A multimodal approach to bridge the Music Semantic Gap

Òscar Celma, Perfecto Herrera, and Xavier Serra

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 multimodal approach that adds explicit and formal semantics to these annotations. We believe that merging both approaches (bottom-up and top-down) can overcome the existing semantic gap in the musical domain.**

Index Terms— **Semantic Gap, Music Information Retrieval, Multimodal processing**

I. 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 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 than yours 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.

Òscar Celma, Perfecto Herrera, and Xavier Serra are with the Music Technology Group, Universitat Pompeu Fabra, Barcelona, SPAIN

II. The Music Information Plane

Due to the inherent complexity to describe multimedia objects, a layered approach with different levels of granularity is needed when designing an ontology for a particular domain. Depending on the requirements, one might choose the appropriate level of abstraction. In the multimedia field and, in concrete, in the music field we foresee three levels of abstraction: low-level (physical and basic *semantic*) features, mid-level semantic features, and human understanding and reasoning. The first level includes physical features of the objects, such as the sampling rate of an audio file, as well as some basic features like the spectral centroid of an audio frame, or even the predominant chord in a sequential list of frames. A higher-level of abstraction aims at describing concepts such as a guitar solo, or tonality information (e.g key and mode) of a music title. Finally, the reasoning level uses inference methods and semantic rules to retrieve, for instance, several audio files with similar guitar solos over the same key.

Similarly, 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 dimension is the level of abstraction in the information extraction process of this data (see Fig.1). 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: energy, frequency, mel frequency cepstral coefficients, 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 when interacting with music knowledge.

III. Pushing the current limits

The main problem, then, is how to push automatic media-based descriptions up to the human understanding. We believe that this process can not be achieved if we focus in only one direction (say, a bottom-up approach). For many years

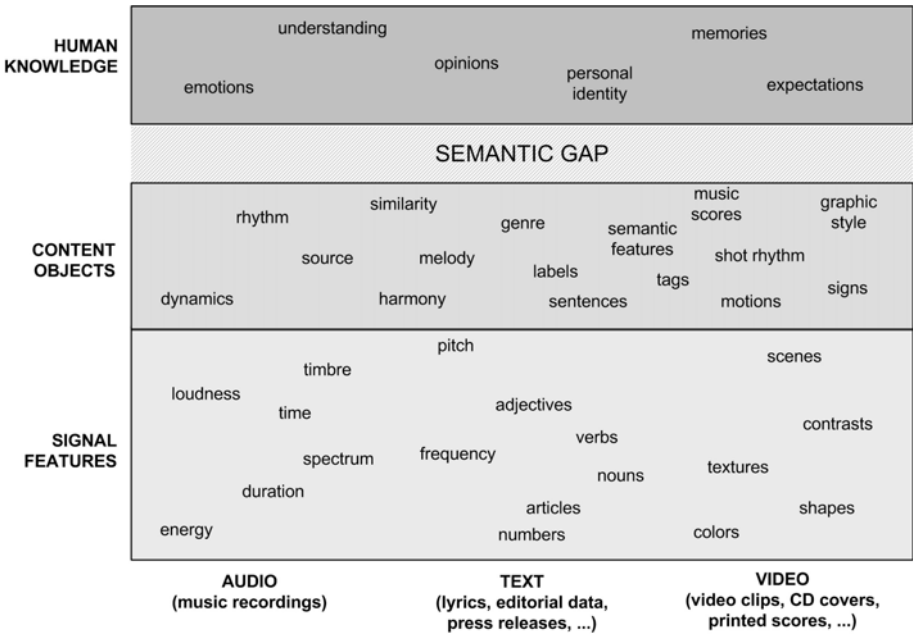

Fig. 1. The music information plane and its semantic gap between content objects and human understanding.

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 believe 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, or even Computational Neuroscience and Computational Musicology.

Regarding ontologies and basic reasoning rules, in [1] we have proposed a general multimedia ontology based on MPEG-7, described in OWL[1] language, that allows to formally describe the automatic annotations from the audio (and, obviously, more general descriptions of multimedia assets). The approach contributes a complete and automatic mapping of the whole MPEG-7 standard to OWL. It is based on a XML Schema to OWL mapping that tries to be as transparent as possible. The previous mapping is complemented with an XML metadata instances to RDF mapping that completes a tool set to transfer metadata from the XML to the Semantic Web domain.

Once all the multimedia metadata —not only automatic acoustic annotations from audio files, but editorial and cultural data too [2]— 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, such as propagation of music annotations based on audio similarity or detect inconsistencies in editorial metadata.

IV. Conclusions

We have presented the music information plane and the existing semantic gap that occurs between content object level and human understanding. Thus, we foresee that a mixing approach (both bottom-up and top-down) can help to reduce the existing semantic gap in the music field.

Moreover, we are now viewing an explosion of the practical applications coming out from the Music Information Retrieval research: Music Identification systems, Music Recommenders, Playlist Generators, Music Search Engines, Music Discovery and Personalization systems, and this is just the beginning[2]. 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.

Acknowledgment

The reported research has been funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). Additional information can be found at the project website*http://www.semanticaudio.org*.

References

[1] Garcia, R. and Celma, O., *Semantic Integration and Retrieval of Multimedia Metadata*, 2005, Proceedings of 4rd International Semantic Web Conference, Galway, Ireland.

[2] Pachet, F., *Knowledge Management and Musical Metadata*, 2005, Encyclopedia of Knowledge Management.

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

[2] A detailed list of MIR systems are available at http://mirsystems.info/