Bridging the Music Semantic Gap

Òscar Celma[1], Perfecto Herrera[1], and Xavier Serra[1]

Music Technology Group, Universitat Pompeu Fabra, Barcelona, SPAIN
http://mtg.upf.edu

Abstract. In this paper we present the music information plane and the different levels of information extraction that exist in the musical domain. Based on this approach we propose a way to overcome the existing semantic gap in the music field. Our approximation is twofold: we propose a set of music descriptors that can automatically be extracted from the audio signals, and a top-down approach that adds explicit and formal semantics to these annotations. These music 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. We belive that merging both approaches (bottom-up and top-down) can overcome the existing semantic gap in the musical domain.

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,

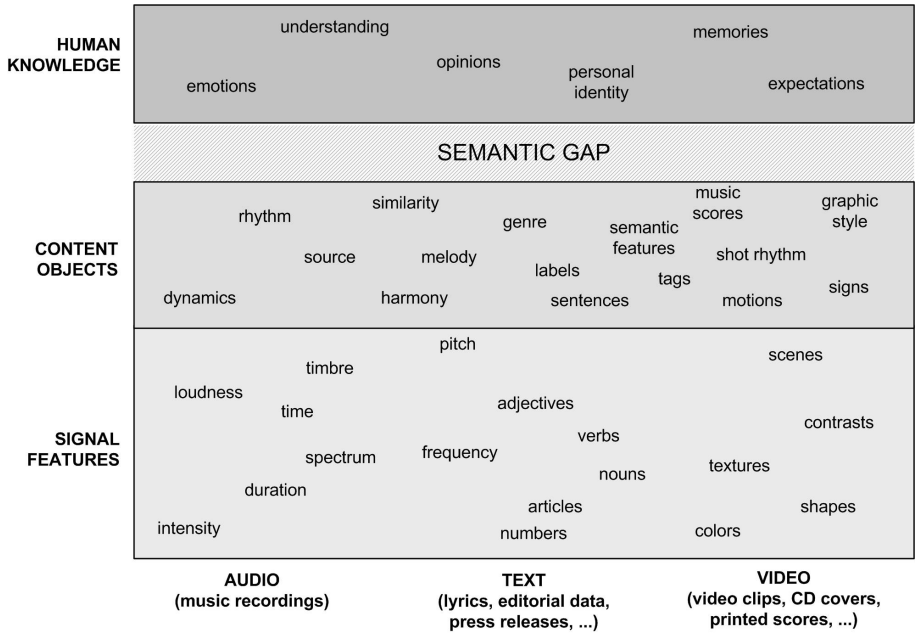

Fig. 1. The music information plane and its semantic gap

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.

This paper is structured as follows: section 2 will present the music information plane. This section will give a general overview of the different levels of information extraction that exist in the musical domain. Then, section 3 will present the work and developments made in a EU-project, named SIMAC[1] (Semantic Interaction with Music Audio Contents), to bridge the semantic gap and to enhance the music enjoyment experience. We will introduce several semantic descriptors of music contents, developed for different musical facets (rhythm, harmony, timbre, etc.). Before concluding, section 4 presents a multimodal approach to overcome the existing semantic gap in the musical domain. Finally, a discussion on future trends and open issues that deserve further research will conclude the paper.

2 The Music Information Plane

We describe the music information plane in two dimensions. One dimension takes into account the different media types that serve as input data. The other

[1] http://www.semanticaudio.org

dimension is the level of abstraction in the information extraction process of this data.

The input media types include data coming from: audio (music recordings), text (lyrics, editorial text, press releases, etc.) and image (video clips, CD covers, printed scores, etc.). On the other side, for each media type there are different levels of information extraction. The lowest level is located at the signal features. This level lay far away from what an end-user might find meaningful. Anyway, it is the basis that allow to describe the content and to produce more elaborated descriptions of the media objects. This level includes basic audio features, such as: pitch, frequency, timbre, etc., or basic natural language processing for the text media. At the mid-level (the content objects level), the information extraction process and the elements described are closer to the end-user. This level includes description of musical concepts (e.g. rhythm, harmony, melody), or named entity recognition for text information. Finally, the higher-level, the Human Knowledge, includes information tightly related with the human beings. Figure 1 depicts the music information plane.

Next section (section 3) describes some features that can be automatically extracted from the audio signal and that constitutes the content objects — located at the mid-level of abstraction. Then, section 4 gives a vision of the multimodal approach and presents a music ontology that describes some of the features automatically extracted from the audio. We belive that merging both approaches (bottom-up and top-down) can overcome the existing semantic gap in the musical domain.

3 Semantic Description of Music Content Objects

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].

Alternatively, our 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 .

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 our contributions to all these aspects.

3.1 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 . 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. 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 , or that are based on predicting the phase and energy of signal components in the complex domain , 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 , 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 , 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 (see [3]) 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 , 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 Music Information Retrieval (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 [4].

3.2 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 [6].

Harmonic based retrieval has not been extensively explored before. A successful approach at identifying harmonic similarities between audio and symbolic data was presented in . 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 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.

We have proposed a state-of-the-art approach to tonality estimation by correlating chroma distributions with key profiles derived from music cognition studies [7]. 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 . 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 [8]. This methodology has already shown great promise for both chord recognition and structural segmentation.

3.3 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 [2].

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 . 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 . Although they can be computed after some source separation , reasonable approximations can be achieved using simpler sound classification approaches that do not attempt separation , .

Additionally, our research in the area of instrumentation has contributed to the current state of the art in instrument identification of mono-instrumental music , 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.

3.4 Intensity

Subjective intensity, or the sensation of energeticness we get from music, is a concept commonly and easily 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 [9], 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. 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.

3.5 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 . 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%.

Next example shows the description of an automatically annotated audio file, based on some of the descriptors presented in this section. There are descriptors that have an enumerated value as output —usually a label—, whereas other descriptors' values are numeric (e.g. floats or integers).

```
<?xml version='1.0' encoding='UTF-8'?>
<DescriptorsPool>
  <ScopePool name='Song' size='1'>
    <!-- Rhythm descriptors -->
    <AttributePool name='Tempo'>62</AttributePool>
    <AttributePool name='Measure'>
      <Enumerated>Binary</Enumerated>
    </AttributePool>
    <!-- Tonality descriptors -->
    <AttributePool name='Key'>
      <Enumerated>B</Enumerated>
    </AttributePool>
    <AttributePool name='Mode'>
      <Enumerated>Minor</Enumerated>
    </AttributePool>
    <AttributePool name='Key-Strength'>0.8412</AttributePool>
    <!-- Intensity descriptor -->
    <AttributePool name='Intensity'>
      <Enumerated>Soft</Enumerated>
    </AttributePool>
    <AttributePool name='Danceability'>
      <Enumerated>Few</Enumerated>
    </AttributePool>
    ...
  </ScopePool>
```

</DescriptorsPool>

Listing 1.1. Example of an automatically annotated audio file.

4 Pushing the current limits

In section 3 we have introduced some mid-level music descriptors, but we still lack of formal semantics to describe the audio context (note that the annotation description file presented in the previous section is just a plain XML file).

The main problem, then, is how to push automatic media-based descriptions up to the human understanding. We belive that this process can not be achieved if we focus in only one direction (say, a bottom-up approach). For many years Signal Processing has been the main discipline used to automatically generate music descriptors. More recently Statistical Modeling, Machine Learning, Music Theory and Web Mining technologies (to name a few) have also been used to push up the semantic level of music descriptors. Anyway, we belive that the current approaches to automatic music description, which are mainly bottom-up, will not allow us to bridge the semantic gap. Thus, we need an important shift in our approach. The music description problem will not be solved by just focusing on the audio signals; a Multimodal Processing approach is needed. We also need top-down approaches based on Ontologies, Reasoning Rules, Music Cognition, etc. Figure 2 shows how the multimodal approach can help to overcome the current semantic gap in the music field.

Regarding ontologies and basic reasoning rules; in [5] we propose a general multimedia ontology based on MPEG-7, described in OWL[2], that allows to formally describe the automatic annotations from the audio (and, obviously, more general descriptions of multimedia assets). Table 1 shows some mappings from SIMAC ontology to the macro MPEG-7 OWL ontology. Once all the multimedia metadata —not only automatic annotations from audio files, but editorial and cultural data— has been integrated in a common framework (that is, in our case, in the MPEG-7 OWL ontology) we can benefit from the, now, explicit semantics. Based on this framework, we foresee some usages of the ontology to help the process of automatic annotation of music. The following two sections present some ideas.

4.1 Entity integration and Duplicate detection

A typical scenario that shows the usefulness of the duplicate detection could be the following: an Internet crawler is looking for audio data and it downloads all the files. Getting editorial and related information for these audio files can be achieved reading the information stored in the ID3 tag. Unfortunately, sometimes there is no basic editorial information like the title of the track, or the performer. However, content-based low-level descriptors can be computed for these files,

[2] http://www.w3.org/2004/OWL/

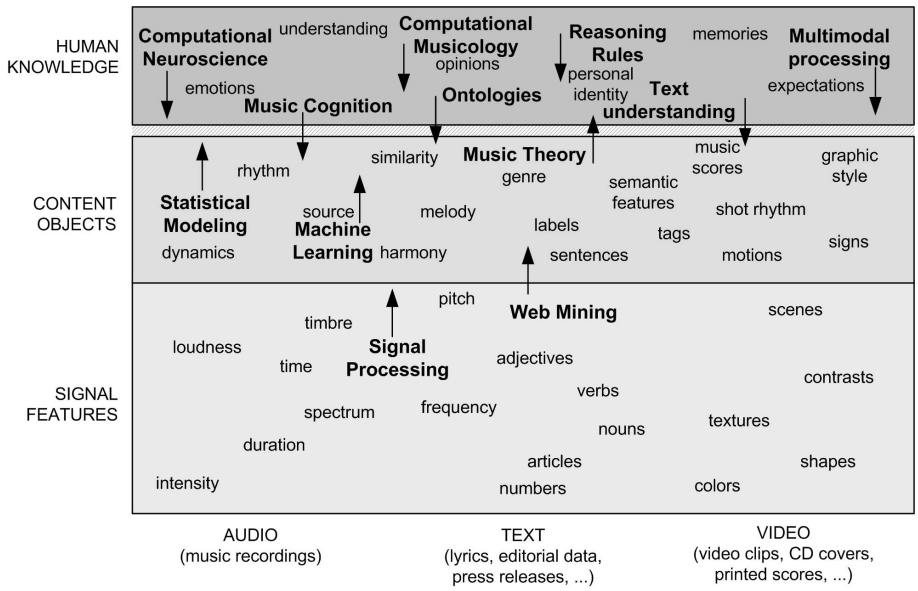

Fig. 2. Multimodal approach the bridge the semantic gap in the music information plane

including its MusicBrainz fingerprint, a string that uniquely identifies each audio file based on its content, named PUID. The next example shows an RDF/N3 description for a track with the calculated tempo and fingerprint:

```
<http://example.org/track#1> a simac:Track;
 simac:tempo "74";
 musicbrainz:puid "3c41bc1-4fdc-4ccd-a471-243a0596518f".
```

On the other hand, MusicBrainz database has the editorial metadata —as well as the fingerprint already calculated— for more than 3 millions of tracks. For example, the RDF description of the song "Blowin' in the wind", composed by Bob Dylan, is:

```
<http://example.org/track#2> a musicbrainz:Track;
 dc:title "Blowin' in the wind";
 dc:author [musicbrainz:sortName "Bob Dylan"];
 musicbrainz:puid "e3c41bc1-4fdc-4ccd-a471-243a0596518f".
```

A closer look to both examples should highlight that the two resources are sharing the same MusicBrainz's fingerprint. Therefore, it is clear that, using a simple rule, one can assert that both audio files are actually the same file, that is to say the same instance in terms of OWL, $owl : sameIndividualAs$. Figure 3 shows a possible rule that detects duplications of individuals.

$simac : Artist \subseteq mpeg7 : CreatorType$
$simac : name \equiv mpeg7 : GivenName$
$simac : Track \subseteq mpeg7 : AudioSegmentType$
$simac : title \subseteq mpeg7 : Title$
$simac : duration \equiv mpeg7 : MediaDuration$
$simac : Descriptor \equiv mpeg7 : AudioDSType$
$simac : mode \equiv mpeg7 : Scale$
$simac : key \equiv mpeg7 : Key$
$simac : tempo \equiv mpeg7 : Beat$
$simac : meter \equiv mpeg7 : Meter$

Table 1. Simac music ontology to MPEG-7 OWL ontology mappings.

$mpeg7 : AudioType(track1) \wedge mpeg7 : AudioType(track2) \wedge musicbrainz : puid(track_1, puid_1) \wedge musicbrainz : puid(track_2, puid_2) \wedge (puid_1 = puid_2)$
$\Rightarrow$
$owl : sameIndividualAs(track_1, track_2)$

Fig. 3. Simple rule to assert that two individuals ($track_1$ and $track_2$) are the same.

From now on, we have merged the metadata from both sources and we have deduced that the metadata related with both tracks is, actually, referred to the same track. This data integration (at the instance level) is very powerful as it can combine and merge context-based data (editorial, cultural, etc.) with content-based data (extracted from the audio itself).

4.2 Propagation of annotations

Another interesting usage is the propagation of annotations. That is, when we have information from one source (i.e an audio file) and we want to propagate some of the annotations to another source.

Given a song ($track_1$) with a set of high-level annotations (either supervised by a musicologist, or gathered through a process of web mining, for instance), and a song ($track_2$) that lacks some of these high-level descriptions, then we can apply a set of rules that can propagate part of the annotations of $track_1$ to $track_2$. To decide whether we can propagate this information, we need an extra component in the system that tell us how similar —based on automatically extracted audio features— are songs $track_1$ and $track_2$. If they are close together, then it makes sense to propagate some annotations from one song to another. Figure 4 exemplifies this case.

This process could be supervised by an expert. Thus, the process of annotating would be, now, to check whether this propagated annotations make sense or not. This can clearly improve the time spent for an expert to annotate a given song.

$mpeg7 : AudioType(track_1) \wedge mpeg7 : AudioType(track_2) \wedge similars(track_1, track_2))$
$\Rightarrow$
$propagateAnnotations(track_1, track_2)$

Fig. 4. Simple rule to propagate annotations from one song ($track_1$) to another ($track_2$).

5 Conclusions

We have presented the music information plane and the existing semantic gap. To overcome this gap we have presented a set of mid-level music descriptors that allow to describe the music audio files with a reasonable level of detail. Moreover, we have proposed a mixing approach (both bottom-up and top-down) that we belive it can help to reduce the existing semantic gap in the music field.

Most of the problems addressed in the SIMAC project 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 for 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.

Moreover, we are now viewing an explosion of the practical applications coming out from the MIR research: Music Identification systems, Music Recommenders, Playlist Generators, Music Search Engines, Music Discovery and Personalization systems, and this is just the beginning. If we succeed in bridging the semantic gap there will be no limit to the applications that could be developed. At this stage, we might be closer in bridging the semantic gap in music than in any other multimedia knowledge domain. Music was a key factor in taking Internet from its text-centered origins to being a complete multimedia environment. Music might do the same for the Semantic Web.

6 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? In *Proceedings of the 3rd ISMIR Conference*, pages 157–163, Paris, France, 2002.
2. J.-J. Aucouturier and F. Pachet. Improving timbre similarity: how high's the sky. In *Journal of Negative Results in Speech and Audio Science*, 2004.

3. S. Dixon. F. Gouyon. Dance music classification: A tempo-based approach. *Proceedings of the 5th International ISMIR Conference*, 2004.
4. S. Dixon. F. Gouyon. A review of automatic rhythm description systems. *Computer Music Journal*, 29:34–54, 2005.
5. R. Garcia and O. Celma. Semantic integration and retrieval of multimedia metadata. In *Proceedings of 4rd International Semantic Web Conference. Knowledge Markup and Semantic Annotation Workshop*, Galway, Ireland, 2005.
6. E. Gomez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing, Special Cluster on Computation in Music*, 18(3), 2006.
7. P. Herrera. Gomez, E. Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. *Proceedings of the 5th ISMIR Conference, (2004).*, 2004.
8. J. Pickens P. Bello. A robust mid-level representation for harmonic content in music signals. In *Proceedings of the 6th ISMIR Conference*, London, UK, 2005.
9. A. Zils and F. Pachet. Extracting automatically the perceived intensity of music titles. *Proceedings of DAFX-03*, 2003.