

Open Position: Multilingual Orchestra Conductor. Lifetime opportunity.

Eloi Batlle
Audiovisual Institute
Universitat Pompeu Fabra
Barcelona, Spain
eloi@iaa.upf.es

Eric Gaus
Audiovisual Institute
Universitat Pompeu Fabra
Barcelona, Spain
egaus@iaa.upf.es

Jaume Masip
Audiovisual Institute
Universitat Pompeu Fabra
Barcelona, Spain
jmasip@iaa.upf.es

ABSTRACT

In this paper we show the need to see the Music Information Retrieval world from different points of view in order to make any progress. To help the interaction of different languages (engineers, musicians, psychologists, etc.), we present a tool that tries to link all those backgrounds.

1. INTRODUCTION

When we started research in Music Information Retrieval some years ago, we start discussing different ideas, projects and applications dealing with MIR. It was really a funny discussion because, in fact, it became a brainstorming meeting. But after one hour and a half, we realized that different members of our group were talking in different languages. While some of us were talking about research results using Hidden Markov Models and Zero Crossings of one specific input waveform, others were talking about the expressiveness and vivacity of that specific audio. Thus, the discussion turned over into a new direction: Should the Music Information Retrieval community be able to build a dictionary to translate the information in both senses? We conclude the discussion with an unanimous conclusion: Yes, it should! In this paper, we present a simple tool that could help in the construction of this dictionary although it is not the goal of this paper to explain it exhaustively. We just present it, as another contribution to the existing tools, in order to better define their features and requirements.

2. HUMAN ANALYSIS

A song or a musical piece can be analyzed, by humans, from many different points of view. Melodic analysis is the most intuitive one. The melody of a song can be easily defined, according to Leonard Bernstein, as the part of the music that can be whistled. In most cases, the melody is played by a singing voice, although it can be mixed with many different instruments. When different voices are singing together, the melody is often associated with the voice with higher

pitch. If the musical piece has no singing voice, the melody is usually played by a specific instrument but, if there is no dominant instrument, the melody can be found wherever the listener wants. Percussive music has no melodies.

We can see that it is not easy to define what the melody is. Different musical genres can interpret the melody concept in many different senses as well as different socio-cultural aspects can strongly affect our own perception of the melody, too. Any attempt to define what the melody is seems to be a very difficult task for humans. Thus, nobody can expect this difficult task be accomplished by computers. The Bernstein definition, the best one from our point of view, becomes completely useless for the actual known programming techniques.

Music can also be analyzed from a rhythmical point of view. According to simple definitions, one could think that rhythm is whatever you can follow just hitting your leg with your hand. Unfortunately, this definition is not valid here, because it covers only a narrow subset of the whole meaning of the word. Rhythm can be analyzed in three different levels [3]: The first level is the macro-level rhythm analysis. This kind of analysis studies the structure of the piece, that is, the chorus and solos in a song or the different acts in an opera. With the mid-level rhythm analysis one can distinguish between different phrases in a song and observe how they can be responded. Finally, at the micro-level rhythm analysis, the note durations and rhythmic bases are studied.

There are lots of Drum Loops collections. Disk Jockeys will play, mix and modify them in order to create different rhythmic patterns in their performances. Also, classical music composers know that symphonies are divided into three sections or movements. Both of them are working with rhythm, but at different cognition levels. When computers manage rhythmic information, they must take into account all these three levels.

Harmony is also quite important in a musical analysis. While melody involves the evolution in time of one specific part (instrument) of the music, harmony involves all the instruments and notes played together at a given time lapse, that is, the chords. The evolution of these chords is also studied, as well as voicing and sub-melodies created by the time evolution of the different notes in the chords. Harmonic analysis work is very important for composers and performers, but not so

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

important for average listeners: nobody remembers a song for a specific $\text{II } m7 - V 7 - I \text{ maj}7$ succession! Then, we can consider that harmonic analysis is not relevant for a lot of applications of Music Information Retrieval.

Timbre analysis has become more and more important in the last twenty years. The evolution of music has been very important in the last centuries, but not until recently timbre has grown as a major study subject. Different genres have been created, but music has been played with almost the same instruments. In the last years, with the fast growth of analog electronic (60's and 70's) and digital (80's and 90's) technologies, a lot of new instruments have been created. These new instruments create new timbres. Furthermore, sometimes the timbre can exactly identify a specific musical piece or composer. Some of these new instruments are physical instruments while the other ones are "virtual". Nowadays, timbre characteristics are really important in the aesthetic aspects of new music [4]. With virtual instruments one can create texture-based music: music without melody and without rhythm, just playing with timbres.

3. AGAIN, DOES IT MAKE SENSE TO USE A COMPUTER?

This discussion is not new and other areas also opened similar questions: Should we use computers to solve subjective problems? Or even more: Can we do it? AI researchers work very hard to make the answer yes, but unfortunately there are some problems that it is not clear at all the path we should follow (if there is any).

In this section we introduce a method that is able to show some interesting results when dealing with music similarity and retrieval. We may say that these results are better when the objectivity of the query is high, but with this statement a new question arises: if the answer is subjective how can we say if it is right or wrong? Therefore the computer should see life as a gradient in gray and not only black and white. This method is described in more detail in [2]. The idea behind it lies on the fact that music (and all audio in general) can be seen as a sequence of acoustic events. Since music has a strong meaning in its temporality, the similarity system will exploit this fact to process the audio.

Let's imagine the following situation: we have a song played by a guitar, then a violin, then a piano and finally again a guitar. We can describe this piece of music (up to some level of abstraction) using its players, that is guitar \rightarrow violin \rightarrow piano \rightarrow guitar. Now we are given another piece of music and we are asked to find its similarity with the former song. The question to answer is: can this second music piece be performed with the sequence of players guitar \rightarrow violin \rightarrow piano \rightarrow guitar? Or, what is the same, as a conductor of an orchestra, can I reproduce the given music if I conduct the players in the order guitar \rightarrow violin \rightarrow piano \rightarrow guitar? Will it sound more or less the same? If the answer is yes we have got it: they are similar. Most MIR approaches work the other way around. The classical approach is to have a list of descriptions of all the music in the database. Then extract a description of the unknown music and find the closest descriptions from the original database. Our approach never extracts a description of the unknown audio. We simply try to play the new piece with the players of the known songs.

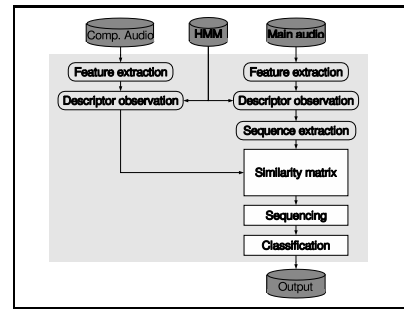


Figure 1: Data flow within the system process.

Of course, in the real world, music cannot be described with simple instruments and songs are usually made of complex sounds. Therefore, the "players" we will use to describe our music are going to be abstract and with no physical meaning. So now the name of the game is to find these abstract players from the music. We can do it using Hidden Markov Models (HMM) [1] and based on their property as a double embedded stochastic process: one that can be seen (the music itself) and one that is hidden (the orchestra). We will use this system based on the source generation of the sound rather than its description with a sequence of feature and parameters like MFCC, spectral flatness, etc. to see and discuss the advantages, disadvantages, uses and limitations of automatic music information retrieval systems. (ok, writing down the sequence of generator is like writing a sequence of features, but we have to look at the inner philosophy of this sentence). Since the generators of the audio are also descriptors, we will use the two words indiscriminately. One of the main features of these HMM is that they can describe the generation

While analyzing fragments of music, we can refer to self-similarity and cross-similarity. We will talk about self-similarity when the analysis is with the audio against the audio itself. This will be very useful when trying to find musical structures, chorus, repetitions, etc. On the other hand, we have cross-similarity when the analysis is performed against other pieces, useful to find clusters with the same musical style, music browsing, etc.

3.1 System structure

The system inputs are the main audio track, the to-be-compared audio track and the audio descriptors. The system output is the similarity structure between both audio inputs. Both audio tracks are first transformed into feature vector parameters and then observed by the audio descriptors. The main track observations are used to generate a fingerprint sequence of audio descriptors through a Viterbi algorithm as already described in the papers cited before and then a similarity matrix is built by matching the compared audio descriptor observations against the main audio track sequence of descriptors. Finally, the correspondence blocks determines which audio segments could have been generated by the "same" sequence of observers and the classifier filters out the sequences and finds similarity structures. Figure 1 shows the data flows within the system process.

3.2 Similarity matrix

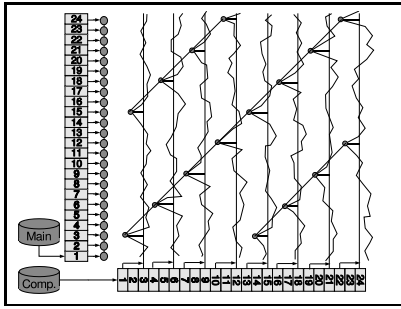


Figure 2: Similarity matrix.

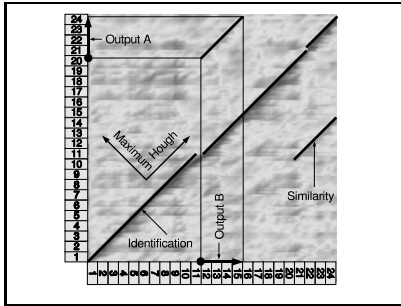


Figure 3: Extraction of similar segments.

The similarity matrix is built as sketched in Figure 2. First column shows the sequence of feature vectors extracted from the main audio track converted into a sequence of audio descriptors in the second column. The compared audio track feature vectors are represented as a row at the figure at the bottom. Each column of the graph shows the distance metric of three feature vectors against all the descriptor observers. In the example of Figure 2, the main and the compared audio tracks are the same which yields to a diagonal of maximum scores between the columns graphs. This metric is used for identification purposes which is out of the scope of this paper. Moreover, this example assumed that the audio track is composed with a theme repeated twice. As shown in the same example, the repetition produces two secondary diagonals, the top-left diagonal indicates that the second section is similar to the first, while the bottom-right indicates that the first section is similar to the second. We can conclude therefore that similarity correspondences between two audio segments can be inferred from the matrix as continuous high score diagonal lines.

3.3 Correspondences

The correspondence extraction block is in charge of extracting similar segment pairs from the similarity matrix. The algorithm combines a Hough transform [7] in the (+1,+1) direction vector with maximum detection in the (-1,+1) direction vector. Diagonal lines of Hough maximums are first characterized with their starting point and line length. Then, simple heuristics are applied to interpolate lines with discontinuities smaller than a certain maximum threshold. The output of the correspondence block is a sequence of similar sequence pairs that shows the starting point in the main audio, the starting point in the compared audio and the length of both segments. Figure 3 shows an example of extracting a similar segment pair from the similarity matrix.

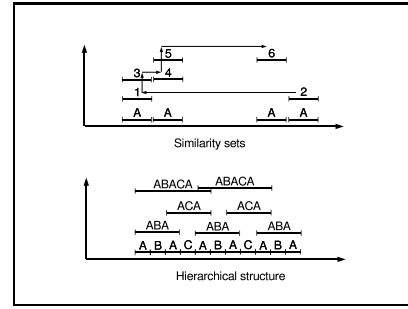


Figure 4: Classification of similar segments.

3.4 Classification

The last block is shown in Figure 4 and classifies similar segment pairs and builds the overall similarity structure between both audio tracks. All similar segment pairs are matched between themselves by applying accuracy thresholds in the segment margins. This matching algorithm generates sets with similar themes at different levels of granularity. Finally, the similarity sets are matched among themselves to generate hierarchical structures with increasing granularity.

4. OPEN DISCUSSION

It is really pleasant to realize that the Music Information Retrieval community has grown due to people coming from many different cognition areas. For instance, musicians and engineers can talk about music, about their sensations when listening Mozart or U2, about their intentions when creating different textures and so on. But we should not be misled by this fact. Although all of them talk about music, and their contribution to it always comes in handy, they often speak in different languages. We know it is a well known problem, and it is not our intention to discuss that [5].

But from the computer science point of view, we think that we should focus our efforts through two different paths. The first one is related to the so called *objective parameters of music*, that is structure, rhythm or timbre description. It's fairly easy to extract many parameters from music. Some techniques need a lot of improvements, but in the next few years, we expect those techniques to be almost perfect. Therefore, let us imagine that we have algorithms that are able to extract as many features of any audio signal as we want (or we need). What do we do with all this data? We can use it to classify genres, to find similarities between clarinet solos and so on. Do the computer science community take some important decisions to distinguish between genres? They sometimes do, but they often lack the background. Musicologists should be included in the research groups.

On the other hand, we should focus our studies in the perceptual aspects of music [6]. Why is the tuba sound generally associated to weight feelings? The main problem to study this is we don't know exactly how the human brain works. If we don't know this, how do we expect a computer performing this task? It is impossible. Psychologist should be included in the research groups.

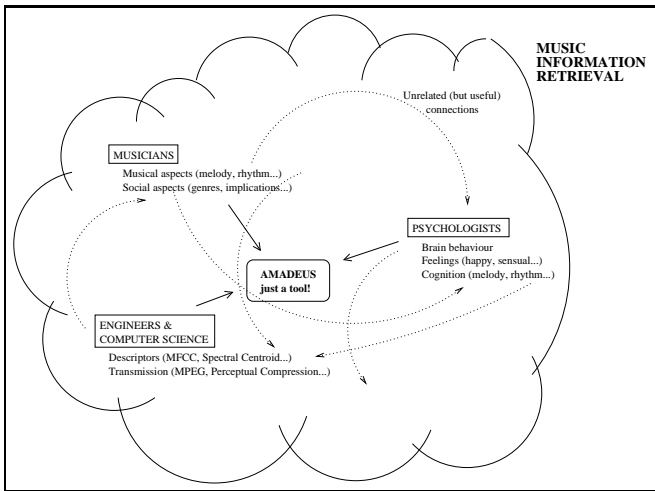


Figure 5: General Overview of Music Information Retrieval

The system described in Figure 4 should be a small contribution to that meeting point that sometimes seems to be inexistent. We can extract different timbre, rhythmical and (still not available) melodic structures. Note that we are managing musical concepts, not low level engineering concepts such as MFCC, Spectral Centroid, etc. These last concepts are used but, never shown. Musical and Psychological knowledge is the main architect of this transformation.

As mentioned above, we have to think in a gray scale. Psychologist and Musicians have to transform this picture into a color landscape.

5. REFERENCES

- [1] Batlle, E., Cano, P. "Automatic Segmentation for Music Classification using Competitive Hidden Markov Models", *Proceedings of the International Symposium on Music Information Retrieval*, Plymouth, USA, 2000.
- [2] Batlle, E., Masip, J., Gaus, E. "Automatic Song Identification in Noisy Broadcast Audio", *Proceedings of the International Conference on Signal and Image Processing*, Kauai, USA 2002.
- [3] Stephen Handel. *Listening: An Introduction to the Perception of Auditory Events*. The MIT Press. 1991. 2nd. Edition.
- [4] Javier Blázquez y Omar Morera. Prólogo de Simon Reynolds. *Loops una historia de la música electrónica*. Ed. Mondadori. Barcelona, 2002.
- [5] Overview on the Sound Modeling Panel. *Workshop on Current Research Directions in Computer Music*. Barcelona, Nov 15-16-17, 2001.
- [6] Vinet, H. Herrera, P. Pachet, F.. *The CUIDADO Project*. Proceedings of ISMIR 2002 - 3rd International Conference on Music Information Retrieval, Paris, France, 2002.

[7] Illingworth, J., Kittler, J., "A survey of the Hough transform," , *Computer Vision, Graphics, and Image Processing*, vol. 44, pp. 87-116, 1988.