

Automatic genre classification of music content: a survey

N. Scaringella, G. Zoia, *Member, IEEE*, D. Mlynek, *Member, IEEE*

Abstract— The creation of huge databases coming from both restoration of existing analog archives and new content is demanding more and more reliable and fast tools for content analysis and description, to be used for searches, content queries and interactive access. In that context, musical genres are crucial descriptors since they have been widely used for years to organize music catalogues, libraries and music stores. Despite their use, musical genres remain a poorly defined concept, which make of the automatic classification problem a non-trivial task. In this article, we review the state-of-the-art in automatic genre classification and present new directions in automatic organization of music collections.

Index Terms—Musical genres, music information retrieval, feature extraction, machine learning.

I. INTRODUCTION

Musical genres are the main top-level descriptors used by music dealers and librarians to organize their music collections. Though they may represent a simplification of one artist's musical discourse, they are of a great interest as summaries of some shared characteristics in music pieces.

With Electronic Music Distribution (EMD), music catalogues tend to become huge (the biggest online services propose around 1 million tracks); in that context, associating a genre to a musical piece is crucial to help users finding what they are looking for.

In fact, the amount of digital music urges for efficient ways to browse, organize and dynamically update collections: it definitely requires new means for automatic annotation. In the case of music genre annotation, Weare [1] reports that the manually labeling of a hundred thousand songs for Microsoft's MSN Music Search Engine needed about 30 musicologists for one year.

Manuscript submitted on November 1st, 2005.

N. Scaringella is with the Signal Processing Institute EPFL-STI-ITS-LTS-3, ELB 116, Station 11, CH-1015 Lausanne, Switzerland (phone: +41 21 693 09 29; fax: +41 21 693 46 63; e-mail: nicolas.scaringella@epfl.ch).

G. Zoia, is with the Signal Processing Institute EPFL-STI-ITS-LTS-3, ELB 116, Station 11, CH-1015 Lausanne, Switzerland (phone: +41 21 693 69 81; fax: +41 21 693 46 63; e-mail: giorgio.zoia@epfl.ch).

D. Mlynek, is with the Signal Processing Institute EPFL-STI-ITS-LTS-3, ELB 121, Station 11, CH-1015 Lausanne, Switzerland (phone: +41 21 693 46 81; fax: +41 21 693 46 63; e-mail: daniel.mlynek@epfl.ch).

At the same time, even if terms such as *Jazz*, *Rock* or *Pop* are widely used, they often remain loosely defined so that the problem of automatic genre classification becomes a non-trivial task. In the second section of this survey, we discuss the importance of musical genres with their definitions and hierarchies. The third section presents techniques to extract meaningful information from audio data to characterize musical excerpts. We then review the state of the art in genre classification through three main paradigms: expert systems, unsupervised classification, and supervised classification; some recent results are reported in the final section which is devoted to new emerging research fields and techniques that investigate the proximity of musical genres.

II. MUSICAL GENRES

Musical genres are categories that have arisen through a complex interplay of cultures, artists and market forces to characterize similarities between musicians or compositions and organize music collections. Yet, the boundaries between genres still remain fuzzy as well as their definition making the problem of automatic classification a non-trivial task.

The music genre classification problem asks for a taxonomy of genres i.e. a hierarchical set of categories to be mapped onto a music collection. Pachet and Cazaly [2] studied a number of musical genre taxonomies used in industry and on the Internet and showed that it is not straightforward to build up such a hierarchy of genres. As a good classification relies on a carefully thought taxonomy, we start here from a discussion on a number of critical issues.

A. *Artists, albums or titles?*

One basic question to be raised is to what kind of musical item genre classification should apply: to a title, to an album or to an artist. If we suppose that one song can be classified into one genre (which is already questionable), it is not that simple anymore for an album, which may contain heterogeneous material. The same applies to artists; some of them have covered such a wide range of genres during their career that it does not make too much sense to try to associate them to a specific class.

B. *Non-agreement on taxonomies*

Pachet and Cazaly [2] showed that a general agreement on genre taxonomies does not exist. Taking the example of

well-known web sites like Allmusic¹ (531 genres), Amazon² (719 genres), and Mp3³ (430 genres), they only found 70 terms common to the 3 taxonomies. They notice that some widely used terms like *Rock* or *Pop* denote different sets of songs and those hierarchies of genres are differently structured from one taxonomy to the other.

C. Ill-defined genre labels

Taking a close look at some specific and widely used musical genres, we observe how varied can be the criteria that define a specific genre; some examples:

- *Indian* music is geographically defined;
- *Baroque* music is related to an era in history (while encompassing a wide range of styles and a wide geographic region);
- *Barbershop* music is defined by a set of precise technical requirements;
- *Post-rock* is a term devised by music critic Simon Reynolds

Pachet and Cazaly [2] argue that this semantic confusion within a single taxonomy can lead to redundancies that may not be confusing for human users but that may hardly be solved by automatic systems. Furthermore, genre taxonomies may be dependant on cultural references. For example, a song by the French singer Charles Aznavour would be considered as *Variety* in France but would be filed as *World Music* in the UK.

D. Scalability of genre taxonomies

Hierarchies of genres should also consider the possibility to add new genres to take into account music evolution. New genres appear frequently and are typically the result of some merging of different genres (*Psychobilly* can be seen as the merging of *Rockabilly* and *Punk*) or the splitting of one genre into subgenres (original *Hip-Hop* has lead to different subgenres such as *Gangsta Rap*, *Turntablism*, *Conscious Rap*). This is a major issue for automatic systems. Adding new genres and subgenres to a taxonomy is easy but having an automatic system requiring supervised training able to adapt itself is somehow tricky.

E. Local conclusion

Due to the difficulty of defining a universal taxonomy, more reasonable goals must be considered. In fact, Pachet

¹ <http://www.allmusic.com>

² <http://www.amazon.com>

³ <http://www.mp3.com>

and Cazaly eventually gave up their initial goal to define a general taxonomy of musical genres [2] and Pachet and al. decided to use simple two-level genre taxonomy of 20 genres and 250 subgenres in the context of the Cuidado music browser [3].

III. FEATURE EXTRACTION

In the digital media world, generic audio information is mostly represented by bits allowing a direct reconstruction of an analogue waveform. But accepting to decrease generality (e.g. to common western music) music information can also be described more or less accurately by some higher-level model-based representations – typically event-like formats such as MIDI or symbolic formats such as MusicXML.

However, in real world applications a precise symbolic representation of a (new) song is rarely available and one has to deal with the most straightforward form, i.e. audio samples. Audio samples, obtained by sampling the exact sound waveform, can not be used directly by automatic analysis systems because of the low level and low “density” of the information they contain; put in another way, the amount of data is huge and the information contained in audio samples taken independently is too small to deal with humans at the perceptual layer (as opposite to sensorial). The first step of analysis systems is thus to extract some *features* from the audio data to manipulate more meaningful information and to reduce the further processing.

Extracting features is the first step of most pattern recognition systems. Indeed, once significant features are extracted, any classification scheme may be used. In the case of audio signals, features may be related to the main dimensions of music including melody, harmony, rhythm, timbre and spatial location.

A. *Timbre*

Timbre is currently defined in literature as the perceptual feature that makes two sounds with the same pitch and loudness sound different. Features characterizing timbre analyze the spectral distribution of the signal though some of them are computed in the time domain. These features are global in the sense that they integrate the information of all sources and instruments at the same time.

1) *Timbre features:*

An exhaustive list of features used to characterize timbre of instruments may be found in [4]. Most of these descriptors have been used as well in the context of music genre recognition, though some features are more adapted

to characterize monophonic instrument rather than polyphonic mixtures. These descriptors are usually referred to as being *low-level* since they describe sound on a fine scale (they are typically computed for slices of signal between 10 and 60 seconds). Some of these descriptors have been normalized in the MPEG-7 Audio standard [30] i.e. their extraction algorithm is normative.

We summarize here the main low-level features used in genre characterization applications:

- **Temporal features:** features computed from the audio signal frame (zero-crossing rate, linear prediction coefficients, etc.).
- **Energy features:** features referring to the energy content of the signal (Root Mean Square energy of the signal frame, energy of the harmonic component of the power spectrum, energy of the noisy part of the power spectrum, etc.).
- **Spectral shape features:** features describing the shape of the power spectrum of a signal frame: centroid, spread, skewness, kurtosis, slope, roll-off frequency, variation, Mel-Frequency Cepstral Coefficients (MFCCs).
- **Perceptual features:** features computed using a model of the human hearing process (relative specific loudness, sharpness, spread).

Transformations of features such as first and second-order derivatives are also commonly used to create new features or to increase the dimensionality of feature vectors.

The importance of psycho-acoustic transformations for effective audio feature calculation is studied in [27] in the context of genre recognition. It is suggested that transforming spectrum energy values into the logarithmic decibel scale, calculating loudness levels through incorporating equal-loudness in the Phone scale and computing specific loudness sensation in terms of the Sone scale are crucial for the audio description task.

2) *Texture window:*

Most of these descriptors are computed at regular time intervals, over short windows of typical length between 10 to 60 ms. In the context of classification, timbre descriptors are then often summarized by evaluating low-order statistics of the descriptors' distribution over larger windows commonly called *texture windows* [5]. Modeling timbre on a higher time scale not only reduces computation further but it is also perceptually more relevant as the short frames of signal used to evaluate features are not long enough for human perception. It is suggested in [6] that better

classification results may be obtained by modeling feature evolution over a texture window with an autoregressive model rather than with simple low-order statistics.

The impact of the size of the texture window over classification accuracy has been studied in [5]. It is shown that indeed the use of a window increases significantly the classification accuracy compared to the direct use of the analysis frames. The conclusion is that texture windows of 1 second are a good compromise since no significant gain in classification accuracy is obtained by taking larger windows while the accuracy decreases (almost linearly) as the window is shortened.

Rather than using texture windows with constant size and arbitrary positions, some authors try to associate windows to actual musical events. West and Cox [7] segment the audio stream with an onset detector whereas Scaringella and Zoia [8] use a musical beat tracker. The extracted segments are then used as the usual texture windows with timbre information supposed to be more coherent.

B. Melody, Harmony

Harmony may be defined as the use and study of pitch simultaneity and chords, actual or implied, in music. On the contrary, melody is a succession of pitched events perceived as a single entity. Harmony is sometimes referred to as the vertical element of music with melody being the horizontal element. Melodic and harmonic analysis having been for a long time used by musicologists to study musical structures, it is tempting to try to integrate such analysis when modeling genre.

A good overview of melody description and extraction in the context of audio content processing can be found in [9]. For the estimation of multiple fundamental frequencies of concurrent musical sounds, one may refer to [10] while chord extraction is addressed e.g. in [11].

In any case, for the time being, melodic and harmonic content are more robustly described by lower level attributes than notes or chords. To our knowledge, there has been only one attempt to use such features when modeling genres of audio signals [5] while they have been used more intensively in the context of semantic segmentation and summarization of music [31]. The basic idea is to use a function characterizing pitch distribution of a short segment like most melody/harmony analyzers; the difference is that no decision on the fundamental frequency, chord, key or other high-level feature is undertaken. On the contrary, a set of descriptors are computed from this function including amplitude and positions of its main peaks, interval between peaks, sum of the detection function and possibly any

kind of statistical descriptor of the distribution of the pitch content function. Two versions of the pitch function are typically used: an *unfolded* version that contains information about the pitch range of the piece and a *folded* one, in which all pitches are mapped to a single octave giving a good description of the harmonic content of the piece.

C. *Rhythm*

A precise definition of rhythm does not exist. Most authors refer to the idea of *temporal regularity*. As a matter of fact, the perceived regularity is distinctive of rhythm and distinguishes it from *non-rhythm*. More generically, the word rhythm may be used to refer to all of the temporal aspects of a musical work.

Intuitively, it is clear that rhythmic content may be a dimension of music to consider when discriminating between straight-ahead *Rock* music from rhythmically more complex *Latin* music, or when isolating some *Classical* music in which the sensation of pulse is not so evident and the expressive rhythm variation more common.

A review of automatic rhythm description systems may be found in [12]. These automatic systems may be oriented towards different applications: tempo induction, beat tracking, meter induction, quantization of performed rhythm, or characterization of intentional timing deviations. Yet, since state-of-the-art rhythm description systems have still a number of weaknesses, a lower level approach is used in genre recognition system (for example, tempo and beat tracking algorithms typically make errors of metrical levels so that they give unreliable information for machine learning algorithms),

Following the same approach as the one introduced for low-level pitch attributes, descriptors may be extracted from a function measuring the importance of periodicities in the range of perceivable tempi (typically 40 to 200 bpm in genre classification applications). Such function may be obtained by autocorrelation-like transform of features over time (interesting features being usually energies in different frequency bands); it is also possible to use FFT to evaluate modulations of features (typically over windows of 6 seconds) or to build a histogram of inter-onset intervals. Gouyon et al. [13] give an in-depth study on low-level rhythmic descriptors extracted from different periodicity representations. In particular, they obtain encouraging results with a set of MFCCs-like descriptors extracted from a periodicity function (rather than from a spectrum).

D. *Extracting features from semantically significant audio segments*

The descriptors presented earlier may be extracted for the complete audio signals. Yet, in many classification tasks, a small segment of audio is used as it may contain sufficient information to characterize the content of a complete

song because in many musical genres repetitions are observed inherently to the musical structure. This idea is even more relevant since the required computation may be greatly reduced considering only a small part of the signal.

Most of the proposed algorithms for musical genre classification indeed use one small segment of audio per title: typically a 30-second long segment starting 30 seconds after the beginning of the piece to avoid introductions that may not be representative of the whole piece.

In the context of artist identification, Berenzweig et al. [14] have proposed to detect automatically singing segments and have obtained improved results by analyzing only the singing part: it may indeed be easier to identify artists by listening to their voices rather than their music.

E. Local conclusion

Table I summarizes the types of features currently used in music information retrieval applications.

TABLE I
TYPICAL FEATURES USED TO CHARACTERIZE MUSIC CONTENT

Timbre	Melody/Harmony	Rhythm
Texture Model: model of features over texture window:	Pitch Function: measure of the energy in function of music notes	Periodicity Function: measure of the periodicities of features
1. Simple modeling with low-order statistics	1. Unfolded function: describes pitch content and pitch range	1. Tempo: periodicities typically in the range 0.3 to 1,5s (i.e. 200 to 40 bpm)
2. Modeling with autoregressive model	2. Folded function: describes harmonic content	2. Musical pattern: periodicities between 2 and 6 seconds (corresponding to the length of one or more measure bar)
3. Modeling with distribution estimation algorithms (for example, EM estimation of a GMM of frames)		

Extraction of high-level descriptors from unrestricted polyphonic audio signals is not yet state of the art. Thus most approaches focus on timbre modeling based on combinations of low-level descriptors. Timbre may contain sufficient information to roughly characterize musical genres as research demonstrated that humans with little to moderate musical training were able to perform a correct classification of music (among 10 genres) in 53 % of the cases after listening to only 250 milliseconds and in 72% of cases based on only 3 seconds of audio [15]. This suggests that no high-level understanding of music is needed to characterize genres as 250 milliseconds and in a lesser manner 3 seconds are too little time to recognize a musical structure.

Aucouturier and Pachet [16] have a more pessimistic point of view. They have studied the correlation between timbre similarity and genre. They used a state-of-the-art timbre similarity measure [17] and a database of 20000 titles distributed over 18 genres. Their results show that there is only little correlation between timbre and genres suggesting that classification schemes based solely on timbre are intrinsically limited. They also suggest that such classification schemes may hardly scale in both the number of titles and in the number of genre classes.

Arguing that there may not be enough information in audio signals to characterize the musical genre of a title, they proposed to take some cultural features into account [16] by mining the web to extract relevant keywords associated to music titles. Indeed, when one tries to derive genre from audio only, the basic assumption is that genre is an intrinsic attribute of a title as its tempo for example, which is definitely questionable (see section II).

IV. EXPERT SYSTEMS

Expert systems explicitly implement sets of rules. For the genre classification task, this would be equivalent to enumerate a number of rules that would precisely and uniquely characterize a genre. As far as we know, no model based on expert systems has been proposed to characterize musical genres. The work by Pachet and Cazaly [2] for a taxonomy of musical genres can be compared to an expert system approach though it did not lead to an actual implementation; yet it is worth mentioning it as it allows a deeper comprehension of the difficulties of music genres classification.

Pachet and Cazaly have tried to define characteristics of genres and their relations. They have formally stated differences among genres with a language based on descriptors such as the instrumentation, the type of voice, the type of rhythm, the tempo of the song (for example, *Ska* is derived from *Mento* and it is different because it has a faster tempo and a brass section). This implies that these descriptors must be detailed enough to characterize differences among subgenres.

This approach, if possible at all, is not appropriate for genre classification given the complexity of the task and the difficulty to objectively describe very specific subgenres. Moreover it requires a (automatic) manner to obtain reliable high-level descriptors from the audio signal, which is not state-of-the-art as seen in the previous section.

Expert systems, though they incorporate deep knowledge of their subject, are expensive to implement and to maintain. As the number of manually generated rules grows, they may yield unexpected interactions and side effects, so that software engineering issues become increasingly important. In the last few years, the machine learning

approach has garnered increasing interest. From the point of view of related disciplines, the machine learning approach has come to dominate similar areas of natural language processing and pattern recognition such as automatic speech recognition or face recognition.

V. THE UNSUPERVISED APPROACH

While some approaches tend to classify music given an arbitrary taxonomy of genres, another point of view is to cluster data in a non-supervised way so that a classification will emerge from the data themselves based on objective similarity measures. The advantage is to avoid the constraint of a fixed taxonomy, which may suffer from ambiguities and inconsistencies as it has been seen earlier. Moreover some titles may simply not fit into a given taxonomy.

In the unsupervised approach, an audio title is represented by a set of features as seen in section III and a similarity measure is used to compare titles among each others. Unsupervised clustering algorithms take then advantage of the similarity measure to organize the music collection with clusters of similar titles.

A. Similarity measures

The simplest choice to measure distance between two feature vectors is, for example, to use a Euclidean distance or a cosine distance. However these distances will only make sense if the feature vectors are time-invariant. Otherwise two perceptually similar titles may be distant according to the measure if the similar features are time-shifted. To build a time-invariant representation of a time series of feature vectors, one usually builds a statistical model of the distribution of the features and then uses the distance to compare these models directly.

Typical models include Gaussian and Gaussian mixtures (GMMs) (GMMs have been used to build song timbre models in [17], [18] and [19]). The Kullback-Leibler divergence or relative entropy is the natural way to evaluate distance between probability distributions but it is not suited for GMMs. Alternative measures include sampling, Earth's Mover distance and the Asymptotic Likelihood Approximation (see [17]).

Considering the fact that, unlike most classic pattern recognition problems, the data to be classified are time series data, Shao et al. [20] use Hidden Markov Models (HMMs) to model the relationship between features over time. One interest of HMMs is that they provide a proper distance metric so that once each piece is characterized by its own HMM, the distance between any pieces of the database can be computed.

B. Clustering algorithms

K-means is probably the simplest and most popular clustering algorithm. It allows partitioning a set of vectors into K disjoint subsets. One of its weaknesses is that it requires the number of clusters (K) to be known in advance.

Shao et al. [20] cluster their music collection with the Agglomerative Hierarchical Clustering, a clustering algorithm that starts with N singleton clusters (where N is the number of titles of the database) and that forms a sequence of clusters by successive merging.

The Self-Organizing Map (SOM) and the Growing Hierarchical Self-Organizing Map (GHSOM) are used to cluster data and organize them on a 2-dimensional space in such a way that similar feature vectors are grouped close together. SOMs are unsupervised artificial neural networks that map high dimensional input data onto lower-dimensional output spaces while preserving the topological relationships between the input data items as faithfully as possible. GHSOMs are a special case of SOMs which make use of a hierarchical structure with multiple layers where each layer consists of a number of independent SOMs. Rauber et al. [21] use an output space of dimension 2 to allow a visual representation of a music collection with a GHSOM.

In some terms, the major drawback of unsupervised techniques may be that the obtained clusters are not labeled. In any case, these clusters do not always reflect genre hierarchies, rather similarities dependent on the type of features (rhythmical similarities, melodic similarities, etc.). Rousseaux and Bonardi [22] argue that e.g. in the context of EMD the notion of genre may disappear in favour of the development of an *ad-hoc* organization of audio samples centred on prototypes and similarity.

VI. THE SUPERVISED APPROACH

The supervised approach to music genre classification has been studied more extensively. The methods of this group suppose that a taxonomy of genres is given and they try to map a database of songs into it by machine learning algorithms. As a first step, the system is trained with some manually labeled data, and then it is used to classify unlabelled data. The major interest of supervised classification compared to the expert system approach is that one does not need to explicitly describe musical genres: the classifier attempts to form automatically relationships between the features of the training set and the related categories.

We describe here a number of commonly used supervised machine learning algorithms. We do not pretend to make an exhaustive list of such algorithms but focus on those that have been used in the context of music genre

classification. Then we present the results obtained with these algorithms in literature.

A. Supervised classifiers

- **K-Nearest Neighbor (KNN):** it is a non-parametric classifier based on the idea that a small number of neighbours influence the decision on a point. More precisely, for a given feature vector in the target set, the K closest vectors in the training set are selected (according to some distance measures) and the target feature vector is assigned the label of the most represented class in the K neighbour (there is actually no other training than storing the features of the training set). KNNs are evaluated in the context of genre classification in [5] and [18].
- **Gaussian Mixture Models (GMM):** GMMs model the distribution of feature vectors. For each class, we assume the existence of a probability density function expressible as a mixture of a number of multi-dimensional Gaussian distributions. The iterative Expectation Maximization (EM) algorithm is usually used to estimate the parameters for each Gaussian component and the mixture weights. GMMs have been widely used in the music information retrieval community, notably to build timbre models as seen in Section (V. A.). They can be used as classifiers by using a maximum likelihood criterion to find the model best suited to a particular song. They have been used to directly model musical genres in [5]. In [23], a tree-like structure of GMMs is used to model the underlying genre taxonomy: a divide-and-conquer strategy is used to first classify items on a coarse level and then on successively finer levels. The classification decision is thus decomposed into a number of local routing or refinement decisions in the taxonomy. In addition, feature selection at every refinement level allows optimizing classification results. West and Cox [7] use a Maximal Classification Binary Tree built by forming a root node containing all the training data and then splitting that data into two child nodes by using single Gaussian classifier with Mahalanobis distance measurements. In order to split a node, all possible combinations of classes are formed and the combination of classes yielding the best split is chosen (notice that the creation of the tree is unsupervised whereas the classifiers used for splitting on each node are trained in a supervised manner).
- **Hidden Markov Model (HMM):** Hidden Markov Models (HMMs, already introduced in section V. A. as a similarity measure) can also be used for classification purposes. They have been extensively used in speech recognition because of their capacity to handle time series data. HMMs may be seen as a double embedded

stochastic process: one process is not directly observable (hidden) and can only be observed through another stochastic process (observable) that produces the time set of observations. Though they may be well suited to modeling music, to our knowledge, HMMs have only been used in [8] and [24] for genre classification of audio content (they have been used in [20] as well but in the case of unsupervised organization of a music collection).

- **Linear Discriminant Analysis (LDA):** The basic idea of LDA is to find a linear transformation that best discriminates among classes and to perform classification in the transformed space based on some metric such as Euclidean distance. LDA with Adaboost has been used in [25]. Adaboost, for Adaptive Boosting, is used in conjunction with other learning algorithms (LDA in this case) to improve their classification and generalization performances. Adaboost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. In [7], a Fishers Criterion Multi-class LDA is used to reduce dimensionality of the classification problem before modeling with a Gaussian distribution.
- **Support Vector Machines (SVM):** SVMs are based on two properties: margin maximization (which allows for a good generalization of the classifier) and nonlinear transformation of the feature space with kernels (as a data set is more easily separable in a high dimensional feature space). SVMs have been used in the context of genre classification in [8] and [27]. In [19], SVMs are used for genre classification with a Kullback-Lleiber divergence-based kernel to measure the distance between songs. In [28], genre classification is done with a mixture of SVM experts. A Mixture of Experts solves a classification problem by using a number of classifiers to decompose it into a series of sub-problems. Not only does it reduce the complexity of each single task but it also improves the global accuracy by combining the results of the different classifiers (experts). Of course, the number of needed classifiers is increased, yet, by having each of them handle a simpler problem, the overall required computational power is reduced.
- **Artificial Neural Network (ANN):** An ANN is composed of a large number of highly interconnected processing elements (neurons) jointly working to solve specific problems. The most widely used supervised ANN for pattern recognition is the Multi-Layer Perceptron (MLP). It is a very general model that can in principle approximate any non-linear function. MLPs have been used in [14] in the context of artist identification. Neural Networks as well as the other reviewed architectures (except HMMs), can only handle

static patterns. This weakness is partly overcome in [14] by inputting a number of adjacent feature vectors into the network so that contextual information is taken into account: this strategy corresponds to the so called Feedforward Time-Delay Neural Network (TDNN). Other paradigms oriented towards the processing of temporal sequences have been proposed (Recurrent Networks such as the Elman-Network) but have not been used yet in the context of music genre classification. Soltau et al. [24] have introduced in the context of recognition of music genres an original method for explicit time modeling of temporal structure of music (ETM-NN): a MLP is trained to recognize music genres but, rather than considering its output, the activation of its hidden neurons is considered as a compact representation of the input feature vector (it is known indeed that the first half of a feed-forward network performs a specific nonlinear transformation of the input data into a space in which the discrimination should be simpler). Each hidden neuron can be seen as an abstract musical event – not necessarily related to an actual musical representation. The sequence of abstract events over time is then analysed to build one single feature vector which is fed to a second network that implements the final decision about the genre of the musical piece. The ETM-NN architecture is evaluated versus other classifiers in [8].

B. Classification results

The taxonomy and data collections used in state-of-the-art works on genre classification are often very simple and incomplete (typically between 2 and 10 genres and rarely more than 2000 songs) and are usually more reflective of the data available to the authors than of a rigorous analysis of genres; it is consequently rather difficult to compare the different approaches. The Music Information Retrieval Evaluation eXchange (MIREX)⁴ is trying to unify efforts and give a rigorous comparison of algorithms by organizing an evaluation contest of state-of-the-art algorithms dedicated to various MIR applications including genre classification.

For the last edition of the MIREX genre classification contest⁵, two databases (from two different sources) were set up to produce a reasonably challenging problem according to the available data. The first database is composed of 1515 songs over 10 genres (*Classical, Ambient, Electronic, New-Age, Rock, Punk, Jazz, Blues, Folk, Ethnic*; 1005 training files, 510 testing files) and the second of 1414 songs over 6 genres (*Rock, Hip-Hop, Country, Electronic, New-Age, Reggae*; 940 training files, 474 testing files).

⁴ http://www.music-ir.org/mirexwiki/index.php/MIREX_2005

Table II gives a summary of the classification accuracies obtained on the 2 databases by the 12 algorithms that were submitted by 10 different authors. Results both in terms of normalized and non-normalized accuracies are shown. Normalized accuracy corresponds to the case when results are normalized according to the number of songs per class (since classes have not the same number of songs). Differences between normalized and non-normalized results are due to the fact that in the latter case, results are influenced by the prior probabilities of having a class. For more details on algorithms, evaluation method and results please refer to the MIREX 2005 audio genre classification contest website⁵.

TABLE II
CLASSIFICATION ACCURACIES OF THE ALGORITHMS SUBMITTED AT MIREX 2005

	Dataset 1 Normalized	Dataset 2 Normalized	Dataset 1	Dataset 2
Max Accuracy	73.04%	82.91%	77.75%	86.92%
Min Accuracy	53.47%	49.89%	55.29%	47.68%
Mean Accuracy	67.28%	72.61%	68.38%	75.88%

Though these experiments were performed on a rather limited scale, the obtained results appear to be in accordance with Aucouturier and Pachet (see part III E and [16]): whereas both datasets have a comparable total number of files, the first one has more classes than the second and the results obtained on the first dataset are significantly lower. It seems confirmed that such classification schemes may hardly scale in the number of genre classes.

Table III shows the confusion matrix obtained on the first dataset with the algorithm submitted to MIREX 2005 by the authors [28]. Looking at this matrix, it is noticeable that classification errors make sense. For example, 29.41% of the *Ambient* songs were misclassified as *New-Age* and these two classes seem to clearly overlap when listening to the audio files. In the same way, 14.71% of the *Blues* examples were considered as *Rock* by the algorithm. From these results, it seems reasonable that relaxing the strict classification paradigm and allowing a file to be labeled with multiple classes can be a way to implement a realistic classification system.

TABLE III
CONFUSION MATRIX FOR THE DATASET I AND FOR THE ALGORITHM SUBMITTED BY THE AUTHORS TO MIREX 2005

Truth Prediction	Ambient	Blues	Classic	Electronic	Ethnic	Folk	Jazz	New-Age	Punk	Rock
Ambient	52.94%	0.00%	0.00%	7.32%	4.82%	0.00%	0.00%	26.47%	0.00%	5.95%
Blues	0.00%	76.47%	0.00%	0.00%	0.00%	4.17%	0.00%	0.00%	0.00%	3.57%

⁵ <http://www.music-ir.org/evaluation/mirex-results/audio-genre/index.html>

Classic	2.94%	0.00%	100.00%	0.00%	8.43%	0.00%	0.00%	0.00%	0.00%	0.00%
Electronic	5.88%	0.00%	0.00%	53.66%	6.02%	4.17%	4.55%	5.88%	0.00%	19.05%
Ethnic	2.94%	0.00%	0.00%	7.32%	59.04%	12.50%	4.55%	20.59%	0.00%	0.00%
Folk	0.00%	5.88%	0.00%	1.22%	3.61%	62.50%	0.00%	2.94%	0.00%	2.38%
Jazz	0.00%	2.94%	0.00%	3.66%	6.02%	4.17%	81.82%	8.82%	0.00%	5.95%
New-Age	29.41%	0.00%	0.00%	4.88%	4.82%	8.33%	4.55%	32.35%	0.00%	5.95%
Punk	0.00%	0.00%	0.00%	0.00%	0.00%	4.17%	0.00%	0.00%	100.00%	4.76%
Rock	5.88%	14.71%	0.00%	21.95%	7.23%	0.00%	4.55%	2.94%	0.00%	52.38%

VII. FUTURE DIRECTIONS

Table IV summarizes the advantages and drawbacks of the three main paradigms reviewed for music collection organization.

TABLE IV
PARADIGMS AND CLASSIFICATION METHODS

Expert Systems	Unsupervised Clustering	Supervised Classification
1. Uses a taxonomy	1. No taxonomy: classification emerges from the data	1. Uses a taxonomy
2. Each class is defined by a set of explicit high-level characteristics	2. Organization according to similarity between excerpts	2. The learning algorithm maps features to classes without describing rules explicitly
Impracticable since:	3. Typical clustering algorithms: K-Means, Agglomerative Hierarchical Clustering, Self-Organizing Map and Growing Hierarchical SOM	3. Typical supervised learning algorithms: KNN, Neural Networks, LDA, SVMs...
1. Extraction of high level descriptors is not state-of-the-art		
2. Difficult to objectively describe music genres		

New problems, relying on similar techniques, are emerging in the field of music information retrieval as new markets and applications eventually take off. Many of the previously introduced algorithms can be applied with minor changes to these new applications while results coming from innovative research fields can provide useful feedbacks on genre classification techniques.

A. Classification into perceptual categories

While most work on music classification has focused on musical genres, some authors have proposed other labeling focused on *perceptual* categories of music. The corresponding categories are usually referred to as *moods* (*contentment, depression, exuberance, anxious...*) or *emotions* (*cheerful, delicate, dark, dramatic...*) but may be associated to any kind of adjective (*funky, quiet, loud, lonesome...*). Other interesting dimensions of music can be

considered such as perceived complexity which may be loosely defined as the effort a listener has to put into analyzing the music in order to capture its main characteristics and components.

An overview of classification into perceptual categories can be found in [29]. They conclude that the classification results are hardly over the baseline, which seems to confirm the negative results of [16] suggesting the need for extra-musical information.

B. Novelty detection

Novelty detection is the identification of new or unknown data or signal that a machine learning system is not aware of during training. It is an essential part of any realistic music classification tool since some songs may not correspond to any of the classes supported by the system – in this case it may make more sense to identify the type of the song as unknown rather than giving it an improper label. As far as it is known, there has been only one attempt to apply novelty detection to music signals (see [26]).

C. Classification with multiple labels

The classification paradigms shortly introduced in the previous sections are usually thought for strict classification: one excerpt must belong to one genre. Yet it may be hard to fit unambiguously one song into one box. Taking into account ambiguity or in other words, allowing multi-genres classification is probably closer to the human experience in general, for sure to the artist's point of view. Artists usually produce music without concerning themselves in which genre they are working. Furthermore in most Internet based classifications, artists, albums or titles are typically associated to a number of genres. Even in more conventional record shops, one may find some discs in different areas.

As far as it is known, no algorithm has been proposed yet to associate multiple genre labels to one song. The lack of work in this area is easily understandable as state-of-the-art algorithms still have difficulties to associate unambiguous general labels to songs while multiple labels may be appropriate to precise sub-genres. In any case, it is clearly a direction to follow to build a realistic classification system.

D. From taxonomies to folksonomies

The target audience is a crucial point to take into account in the design of an automatic classification strategy. Traditional taxonomies work by establishment of a clear view and organization of the corpus on which users have to

agree in order to properly use the classification scheme. As it has been shown in section II *B*, Internet based music genre taxonomies are often very complex and the corresponding genre labels may only make sense for expert users.

In more recent years, Web publishing has approached to the mass market thanks to continuously falling technology cost and barriers (notably with weblogs and WIKIs). From this situation new and different classification strategies have emerged, such as the so-called *folksonomies* that can be loosely defined as user-generated classification schemes specified through bottom-up consensus.

In the case of music classification, letting consumers define their own personal taxonomies would allow for a better confidence and experience since the organizer of the information becomes its primary user. However, such a scenario raises a number of issues in the design of a classification tool. Users should notably have the possibility to train the classification tool incrementally to show new examples to the system in order to refine its judgement. Moreover, one should have the possibility to expand its own taxonomy of genres both in width (new root genre) and depth (new subgenre). Hierarchical systems (like in [23]) should be favored since in that case adding genres is equivalent to adding a new classification tree or new leaves to an existing tree. Another advantage of hierarchical systems is that different features may be used at each level so that features optimized for the discrimination of some specific genres can be used.

VIII. CONCLUSION

In this article, we highlight how convoluted the definitions of musical genres are in spite of their relevance in our historical and cultural background. We reviewed typical feature extraction techniques used in music information retrieval for the different music elements (see Table I); given these features the three main paradigms for audio genre classification were presented with their advantages and drawbacks (see Table IV). State-of-the-art results obtained during the music genre classification contest of MIREX 2005 are presented and discussed (see Table II and III). Finally, we introduced new emerging research fields and techniques that investigate the proximity of musical genres, such as folksonomies and perceptual categories.

Overall, we find that research is evolving from purely objective machine calculations to techniques where learning phases, training datasets, preliminary knowledge, etc. strongly influence performance and results. This is particularly comprehensible for musical genre classification, which has always been influenced by experience, background and sometimes personal feeling. But even in several other classification domains, music related or not, many outstanding

solutions exist where machine learning plays a fundamental role, complementary to signal processing.

REFERENCES

- [1] R. Dannenberg, J. Foote, G. Tzanetakis, C. Weare, “*Panel: new directions in music information retrieval*”, in *Proc. Int. Computer Music Conf.*, Habana, Cuba, September 2001.
- [2] F. Pachet, D. Cazaly, “*A taxonomy of musical genres*”, in *Proc. Content-Based Multimedia Information Access (RIAO)*, Paris, France, 2000.
- [3] F. Pachet, J.J. Aucouturier, A. La Burthe, A. Zils, A. Beurive, “*The Cuidado Music Browser: an end-to-end Electronic Music Distribution System*”, in *Multimedia Tools and Applications*, 2004. Special Issue on the *CBMI03* Conference, Rennes, France, 2003.
- [4] G. Peeters, “*A large set of audio features for sound description (similarity and classification) in the CUIDADO project*”, CUIDADO I.S.T. Project Report, 2004.
- [5] G. Tzanetakis, P. Cook, “*Musical genre classification of audio signals*”, in *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, July 2002.
- [6] A. Meng, P. Ahrendt, J. Larsen, “*Improving music genre classification by short-time feature integration*”, in *Proceedings of the 6th Int. Symposium on Music Information Retrieval*, London, UK, 2005.
- [7] K. West, S. Cox, “*Finding an optimal segmentation for audio genre classification*”, in *Proceedings of the 6th Int. Symposium on Music Information Retrieval*, London, UK, 2005.
- [8] N. Scaringella, G. Zoia, “*On the modeling of time information for automatic genre recognition systems in audio signals*”, in *Proceedings of the 6th Int. Symposium on Music Information Retrieval*, London, UK, 2005.
- [9] E. Gomez, A. Klapuri, B. Meudic, “*Melody description and extraction in the context of music content processing*”, in *Journal of New Music Research*, Vol. 32(1), 2003.
- [10] A. Klapuri, “*Multiple fundamental frequency estimation based on harmonicity and spectral smoothness*”, in *IEEE Trans. Speech and Audio Proc.*, 11(6), 804-816, 2003.
- [11] G. Zoia, R. Zhou, D. Mlynek, “*A multi-timbre chord/harmony analyzer based on signal processing and neural networks*”, in *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, Siena, Italy, 2004.
- [12] F. Gouyon, S. Dixon, “*A review of automatic rhythm description system*”, in *Computer Music Journal*, vol. 29, pp. 34-54, 2005.
- [13] F. Gouyon, S. Dixon, E. Pampalk, G. Widmer, “*Evaluating rhythmic descriptors for musical genre classification*”, in *AES 25th International Conference*, London, England, 2004.
- [14] A. Berenzweig, D. Ellis, S. Lawrence, “*Using voice segments to improve artist classification of music*”, in *Proc. of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.
- [15] D. Perrott, R. O. Gjerdingen, “*Scanning the dial: an exploration of factors in the identification of musical style*”, Research Notes, Department of music, Northwestern University, Illinois, 1999.
- [16] J.J. Aucouturier, F. Pachet, “*Representing musical genre: a state of the art*”, in *Journal of New Music Research*, vol. 32, no. 1, pp. 83-93, 2003.

- [17] J.J. Aucouturier, F. Pachet, “*Music similarity measures: what’s the use?*”, in *Proc. of the 3rd Int. Symposium on Music Information Retrieval*, 2002.
- [18] E. Pampalk, A. Flexer, G. Widmer, “*Improvements of audio based music similarity and genre classification?*”, in *Proc. of the 6th Int. Symposium on Music Information Retrieval, London, UK*, 2005.
- [19] M. Mandel, D. Ellis, “*Song-level features and support vector machines for music classification*”, in *Proc. of the 6th Int. Symposium on Music Information Retrieval, London, UK*, 2005.
- [20] X. Shao, C. Xu, M. Kankanhalli, “*Unsupervised classification of musical genre using hidden Markov model*”, in *IEEE Int. Conf. of Multimedia Explore (ICME)*, Taipei, Taiwan, China, 2004.
- [21] A. Rauber, E. Pampalk, D. Merkl, “*Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity*”, in *Proc. of the 3rd Int. Conf. on Music Information Retrieval*, Paris, France, 2002.
- [22] F. Rousseaux, A. Bonardi, “*Reconcile art and culture on the web: lessen the importance of instantiation so creation can better be fiction*”, in *Proc. First Int. Workshop on Philosophy and Informatics*, Cologne, Germany, 2004.
- [23] J.J. Burred, A. Lerch, “*A hierarchical approach to automatic musical genre classification*”, in *Proc. Of the 6th Int. Conf. on Digital Audio Effects (DAFx)*, London, UK, 2003.
- [24] H. Soltau, T. Schultz, M. Westphal, A. Waibel, “*Recognition of music types*”, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, USA, 1998.
- [25] N. Casagrande, D. Eck, B. K., “*Geometry in sound: a speech/music audio classifier inspired by an image classifier*”, in *Proc. Of the Int. Computer Music Conference (ICMC)*, 2005.
- [26] A. Flexer, E. Pampalk, G. Widmer, “*Novelty detection based on spectral similarity of songs*”, in *Proc. of the 6th Int. Symposium on Music Information Retrieval, London, UK*, 2005.
- [27] T. Lidy, A. Rauber, “*Evaluation of feature extractors and psycho-acoustic transformations for music genre classification*”, in *Proc. of the 6th Int. Symposium on Music Information Retrieval, London, UK*, 2005.
- [28] N. Scaringella, D. Mlynek, “*A mixture of support vector machines for audio classification*”, Music Information Retrieval Evaluation exchange (MIREX) website, http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/scaringella.pdf, 2005.
- [29] T. Pohle, E. Pampalk, G. Widmer, “*Evaluation of frequently used audio features for classification of music into perceptual categories*”, in *Proc. of the 4th Int. Workshop on Content-Based Multimedia Indexing*, Riga, Latvia, 2005.
- [30] MPEG-7, “*Information Technology – Multimedia Content Description Interface – Part 4: Audio*”, ISO/IEC JTC 1/SC29, ISO/IEC FDIS 15938-4:2002, 2002.
- [31] W. Chai, “*Semantic segmentation and summarization of music*”, in *IEEE Signal Processing Magazine: Special Issue on Semantic Retrieval of Multimedia*, March 2006.

About the Author – NICOLAS SCARINGELLA is Phd student at the Signal Processing Institute of the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, since October 2004. He was awarded Master Degree of Engineering, majoring in Electronics, Telecommunications and Computer Science from the Ecole de Chimie, Physique et Electronique de Lyon (ESCPE Lyon), Lyon, France, in October 2004. His research interests focus on music information retrieval, audio signal processing, machine learning and automatic music transcription.

About the Author – GIORGIO ZOIA is a scientific advisor at the Signal Processing Institute of the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. In April 2001 he received a PhD *es Sciences Techniques* from EPFL with a thesis on fast prototyping of architectures for audio and multimedia. His research interests evolved from digital video, digital design and CAD synthesis optimization in submicron technology to compilers, virtual architectures and fast execution engines for digital audio. Fields of interest in audio include 3D spatialization, audio synthesis and coding, representations and description of sound, interaction and intelligent user interfaces for media control. He has been actively collaborating with MPEG since 1997, with several contributions concerning model-based audio coding, audio composition (Systems) and analysis of computational complexity.

About the Author – DANIEL MLYNEK is a professor at the Signal Processing Institute of the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. He gained a wide experience from the semiconductor industry. His developments dealt with the integrated circuits used for signal processing, particularly in the telephony and the audio-visual fields. He was responsible for the Digital TV project with ITT Semiconductors the aim of which was to digitize the functions of TV receivers. Professor Mlynek has about 60 patents. He was as Technical Director World Wide, responsible for 150 people and 3 design centres in USA, Europe and Japan. At ITT Semiconductors he introduced the 1.5 micron, 1.2 micron and 0.8 micron technologies into production, being also responsible for the development of those technologies. His current main fields of interest are Telecom Systems especially for data acquisition and transport, multimedia systems including MPEG2, MPEG-4 and HDTV, Design and Test of complex ASICs, Intelligent Systems with applications in different areas. His latest publications are the WEB Course on Basics in Electronics and VLSI Design, Fuzzy Logic Systems (Wiley), Intelligent Systems and Interfaces (KLUWER), Fuzzy and Neuro-Fuzzy System in Medicine (CRC).