Independent features for content-based music genre classifiaction

L. Fergani

University of Sciences and Technology, BP14, El-Alia, Bab Ezzouar, Algiers,Algeria, 16111 Algiers, Algeria
lamifer@msn.com

In this paper, we consider the problem of finding latent structure in high dimensional data. It is assumed that the observed data (here music signals) are generated by unknown latent variables and their interactions. The task is to find the latent variables and the way they interact, given the observed data only. A novel method for solving the above problem is independent component analysis (ICA). Furthermore, ICA achieves independency of the features components which reduces redundancy of information. The new independent features are then used for genre classification through Support Vector Machine (SVM) classifier. We will show through different experiments that this approach gives better accuracy rates than classical feature sets such as wavelet based, spectral, temporal or MFCC feature sets associated with different classifiers such as Multiclass SVM, Multilabel SVM. These results are obtained with a database of 64 songs issued from the database of the Algerian radio. We thus obtain scores of 93% to 95% for six genres. Interesting comparative results are reported and commented.

1 Introduction

Progress in networking transmission, compression of audio, and protection of digital data allows now or in the near future to deliver quickly and safely music to users in a digital format through networks, either Internet, or digital audio broadcasting. Further, digitalization of data makes it possible for users, to access huge catalogues of musical titles. This situation demands for tools able to ease searching, retrieving, and handling such a huge amount of data. Among those tools, automatic musical genre classifiers (AMGC) can have a particularly important role, since they could be able to automatically index and retrieve audio data in a human-independent way. Music genre is very useful for music indexing and content-based music retrieval. The research field of automatic music genre classification has got increasing importance in the last few years. Music Information Retrieval (MIR) is the most important application of AGC, but it is not the only one. The automatic analysis of music stored in audio format is one of the important topics of MIR [3]. The majority of such audio analysis techniques make use of numerical features, called descriptors, that attempt to capture information about musical content. Many different sets of descriptors have been proposed so far. A large number of them are mainly originating from speech recognition or signal processing area. They can be divided generally into time-domain. or spectral-domain features [1]. Wavelet based features have also been introduced and have proven their efficiency. In this work, we propose a new set of independent features, which are obtained by applying Independent Component Analysis over a concatenation of basic features. It will be compared to different classical sets of features obtained either by concatenation of basic features (spectral, temporal and MFCC) or built on statistics of wavelet coefficients. The performances and importance of the proposed feature sets are evaluated by training three pattern recognition classifiers: a Multi-Class Support Vector Machine (MC-SVM), a Multi-Label Support Vector Machine (ML-SVM), and an Artificial Neural Network (ANN). We used, at this purpose, music collections from different physical supports old disks, compact disks, radio and the web. The chosen taxonomy for musical pieces is defined by the Algerian radio listeners and accepted by most musicologists.

The paper is structured as follows. A short state of the art on music features extraction is provided in section 2. Section 3 deals with Independent Component Analysis. Section 4 defines and describes how to use ICA to built independent features.The evaluation of the performances for the proposed feature sets via automatic classification is described in section 5. Experiments and interested results are reported. The last section is devoted to conclusions and future directions

2 Feature extraction

Two different approaches have been used to extract audio features [4]. 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. 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.

2.1 Time or Spectral domain features

These features are computed over the time or spectral representation of the signal [1]. Among all the huge quantity of existing features, we have chosen the following ones:

- *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.

$$Z_t = \frac{1}{2}\sum_{n=1}^{N}\left|sign(x(n) - sign(x-1)\right| \quad (1)$$

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

- *Spectral Centrod:* the spectral centroid is the gravity centre of the magnitude spectrum of the Short Time Fourier Transform (STFT). It measures the spectral brightness of a sound

$$C_t = \frac{\sum_{n=1}^{N} M_t(n) * n}{\sum_{n=1}^{N} M_t(n)} \quad (2)$$

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

- *Spectral Rolloff:* Spectral rolloff is a measure of spectral shape. It's defined as the frequency F below which 85% of the magnitude distribution is concentrated.

$$\sum_{n=1}^{R_t} M_t(n) = 0.85 * \sum_{n=1}^{N} M_t(n) \quad (3)$$

- *Spectral Flux:* Spectral flux is a measure of local spectral changes in the signal.

$$F_t = \sum_{n=1}^{N} (N_t(n) - N_{t-1}(n))^2 \quad (4)$$

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.

- *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.

2.2 Wavelet features

The Wavelet Transform (WT) and more particularly the Discrete Wavelet Transform (DWT) is a relatively recent and computationally efficient technique for extracting information about non stationary signals like music. It provides a compact time-frequency representation of a signal. The DWT analysis can be performed using a fast, pyramidal algorithm. In the pyramidal algorithm the signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive high pass and low pass filtering of the time domain signal and is defined by the following equations:

$$d(k) = \sum_n x(n) g(k - 2n) \quad (5)$$

$$a(k = \sum_n x(n) h(k - 2n) \quad (6)$$

where d(k) and a(k) are respectively the detail information and the coarse approximation of the signal (i.e. outputs of the high pass and low pass filters g and h). The filters must be chosen carefully and there are a variety of different wavelet families that have been proposed so far : Coiflet, Symlet, Meyer and Daubechies [5]. The properties of the wavelet condition the quality of the wavelet analysis. In order to further reduce the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients are used. That way the statistical characteristics of the "texture" or the "music surface" of the piece can be represented [4].

We built the following features: the mean of the absolute value of the coefficients in each subband, the standard deviation of the coefficients in each subband, the ratios of the mean values between adjacent subbands. These features provide information either about frequency distribution or amount of change of this distribution.

2.3 Mel Frequency Cepstral coefficients

The MFCC represent the shape of the spectrum with very few coefficients. The cepstrum is the Fourier Transform (or Discrete Cosine Transform DCT) of the logarithm of the spectrum . The Mel-cepstrum is the cepstrum computed on the Mel-Bands instead of the Fourier spectrum. The use of Mel scale allows to better take into account the mid-frequencies part of the signal. The MFCC are the coefficients of the Mel cepstrum. The first coefficient is being proportional to the energy is not stored, The five next ones are stored for each frame. We can also store the Delta-MFCC and Delta-Delta MFCC which are the first and second order derivative of the MFCC along time. The MFCC are computed following the scheme represented on Fig.1

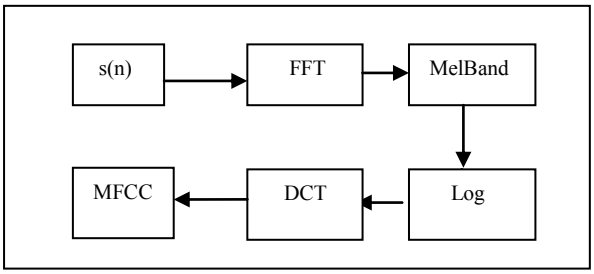

Figure 1: MFCC computation

3 Latent variables estimation

Here, the term latent means hidden, unknown or unobserved. The term structure refers to some regularities in the data which consist in feature vectors described in section 2. It will be assumed that these feature vectors ,that we can also name observed data are generated by interactions between latent variables. The objective is to find out what these latent variables are and how they interact. Depending on the point of view, the "structure" in the data is either due to the values taken by the latent variables or due to the way the latent variables interact. We will assume that there are no inherent dependencies between the latent variables. Independent Component Analysis is a well-known method of finding latent structure In data. ICA is a statistical method that expresses a set of multidimensional observations as a combination of unknown latent variables. These underlying latent variables are called sources or independent components and they are assumed to be statistically independent of each other.

There are two schools of thought with respect to what actually is the aim in estimating the independent components in the data. First, one may be regard the data being generated by a combination of some existing but unknown independent source signals s_i, and the task is to estimate them. This viewpoint is chosen in the so called blind source separation (BSS) framework. Another point of view is to regard ICA as a method of presenting the data in a more comprehensible way by revealing the hidden structure in the data and often reducing the dimensionality of the representation. According to his latter school of

thought, it might well be that there are no "true" source signals generating the data but it still pays to represent the data as a combination of few latent factors that are statistically as independent as possible. It is a data mining approach of the problem.

3.1 Independent Component Analysis

The Independent Component Analysis (ICA) allows the separation of sources under the hypothesis of statistical independence. Some estimation algorithms of the ICA model already exist. These different versions change according to the selected contrast function and scheme adopted for its optimization. Their performances differ in their stability, convergence speed and their memory need [12]. The ICA of a random real vector **x** consists in finding a linear transformation **s** = W**x** so that the components would be as independent as possible in the way to maximize a function $\Omega(s_1, s_2 \ldots s_m)$ which measures the independency. Ω is called contrast function or cost function. The ICA model as defined previously presents in addition to the condition of sources statistical independency the following restrictions:

- At the most one among the sources conforms to a gaussian distribution.

- Two kinds of indeterminations are generated by the ICA model: an indetermination on the estimated components order and an indetermination on the amplitude of the estimated sources.

3.2 Fast Independent Component Analysis algorithm

The Fast Independent Component analysis (FASTICA) algorithm developed by Hyvarinen and Oja [13] uses the negentropy (non negative entropy) as contrast function. The negentropy J(**x**) associated to $p_x(u)$ will be defined as:

$$J(\mathbf{x}) = \int_{-\infty}^{+\infty} p_x(u) \log \frac{p_x(u)}{\Phi_\upsilon(u)} du \qquad (7)$$

Where $p_x(u)$ is the probability density function of the random vector **x** and ν is the gaussian centred random variable, with the same variance than **x**, which admits $\Phi_\nu(u)$ as probability density function. A robust estimator of the negentropy has been proposed by Hyvarinen in [4]. The FASTICA algorithm uses also the Newton-Raphson method to maximize this measure. A learning rule searches the direction (i.e. a line vector w from the separation matrix W) for which the w^Tx projection maximizes the non Gaussianity in the meaning of the negentropy $J(w^Tx)$. The w^Tx variance must be equal to the unit, thus for whitened data this is equivalent to restrain the norm of w to unity. The maxima of $J(w^Tx)$ are obtained for some optima of $E\{G(w^Tx)\}$. Then, according to the Kuhn-Tucker [4] condition, the optima of $E\{G(w^Tx)\}$ under the constrain:

$$E\left\{G\left(w^T x\right)^2\right\} = \|w\|^2 = 1 \qquad (8)$$

are obtained at the points where :

$$E\left\{xg\left(w^T x\right)\right\} - \beta w = 0 \qquad (9)$$

g is the derivative of the G function

3.3 Sources estimation

Equation (4) is resolved by the Newton method, this comes to calculate each line w of the matrix W by formula:

$$w(k) = E\left\{xg\left(w^T(k-1)x\right)\right\} - E\left\{g'(w^T(k-1)x\right\}w(k-1) \qquad (10)$$

The algorithm takes then the following form:

1) Take a random initial vector w(0)

2) Do $w(k) = E\left\{xg\left(w^T(k-1)x\right)\right\} - E\left\{g'(w^T(k-1)x\right\}w(k-1)$

3) Normalize w(k)

4) If $\|w^T(k)w(k+1)\|$ is not enough near to 1, then go back 2, else pull out w(k).

The resulting vector is a column of W allowing to separate one among the non-Gaussian sources with $w^T(k)x$ is one of the sources and x is an observation. In order to estimate several components, the previous algorithm must be used and to avoid the convergence of the vectors to the same maximum, it is necessary to decorrelate the projections $w^T(1)x$,……., $w^T(n)x$ after each iteration. Two approaches are possible: a deflation approach or a symmetric decorrelation approach [4].

4 Evaluation and Results

In order to evaluate the proposed feature sets, we trained a Multi-Class Support Vector Machine (MC-SVM) classifier using real Algerian music pieces.

4.1 Datasets

Training the classifier 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 which associates to each example a unique label. The files were stocked at 22050 Hz, 16 bits, mono audio files.

4.2 Classification

The database is divided into 80% of the files for the learning process and 20% for the test. We have implemented a Multi-Class Support Vector Machine to achieve the classification task, with "one against one" strategy. The necessary learning step uses the Sequential Minimal Optimization (SMO) algorithm We have chosen the RBF kernel and a grid search allows during the learning step to optimize the pair of parameters C (penalty parameter) and σ (gaussian kernel width).

4.3 Feature sets

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 windows

minimizes edge effects and successive analysis windows overlap each other every 10ms. We then compute for each frame the descriptors cited in the previous section.

The means and variances of spectral centroid, rolloff, flux, zerocrossings and five first MFCC 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 resulting feature vector (TSPMFCC) is a 19 dimensional vector. Applying ICA to extract the latent structure leads to a new feature vector (TSPMFCC-IND). For wavelet feature vectors the computation is slightly different; For each music piece, a discrete wavelet transform is computed over a segment of 3s. The statistics of the coefficients of the DWT are then used to build acoustic vector for each signal. Their dimension depends on the chosen resolution, the type of filter and the statistics used. Three sets of feature vectors are extracted : the mean of the absolute value of the coefficients in each subband (MVPEA), the standard deviation of the coefficients in each subband (STDPEA) and the ratios of the mean values between adjacent subbands (MCPEA). We used the following wavelet filters : Daubechies1(db1), Daubechies5 (db5), Meyer (Meyer), Symlet2 (Sym2) and Coiflet1 (Coif1).

4.4 TSPMFCC-IND features

Table I represents the grid-search of the learning step for the MC-SVM. The feature set used is the TSPMFCC-IND. The RBF kernel has been chosen. We can see that for $C=2^8$ and $\sigma= 2^{-9}$, the classification accuracy percentage reachs 95% which means that the learning step was succesfull. Table II gives the confusion matrix for the test step. It shows the classification scores for six genres. We can see that these scores are of 100% for five genres while they reach 75% for the last one leading to a global score of 95.8%.

Table I: Grid-Search (C, σ), Learning step

C	2^0	2^2	2^4	2^6	2^8
2^0	0.8	0.8	0.75	0.75	0.75
2^{-1}	0.8	0.9	0.9	0.85	0.85
2^{-3}	0.75	0.9	0.85	0.8	0.8
2^{-5}	0.7	0.8	0.9	0.85	0.85
2^{-7}	0.6	0.7	0.8	0.9	0.9
2^{-9}	0.6	0.6	0.7	0.8	**0.95**

Table II: Confusion matrix, Test step

	Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	Raï
Chaabi	**1**	0	0	0	0	0
Kabyle	0	**1**	0	0	0	0
Chaoui	0	0	**1**	0	0	0
Haouzi	0	0	0	**1**	0	0
Staïfi	0	0	0	0	**1**	0
Raï	0	0	0	0.25	0	**0.75**

4.5 TSPMFCC-IND versus TSPMFCC

In order to compare between the TSPMFCC and the TSPMFCC-IND features, Figure 2 shows the accuracy percentages for different TSPMFCC and TSPMFCC-IND feature vectors. These vectors are composed of different combinations of simple descriptors, but their dimension is the same: 26 descriptors. We have Class I , Class II, Class III and Class IV feature vectors. All the combinations contain MFCC descriptors. The classifier is the MC-SVM with RBF kernel. We can see that the best score is obtained for TSPMFCC-IND with 85%. The feature set for this score is composed of 13 MFCC and 13 Delta-MFCC.

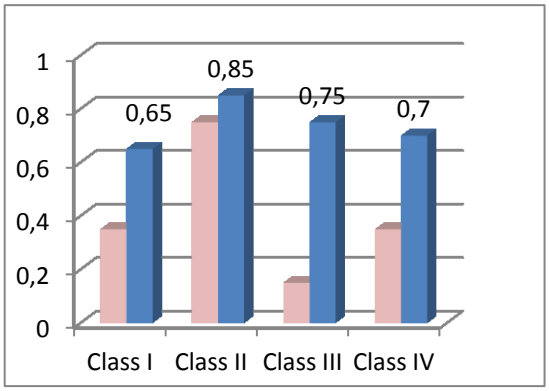

Figure 2: TSPMFCC-IND versus TSPMFCC

4.6 TSPMFCC-IND versus others

In order to compare between our TSPMFCC-IND and other feature vectors, Figure 3 gives accuracy percentages for two classifiers and for different acoustic vectors. We used MC-SVM and Multi-Label SVM classifiers. Two categories of feature vectors have been used: TSPMFCC and Wavelets. TSPMCC features are labeled Class (for classic), IND (for independent), Inst (for instantaneous) or Fus (for fusion). Wavelet features are labeled MCPEA, MVPEA, STDPEA and STDPEA-Multilabel (STDPEA with Multilabel SVM). We can see that the best score is obtained for the TSPMFCC-IND features and reachs 95% followed by the MCPEA-Multilabel features with a score of about 93%..

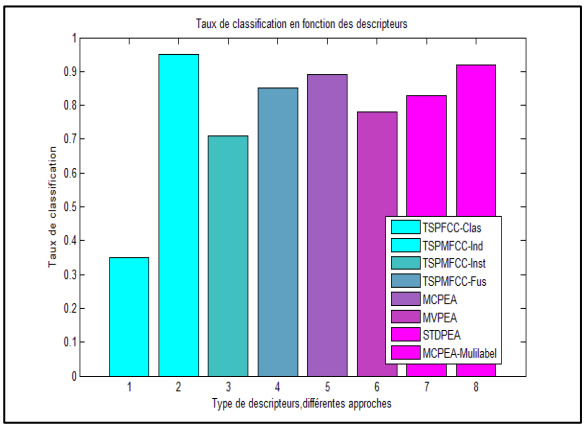

Figure 3: TSPMFCC-IND versus others

5 Conclusions

We can see that introducing Independent Component Analysis to retrieve latent structure in our data improved the scores of genre classification. 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. New genres not depicted in our actual work (such as gnawi music must be included; 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. The number of excerpts for each genre should also be augmented; this would for sure lead to better classification rates. Another way of research may concern reducing the dimension of feature vectors which may be done in parallel with Independent Component Analysis.

REFERENCES

[1] G. Tzanetakis, P. Cook, "Musical genre classification of audio signals", in IEEE Trans. On Speech and Audio Processing, 2002, pp. 293-302.

[2] R. O. Duda, P. E. Hart, Pattern recognition, John Wiley and sons, 2001.

[3] M. Mandel, D. Ellis, "Song-level features and support vector machines for music classification", in Proc. Int. Conf. on Music Information Retrieval (ISMIR'05), London, 2005, pp. 594–599.

[4] T. Lambrou, P. Kudumakis and all, "Classification of audio signals using statistical features on time and wavelet transform domains" in Proc. Int. Conf. on acoustics, speech and signal processing (ICASSP98), USA, pp. 3621-3624.

[5] J.J. Aucouturier, F. Pachet, "Musical genre: a survey", in Journal of new music research,2003, pp. 1234-1265.

[6] M.R. Boutell, J. Luo, X. Shen and C.M. Brown, "Learning multi-label scene classification", in Pattern Recognition, 2004, pp. 1757-1771.

[7] K. Brinker, J. Furnkranz and E. Hullermeier, "A unified model for multilabel classification and ranking" in Proc. European Conf. on Artificial Intelligence (ECAI'06), Italy, 2006, pp. 489-493.

[8] V. Vapnik, Statistical Learning Theory, John Wiley and sons, 2001.

[9] T. Li, C. Zhang and S. Zhu, "Empirical studies on multilabel classification" in Proc. Int. Conf. on Tools with Artificial Intelligence, USA, 2006, pp. 86–92.

[10] G. Tsoumakas, I. Katakis, "Multi-label classification: An overview", in International Journal of Data Warehousing and Mining, 2007, pp.1–13

[11] C. Jutten, J. Herault "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture" in Signal Processing, 1991, pp.1-10

[12] A. Hyvärinen "Fast and robust fixed-point algorithms for independent component analysis" in IEEE Trans. On Neural Networks, 1999, pp.626-634