

Extraction of syntactic cues from polyphonic music

Research proposal of Benoit Catteau (ELIS-UGent)

Introduction

At the time of the Graduate School I will be engaged in a project, called GOA-SEMA. The main objective of this project is to develop a perception-based instrumental musical content analysis theory, and to validate that theory in representative application domains such as audio mining, interactive multimedia and brain research. The project promoters are Marc Leman (IPEM-UGent) and Jean-Pierre Martens (ELIS-UGent). The official start of the 6 years project was January 1, 2004, but in October I will be among the first researchers to start working on the project.

Planned research

The objective of the work package I will be involved in is : development of new perceptually motivated tools for the extraction, directly from the audio signal, of objective/syntactical cues that can act as an intermediate representation between the audio signal (a physical object) and the subjective/affective qualities evoked by this signal.

At present, there is no agreement on which structural audio features are most appropriate for computational modeling, nor is it clear how to extract these features from the audio signal. Thus, there is a great need for robust and powerful statistical signal processing techniques that can be applied in an elegant way to polyphonic music analysis. In the past, ELIS and IPEM have adopted a rather unique perception-based approach to music analysis. It is the intention to draw on this past experience and to extend the available modeling capacity significantly.

By taking known constraints of the human auditory system into account at all levels of the audio signal analysis, one should be able to generate low-level feature representations of the acoustic signal that can - better than any features emerging from a non-perceptually motivated analysis - be mapped to the psychophysical sensations evoked by this signal. There is experimental evidence from the speech processing domain for instance that the perceptually motivated MFCCs (Mel-Frequency Cepstral Coefficients) give a benefit over non-perceptually motivated parametric representations of the speech signal: MFCCs have now become the dominant speech representation in modern speech recognition systems.

At the medium level, we believe that by properly analyzing the temporal modulations (with frequencies ranging from 10 to 300 Hz) observed in the peripheral auditory model outputs, will emerge in features that can contribute to the computation of subjective qualities related to loudness fluctuations, roughness, pitch (virtual as well as spectral pitch), onsets, consonance, etc.

Once our auditory model will be completed, its outputs will be analyzed at a larger time scale so as to describe the different structures in music. To that end we will have to develop new and reliable methods for (i) the detection of events (e.g. beats, pitches, melodies, phrase boundaries), (ii) the description of gestures related to melodic patterns, harmonic progression, tonality or key patterns, rhythmic patterns, (iii) the location and identification of voices (musical instruments, human voices), and (iv) the specification of particular information theoretic measures such as complexity, redundancy etc. The extracted features will be extremely important in Music Information Retrieval (MIR) because they reveal the patterns that can be matched to similar patterns of other musical pieces. They can also be used to classify musical pieces according to genre (folk, rock, classical music, etc.), sub-genre (symphony, concerto, opera, musical), performance style (jazzy, classical), etc. Most algorithms that were conceived in the past rely on deterministic signal processing techniques (e.g. the use of filters, spectral distance measures, explicit segmentation-before-classification algorithms, etc). However, recently more and more research is directed towards the application of data-driven statistical modeling techniques as they have been applied before in e.g. speech processing.

There is already a lot of experience with statistical models (e.g. Hidden Markov Models, Neural Networks, etc.) in the pattern recognition and speech recognition domains, but in the polyphonic music recognition domain, they are far less developed. It is our aim to investigate what the fundamental differences between speech and polyphonic music are, and how to translate these into appropriate modifications of the present statistical modeling strategies.