SIMAC: SEMANTIC INTERACTION WITH MUSIC AUDIO CONTENTS

Perfecto Herrera[1], Juan Bello[2], Gerhard Widmer[3], Mark Sandler[2], Òscar Celma[1], Fabio Vignoli[4], Elias Pampalk[3], Pedro Cano[1], Steffen Pauws[4], Xavier Serra[1]

[1]Universitat Pompeu Fabra, Barcelona; [2]Queen Mary University, London;
[3]Austrian Research Institute for Artificial Intelligence, Vienna; [4]Philips Research, Eindhoven

Keywords: Semantic Audio, Music Information Retrieval, Music Similarity, Music Recommendation, Music Description.

Abstract

The SIMAC project addresses the study and development of innovative components for a music information retrieval system. The key feature is the usage and exploitation of semantic descriptors of musical content that are automatically extracted from music audio files. These descriptors are generated in two ways: as derivations and combinations of lower-level descriptors and as generalizations induced from manually annotated databases by the intensive application of machine learning. The project aims also towards the empowering (i.e. adding value, improving effectiveness) of music consumption behaviours, especially of those that are guided by the concept of similarity.

1 Introduction

In recent years the typical music consumption behaviour has changed dramatically. Personal music collections have grown favoured by technological improvements in networks, storage, portability of devices and Internet services.The amount and availability of songs has de-emphasized its value: it is usually the case that users own many music files that they have only listened to once or even never. It seems reasonable to think that by providing listeners with efficient ways to create a personalized order on their collections, and by providing ways to explore hidden ``treasures'' inside them, the value of their collection will drastically increase.

Beside, on the digital music distribution front, there is a need to find ways of improving music retrieval effectiveness. Artist, title, and genre keywords might not be the only criteria to help music consumers in finding music they like. This is currently mainly achieved using cultural or editorial metadata ("this artist is somehow related with that one") or exploiting existing purchasing behaviour data ("since you bought this artist, you might also want to buy this one, as other customers with a similar profile did"). A largely unexplored (and potentially interesting) alternative is using semantic descriptors automatically extracted from the music audio files. These descriptors can be applied, for example, to organize a listener's collection, recommend new music, or generate playlists.

In the past twenty years, the signal processing and computer music communities have developed a wealth of techniques and technologies to describe audio and music contents at the lowest (or close-to-signal) level of representation. However, the gap between these low-level descriptors and the concepts that music listeners use to relate with music collections (the so-called "semantic gap") is still to a large extent waiting to be bridged.

The remaining sections of this paper will present the work and developments made in the SIMAC project to bridge the semantic gap and to enhance the music enjoyment experience. We will first introduce several semantic descriptors of music contents, developed for different musical facets (rhythm, harmony, timbre, etc.). Music similarity will be then discussed and the evaluation of a complete system will be presented. Three prototypes, incorporating semantic descriptors and similarity metrics will be then outlined. A discussion on future trends and open issues that deserve further research will conclude the paper.

2. Semantic Description of Music Contents

Music content processing systems operating on complex audio signals are mainly based on computing low-level signal features. These features are good at characterising the acoustic properties of the signal, returning a description that can be associated to texture, or at best, to the rhythmical attributes of the signal [1], [42].

Alternatively, the SIMAC approach proposes that music content can be successfully characterized according to several "musical facets" (i.e. rhythm, harmony, melody, timbre) by incorporating higher-level semantic descriptors to a given feature set. Semantic descriptors are measures that can be computed directly from the audio signal, by means of the combination of signal processing, machine learning techniques, and musical knowledge. Their goal is to emphasise the musical attributes of audio signals (e.g. chords, rhythm, instrumentation), attaining higher levels of semantic complexity than low-level features (e.g. spectral coefficients, Mel frequency cepstral coefficients, and so on), but without being bounded by the constraints imposed by the rules of music notation. Describing musical content according to this

view does not necessarily call for perfect transcriptions of music, which are outside the scope of existing technologies, even though recent outstanding progress has been reported [22].

Our view is that several of the shortcomings of the purely data driven techniques can be overcome by applying musical knowledge. The richness of the description that can be achieved is well beyond that from existing music downloading and retrieval prototypes. Our results also suggest that the use of meaningful descriptors pushes the "glass ceiling" for music classification to levels higher than originally anticipated for previous data-driven approaches.

Our proposed description scheme can be seen as a function of musical dimensions: rhythm, harmony, timbre and instrumentation, long-term structure, intensity, and complexity. The following sections are devoted to outlining SIMAC contributions to all those aspects.

2.1 Rhythm

In its most generic sense, rhythm refers to all of the temporal aspects of a musical work, whether represented in a score, measured from a performance, or existing only in the perception of the listener [16]. In the literature the concept of "automatic rhythm description" groups a number of applications as diverse as tempo induction, beat tracking, rhythm quantisation, meter induction and characterisation of timing deviations, to name but a few. In SIMAC, we have investigated a number of these different aspects, from the low-level of onset detection, to the characterization of music according to rhythmic patterns.

At the core of automatic rhythmic analysis lies the issue of identifying the start, or onset time, of events in the musical data. As an alternative to standard energy-based approaches we have proposed methodologies that work solely with phase information [3], or that are based on predicting the phase and energy of signal components in the complex domain [4], greatly improving results for both percussive and tonal onsets. However, there is more to rhythm than the absolute timings of successive musical events. For instance, we have proposed a general model to beat tracking [9], based on the use of comb filtering techniques on a continuous representation of "onset emphasis", i.e. an onset detection function. Subsequently, the method was expanded to combine this general model with a context-dependent model [10], by including a state space switching model. This improvement has been shown to significantly improve upon previous results, in particular with respect to maintaining a consistent metrical level and preventing phase switching between off-beats and on-beats.

Furthermore, in our work we demonstrate the use of high-level rhythmic descriptors for genre classification of recorded audio. An example is our research in tempo-based classification [17], [15], showing the high relevance of this feature while trying to characterize dance music. However, this approach is limited by the assumption that, given a musical genre, the tempo of any instance is among a very limited set of possible tempi. To address this, in [11], an approach is proposed that uses bar-length rhythmic patterns for the classification of dance music. The method dynamically estimates the characteristic rhythmic pattern on a given musical piece, by a combination of beat tracking, meter annotation and a k-means classifier. Genre classification results are greatly improved by using these high-level descriptors, showing the relevance of musically-meaningful representations for MIR tasks. For a more complete overview of the state of the art on rhythmic description and our own contributions towards a unified framework see [16].

2.2 Harmony

The harmony of a piece of music can be defined by the combination of simultaneous notes, or chords; the arrangement of these chords along time, in progressions; and their distribution, which is closely related to the key or tonality of the piece. Chords, their progressions, and the key are relevant aspects of music perception that can be used to accurately describe and classify music content [13].

Harmonic based retrieval has not been extensively explored prior to SIMAC. A successful approach at identifying harmonic similarities between audio and symbolic data was presented in [32]. It relied on automatic transcription, a process that is partially effective within a highly constrained subset of musical recordings (e.g. mono-timbral, no drums or vocals, small polyphonies). To avoid such constraints we adopt the approach where we describe the harmony of the piece, without attempting to estimate the pitch of notes in the mixture. Avoiding the transcription step allows us to operate on a wide variety of music.

This approach requires the use of a feature set that is able to emphasise the harmonic content of the piece, such that this representation can be exploited for further, higher-level, analysis. The feature set of choice is known as a Chroma or Pitch Class Profile, and they represent the relative intensity of each of the twelve semitones of the equal-tempered scale. This feature is related to one of the two dimensions of the pitch helix [36] that is related to the circularity of pitch as you move from one octave to another, and that can be accurately estimated from raw audio signals.

In SIMAC, we have proposed a state-of-the-art approach to tonality estimation [14] by correlating *chroma* distributions with key profiles derived from music cognition studies [24]. Results show high recognition rates for a database of recorded classical music. In our studies, we have also concentrated on the issue of chord estimation based on the principled processing of chroma features, by means of tuning, and a simple template-based model of chords [19]. Recognition rates of over 66% were found for a database of recorded classical music, though the algorithm is being used also with other musical genres. A recent development includes the generation of a harmonic representation by

means of a Hidden Markov Model, initialized and trained using musical theoretical and cognitive considerations [5]. This methodology has already shown great promise for both chord recognition and structural segmentation.

2.3 Timbre and instrumentation

Another dimension of musical description is that defined by the timbre or instrumentation of a song. Extracting truly instrumental information from music, as pertaining to separate instruments or types of instrumentation implies classifying, characterizing and describing information which is buried behind many layers of highly correlated data. Given that the current technologies do not allow a sufficiently reliable separation, work has concentrated on the characterization of the "overall" timbre or "texture" of a piece of music as a function of low-level signal features. This approach implied describing mostly the acoustical features of a given recording and gaining little abstraction about its instrumental contents.

Even though it is not possible to separate the different contributions and "lines" of the instruments, there are some interesting simplifications that can provide useful descriptors. Examples are: lead instrument recognition, solo detection, or instrument profiling based on detection without performing any isolation or separation [20]. The recognition of idiosyncratic instruments, such as percussive ones, is another valuable simplification. Given that the presence, amount and type of percussion instruments are very distinctive features of some music genres and, hence, can be exploited to provide other natural partitions to large music collections, we have defined semantic descriptors such as the percussion index or the percussion profile [21]. Although they can be computed after some source separation [18], reasonable approximations can be achieved using simpler sound classification approaches that do not attempt separation [43], [34].

Additionally, our research in the area of instrumentation has contributed to the current state of the art in instrument identification of mono-instrumental music [7], using line spectral frequencies (LSF) and a k-means classifier. An extension to this work is currently exploring the possibility of enhancing this approach with a source separation algorithm, aiming at selective source recognition tasks, such as lead instrument recognition.

2.4 Intensity

Subjective intensity, or the sensation of *energeticness* we get from music, is a concept commonly and naïvely used to describe music content. Although intensity has a clear subjective facet, we hypothesized that it could be grounded on automatically extracted audio descriptors.

Inspired by the findings of Zils and Pachet [44], our work in this area has resulted in a model of subjective intensity built from energy and timbre low-level descriptors extracted from the audio data [35]. We have proposed a model that decides among 5 labels (*ethereal*, *soft*, *moderate*, *energetic*, and *wild*), with an estimated effectiveness of nearly 80%. The model has been developed and tested using several thousands subjective judgements.

2.5 Structure

Music structure refers to the ways music materials are presented, repeated, varied or confronted along a piece of music. Strategies for doing that are artist, genre and style-specific (i.e. the A-B themes exposition, development and recapitulation of a sonata form, or the intro-verse-chorus-verse-chorus-outro of "pop music"). Detecting the different structural sections, the most repetitive segments, or even the least repeated segments, provide powerful ways of interacting with audio content by means of summaries, fast-listening and musical gist-conveying devices, and on-the-fly identification of songs.

The section segmenter we have developed extracts segments that roughly correspond to the usual sections of a pop song or, in general, to sections that are different (in terms of timbre and tonal structure) from the adjacent ones. The algorithm first performs a rough segmentation with the help of change detectors, morphological filters adapted from image analysis, and similarity measurements using low-level descriptors. It then refines the segment boundaries using a different set of low-level descriptors. Complementing this type of segmentation, the most repetitive musical pattern in a music file can also be determined by looking at self-similarity matrices in combination with a rich set of descriptors including timbre and tonality (i.e. harmony) information [26]. Ground-truth databases for evaluating this task are still under construction, but our first evaluations yielded an effectiveness of section boundary detection higher than 70%.

2.6 Complexity

We define music complexity as the property of a musical element that determines how much effort the listener has to put into following and understanding that element. Music complexity in this context is understood as a multifaceted, semantic descriptor of musical audio content, which can be decomposed into timbral, rhythmic, structural, tonal and other facets. It is interesting to note the relationship between complexity and preference that has been put forward in the theory of Arousal Potential [6], which states that an individual's preference for a certain piece of music is related to the amount of activity it produces in the listener's brain, to which he refers as the arousal potential. According to this theory, there is an optimal arousal potential that causes the maximum liking, while a too low, as well as a too high, arousal potential results in a decrease of liking (i.e., it follows an inverted U-shaped curve). Since then, many experiments have been conducted showing clear interdependence between

the complexity of musical instances and the preference for them. Therefore, we can assume that complexity might be of relevance for systems dealing with music recommendation.

A complexity-related descriptor, computed by means of applying detrended fluctuation analysis (DFA), which reveals correlations within data series across different time scales, was found to be tightly correlated to semantic concepts linked to danceability [37]. Descriptors related to acoustic complexity (i.e. disparity of left and right channels, and dynamic changes), to timbre complexity (by means of exploiting data compression algorithms), and to harmonic complexity (from the existing tonal descriptors of section 2.2) have also been developed and included in an exploration prototype. Ground-truth databases for testing the effectiveness of musical complexity algorithms require massive listening tests which are currently being prepared.

3. Music Similarity

Finding 'similar' songs, albums, or artists is one of the most appreciated features for music playing systems and devices capable to get access to large music collections [38]. From a user perspective, judging similarity of songs either involves the comparison of two songs or the comparison of a set of alternative songs to a referent or ideal (e.g., a seed song). Simply stating that two songs are similar is not sufficient: we need to say that two songs are similar because of their instrumentation, their compositional style, their performers, the subject of their lyrics, etc. Evidently, similarity needs to be explained with respect to a feature or a set of features, which may change according to user education and preferences, listening context, attentional and cognitive limitations, and even depending on the songs that have to be compared at a given moment. A straightforward method is to list all features of the songs involved and find the overlap in features. The reality is more complicated: similarity judgements seem to come down to the computation of a 'psychological function' of shared, distinctive, and comparable features of the objects involved (i.e., songs, in our case). In order to perform this comparison, cognitive processes and reasoning using knowledge and conventions from the real world play an important role. Music psychology has already pointed out that besides instrumentation, at least tempo and genre information are relevant for generating similarity judgments of music [8]. We believe that, besides the involvement of various features, the contribution of each individual feature to the overall similarity needs to be weighted. Given that the importance of features is heavily dependent on the context and the listening intention at hand, the user should be empowered to have total control on this weighting procedure.

In this section we consider two sources from which similarity can be computed: (1) the audio signal and (2) information on the web. Audio-based similarity is usually based on low-level audio statistics and therefore it disregards most of the truly "musical" facets and, not less important, all the cultural background where music is generated and "consumed". This lack of cultural information can be addressed by web-based approaches which use, e.g., Google to find web-pages related to an artist and extract relevant information (e.g. word lists) from these pages. Based on these word lists, the similarity of two artists can be computed. In the remainder of this section, we briefly describe one approach for each source.

3.1 Audio-Based Similarity

As we stated in Section 2, there are a number of interesting descriptors that can be extracted directly from the audio signal. However, for a simple playlist generator [31] which requires minimum user interaction we have implemented a similarity measure based on simple low-level audio statistics [29]. In particular, we use a combination of two techniques: spectral similarity and fluctuation patterns.

Spectral similarity reflects to some extent timbre characteristics [25], [2],[25] which, additionally, are assumed to correlate with instrumentation. Numerous research efforts have already been devoted to timbre similarity in music [2], [12], [23], [25]. Unquestionably, timbre similarity is grounded by perception; non-musicians rather choose instrumentation over correct melody and harmony in similarity judgement of music by mere listening [41]. The basic idea is to summarize tracks by the typical spectra that occur in them. These summaries are then compared to each other to obtain a similarity value. In particular this is done by dividing the track into many very short (e.g. 20ms) frames. For each frame, the Mel frequency cepstral coefficients (MFCCs) are computed. These MFCCs are then clustered to find the most typical spectral shapes that occur. Once these cluster models are computed there are different ways to compare them (see Figure 1). One approach is to compute the likelihood of generating samples from one cluster by the other one.

To complement the spectral similarity we analyze periodic fluctuations in the loudness over time. These fluctuations are related to the beats, tempo and rhythm [27]. However, it is important to consider that these fluctuation patterns (FPs) are not comparable to what is know as rhythm patterns in music [12]. FPs are computed by cutting the track into 6 second segments. For each of these segments the spectrogram is computed and psychoacoustic transformations are applied (e.g. MFCCs). For each frequency band we compute the loudness amplitude modulations with a FFT. Thus, if there is an event periodically reoccurring every 500ms (2Hz or 120bpm) then this will be reflected in the fluctuation pattern. For each segment we thus obtain a 2-dimensional FP matrix where each row represents one frequency bands, each column a specific modulation frequencies and the values in the cells are the strength of the fluctuation. To summarize the multiple FPs per piece we compute the median of each cell and thus obtain a single pattern.

From the FPs we extract two descriptors: *Focus* (FP.F) and *Gravity* (FP.G). Focus relates to the clearness of the beats. Gravity is related to the perceived tempo. Focus is computed

as the mean of the fluctuation pattern after normalizing the pattern so that the maximum value equals 1. Gravity is computed as the centre of gravity on the modulation frequency axis minus the theoretical centre of gravity. The distance between two FPs is computed by interpreting the matrices as vectors and computing the Euclidean distance. The distance between the single value descriptors (FP.F, FP.G) is computed as the absolute difference. All distances are combined by linearly weighting their values and summarizing them to one value.

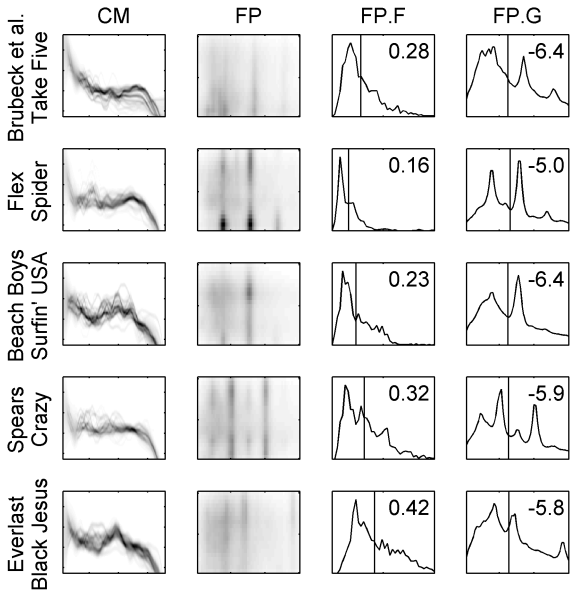

Figure 1. Visualization of the features used for audio-based similarity computations for five songs. On the y-axis of the cluster model (CM) is the loudness (dB-SPL), on the x-axis are the Mel frequency bands. The plots show typical spectral shapes and their variances on top of each other. On the y-axis of the FP are the Bark frequency bands, the x-axis is the modulation frequency (in the range from 0-10Hz). The y-axis on the FP.F histogram plots are the counts, on the x-axis are the values of the FP (from 0 to 1). The y-axis of the FP.G is the sum of values per FP column; the x-axis is the modulation frequency (from 0-10Hz). In the FPs, vertical lines indicate reoccurring periodic beats.

3.2 Web-based Similarity

A very different approach to compute the similarity of artists is to analyze the content of webpages that contain their names. In particular, an artist name and some constraints (e.g. "music" and "review") are used to query the World Wide Web through a search engine such as Google. The top ranked pages (e.g. the first 50) are retrieved and parsed. For each artist we obtain a long list of word occurrences [40]. Lists from different artists are then compared to each other using standard text retrieval techniques [33]). Using this type of similarity we can either classify artists into genres [23] or develop interfaces to browse music collections on the artist level [30]. In particular, the word occurrences can be used to automatically generate text summaries that describe an artist.

3.3 Evaluation

Direct evaluation of a similarity measure is rather difficult as it would require extensive listening tests (to annotate a music collection), or would require running a listening test for every variation of the algorithm (that is, for all parameter settings of interest). Alternatively, we can measure the performance indirectly through genre classification, for example, by using a nearest neighbour classifier, and assuming that (very) similar pieces (or artists) belong to the same genre.

In order to avoid unrealistic results due to overfitting the data, these techniques need to be correctly applied to several independent music collections, splitting them into training and testing sets, and ensuring the exclusion of the same artist in the training and in the testing (this is required to block the effect of having similar production effects, or artist voices in both sets).

A different type of evaluation can be done by studying in practice the results of different similarity functions. To achieve this goal, we have developed a system in which the similarity is based on a weighted combination of timbre, genre, tempo, year, and mood. The end-user can specify her personal definition of similarity by weighting these aspects on a graphical user interface. A conclusive user evaluation was conducted to assess the usability of the system (User Defined Similarity or UDS) in comparison to two control systems (CTRL1 and CTRL2) in which the user control on defining the similarity function was diminished. When a user asks for songs similar to a seed song in the UDS system, the jukebox displays the screen shown in Figure 2. The similarity components are represented by adapters. These adapters can be dragged on the bull's eye (as shown on the right-hand side of the screen). The radial distance of an adapter to the centre determines the weight of its corresponding similarity component in the similarity function. In this way, a user has the possibility to change the similarity function that is applied to the music collection. The list of songs on the left-hand side of the screen is sorted according to the degree of similarity; the songs that are closest to the seed are positioned at the top of the list. When using the two control systems (CTRL1 and CTRL2) the user interaction with of the system was the same, although users could not manipulate the definition of similarity. The similarity in CTRL1 was set accordingly to timbre only, in CTRL2 it was set to a fixed combination of timbre, genre and tempo. A thorough description of the experiment is provided in [39]. The user evaluation involved twenty-two participants, who were invited to use the three systems. The task was to compile a playlist of ten songs and the only criterion given was the quality of the playlist. The performance of the participants was analysed with respect to two types of measures: objective measures such as time spent and number of actions executed and subjective measure such as perceived quality and ease of use of the system and system preferences. From the results of the three

runs, two groups of people with different behaviour were identified. The "slow" group spent much time to explore the possibilities offered by the systems, while the "fast" group tried to minimize the time spent with the system. The user evaluation revealed also that some additional effort was needed to learn to work with the user-driven similarity function during first-time use, and therefore most users find the proposed system somewhat less easy to use than the other systems.

In conclusion, providing users with complete control on their personal definition of music similarity is found to be more useful and preferred than providing closed definitions that allowed no control on the involved musical facets.

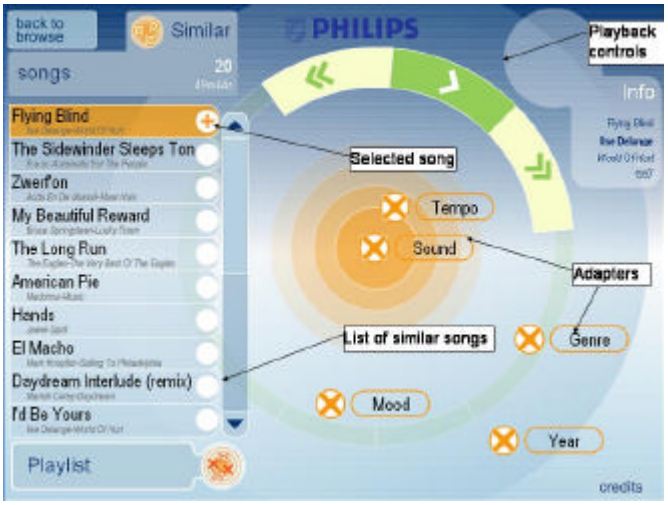

Figure 2. The E-Mu jukebox in "Similar Songs" mode. The user can define the similarity function applied to the music collection by dragging the sound/tempo/mood/genre/year adapters on the screen. An adapter that is close to the center is weighted more than when it is positioned in the periphery.

4. Prototypes

Three software prototypes integrating state-of-the-art automatic audio description and music similarity technology are under development:

The Music Annotator is an environment for the annotation and generation of music metadata at different levels of abstraction. It is composed of three tiers: an annotation client that deals with micro-annotations (i.e. within-file annotations about note onsets, chords, percussive events, beats, etc.), a collection tagger, which deals with macro-annotations (i.e. across-files annotations), and a collaborative annotation subsystem, which manages large-scale annotation tasks that can be shared among different research centres. The annotation client is an enhanced version of WaveSurfer, a speech annotation tool. The collection tagger includes tools for automatic generation of unary descriptors, invention of new descriptors, and propagation of descriptors across sub-collections or playlists. Finally, the collaborative annotation subsystem makes it possible to share the annotation chores and results between several research institutions, reducing the time and cost of them.

The Music Organizer and Explorer demonstrates the visualization and navigation across existing collections of music titles. 2-D maps are used to map songs according to semantic descriptors, and different similarity distance metrics can be tried in order to find similar music to a given seed song.

The Music Recommender is intended for providing recommendations of music titles that are legally downloadable from the WWW. We believe this is one of the first systems that combines audio generated information and cultural information to produce recommendations. The system, named *Foafing the Music*, relies on user preferences and user listening habits (computed from content analyses of the user's music collection). Tracking of preferences is managed by means of the Audioscrobbler[1] plugin. The system also exploits musical information that has been specially crawled from the Internet and properly structured and mined to generate musical knowledge. Moreover, nowadays, music websites are alerting the user about new releases or artist's related news, mostly in the form of RSS feeds. For instance, iTunes Music Store[2] provides an RSS (version 2.0) feed generator, updated once a week, which publishes new releases of artists' albums.

User profiles are based on the Friend of a friend (FOAF) initiative. The FOAF project provides a framework for representing information about people, their interests, relationships between them and their social connections. The FOAF vocabulary contains terms for describing personal information -name, nick, mailbox, interest, images, group membership…-. FOAF is based on the RDF/XML vocabulary. A FOAF description, then, describes a person in a machine readable format. Currently, the FOAF initiative is one of the big attainments of the Semantic Web.

Music and artists' recommendations, in the *Foafing the Music* system, are generated through the following steps:

1. Get interests from user's FOAF profile,
2. Detect artists and bands from these interests,
3. Select related artists, from artists encountered in the user's FOAF profile, and
4. Rate results by relevance.

The system reads an input FOAF profile -that is, an RDF/XML file-, and extracts user's interests. Then, it queries a music repository in order to detect whether the interest is a music artist (or a band), and selects similar artists to the ones found. To get artists' similarities, a focused web crawled has been implemented to look for relationships between artists (such as: related to, influenced by, followers of, etc.). Moreover, a music similarity distance is used to recommend

[1] http://www.last.fm
[2] http://www.apple.com/itunes

tracks that are similar to tracks composed or played by artists found in the FOAF profile.

Once the related artists have been selected, Foafing the Music filters music related information coming form RSS feeds to:

- Get new music releases,
- Create, automatically, playlists based on audio similarity.
- Download (or stream) audio from MP3-blogs and Podcast sessions, and
- View incoming concerts near to user's city
- Read artists' related news

5. Conclusions and further directions

In this paper we have provided an overview of the scientific achievements of the SIMAC project. The combination of both semantic descriptors addressing multiple musical facets and user-configurable similarity metrics emerge as key elements for successful systems aimed to enhance the interaction with music audio collections.

The reported research has purposely left aside one of the most important elements for that interaction: melody. There are unsolved technological problems regarding the extraction of the principal melody of a piece of music (not to mention the unreliability of user queries when "query by humming" is used). Source separation is another of the "holy grails" in music content processing. The benefits of achieving a separated representation of the concurrent musical streams are obvious but there is still no general approach to achieve that goal in a reliable and usable way. Even though most of our algorithms do not perform source separation, there is a lot of musical information that has been extracted with enough reliability and consistency to be exploited in music exploration, retrieval and recommendation.

Similarity metrics that incorporate musical facets beyond timbre are still to be explored. As this project is providing semantic descriptors which cover tonality, instrumentation, rhythm, intensity, structure or complexity, new similarity metrics will be available soon and will make it possible to decide which of them contribute to the perceived similarity among songs. Other limitations on the current approaches to similarity lie on the facts that similarity judgements are not transitive (i.e. the statement "song A is more similar to B than to C" cannot be held under all circumstances) nor symmetric (symmetric (i.e. artist A may be similar to artist B, but artist B may not be similar to artist A in the case that A is follower of B). Another shortcoming lies in the fact that cultural and individual biases cannot be easily formalized. To conclude, web-based similarity has also text-mining inherent limitations (e.g. artist names are not unique, some of them have multiple meanings, artists that are not very popular have very few pages describing them…).

Most of the problems addressed in SIMAC could be alleviated or would change its focus if music files were enriched with metadata from their own origin (i.e. the recording studio). As this does not seem to be a priority by music technology manufacturers, we foresee a long life to our field, as digital music consumers are asking for the benefits of populating their music collections with a consistent and varied set of semantic descriptors.

Acknowledgements

The reported research has been funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). More than 15 collaborators have to be acknowledged as providing crucial input to the project, but space restrictions make impossible listing all of them. Additional information can be found at the project website http://www.semanticaudio.org.

References

[1] J.-J. Aucouturier and F. Pachet. "Music similarity measures: What's the use?" *Proc. of the 3rd ISMIR Conference*, pp. 157-163, (2002).

[2] J. J. Aucouturier, F. Pachet. "Improving timbre similarity: how high's the sky", *Journal of Negative Results in Speech and Audio Science*, 1 (1), (2004).

[3] J. P. Bello, M. Sandler. "Phase-based note onset detection for music signals". *Proc. of the IEEE ICASSP*, (2003).

[4] J. P. Bello, C. Duxbury, M. E. Davies, and M. B. Sandler. "On the use of phase and energy for musical onset detection in the complex domain". *IEEE Signal rocessing Letters*, 11(6), pp. 553-556, (2004).

[5] J. P. Bello, J. Pickens. "A robust mid-level representation for harmonic content in music signals". *Proc. of the 6th ISMIR Conference*, (2005).

[6] D. E. Berlyne. *Aesthetics and psychobiology*. Appleton-Century-Crofts, New York, (1971).

[7] N. Chetry, M. Davies, M. Sandler. "Musical instrument identification using LSF and k-means". *Proc. 118th AES*, (2005).

[8] G. C., Cupchik, , M. Rickert, J., Mendelson. "Similarity and preference judgment of musical stimuli", *Scandinavian Journal of Psychology*, 23, pp. 273-282, (1982).

[9] M. E. P. Davies, M. D. Plumbley. "Causal Tempo Tracking of Audio". *Proc. of the 5th International ISMIR Conference*, (2004).

[10] M. E. P. Davies, M. D. Plumbley. "Beat tracking with a two state model". *Proc. of IEEE ICASP*, (2005).

[11] S. Dixon, F. Gouyon, and G. Widmer. "Towards characterization of music via rhythmic patterns". *Proc. of the 5th ISMIR Conference*, pp 509-516 (2004).

[12] S. Dixon, E. Pampalk, G. Widmer. "Classification of dance music by periodicity patterns". *Proc. of the 4th ISMIR Conference*, (2003).

[13] E. Gómez, "Tonal description of polyphonic audio for music content processing". *INFORMS Journal of Computing*, 17(11), (to appear).

[14] E. Gómez, P. Herrera. "Estimating The Tonality Of Polyphonic Audio Files: Cognitive Versus Machine Learning Modelling Strategies". *Proc. of the 5th ISMIR Conference*, (2004).

[15] F. Gouyon, S. Dixon. "Dance music classification: A tempo-based approach". *Proc. of the 5th International ISMIR Conference*, (2004).

[16] F. Gouyon, S. Dixon. "A Review of Automatic Rhythm Description Systems". *Computer Music Journal*, 29 (1), pp. 34-54, (2005).

[17] F. Gouyon, S. Dixon, E. Pampalk, G. Widmer. "Evaluating rhythmic descriptors for musical genre classification". *Proc. of 25th Int. AES Conference*, (2004).

[18] D. FitzGerald, J. Paulus. "Unpitched percussion transcription", In A. Klapuri and M. Davy, "Automatic classification of pitched musical instrument sounds", In A. Klapuri, M. Davy (Eds.), *Signal processing methods for music transcription*, Springer, (2006).

[19] C. A. Harte, M. Sandler. "Automatic chord identification using a quantised chromagram". *Proc. of the 118th Convention. of the AES*, (2005).

[20] P. Herrera, A. Klapuri, M. Davy. "Automatic classification of pitched musical instrument sounds", In A. Klapuri, M. Davy (Eds.), *Signal processing methods for music transcription*, Springer, (2006).

[21] P. Herrera, V. Sandvold, F. Gouyon. "Percussion-related Semantic Descriptors of Music Audio Files". *Proc. of 25th International AES Conference*, (2004).

[22] A. Klapuri, M. Davy, *Signal processing methods for music transcription*, Springer, (2006).

[23] P. Knees, E. Pampalk, G. Widmer, G. "Artist Classification with Web-based Data". *Proc. of the 5th ISMIR Conference*, (2004).

[24] C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York, (1990).

[25] B. Logan, A. Salomon, "A music similarity function based on signal analysis", *Proc.of IEEE Int.Conf.on Multimedia and Expo*, (2001).

[26] Ong, B. Herrera, P. 2004. "Semantic Segmentation of Music Audio Contents" *Proc. of the ICMC*, (2005).

[27] E. Pampalk. *Islands of Music: Analysis, Organization and Visualization of Music Archives*. Master's thesis, (2004).

[28] E. Pampalk, S. Dixon, G. Widmer. "On the evaluation of perceptual similarity measures for music". *Proc. of the 6th DAFx Conference*, (2003).

[29] E. Pampalk, A. Flexer, G. Widmer. "Improvements of Audio-Based Music Similarity and Genre Classification", *Proc. of the 6th ISMIR Conference*, (2005).

[30] E. Pampalk, A. Flexer, G. Widmer. "Hierarchical organization and description of music collections at the artist level", *Proc. of the 9th ECDL*, (2005).

[31] E. Pampalk, T. Pohle, and G. Widmer. "Dynamic Playlist Generation Based on Skipping Behaviour" *Proc. of the 6th ISMIR Conference*, (2005).

[32] J. Pickens, J. P. Bello, G. Monti, T. Crawford, M. Dovey, M. Sandler, D. Byrd. "Polyphonic score retrieval using polyphonic audio queries: A harmonic modelling approach". *Proc. of the 3rd ISMIR Conference*, pp. 140-149, (2002).

[33] G. Salton, M. J. McGill. *Introduction to Modern Information Retrieval*, McGraw-Hill, NY, (1983).

[34] V. Sandvold, F. Gouyon, P. Herrera. "Drum sound classification in polyphonic audio recordings using localized sound models". *Proc. of the 5th ISMIR Conference*, (2004).

[35] V. Sandvold, P. Herrera. "Towards a Semantic Descriptor of Subjective Intensity in Music". *Proc. of the ICMC*, (2005).

[36] R. Shepard. "Circularity in judgments of relative pitch". *Journal of the Acoustical Society of America*, 35, pp. 2346-2353, (1964).

[37] S. Streich, P. Herrera. "Detrended Fluctuation Analysis of Music Signals: Danceability Estimation and further Semantic Characterization", *Proc. of 118th AES Convention*, (2005).

[38] F. Vignoli. "Digital Music Interaction concepts: a user study". *Proc. of the 5th ISMIR Conf.*, pp. 415-420, (2004).

[39] F. Vignoli, S. Pauws. "A Music Retrieval System Based on Music Similarity and its Evaluation", *Proc. of the 6th ISMIR Conference*, (2005).

[40] B. Whitman, S. Lawrence. "Inferring descriptions and similarity for music from community metadata". *Proc. of the ICMC*, pp.591-598, (2002).

[41] R. S. Wolpert. "Recognition of melody, harmonic accompaniment, and instrumentation: musicians vs. non-musicians", *Music Perception*, 8 (1), pp. 95-106, (1990).

[42] C. Yang. "MACSIS: A scalable acoustic index for content-based music retrieval". *Proc. of the 3rd ISMIR Conference*, (2002).

[43] K. Yoshii, M. Goto, H. G. Okuno. "Automatic Drum Sound Description for Real-World Music Using Template Adaptation and Matching Methods", *Proc. 5th ISMIR Conference*, (2004).

[44] A. Zils, F. Pachet. "Extracting Automatically the Perceived Intensity of Music Titles", *Proc. of DAFX-03*, (2003).