC (9 or 13) doesn’t ameliorate the classification accuracy. The scores obtained for the LPC coefficients and their combination with time or spectral domain ones are the worse. The proposed feature vectors are the following: TS5MFCC (Time, Spectral and 5MFCC), TS5LPC...
(Time, Spectral and 5LPC), TS (Time and Spectral), 5LPC, 9LPC, 13LPC, 5MFCC, 9MFCC and 13MFCC.

Table 2: Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Chaabi</th>
<th>Haouzi</th>
<th>Kabyle</th>
<th>Malouf</th>
<th>Chaoui</th>
<th>Stafi</th>
<th>Rai</th>
<th>Andalou</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chaabi</td>
<td>62.5%</td>
<td>12.5%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>25%</td>
</tr>
<tr>
<td>Haouzi</td>
<td>0%</td>
<td>87.5%</td>
<td>0%</td>
<td>0%</td>
<td>12.5%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Kabyle</td>
<td>0%</td>
<td>12.5%</td>
<td>75%</td>
<td>0%</td>
<td>0%</td>
<td>12.5%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Malouf</td>
<td>0%</td>
<td>12.5%</td>
<td>0%</td>
<td>62.5%</td>
<td>0%</td>
<td>25%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Chaoui</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>12.5%</td>
<td>87.5%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Stafi</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>25%</td>
<td>50%</td>
<td>0%</td>
<td>25%</td>
</tr>
<tr>
<td>Rai</td>
<td>12.5%</td>
<td>25%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>50%</td>
<td>12.5%</td>
</tr>
<tr>
<td>Andalou</td>
<td>0%</td>
<td>12.5%</td>
<td>12.5%</td>
<td>25%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>50%</td>
</tr>
</tbody>
</table>

5. Conclusion and Future research

Despite the good results obtained for our Algerian datasets, many future researches had to be done. The genre hierarchy has to be expanded: more genres and also many sub genres must be included such as new genres not depicted in our actual work (such as gnawi music); Sub genres has also to be defined for we can imagine to classify Andalou, Chaabi and Haouzi as sub genres of a more general genre labeled Classical for example. An interesting direction for future research is to associate instrument recognition to our classifier, for Algerian genres maybe described by the type of instrument played. We must also add to our classification system a class “other” where non-Algerian music should be placed. The number of excerpts for each genre should also be augmented; this would for sure lead to better classification rates. The proposed feature set could also be used for similarity retrieval, audio thumb nailing. Adding pitch and rhythm descriptors to the feature set and using new classifiers such Structure Vector Machine (SVM) are future directions for research.

2. References:


