Music Similarity Measures: What's the Use?

Jean-Julien Aucouturier
SONY Computer Science Lab.
6, rue Amyot
75005 Paris, France
+33 1 44 08 05 13
jj@csl.sony.fr

Francois Pachet
SONY Computer Science Lab.
6, rue Amyot
75005 Paris, France
+33 1 44 08 05 16
pachet@csl.sony.fr

ABSTRACT

Electronic Music Distribution (EMD) is in demand of robust, automatically extracted music descriptors. We introduce a timbral similarity measures for comparing music titles. This measure is based on a Gaussian model of cepstrum coefficients. We describe the timbre extractor and the corresponding timbral similarity relation. We describe experiments in assessing the quality of the similarity relation, and show that the measure is able to yield interesting similarity relations, in particular when used in conjunction with other similarity relations. We illustrate the use of the descriptor in several EMD applications developed in the context of the Cuidado European project.

1. INTRODUCTION

The domain of Electronic Music Distribution has gained worldwide attention recently with progress in middleware, network and compression. However, the success of EMD depends largely on the existence of robust, perceptually relevant music similarity relations. It is only with efficient content management techniques that the millions of music titles produced by our society can be made available to its millions of users.

The goal of our work is to design and implement music similarity measures that allow users to find quickly interesting music within large catalogues of Popular music. Typical catalogues size is about 20,000 titles. The term Popular music denotes music that is largely distributed, and includes Classical, Pop, Rock World and all sub varieties of musical genres traditionally produced and listened to in the occidental world. The work described here and the experiments were conducted in the framework of the European Project Cuidado (***Content-based Unified Interfaces and Descriptors for Audio*** *and Music Databases available* ***Online*** [1]).

Many dimensions of music have been shown to be perceptually important for characterizing and for making music judgments: tempo, rhythm, voice qualities, etc. Some descriptors have already been proposed for some of these dimensions [2,3]. This work focuses on one particularly important dimension in popular music: timbre.

The goal of this work is two-fold. First we look for a timbre *extractor,* which is an algorithm that produces a perceptually grounded representation of the global timbre quality of a song.

Secondly, and most importantly, we look for meaningful *exploitation schemes* of this descriptor. This second aspect is crucial in our domain: descriptors as such are useless if they cannot yield, in the end, "interesting" relationships between music titles or excerpts thereof. The second part of this paper is therefore devoted to the exploitation of the timbral similarity measure yielded by our descriptor.

2. A MEASURE OF TIMBRE SIMILARITY

We describe here a measure of the similarity of the "global timbre" of music titles, based on the audio signal. Like all similarity relations, this type of similarity is difficult to describe precisely with words, but its aim is straightforward: the timbre descriptor we look for aims at describing a timbral quality that applies to the whole song (as opposed to a particular point in time or instrument). Typical examples are:

- A *Schumann* sonata ("Classical") and a *Bill Evans* piece ("Jazz") are similar because of their common romantic piano sound,
- A *Nick Drake* tune ("Folk"), an acoustic tune by the *Smashing Pumpkins* ("Rock"), a Bossa Nova piece by *Joao Gilberto* ("World") are similar in that they all consist of a simple acoustic guitar and a gentle male voice, etc.

We describe here the similarities using sophisticated textual descriptors (male voice, romantic piano, etc.). Of course these descriptors are not to be deduced by the system – this task is probably impossible to achieve – but are used only as comments to a global similarity that is intrinsically unlabelled.

2.1 State of Art

There has been a large quantity of work about timbre. However most of it has focussed on monophonic simple sound samples, aiming at *Instrument Recognition* ([4]), i.e. identifying if a note is being played on a trumpet or a clarinet. Here, on the contrary, we are concerned with full polyphonic music and complex instrumental textures, for which we want to extract a global timbre description.

Among related work in this domain, *Automatic Genre Classification* ([5]) tries to categorize music titles into genre classes by looking at spectral or temporal signal features. In this approach, the tested song's timbre is matched against pre-computed models of each possible genre. Each genre model averages the timbre of a large number of songs that are known to belong to this genre. There is no matching from one song to another, but rather from one song to a group of songs.

Music title identification ([6]) deals with identifying the title and artist of an arbitrary music signal. This is done by comparing the unlabelled signal's features to a database containing the features of all possible identified songs. In this case, the matching is done from one song to another, but the system only looks for exact matches, not for similarity.

Our approach borrows from both techniques, since it performs approximate matching of one song to another. Our system uses Mel Frequency Cepstrum Coefficients, which are modelled with

Gaussian Mixture models, and compared to yield a similarity measure.

Foote in [7] presents a system that also uses cepstral coefficients as a front-end, but rather uses a supervised algorithm (tree-based vector quantizer) that learns the most distinctive dimensions in a given corpus. Adding one song to this corpus requires to redo the learning of the tree, which is expensive. On the contrary, our system is completely scalable, since it models each song separately.

Welsh in [8] proposes a "query by similarity" system that is also able to match songs according to their "timbre". He uses a large set of features (1248 floating-point per song) which are compared with the euclidian distance. However, his system doesn't address "timbre similarity" explicitly: his features model the pitch/tonal content of a song ("returning songs in the same key"), the noise level ("whether it is pure classical music or noisy, saturated hard rock") and the rhythm. The timbral similarity observed in some results by the author ("a pop, male vocal song produces results where every song in the top 10 is a male vocal with guitar and drum accompaniment") appears therefore as a "side-effect" of the features above, notably those describing the tonal content of the pieces. Our system is both more restrictive, and more precise: notably, the features that we use are meant to be independant of the pitch. We do not try to model "music similarity" at large, but only "timbral similarity". It is only one similarity relationship among many others (rhythm, melody, style, structure, etc.), some of them addressed by Welsh. We will argue that the interestingness of a music retrieval system lies in the confrontation between several such similarity relationships.

Finally, Logan in [9] has recently proposed a very similar approach to ours, which also uses Cepstrum Coefficients, only with a different modelling and a more complex matching algorithm.

2.2 Feature extraction:

The signal is cut into 2048 points frames (50ms), and for each frame, we compute the short-time spectrum. We then use Mel Frequency Cepstrum ([10]) to estimate the spectral envelope of each frame. The spectral envelope of a signal is a curve in the frequency-magnitude space that "envelopes" the peaks of its short-time spectrum. In the widely researched, above-mentioned problem of instrument recognition, it has been demonstrated that this feature explains a large part of the timbre of instruments ([11]).

The cepstrum is the inverse Fourier transform of the log-spectrum.

$$c_n = \frac{1}{2\pi} \times \int_{\omega=-\pi}^{\omega=+\pi} \log(S(e^{j\omega})) \cdot e^{j\omega \cdot n} d\omega \qquad (1)$$

We call mel-cepstrum the cepstrum computed after a non-linear frequency warping onto a psychoacoustic frequency scale (the *Mel* scale). The c_n in (1) are called Mel Frequency Cepstrum Coefficients (MFCCs).

The low order MFCCs account for the slowly changing spectral envelope, while the higher order ones describe the fast variations of the spectrum. Therefore, to obtain a timbre measure that is independent of pitch, we only use the first few coefficients. In [12], we have measured that the optimum dimension of the set was around 10 coefficients. In this work, we shall use the first 8 coefficients.

2.3 Gaussian Mixture Modelling

The feature extraction yields a feature vector of dimension 8 for each frame, which is believed to be a good and compact representation of the "local timbre" of the frame.

A typical 3-minute song is therefore represented with 3600 feature vectors, i.e. 28,800 coefficients, which then have to be compared with data from other songs. In order to reduce both the quantity and variability of the data to be compared, we model the distribution of each song's MFCCs as a mixture of Gaussian distributions over the space of all MFCCs.

A Gaussian Mixture Model (GMM) estimates a probability density as the weighted sum of M simpler Gaussian densities, called components or states of the mixture. ([13]):

$$p(F_t) = = \sum_{m=1}^{M} c_m \mathrm{N}(F_t, \mu_m, \Gamma_m) \qquad (2)$$

where F_t is the feature vector observed at time t, N is a multivariate Gaussian pdf with mean vector μ_m, covariance matrix Γ_m, and c_m is a mixture coefficient (also called state probability).

We initialise the GMM's parameters by k-mean clustering, and train the model with the classic E-M algorithm ([13]). In this work, we use mixtures of M=3 Gaussian distributions, which have proved sufficient to model the MFCC distribution of most songs.

2.4 Distance between models

We can now use these Gaussian models to match the timbre of different songs, which gives a similarity measure based on the audio content of the music. There are 2 ways such a distance can be computed.

2.4.1 Likelihood

One can match one song (A) against the timbre model of another song (B), by computing the "probability of the data given the model" (likelihood), i.e. computing the probability that the MFCCs of song A be generated by the model of B, using the formula given in (2). This is the most precise and logical way to compute a distance, but it requires to have access to song A's MFCC, which are relatively heavy to compute and to store.

2.4.2 Sampling

If we assume that we don't have access to the songs' MFCC when we want to compute the distance, but only to their timbre models, one can also directly match the models.

While it is easy to compute a distance between two Gaussian distributions (M=1), using for instance the classical Kullback-Leibler distance ([13]), it is a trickier problem to evaluate a distance between two *sets* of Gaussian Distributions, like in a GMM (M>1).

The method we have chosen is to sample from one GMM, and to compute the likelihood of the samples given the other GMM. This corresponds roughly to re-creating a song from its timbre model, and applying the likelihood method defined above to this newly created song and the other song's model. The distances are then normalized between 0 and 1 and made symmetric.

This approach is well suited to large musical databases, where it is crucial not to store the MFCCs themselves, but only the GMM parameters. In dimension 8, each Gaussian distribution in the GMM is represented with only 17 floating point numbers (1 mixture coefficient, 8 coefficients for the mean vector, and 8 coefficients for the covariance matrix, which is assumed to be diagonal). This is more than 1,000 times more efficient memory-wise than storing the MFCCS.

3. EVALUATION

Experiments were performed to evaluate the quality of this timbre similarity measure in the context of Cuidado. In this project we have set up a database of 17,075 popular music titles, together with metadata extracted automatically through different techniques. Metadata include information about artists, genres, tempo, energy, … and the timbre models discussed here.

3.1 Examples

Here we give some examples of duplets (or n-plets) of songs that are found similar by our system, i.e. whose timbre models are closely matched to one another. Many more examples can be found on the project web page ([14]).

3.1.1 Same songs

As a benchmark, it is interesting to note that duplicates of a same song (i.e. different mp3 encoding, different radio broadcasting…) are always closely matched. This echoes the work done on music title identification mentioned in the introduction.

3.1.2 Same artist

There are many examples of songs by the same artist that are closely matched by our system (however see 3.2 for a discussion about this).

- Piano pieces: *Franz Schubert Op90- No2 in E flat major* and *Franz Schubert Op90- No4 in A flat major*
- Harpsichord pieces: *Bach - Wohltemperierte Clavier - Fuga II in C minor* and *Bach - Wohltemperierte Clavier - Praeludium IV in C sharp minor*
- Heavy guitar overload: *Therapy - Brainsaw* and *Therapy - Stop it you're killing me*, etc.

3.1.3 Same Genre

Some similar songs have different artists, but show some kind of genre/style similarity (whatever this means, as music genre is a rather ill-defined concept). Here are some typical examples:

- Piano pieces: *Scriabin - Sonate pour Piano no 2*, *Mozart - Sonate pour Piano KV 533-1* and *Weber - Sonate pour Piano opus 49 no 3*
- Harpsichord pieces: *Bach - Das Wohltemperierte Clavier - Praeludium IV in C sharp minor BWV849* and *Couperin - Gavotte*
- "Power Rock": *Therapy - Brainsaw*, *Skunk Anansie - Intellectualise My Blackness*, *Nirvana - Smells Like Teen Spirit.*

3.2 Objective Evaluation

The objective evaluation of the "quality" of our timbral similarity measure is problematic. In the framework of Cuidado, each song is associated with textual metadata, so we could imagine comparing the timbre similarity against a "textual similarity" of artist or genre. However, this approach is not relevant, since two songs of the same artist or same genre do not necessarily have close timbres.

For instance:

- two songs by The Beatles: "*Helter Skelter*" (heavy overloaded guitars), and "*Lucy in the Sky*" (tremolo organ)
- two jazz pieces: "*Ascension*" by John Coltrane (free jazz saxophone), and "*My Funny Valentine*" sung by Chet Baker, etc.

We have conducted a quantitative study of the correlation between timbre and artist/genre similarity using the 17,075 songs in the Cuidado database. This study shows that such examples are not exceptions, but rather are nearly as numerous as examples of the opposite case.

For each title in the database, we compute its "timbral distance" to all the other titles, and compare these distances to the genre of the titles (only the root level of the Cuidado genre taxonomy is used, i.e. 18 genre families: Ambiance, Blues, Classical, Country, Electronica, Folk, Hard, Hip Hop, Jazz, New Age, Pop, Reggae, Rhythm&Blues, Rock, Rock&Roll, Soul, Variety, World).

Results can be seen in Table 1 and Table 2. Both tables show that for a given genre taxonomy, there is a very poor correlation between genre and timbre. In an Information Retrieval point of view, the precision of a query on genre based on timbral distance is very low (14.1%).

Table 1: Average number of closest songs with the same genre as the query

Number of Timbre Neighbors	Average number of songs in the same genre
Closest 1	0.43
Closest 5	1.43
Closest 10	2.56
Closest 20	4.61
Closest 100	18.09

In Table 2, "Overlap on Same Genre" is the ratio

$$\frac{N_{diff<same}}{N_{diff}} \quad (3)$$

where N_{diff} is the total number of songs with a different genre as the query's, and $N_{diff<same}$ is the number of songs in N_{diff} whose timbral distance to the query is smaller than the mean distance to songs of the same genre. Similarly, "Overlap on Different Genre" describes the proportion of songs which have the same genre as the query, but whose distance to the query is larger than the mean distance to songs of the different genre. Both values are high.

Table 2: Measures of the overlap between different genres

Average distance between titles	27.15
Average distance between titles of the same genre	26.91
Average distance between titles of different genres	27.17
Overlap on same genre	57.1%
Overlap on different genre	27.1%
Precision	14.1%
Recall	61.2%

These correlation measures obviously depend on the composition of the database, and on the genre taxonomy that is used. Moreover, it also depends on the artist or the genre: some artists/genres are more "coherent" than others, e.g. pre-war blues guitarists are more "homogeneous" timbre-wise than *The Beatles.* Nevertheless, this study shows that it is hard to base an objective evaluation of timbral similarity with respect to another music similarity measure.

3.3 Subjective Evaluation

Given the difficulty of an objective evaluation of the quality of our timbre distance, we have conducted a limited subjective evaluation.

Users are presented a target song S, and two test songs A and B, and have to decide which test song A or B is the closest to S. We then compare this ordering with the distances from A and B to S. The pair of songs are not chosen at random in our 20,000 title database, because in most cases both alternatives would be highly dissimilar to the target, and being able to predict a user's choice of which is more dissimilar would be very hard. On the contrary, A and B were chosen in such way that A and S are close to one another according to our measure (e.g. A is one of the 20 nearest neighbors to S), while B and S are more distant. We conducted the evaluation on 10 users from our lab. The average result of the test is that about 80% of the songs are well ordered by our system.

Larger scale user-tests are under way in the context of Cuidado. However, these preliminary results have already shown that, with our experimental protocol, the acceptance of the notion of timbral similarity by users is not always systematic. Deciding whether two songs are "similar" can be uncertain, as it is an ill-defined concept. In particular, it is difficult to evaluate similarity based on one attribute (here *timbre* similarity), because our judgment is simultaneously influenced by other attributes (same tempo, same artist, totally different genre…).

3.4 Interestingness

3.4.1 What is an interesting result?

In itself, the timbral similarity measure examined here does not always yield useful results. The examples of matched songs given in 3.1 are rather expected and unsurprising: it is fairly obvious that two songs by the same artist show some kind of similarity. Similarly, it is not much informative for anyone to say that a song by Metallica is closer to a song by ACDC than to a Beethoven string quartet. These similarities only reinforce some "background" cultural knowledge that we all share. Moreover, most of the time, these similarities can be assessed from simple textual metadata available about the songs.

On the other hand, the timbral similarity measure sometimes uncovers genuinely surprising associations between music titles: songs by different artists or genres, but also different dates of production, different cultural backgrounds, etc. Here are some examples of such songs that are found similar by our system, i.e. whose timbre models are closely matched to one another. These, and many more, can again be heard on the project's web page ([14]).

- Piano music:
  - "Classical" and "Contemporary": *Rachmaninov - Moment Musical opus 16 no 2, Gyorgy Ligeti - Concerto for Piano and Orchestra.*
  - "Classical" and "Jazz": *Schumann - Kreisleriana, Op 16-5* and *Bill Evans - I loves you Porgy*
- Orchestral textures:
  - "Jazz" and "Classical": *Orchestre Symponique de Montreux - Porgy and Bess* and *Prokofiev - Celibidache - Symphonie no 5-1 opus 100.*
  - "Classical" and "Pop": *Beethoven - Romanze fur Violine und Orchester Nr. 2 F-dur op.50* and *Beatles - Eleanor Rigby*
  - "Classical" and "Musicals": *Beethoven - Romanze fur Violine und Orchester Nr. 2 F-dur op.50* and *Gene Kelly – Singin' in the rain*
- "Trip Hop" and "Celtic Folk ": *Portishead - Mysterons* and *Alan Stivell - Arvor You*. (same kind of harpy theremin-like ambiance)

Contrary to the examples in 3.1., these associations couldn't be discovered with a non-signal technique. They provoke an exciting feeling of "discovery", comparable to the one that one gets when suddenly recognizing the origin of a sampled bit in a contemporary song, e.g. Stevie Wonder sampled in a hip-hop tune.

The feeling that users have when they gain a sudden insight into previously puzzling phenomena is studied by cognitive scientists under the name of "Aha!". The *interestingness* of music similarity measures for Music Information Retrieval lies in this very phenomenon of musical "Aha".

3.4.2 Towards a measure of interestingness

In Music Information Retrieval in very large databases, we are faced with the following problem: with any similarity measure, in particular here timbral similarity, the number of titles similar to a given query can be very large, but only a few of these songs are likely to be of any interest to the user.

In the related fields of datamining and knowledge discovery, the notion of *interestingness* has been widely researched in an attempt to increase the utility, relevance and usefulness of the patterns generated by such techniques. In the context of music, we can only aim at a *subjective* measure of "interestingness", which depends on the user who examines the result set. A survey of some "interestingness measures" proposed in such contexts can be found in [15].

Notably, Silberschatz in [16] proposes a definition of subjective interestingness based on *unexpectedness*, which seems well suited to formalize the '"aha" phenomenon illustrated above. For Silberschatz, interestingness measures the extent to which a belief is changed as a result of encountering new evidence (i.e. the discovered knowledge). It is given by:

$$I = \sum_{\alpha} \frac{\left| p(\alpha/E,\varepsilon) - p(\alpha/\varepsilon) \right|}{p(\alpha/\varepsilon)} \qquad (4)$$

where α is a belief, E is new evidence, ε is the previous evidence supporting belief α, $p(\alpha/\varepsilon)$ is the confidence in belief α, and $p(\alpha/E,\varepsilon)$ is the new confidence in belief α given the new evidence E.

We can build on this idea to model the "aha" phenomenon illustrated above. "Aha" lies in the contradiction between two evidences: one based on "textual metadata" (same artist, same genre, etc.), and the other based on timbre. The textual evidence ε can be used to compute an "a priori" confidence in the belief α that "these 2 songs may sound the same". One possible confidence function $p(\alpha/\varepsilon)$ is found in table 3. This in fact is yet another music similarity measure, based on text, $S_{text} = 1 - p(\alpha/\varepsilon)$. Similarly, the timbral evidence E is associated with an updated confidence function $p(\alpha/E,\varepsilon) = 1 - S_{timbre}$, where S_{timbre} is the timbral similarity measure introduced earlier.

It follows that the amount of "aha" generated by a duplet of songs (s,t) could be defined as:

$$"AHA"=\frac{\left|S_{timbre}(s,t)-S_{text}(s,t)\right|}{S_{text}(s,t)} \quad (5)$$

Using this with the Cuidado database, one can retrieve the most surprising – i.e. the most interesting - timbre matches to one query. One application of this is described in section 4.

Table 3: a possible confidence function based on textual metadata evidence

Textual evidence	Confidence
Same artist + same title (e.g. alternative takes, live versions)	0.9
Same artist	0.8
Same title (e.g. covers)	0.7
Same genre/subGenre (e.g. Rock/Indie & Rock/Indie)	0.6
Same genre (e.g. Rock/Indie & Rock/Grunge)	0.5
Related genres (e.g. Rock & pop)	0.4
Slightly related genres (e.g. Rock/Latino & World/Latino)	0.3
Different genres (e.g. Rock/Hard & Classical/Chamber Music)	0.1

3.5 Section conclusion

On the one hand, we have shown the limits of the intrinsic validation of a timbre similarity measure. First, its *objective* validation is often impossible: one need to compare it against other similarity measures, yet they are independent, uncorrelated dimensions. Moreover, this may lead to circularity in the assessment process ("Similarity number 1 is valid if similarity number 2 is valid"). Second, its *subjective* evaluation is also difficult. One has to judge different attributes of music separately, and it is often difficult to disregard the fact that, say, two songs are by the same artist while judging their timbral similarity.

On the other hand, we have shown that a similarity measure, even if validated, is sometimes neither interesting nor useful in itself. It may yield a large number of non-relevant, trivial matches. In an attempt to solve this, we have proposed a measure of interestingness, based on musical "aha", which tends to maximize the difference between two similarity measures, timbral and textual. A discussion about the evaluation of this measure of interestingness can be found in the next section.

More generally, this suggests that the usefulness of similarity measures, and music descriptors at large, is not to be found *intrinsically*, but rather lie, *extrinsically*, in their mutual confrontation. There is a crucial need for meaningful *exploitation schemes* of the descriptors. In the next two sections, we present two applications developed for CUIDADO's Music Browser [1], a client-server set of applications for EMD back offices and Internet music portals. Both applications are precisely aiming at finding interesting compromises between contradictory similarity measures or musical descriptors.

4. APPLICATION 1: "AHA" SLIDER

The first application is a music query system that ranks the results simultaneously on two dimensions, corresponding to the two similarity measures used in 3.4.2.: textual and timbral similarity. The user enters a query by entering a text query, or by selecting one song in the database. The system tries to answer his request: "I like this, find me some other songs that sound the same". Figure 1, 2 and 3 are screenshots of the system. In the answer window, the songs in the 17,075 songs' database are ordered by *timbral similarity* to the query. Simultaneously, a slider at the bottom of the window allows the user to filter the result set in real time, according to the *textual similarity* to the query.

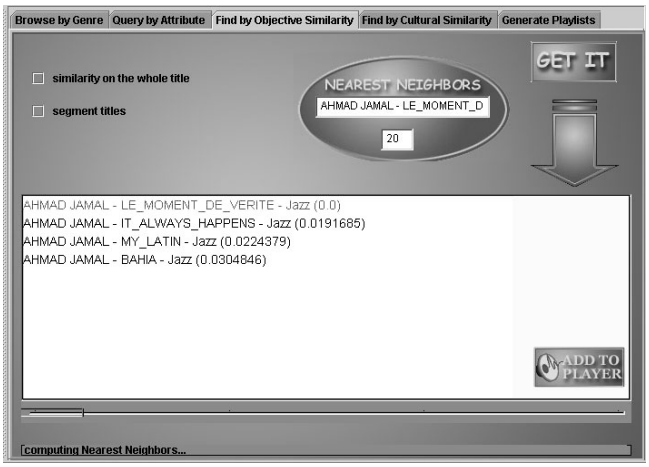

Figure 1: a timbral similarity query with the "aha" slider on the left

In the example in figure 1, 2 and 3, the user has asked the system to return songs similar to a jazz piano tune by Ahmad Jamal, "Le Moment de Verite". In figure 1, the slider is at the left side. This constrains the result set to contain only songs by the same artist: the four Ahmad Jamal songs in the database are ordered by timbral similarity to the query.

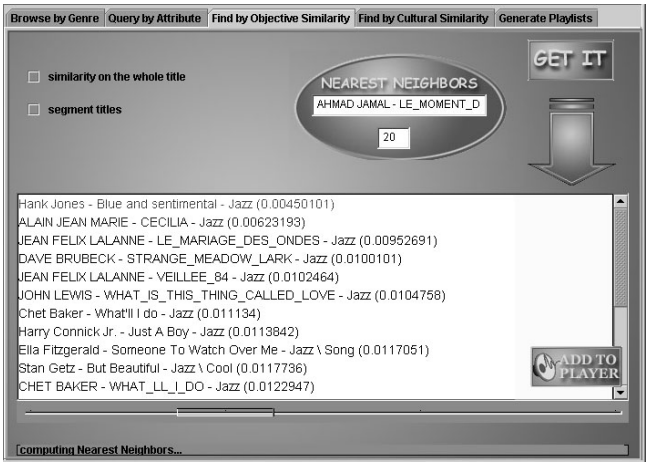

Figure 2: intermediary position of the "aha" slider

In figure 2, the user has moved the slider to an intermediary position. This allows the system to return songs that are more loosely connected to the query, like songs of the same genre: jazz. The results are still ordered by timbre similarity to Ahmad Jamal song, therefore we see that the system has returned other *piano* jazz songs (Alain Jean Marie, Dave Brubeck, John Lewis, Harry Connick Jr, etc.).

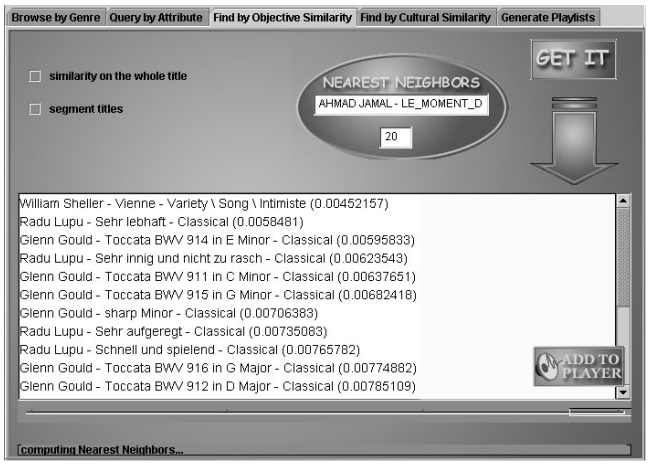

Figure 3: "aha" slider in the rightmost position

Finally, in figure 3, the user has moved the slider on to the rightmost position. This constrains the result set to contain songs textually very different from the query, like songs of non-related genres. Now we see that the system has returned piano songs of genre "classical" (Bach, Schumann) or "variety" (William Sheller). These songs have a high value of "aha", as defined in 3.4.2.

This application attempts to give the user full control over the degree of surprise and freedom in the way the system satisfies his request. A non-exploratory behavior (slider on the left) implies that the system should return exactly the answer to the query, or an answer that is as expected as possible (same title, same artist). An exploratory behavior (slider on the right) consists in letting the system try different regions of the catalogue rather that strictly match the query. In fact, *explorativeness* consists in expecting the system to depart from the query, and return some sorts of "interesting" music proposals.

Systematic studies on users are under way. We are reviewing psychological studies on musical tastes, notably the theory of prototypicality (typical instances of a category – or genre – are usually preferred because they are easier to classify – see [17]) and familiarity (novelty is an important potential source of musical dislike – see [18]).

Preliminary results show some users tend to react negatively towards exploration, at least in the beginning of their interaction. In the long run, however, there seems to be a consensus in the acceptance of this dimension. This can be explained by several factors, but the main one is probably that most users quickly exhaust their capacity in issuing explicit queries: it is only once the well-known artists or hits are queried, in a non exploratory mode, that such a "aha" slider appears as useful.

5. APPLICATION 2: PLAYLIST GENERATION

Most proposals of EMD systems so far have followed a purely database-oriented approach, in that users are proposed individual titles, either queried explicitly (e.g. e-compile, PressPlay, Duet, the "aha" slider above) or through recommendation systems (e.g. Amazon). Departing from these approaches, we have introduced in [19] the idea of producing automatically *sequences* of music titles – play lists, instead of individual titles. These sequences are produced automatically from a set of so-called *global constraints*, which specify properties of the whole sequence, such as:

- the playlist should contain 12 different titles,
- the playlist should not last more than 76 minutes,
- the genre of a title should be *close* to the genre of the next title,
- the playlist should contain at least 60% of *instrumental* titles,
- the sequence should contain titles with increasing tempo, etc.

The problem of generating such playlists given a very large title catalogue with musical metadata, and a set of arbitrary constraints is a NP-hard combinatorial problem. Moreover, in the case of a contradictory set of constraints, there may not be an exact solution. An ideal system should therefore be able to generate good approximate compromises. The Cuidado Music Browser is able to generate such playlists automatically, using a fast algorithm based on adaptive search, and described in [20].

We give here an example of a 10-title playlist with the following constraints:

1- Timbre continuity: the playlist should be timbrally homogeneous, and shouldn't contain abrupt changes of textures.

2- Genre Cardinality: the playlist should contain 30% of Rock pieces, 30% of Folk, and 30% of Pop

3- Genre Distribution: the titles of the same genre should be as separated as possible

Figure 4 shows a screenshot of the playlist generation system.

One solution found by the system is the following playlist:

- `Arlo Guthrie – City Of New Orleans - Genre = Folk/Rock`
- `Belle & Sebastien - The boy done wrong again - Genre = Rock/Alternatif`
- `Ben Harper – Pleasure & Pain - Genre = Pop/Blues`
- `Joni Mitchell - Borderline - Genre = Folk/Pop`
- `Badly Drawn Boy - Camping Next to Water - Genre = Rock/Alternatif`
- `Rolling Stones – You Can't always get what you want - Genre = Pop/Blues`
- `Nick Drake - One of these things first - Genre = Folk/Pop`
- `Radiohead - Motion Picture Soundtrack - Genre = Rock/Brit`
- `The Beatles - Mother Nature's Son - Genre = Pop/Brit`
- `Tracy Chapman - Talkin' about a Revolution - Genre = Rock/Folk`

It is easy to check that the genre cardinality is correct (3 "folk", 3 "pop", 4 "rock"), and the genre distribution constraint is also well satisfied. One can see that the system has also managed to maintain the timbre continuity by selecting the right subgenres ("Folk/Rock" and "Rock/Folk"), and picking songs which mainly consist of acoustic guitar + voice (Nick Drake, Ben Harper, Tracy Chapman, etc.).

It appears that the combination of well-designed sequence properties makes it possible to produce sequences that provide optimum compromises between contradictory constraints. Compared with the "aha slider" which only allows the user to interacts with two dimensions, here one can use an arbitrary big and complex set of constraints, holding on any combination of musical descriptors or similarity measures.

Figure 4. Screenshot of the playlist generation system

6. CONCLUSION AND FUTURE WORK

We have described a measure of the similarity of the "global timbre" of music titles, based on the audio signal. By working on its evaluation, both objective and subjective, we have raised the question whether music similarity measures, and music descriptors at large, are intrinsically useful. We suggest that their validity and their interestingness rather lie, extrinsically, in the confrontation/compromise between several music similarities or descriptors. We introduce the idea of music "aha", which measures the difference between two contradictory similarity measures, timbral and textual. We have reported on two exploitation schemes designed for Cuidado's Music Browser: "aha" slider and playlist generation. Both applications can find interesting compromises between contradictory similarity measures or musical descriptors.

The next step is to conduct a more formal evaluation of these algorithms and interfaces, relying on existing psychological experiments on musical preferences. Future work about our timbral similarity measure also includes studying how to model and match the timbres of very "heterogeneous" songs (e.g. The Beatles – A Day in the Life, Queen – Bohemian Rhapsody, etc.). Future work about interestingness focuses on the design of simple interfaces for playlist generation that allow the specification of complex constraints/interestingness functions graphically.

7. REFERENCES

[1] Pachet, F. Metadata for music and sounds: The Cuidado Project, in proc. of CBMI workshop, Brescia, Italy, 2001.

[2] Scheirer, E. Music-Listening Systems, PhD Thesis, MIT Cambridge, MA USA 2000

[3] Klapuri, A. et al. Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals, in Proc. DAFX-00, Verona, Italy, 2000.

[4] Herrera-Boyer P., et al. Towards Instrument Segmentation for Music Content Description: a Critical Review of Instrument Classification Techniques, in proc. ISMIR 2000.

[5] Tzanetakis, G. Automatic Musical Genre Classification of Audio Signals, in proc. ISMIR 2001.

[6] Allamanche, E. Content-based Identification of Audio Material Using MPEG-7 Low Level Description, in proc. ISMIR 2001

[7] Foote, J. Content-Based Retrieval of Music and Audio, in proc. of SPIE, Vol. 3229, pp. 138-147, 1997.

[8] Welsh, M. et al. Querying large collections of music for similarity. Technical report UC Berkeley Computer Science Division, 1999.

[9] Logan, B. & Salomon, A. A Music Similarity Function Based on Signal Analysis. In proc. of IEEE International Conference on Multimedia and Expo (ICME), August 2001.

[10] Rabiner, L.R. & Juang, B.H., Fundamentals of speech recognition. Prentice-Hall 1993

[11] Schwarz, D. & Rodet, X. Spectral Estimation and Representation for Sound Analysis-Synthesis. In proc. ICMC 1999

[12] Aucouturier, J.J. & Sandler, M. Segmentation of Musical Signals Using Hidden Markov Models, in Proc. 110th Convention of the AES, Amsterdam 2001

[13] Bishop, C. Neural Networks for pattern recognition, Oxford University Press

[14] Project web page : www.csl.sony.fr/~jj/Timbre/timbre.html

[15] Hilderman, R.J. & Hamilton, H.J. Knowledge discovery and interestingness measures: A survey. Technical Report Computer Science, University of Regina, October 1999

[16] Silberschatz, A. & Tuzhilin, A. What makes patterns interesting in knowledge discovery systems. IEEE Trans. On Knowledge And Data Engineering, 8:970--974, 1996

[17] Martindale, C. & Moore, K. Priming, prototypicality, and preference. Journal of Experimental Psychology: Human Perception and Performance, 14, 661-670, 1988.

[18] Hargreaves, D.J. & North, A. C. The Social Psychology of Music.Oxford: Oxford University Press, 1997.

[19] Pachet, F. et al. A Combinatorial approach to content-based music selection, in proc. IEEE International Conference on Multimedia Computing and Systems, Firenze (It), 1999

[20] Aucouturier, J-J. & Pachet, F. Scaling up Playlist Generation Systems, in proc. IEEE International Conference on Multimedia and Expo, Lausanne 2002.