

Introduction to the IPEM Toolbox for Perception-based Music Analysis

Marc Leman, Micheline Lesaffre, Koen Tanghe
IPEM – Dept. Of Musicology, Ghent University, Belgium
<http://www.ipem.rug.ac.be/Toolbox>

Abstract

The motivation for developing a toolbox for perception-based music analysis starts from the observation that much of the world music production (seen over different cultures and time periods) has no score representation, and that if a score representation is available, it still represents but a small percentage of what musical communication really is about. A sonological analysis of musical sounds may provide steps towards a better understanding of the components that contribute to musical information processing but is also insufficient in view of how humans deal with musical signals. Technology has revolutionized our conception of what kind of music research will be possible in the 21st Century. But it is the task of musicologists, in collaboration with other scientists, such as engineers, psychologists, and brain scientists, to make these dreams more concrete. Our basic motivation for developing this toolbox is a pragmatic one: if the musicology aims at understanding music as a social and cultural phenomenon embedded in the physical world and mediated through human faculties of perception, cognition, and processing of expressive communication, then new tools must be developed that allow a fully integrated approach. This paper introduces the main concepts and a state-of-the-art of the modules of the IPEM toolbox for perception-based music analysis. This toolbox represents a bottom-up approach to musical description taking human perception as the basis for musical feature extraction and higher-level conceptualization and representation. The IPEM toolbox provides a collection of MATLAB functions that allow dealing with different aspects of feature extraction as well as a global concept concerning perception-based music analysis. First we present the global internal representation framework. Then we sketch what we have in mind as a global picture and present the modules currently involved. Finally we describe some applications of the toolbox.

1. Introduction

Digital technology and computer modeling of perception and cognition have been increasingly influential on recent developments in musicology. The emergence of digital music on the Internet imposes music retrieval systems based on content extraction and description of acoustical and perceptual features of any possible sound. This research topic opens completely new approaches for musicologists to deal with. The discipline “music analysis” generally relies on scores and on sonological analysis. There is a thorough need to develop tools that go beyond the analysis limited to music that can be represented by a score.

In the past decade several tools for auditory model based music analysis have been proposed and developed:

- Malcolm Slaney's (1993) [6] auditory toolbox extends Matlab's capabilities by providing a number of auditory models. It contains Matlab functions to implement different kinds of auditory models. This toolbox includes Lyon's Passive Long Wave Cochlear Model, Patterson-Holdsworth ERB Filter bank with Meddis Hair cell, Seneff's Auditory Model, MFCC (Mel-scale frequency cepstral coefficients from the ASR world), Spectrogram, Correlogram generation and pitch modeling, simple vowel synthesis.
- Roy Patterson's group [8] at the MRC-APU in Cambridge, UK, has made available a version of their Auditory Image Model (AIM) of peripheral auditory processing that is written in C. It allows different models of auditory perception to be linked together. The AIM is a time-domain model of auditory processing to simulate the auditory images produced by complex sounds. This model simulates how sound waves cause basilar membrane motion in the human cochlea. Subsequently, the results are used as input for a next stage in which the conversion, by hair cells, of motion into a neural activity pattern in the auditory nerve is simulated.
- John Culling [5] has made available PIPEWAVE software for psychoacoustics, a suite of UNIX programs that communicate via UNIX pipes. The programs can perform digital signal processing and auditory modeling. Some of the programs act as generators of waveforms, others perform some transformation on an input waveform and others store or plot the resulting waveforms. The waveforms are passed from one program into the next by pipes, which can chain a number of different programs together. Culling's research is mainly related to speech recognition. He has been investigating "the cocktail party effect" - the ability of a listener to attend selectively to a single voice among many interfering voices.
- The Development System for Auditory Modeling (DSAM) (formerly LUTEar) [7] from the Centre for the Neural Basis of Hearing at Essex is a computational platform and set of coding conventions, which supports a modular approach to auditory system modeling. The LUTEar Core Routines Library, version 2.0.9 (209) (1997), was originally based upon a unified re-interpretation in ANSI C of code produced by workers in the Speech and Hearing laboratory: M. J. Hewitt, T. Shackleton & R. Meddis. The system is written in ANSI-C and works on a wide range of operating systems. LUTEar is a collection of modules, utilities and complete programs made

available to the auditory community for the advancement of the study of signal processing in the auditory system.

The IPEM toolbox aims to provide a foundation for new music analysis. We start from music as audio-signal regardless of its provenances and take human perception as the basis for musical feature extraction at different levels. This means that all kinds of music can be observed including electroacoustic music as well as ethnic music.

The IPEM toolbox uses computer modeling of the human auditory system to provide a foundation of music analysis in terms of human perception. A set of MATLAB functions is provided. In the conceptual framework basic toolbox functions are presented as modules. The concept “module” refers to a well-defined perception-based processing block that may involve just one or several toolbox functions. The toolbox contains feature extraction tools, which apply to different musical parameters such as pitch, roughness, rhythm, and energy.

Modules can be accessed at different levels going out from four different viewpoints. We provide an introductory description, a functional logical description, a signal processing description and an implementation description. In describing the modules involved we first give the *introductory description*. This is a simple verbalization of what a module does. It situates the module in the global concept of the toolbox. The second level concerns the *functional-logical description*, which resembles the way the MATLAB functions are to be used. Then follows the *signal processing description*, which implies a mathematical description of the modules in terms of a functional equivalence model of physiological processes. Finally comes the *implementation description*, which concerns the way in which the MATLAB functions are implemented.

In this paper the modules are presented only by an introductory description.

2. Global Internal Representational Framework

The modules in the toolbox are related to auditory information processing, which means that we start from sound and transform sound into auditory images or inferences. The notions of image, image transformation process, inference and decision define our global internal representational framework.

- ***Auditory images*** reflect features of the sound as internal representations. The content of an auditory image represents features related to the musical signal. The input sound is assumed to be carried by an array of neurons. From the point of view of computer modeling, an image is an ordered array of numbers whose values represent neural activation. We conceived of neural activation in terms of firing rate-code, that is, the probability of neuronal spiking during a certain time interval. A distinction is made between different types of auditory images.
- ***Image transformation processes*** transform sound into images and images into other images. Different memory systems carry different kinds of images. In association with short-term memories and long-term memories we therefore distinguish short-term images and long-term images. Images and image transformation processes are compared with human physiology.

- **Inferences** provide derived information that can be compared to human behavioral responses.
- **Inference processes** compare images, inspect images or extract features from images

This dual validation model is related to human physiology (images and associated processes) and to human behavioral responses (inferences). Figure 1 gives an overview of this internal representational framework.

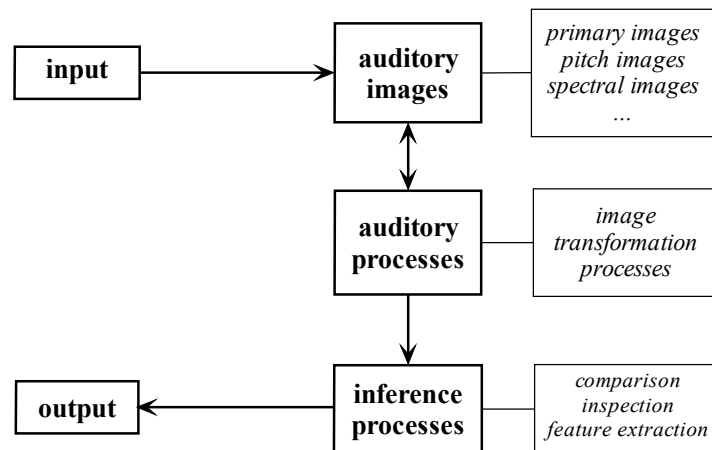


Fig 1: Overview of the internal representational framework

3. Modules

3.1. Global picture

We consider a module as a well-defined process leading to images and (possible) an inference. For now seven modules are included in the toolbox, the aim is to enlarge this amount by developing new modules and inviting users to add their own modules. Figure 2 gives an overview of what we have in mind as a global picture. This is just one possible way to visualize the modules. In this global picture the transformation modules are horizontally organized following the distinction between different description levels. It is a bottom up approach going from sensory, over perceptive to cognitive processing. Vertically the modules are organized from texture (top) to structure (bottom).

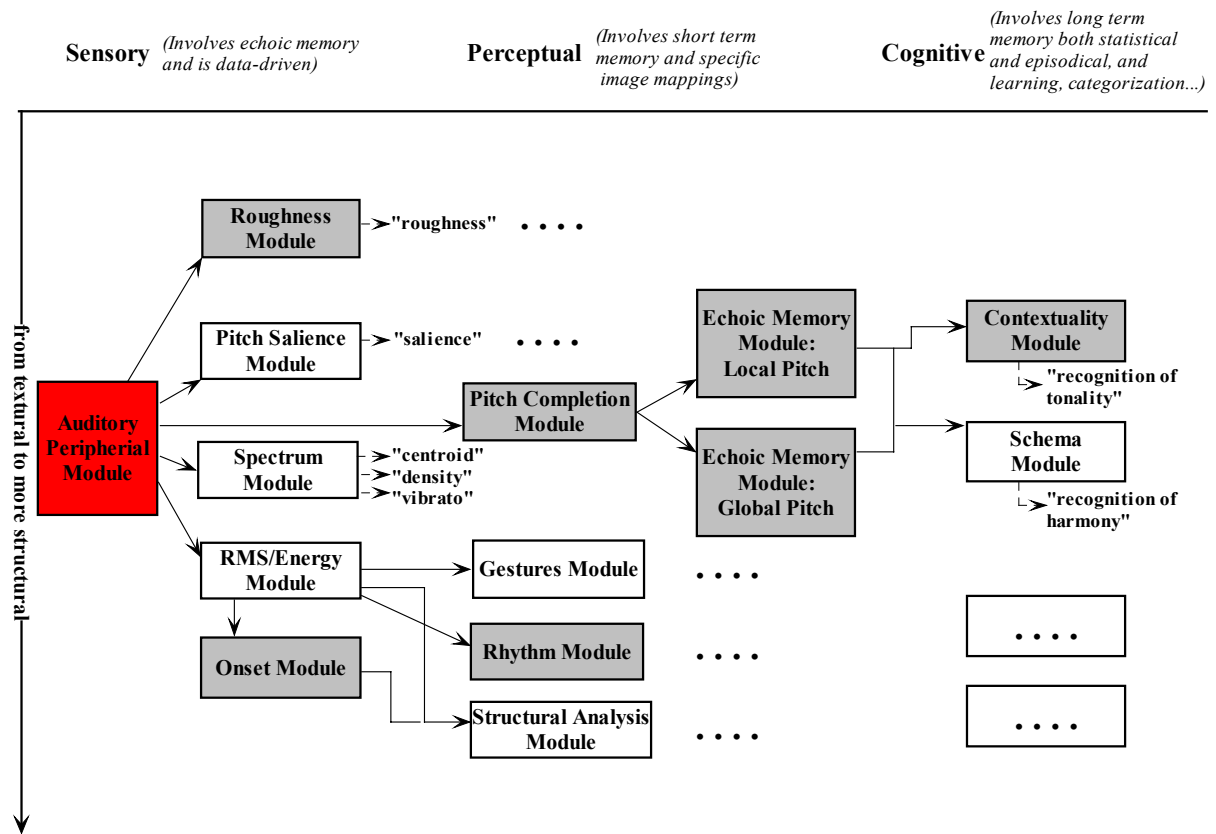


Fig 2: Global picture of the IPeM toolbox modules. The fill color indicates the modules integrated for now; the other modules are under development and will be added in due course

3.1.1. The sensory processing description level

The *Auditory Peripheral Module*, based on a model of the auditory periphery, is to be considered as the basis module in our concept in which musical signals are sampled and then processed by an auditory model into auditory primary images.

Sensory modules involve echoic memory and are based on stimulus driven processing. We assume that they are located in the periphery of the auditory system. At this level we include modules for *Roughness*, *Pitch Saliency*¹*, *Spectrum**, *Energy** and *Onset*.

3.1.2. The perceptual processing description level

Perception modules involve short-term memory and specific auditory image mapping from the temporal domain into the spatial domain. We assume that they are located in the brain stem. At the perceptive level we distinguish modules for *Pitch*, *Gestures**, *Rhythm* and *Structural Analysis**.

¹ The modules marked with an * will be included in the next version of the IPeM toolbox

3.1.3. The cognitive processing description level

Cognitive modules involve long-term memory, learning and categorization. We assume that they are located in the cortex. We consider modules for *Contextuality* and *Schema**.

3.2. Modules at the sensory level

3.2.1. Auditory Peripheral Module (APM)

The APM takes a sound as input and gives as output the *auditory primary image or auditory nerve image*. The musical signal is decomposed in different sub bands (40) and represented as neural patterns of neural firing rate-codes in a number of auditory nerves. The model of auditory periphery that we use is an adapted version of the model by Van Immerseel en Martens (1992) [4]. The processing from sound into auditory primary image involves three processing stages:

- Simulation of the filtering of the outer and middle ear
- Simulation of basilar membrane resonance in the inner ear
- Simulation of a hair cell model

Image Transformation Process

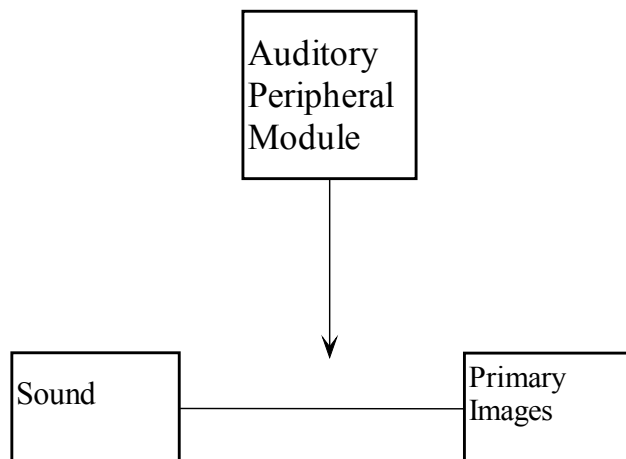


Fig 3: The auditory peripheral image transformation module

Figure 4 shows the auditory nerve image obtained as the result of processing a short excerpt (first four measures) of Schumann's *Kuriose Geschichte*.²

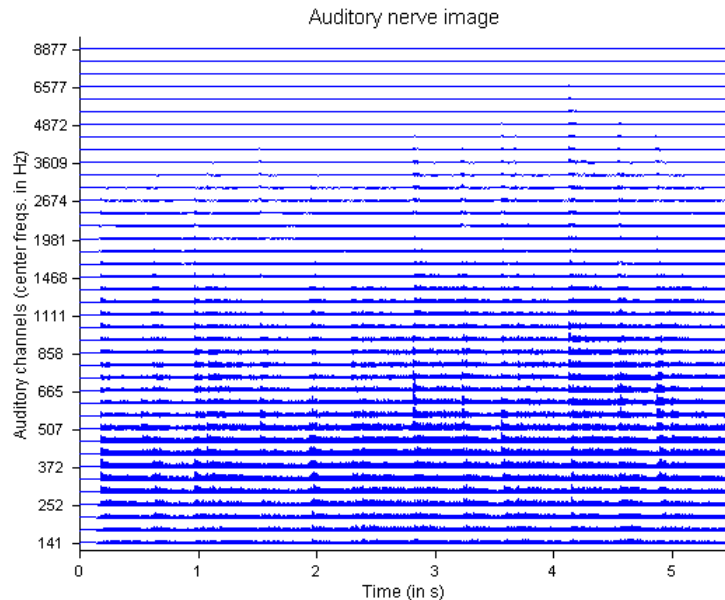


Fig 4: Auditory Nerve Image (ANI)

3.2.2. Roughness Module (RM)

The RM calculates the roughness or sensory dissonance of a sound. The module takes sound as input and produces *roughness estimation*. Roughness is considered to be a sensory process highly related to texture perception. The estimation should be considered an inference, but the module offers more than just an inference. The calculation method of this module is based on Leman's Synchronization Index Model (2000) [3], where roughness is defined as the energy provided by the neuronal synchronisation to relevant beating frequencies in the auditory channels. This model is based on phase locking to frequencies that are present in the neural patterns. It assumes that neurons somehow extract the energy of the beating frequencies and form internal images on which the inference is based. The concept of synchronization index refers to the amount of neural activation that is synchronized to the timing of the amplitudes of the beating frequencies in the stimulus.

² Robert Schumann "Kinderszenen-Kreisleriana",
played by Martha Argerich (Deutsche Grammophon 410 653-2, 1984)

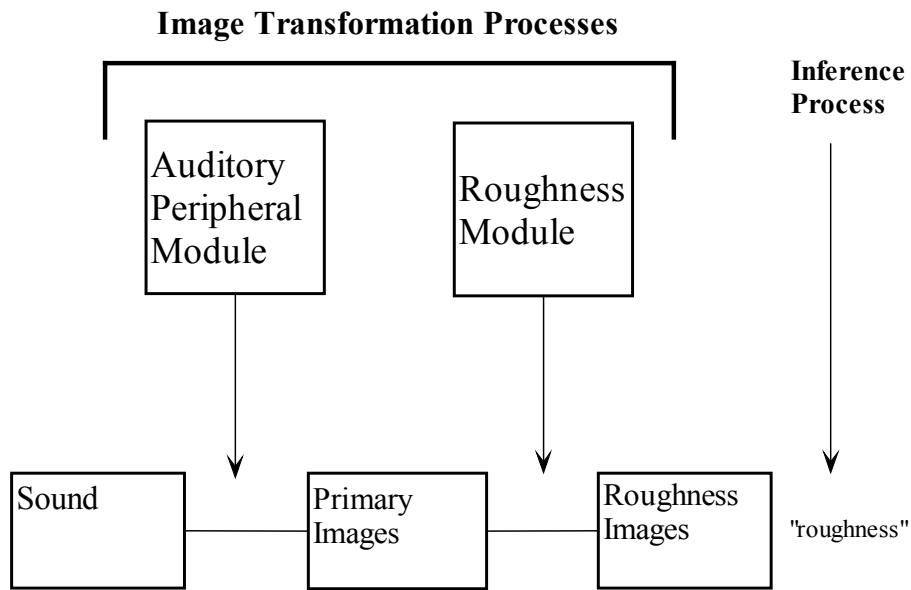


Fig 5: Processes from sound to roughness using the APM and RM

Figure 6 shows the results of calculating roughness of the excerpt from Schumann's *Kuriose Geschichte*

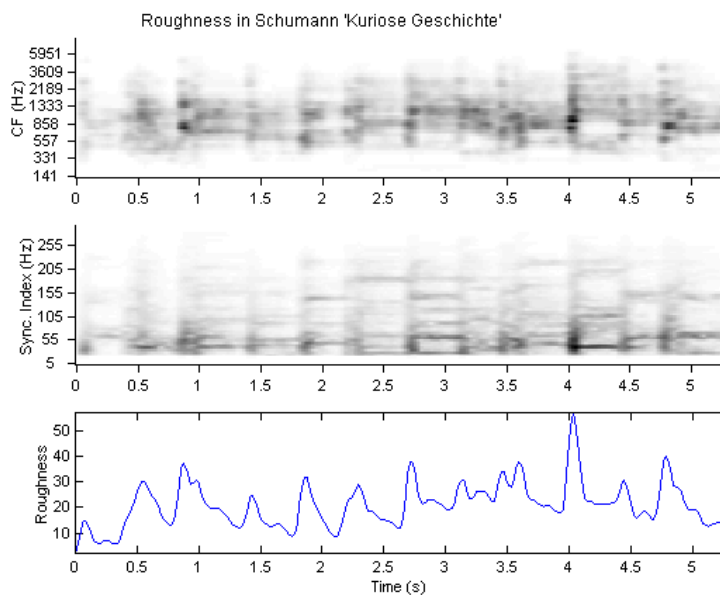


Fig 6: The top panel show the energy as distributed over the auditory channels, the middle panel shows the energy as distributed over the beating frequencies, the lower panel shows the roughness

3.2.3. Onsets Module (OM)

The OM finds the onsets of sound events. It takes sound as input and produces the moments where a new note, chord or sound event is triggered in the musical signal. The onsets module analyzes the energy in the different channels, extracts the relevant peaks and combines the results for each channel to an overall *onset estimation*.

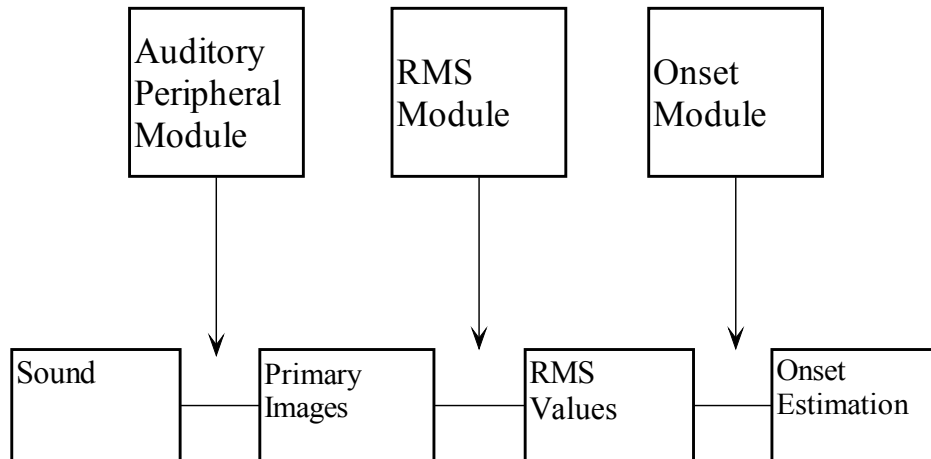


Fig 7: Modules involved for onset detection

Figure 8 shows the extracted onsets from the excerpt of Schumann's *Kuriose Geschichte*

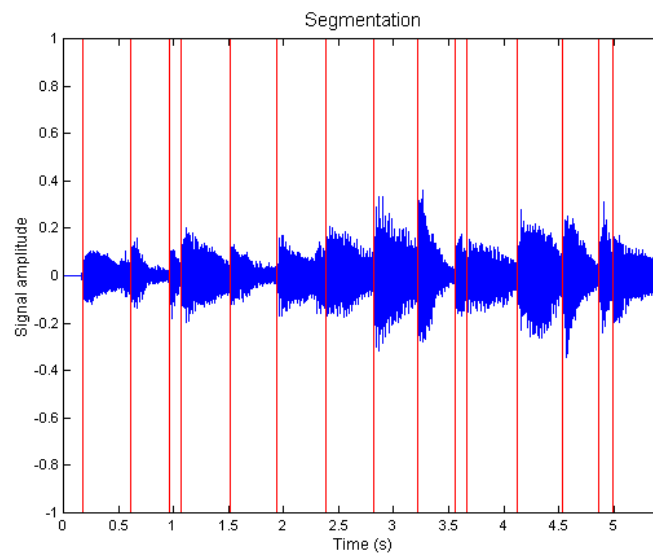


Fig 8: Segmentation of the original sound signal using the onsets module

3.3.Modules at the perceptual level

3.3.1.Pitch Completion Module (PCM)

The PCM takes the neural rate code of the auditory nerve image as an input and performs a periodicity analysis of primary images. The output is a *pitch image* or completion image. The focus of our attention for the periodicity analysis is between 80 Hz and 1250 Hz. The lower limit of 80 Hz accounts for the fact that for smaller frequencies, the sensation of pitch becomes more a sensation of textural properties. This is a shift of perceptual categorization that is taken into account in our modeling. Indeed, our model of roughness took into account a range of frequencies between 5 and 300 Hz but the focus was on frequencies between 50 and 70 Hz

The higher limit of 1250 Hz is related to the limits of neural synchronization. Beyond about 1250 Hz, the neurons are no longer able to follow the exact period of the signal very accurately, and periodicity pitch becomes unreliable.

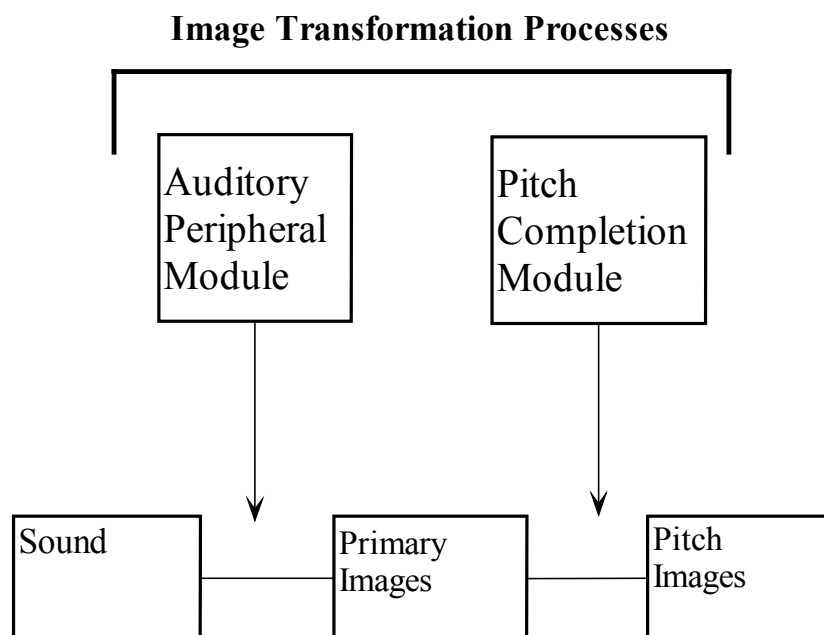


Fig 9: Processes from sound to pitch images using the APM and PCM

Figure 10 shows the pitch image as a result of first processing a musical signal with the APM and then processing the primary image with the PCM

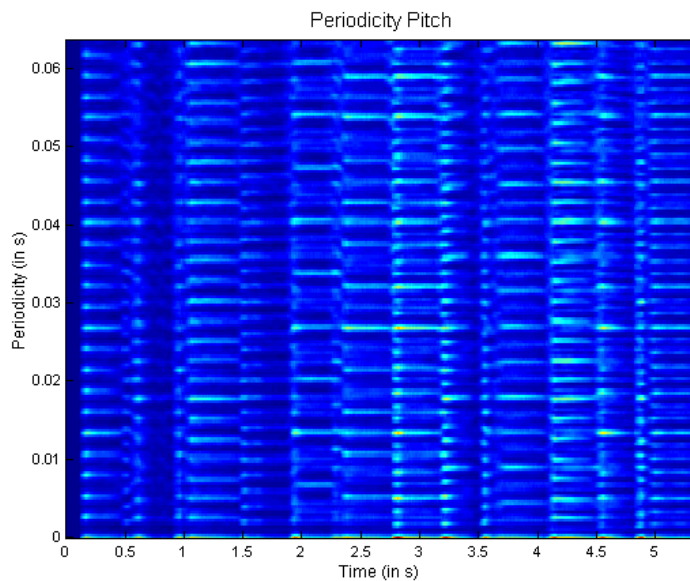


Fig 10: Periodicity Pitch

3.3.2. Rhythm Module (RhM)

The Rhythm Module (RhM) applies the Minimal Energy Change (MEC) algorithm defined by Leman & Verbeke (2000) [2] to the detection of repetition in rhythm patterns. The MEC algorithm calculates the fundamental period of a signal. The RhM takes a sound file as input and the output is an estimation of the fundamental period at each time step. The technique used is a generalization of the Average Magnitude Difference Function (AMDF), a faster alternative for autocorrelation.

The basic ideas behind MEC are that:

- (i) the energy calculated over the period of a repeating pattern is more or less the same at each moment in time
- (ii) minimal changes of this energy point to the period of the repetitive pattern.

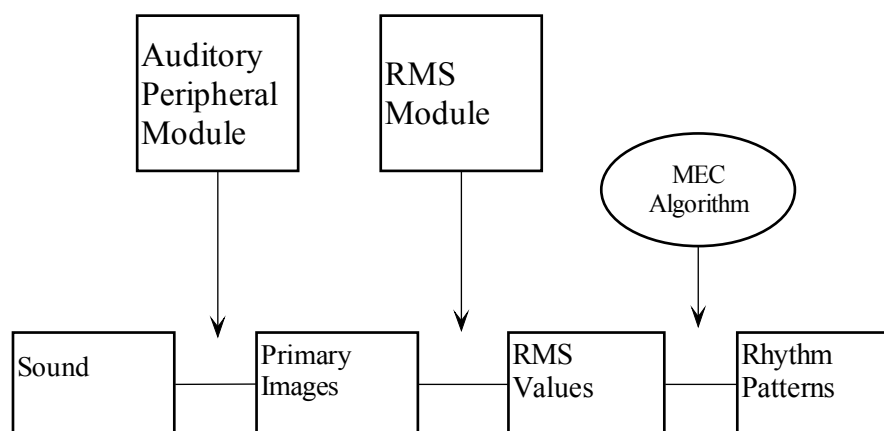


Fig 11: Processes from sound to rhythm patterns

Figure 12 shows the MEC analysis result of a short excerpt of Schumann's *Kuriose Geschichte*. The detected period is about 1.317 s in this case.

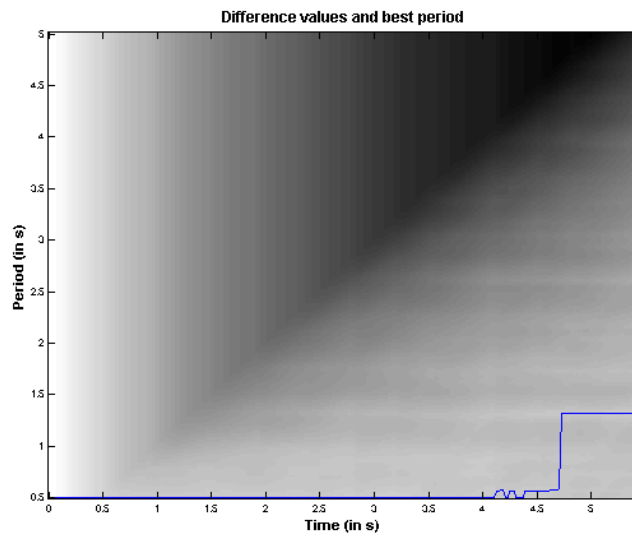


Fig 12: Summed difference values (over all channels) and a plot of the best period on top.

3.3.3. Echoic Memory Module (EMM)

The EMM takes an image as input and gives the leaky integrated image or *echoic image* as output. The images are integrated so that at each time step, the new image is calculated by taking a certain amount of the old image, which is then added with the new incoming image. The EMM can be applied to pitch completion images, for example. The specified echo defines the amount of context that we take into account. With very short half decay time (for example a short echo of 0.1 sec) or almost no context we speak about *Local Pitch Images*. When context is taken into account we use a longer half decay time (for example a long echo of 1.5 sec) and speak about *Global Pitch Images*.

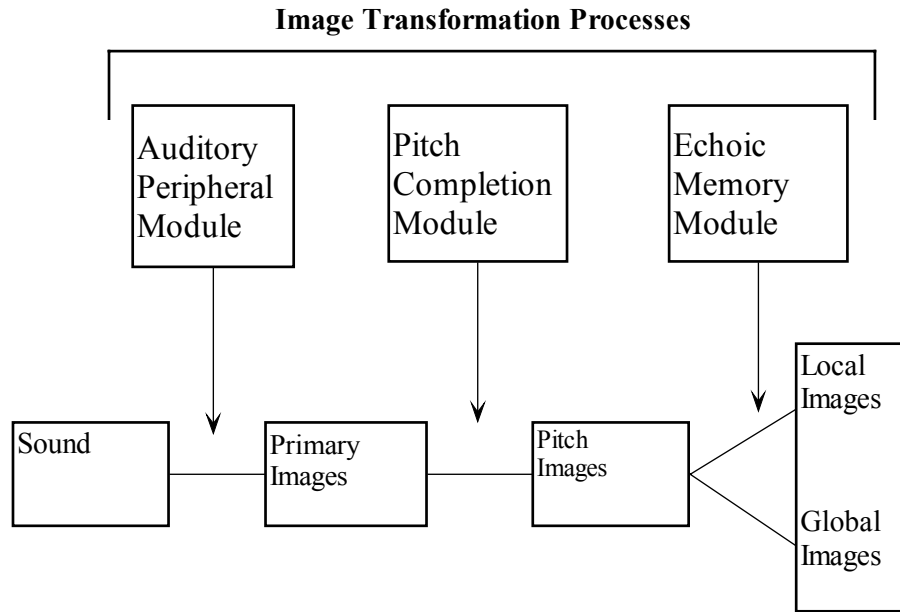


Fig 13: Processes involved from sound to echoic images

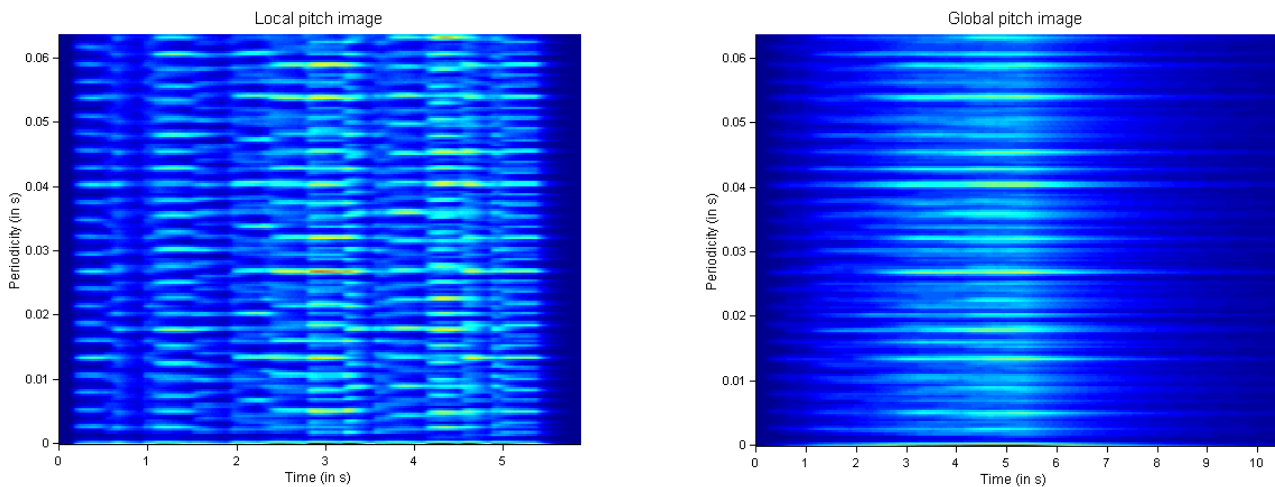


Fig14: Left: local pitch image (echo of 0.1 s) Right: global pitch image (echo of 1.5 s)

3.4. Modules at the cognitive level

3.4.1. Contextuality Module (CM)

The CM measures the pitch commonality between two running pitch images of the same sound, each image having a possible different echo. Contextuality can be used to measure the commonality between local and global pitch images. Two types of measurement are taken into account. The first measurement is based on a method of inspection of the pitch sequence by means of a fixed image or a probe. The second measurement is based on a method of

comparison of the pitch images (each having a possible different echo) at running time. Inspection and comparison of echoic pitch images are useful methods for studying tonal tensions.

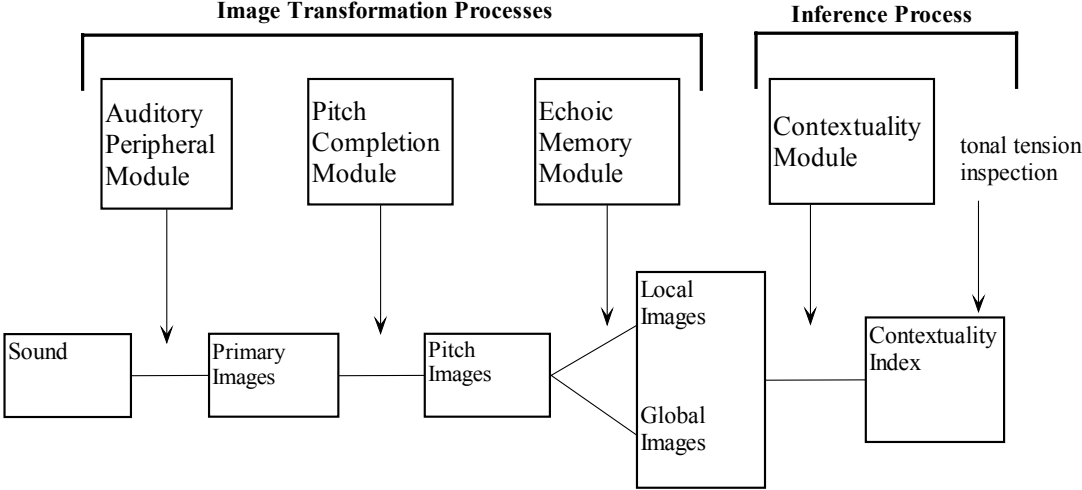


Fig 15: Processes involved from sound to contextuality index

Figures 16 and 17 show the application of contextuality to the short excerpt from Schumann’s Kuriose Geschichte followed by a 0.1 s period of silence and an f sharp Shepard tone.

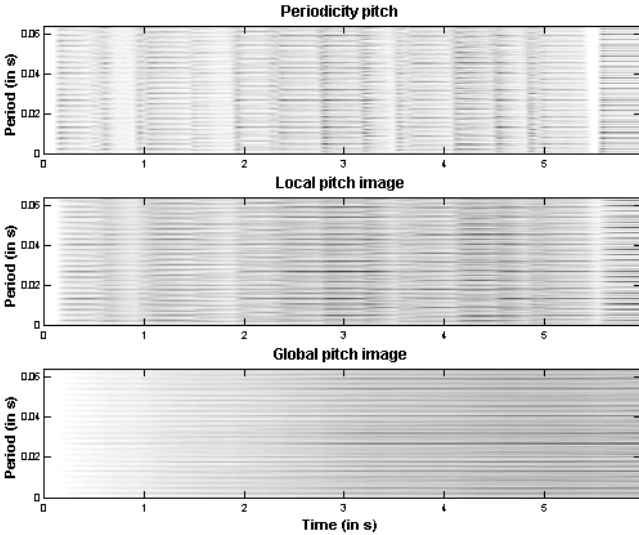


Fig16: Pitch images for the excerpt of Schumann’s Kuriose Geschichte followed by the Shepard tone of f sharp

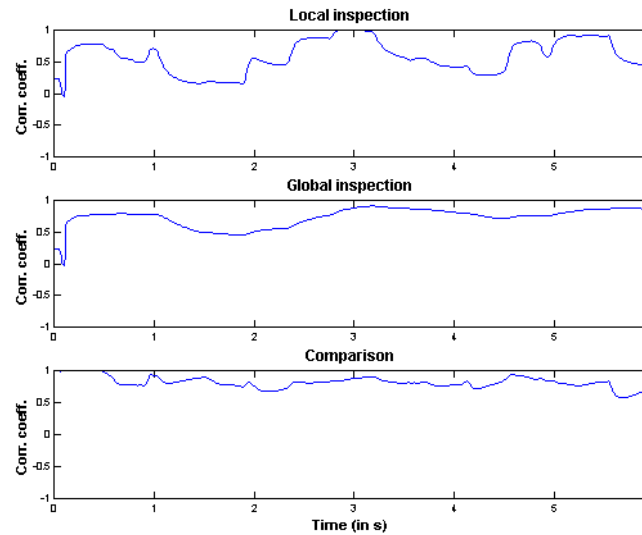


Fig17: Contextuality index for the excerpt of Schumann's Kuriose Geschichte followed by the Shepard tone of f sharp

4.Applications

The IPEM toolbox user can explore the modules by the use of the examples, in the example section of each module, through which the effectiveness of the IPEM functions can be directly perceived. Additionally a set of applications will be added in the future.

Contextuality Module:

In our first example we demonstrate an application of the Contextuality Module.

Contextuality is applied to a melodic sequence in c minor and two chords. The pitch images of the sequence and of the chords are concatenated. First the sequence is followed by the chord c-es-g that fits with the tonality, second by the chord cis-f-gis that does not fit with the tonality of c minor. The location at which the fixed image should be taken is in this example simply at the end of the file. The local and the global echoes are set to 0.1 and 1.5 seconds. Figure 18 shows the pitch images of the sequence in c minor followed by the chord c-es-g.

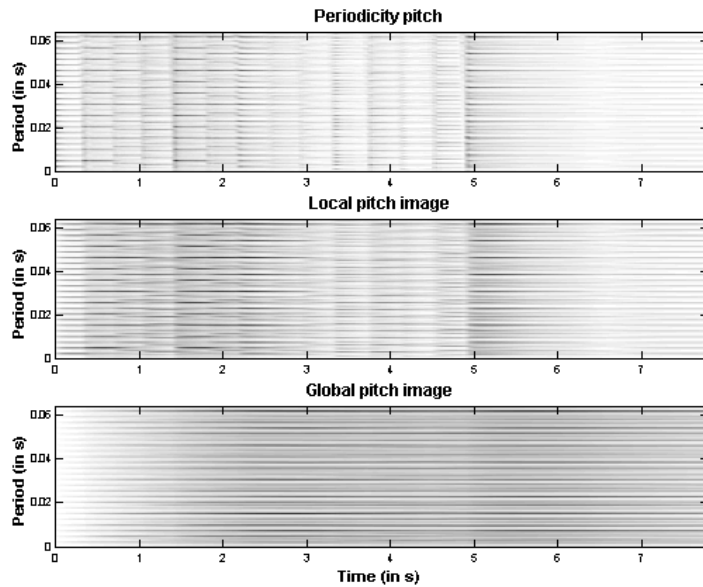


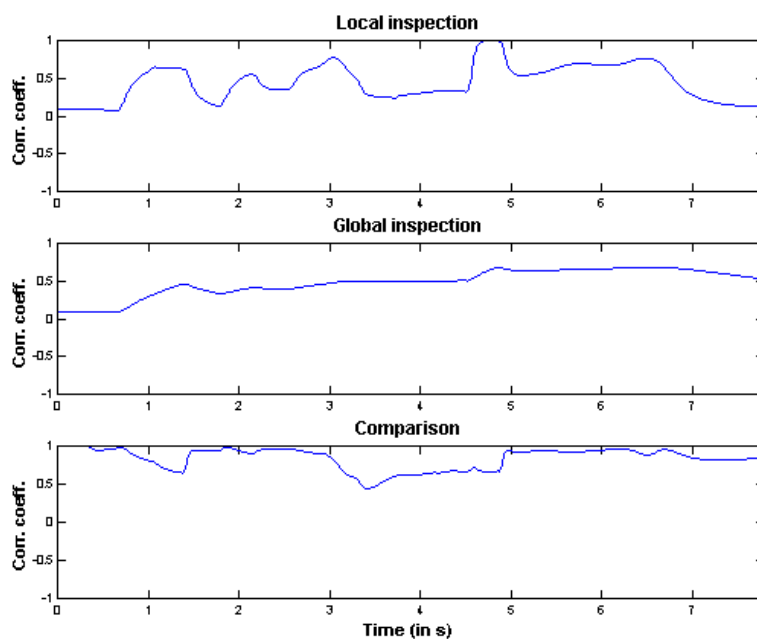
Fig 18: Periodicity Pitch, Local and Global Pitch Images

We apply the method based on inspection to the local and global pitch images. We obtain results by inspecting the local images with the fixed (local) image and inspecting the global images with the fixed (local) image. A graph shows the similarity degree of the chord within the sequence at the local and at the global level.

With the method based on comparison we get results of comparing the local images with the global images over the whole sequence. A graph shows the degree in which the local pitch images fit with the global pitch images.

Figure 19 shows the contextuality index for the sequence in c minor followed by a fitting (a) and a not fitting (b) chord.

(a)



(b)

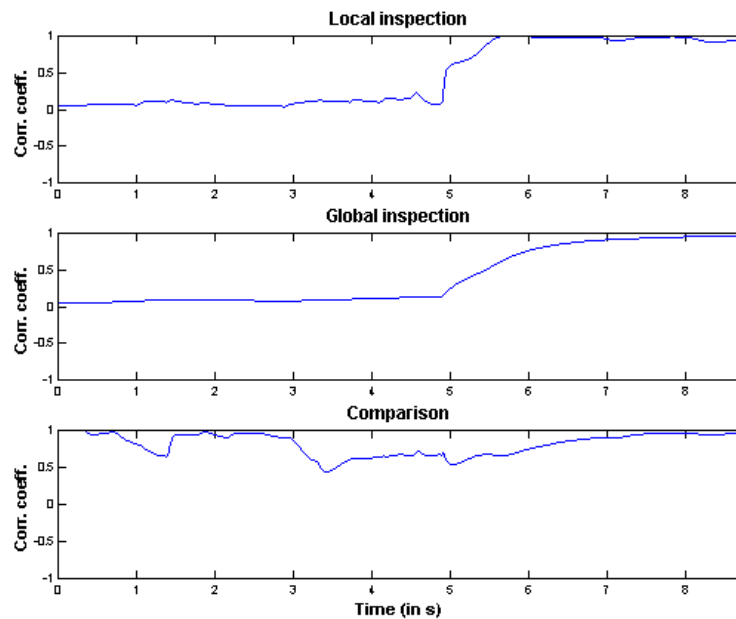


Fig 19: Contextuality index

Rhythm Module:

In the second demonstration the Rhythm Module is used to detect and extract rhythmic patterns from a sound fragment.

The applied method is as follows:

First, an energy signal is calculated from the sound file using RMS calculation.

This energy signal is then further processed using the MEC analysis function. At each moment in time, the minimum of the calculated difference values is selected and the according period is returned as the “best period”.

We can then use this detected period in a sample editor to create a loop with that duration and evaluate this result.

The first example uses a sound fragment from Photek called “The lightening (digital remix)” from the album “Form and function”.

Figure 20 shows the calculated difference values over time, together with the selected “best period”. Since this sound fragment contains quite strict repetition (sequenced drums), the selected period is also quite stable.

Figure 21 shows one intersection of the calculated difference values (at the time $t = 11.240$ s). The minimum in this curve points to the period of the repeating pattern, which is 2.82 s in this case.

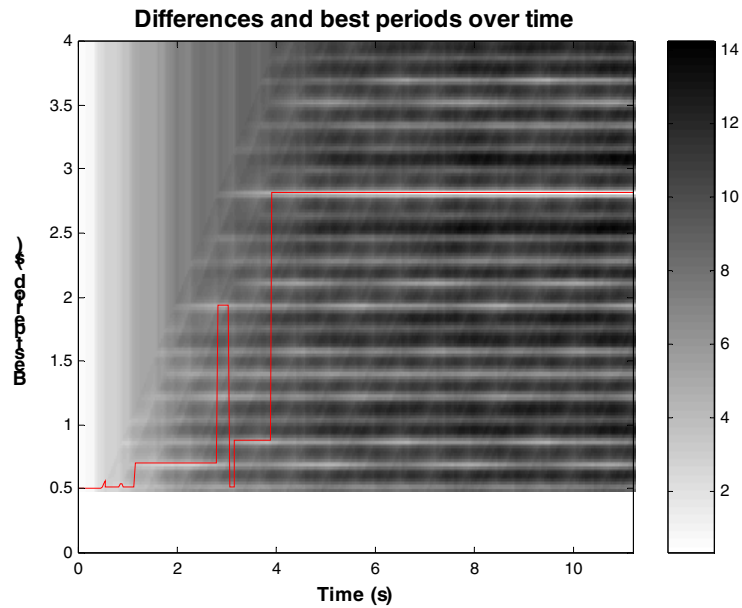


Fig 20: calculated difference values over time and best period

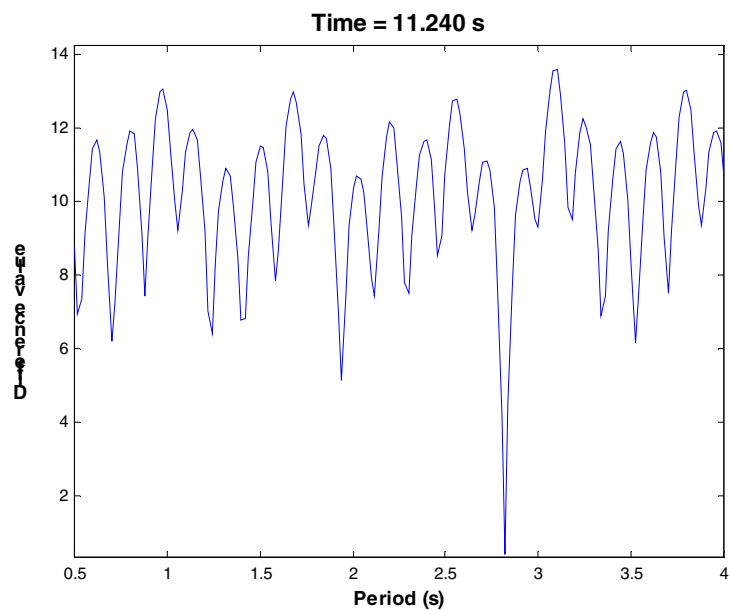


Fig 21: one intersection of the calculated difference values

The second example uses a sound fragment from Tom Waits called “Big in Japan” from the album “Mule variations”. Here, the repetition is not so strict anymore (acoustic recording), and the selected best period exhibits some fluctuation around a period of 2.13 s. Some post processing is needed to use these results in an automatic way. Figures 22 and 23 are the same as for the first example.

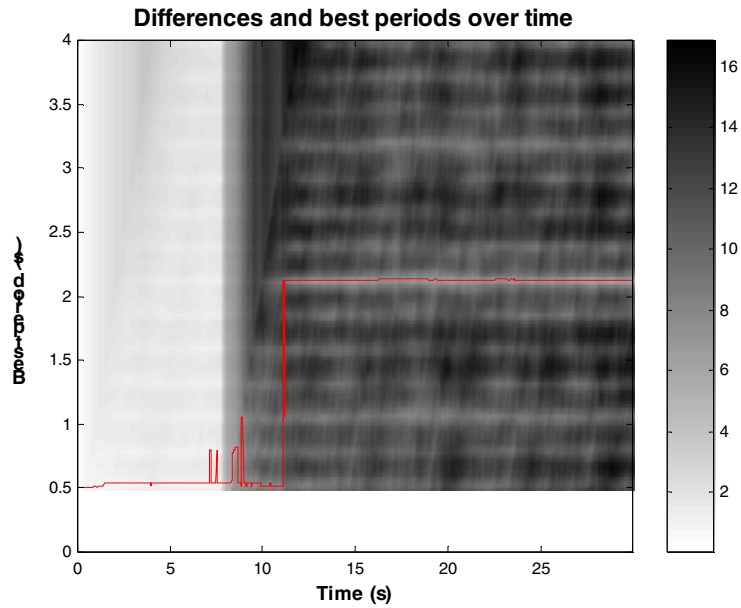


Fig 22: calculated difference values over time and best period

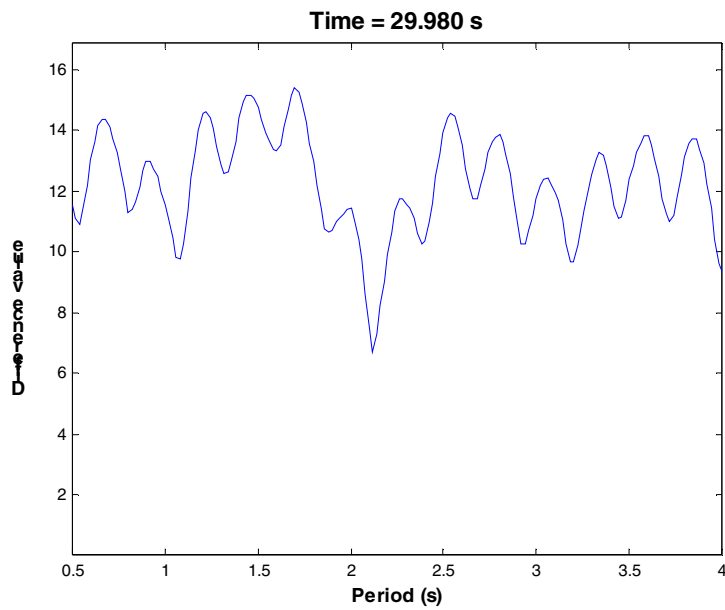


Fig 23: one intersection of the calculated difference values

5. Discussion

We believe that applying computer modeling of the auditory system to provide a foundation of music analysis in terms of human perception delivers deeper insight into the development of a fully integrated approach to music analysis. Here we introduce the version 1.00 (beta) of the IPEM toolbox that has been released in March 2001. This first version consists of a two

parts. The first part contains the presentation of the concepts, introductory descriptions of the modules examples. The second part contains the reference manual that is a practical guide.

A main characteristic and limitation of our current state-of-the-art is that images are basically frame-based and that we do not yet consider an object-based representation of musical content.

The IPEM toolbox will be under constant development. Our focus in the near future will mainly be on the development of new modules in context of the global conceptual picture we have in mind. It is our aim to provide modules and functions that allow researchers to deal with as many as possible aspects of feature extraction in the field of perception-based music analysis.

We are preparing some demonstrations of how the toolbox has been used in our research. These demonstrations will be included in the next version.

We permit consolidation and improvement by inviting users to evaluate the IPEM Toolbox. Users can freely experiment with the modules and build new applications from the basic functions.

Acknowledgement

The toolbox for perception-based music analysis has been developed at IPEM, research center of the department of Musicology at the University of Ghent, with support from the Research Council of the University (BOF) and the Fund for Scientific Research of Flanders (FWO)

References

Papers

[1] M. Leman, "An auditory model of the role of short-term memory in probe-tone ratings", *Music Perception*, vol. 17, n° 4, pp. 435-463

[2] M. Leman and B. Verbeke, "The concept of minimal 'energy' change (MEC) applied to the recognition of repetitive rhythmical patterns in acoustical musical signals", In: K. Jokinen, D. Heylen, and A. Nijholt (Eds.), *Cele-Twente Workshop on Language Technology, Workshop II: Internalising Knowledge*, Ieper, 2000, Twente University, Enschede, pp. 191-200, 2000.

Submitted to *Journal of New Music Research*, Special Issue on Rhythm Perception, Periodicity and Timing Nets

[3] M. Leman, "Visualization and calculation of the roughness of acoustical musical signals using the synchronization index model (SIM), *Proceedings of the COST G-6 Conference of Digital Audio Effects (DAFX-00)*, Verona, 2000

[4] L. Van Immerseel and J. P. Martens, "Pitch and voiced/unvoiced determination with an auditory model", *The Journal of the Acoustical Society of America*, vol.91, pp. 3511-3526, 1992.

Packages

[5] |WAVE

J. F. Culling, "Signal processing software for teaching and research in psychoacoustics under UNIX and X-windows", *Behavior Research Methods Instruments and Computers*. 28, pp. 376-382, 1996

<http://www.cf.ac.uk/psych/CullingJ/pipewave.html>

[6] Auditory Toolbox

M. Slaney, Auditory Toolbox, Technical Report 1998-10

<http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/>

[7] The Development System for Auditory Modeling (DSAM)

<http://www.essex.ac.uk/psychology/hearinglab/lutear/>

[8] The Auditory Image Model (AIM)

Roy D. Patterson, Mike H. Allerhand and Christian Giguere, "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," *Journal of the Acoustical Society of America*, vol 98, pp 1890-1894, 1995

<http://www.mrc-cbu.cam.ac.uk/personal/roy.patterson/aim/>