Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

Comprehensive and Complex Modeling of Structural Understanding, Studied of an Experimental Improvisation

Olivier Lartillot,[*1] Mondher Ayari[#2]

[*]*Finnish Centre of Excellence in Interdisciplinary Music Research, Finland*

[#]*IRCAM-CNRS / University of Strasbourg, France*

[1]olartillot@gmail.com, [2]Mondher.Ayari@ircam.fr

ABSTRACT

Music perception and cognition are ruled by complex interdependencies between bottom-up and top-down processes at various cognitive levels, which have not been fully understood and described yet. Cognitive and computational descriptions of particular facets of music listening remain insufficient if they are not integrated in a comprehensive modeling. In the long term, we aim at proposing a comprehensive and complex cognitive modeling of the emergence of structures in music listening and to test its potential by running a computational implementation on elaborate music.

The study presented in this paper is part of a broader project, whose general aim is to collect an experimentally controlled jazz improvisation with the view to study jazz listeners' understanding of that piece. An eminent jazz guitarist, Teemu Viinikainen, was invited to play an original improvisation while following a few general heuristics that we defined beforehand, concerning the use of pauses, repetitions, accentuations and of various ways of evolving the modal discourse. During a subsequent interview, while listening progressively to the recording, the musician gave a detailed a posteriori analysis that was recorded as well, talking and playing examples on his guitar.

A systematic analysis was performed exhaustively on the piece, starting from a manual transcription of the piece, followed by motivic, harmonic, rhythmical and structural analyses. Our previous cognitive complex modeling of structural analysis of music has been extended further and implemented in the Matlab programming environment. This extended model starts from the audio recordings, and performs altogether transcription and higher-level analyses, with bottom-up and top-down interactions between low-level and high-level processes. The study challenges the traditional dichotomy between transcription and structural analysis and suggests instead a multi-layer structuring of events of various scales (notes, gestures, motifs, chords, phrases, etc.), where higher-level structures contextually guide the progressive discovery of lower-level elements. The model will be further validated and enriched through a comparison with the musician's analysis and with jazz listeners' annotation of the piece collected experimentally.

I. INTRODUCTION

Music perception and cognition are ruled by complex interdependencies between bottom-up and top-down processes at various cognitive levels, which have not been fully understood and described yet. Cognitive and computational descriptions of particular facets of music listening remain insufficient if they are not integrated in a comprehensive modelling. In the long term, we aim at proposing a comprehensive and complex cognitive modeling of the emergence of structures in music listening and to test its potential by running a computational implementation on elaborate music.

II. MUSICAL MATERIAL

The study presented in this paper, focused particularly on the problem of transcription and computational modeling of music analysis is part of a broader project. The general aim of the project is to collect an experimentally controlled jazz improvisation with the prospective to study in detail how expert jazz listeners understand that piece through listening tests.

A. An experimentally controlled jazz improvisation

An eminent jazz guitarist, Teemu Viinikainen, was asked to play an original improvisation, while following a certain number of rules that we specified beforehand. The objective of this partial control of the improvisation is to test several cognitive principles ruling music perception such as discontinuity, parallelism, event stability, accentuation, change of modal scales, etc (Ayari, 2008).

More precisely, the heuristics proposed to the musician are:

- the use of longer silence to separate musical phrases (whereas shorter silences are naturally part of the musician's expression)
- the use of accentuations either on hierarchically important notes or on secondary notes. The musician is asked to focus on one particular pitch, and then on several ones.
- the use of particular scales of the musician's choice, followed by a focus on a particular degree of one scale.
- the use of a particular motif of the musician's choice, which should be freely repeated throughout the piece, but easily recognizable by the listener. This motif, mainly rhythmical, can be transformed in many ways.
- the use of modulations, both by focusing on various degrees of a single scale, and by modifying the intervals used in the scale through alterations.

These heuristics were explained to the musician using concrete simple examples and schematic drawings with simple music notations. Apart from those basic principles, the musician was free to build the improvisation according to his taste. The musician was asked to play freely, so that the improvisation does not follow a known melody, chord sequence or even rigid pulsation or harmony, but on the

contrary get organized intrinsically through a development of original thematic materials and a succession of phases.

The obtained piece is clearly divided into four parts, separated by long silences, and each characterized by specific musical material. From the third and last parts, the silence is much shorter, but the transition is marked by the apparition of a new musical material.

B. The musician's own analysis of the piece

The performance was followed by a filmed interview, in which the musician provided a detailed a posteriori analysis of the improvisation. He explained his decisions while designing the piece and how he would understand the piece on his own point of view. He formulated these explanations in various ways: talking, playing examples on his guitar, and writing a modal analysis on a piece of paper. We endeavor to take a maximum benefit of Teemu Viinikainen's extensive knowledge in many jazz repertoires and styles, and his rich expertise in live jazz improvisation.

At the end of the interview, we asked the musician to play a modal interpretation of the improvisation, following the modal analysis he wrote. This performed modal analysis was recorded as accompaniment track superposed to the initial melodic improvisation track.

C. Longer-term perspective of the project

This music material will be used for listening tests with jazz experts. Listeners' reactions will be compared with the compositional strategies as explained by the musician himself. Both the musician's and the listeners' points of view will be compared with the predictions given by the comprehensive analysis resulting from the computational model. Congruencies between the musician's explanations, the listeners' reactions and the computational predictions enable to validate the correctness of the computational model. Besides, we expect that incongruencies, on the other hand, might help to reveal particular compositional and listening heuristics strategies. It is also hypothesized that integrating these additional heuristics and strategies in the computational modeling should enrich the overall quality of the complex model, and the correctness of the computational predictions.

III. COMPUTATIONAL MODELING

A systematic computational analysis was performed on the piece, starting from a transcription of the piece, followed by motivic, harmonic, rhythmical and structural analyses. Our previous cognitive complex modeling of structural analysis of music (Lartillot and Ayari, 2011) has been extended further and implemented in the Matlab programming environment. The computational model starts from the audio recordings, and performs altogether transcription and higher-level analyses, with bottom-up and top-down interactions between low-level and high-level processes.

This objective of modeling both transcription and higher-level musicological analysis as one single complex system is motivated by the fact that despite the apparent separation between these two processes, when looking in more details, we can notice that transcription itself requires a large amount of musicological information.

A. Basic Transcription

1) Frequency estimation. We propose a method for pitch extraction where two strategies are carried out in parallel.

The first strategy based on autocorrelation function focuses on the fundamental component of harmonic sounds, and can track multiple harmonic sources at the same time (Tolonen and Karjalainen, 2000). The audio signal is decomposed using a two-channels filterbank, one for low frequencies below 1000 Hz, and one for high frequencies over 1000 Hz. On the high-frequency channel is performed an envelope extraction using a half-wave rectification and the same low-pass filter used for the low-frequency channel. The periodicity corresponding to note pitch heights is estimated through the computation of an autocorrelation function using a 46.4 ms-long sliding Hanning window moving every 10 ms. Side-border distortion intrinsic to autocorrelation function is neutralized by dividing the autocorrelation with the autocorrelation of its window (Boersma, 1993). A magnitude compression of the amplitude decreases the width of the peaks in the autocorrelation curve, suitable for multi-pitch extraction. After summing back the two channels, the sub-harmonics implicitly included in the autocorrelation function are filtered out from the halfwave-rectified output by subtracting time-scaled versions of the output. A peak picking frame by frame of this representation results in a pitch curve showing the temporal evolution of the fundamental components of the successive notes played by the musical instruments. One drawback of the use of autocorrelation function for pitch extraction is that the detected frequency is not clearly stabilized on each note, showing fluctuations.

The second strategy for pitch extraction is simply based on the computation of a spectrogram using the same frame configuration as for the first method. In this representation, the curve of the fundamental component is indicated with better accuracy and less fluctuation, but harmonics are shown as well, so the fundamental curve cannot be tracked robustly. The advantages of the two methods are combined by multiplying point by point the two matrix representations, so that the fundamental curve is clearly shown and the harmonics are filtered out (Peeters, 2006).

Global maxima are extracted from the combined pitch curve for each successive frame within the frequency region 50 Hz – 1500 Hz. Peaks that do not exceed 3% of the highest autocorrelation value across all frames are discarded: the corresponding frames do not contain any pitch information, and will be considered as silent frames. The actual frequency position of the peaks is obtained through quadratic interpolation.

The frequency axis of the pitch curve is represented in logarithmic domain and the values are expressed in cents, where octave corresponds to 1200 cents, so that 100 cents correspond to the division of the octave into 12 equal semi-tones.

2) Pitch curve segmentation. Each gesture is further decomposed into notes based on pitch gaps. We need to detect changes in pitch despite the presence of frequency fluctuation in each note, due to vibrato, which can sometimes show very large amplitude. We propose a method based on a single chronological scan of the pitch curve, where a new note is started after the termination of each note. In this method,

notes are terminated either by silent frames, or when the pitch level of the next frame is more than 65 cents away from the mean pitch of the note currently forming.

We restrict this study to predetermined tuning. This means that a given reference pitch level is assigned to a given degree in the 12-tone scale. In the musical example considered in this study, the degree E is associated with the tuning frequency 328 Hz. The other degrees are separated in pitch with a respective distance of 100 cents. In this simple case, to each note segmented in the pitch curve can be assigned the degree on the scale that is closest to the mean pitch measured for that note.

Very short notes are filtered out, when their length is shorter than 3 frames, or, when there is silent frame before and after the note, when the length is shorter than 9 frames. These short notes are fused to neighbor notes, if they have same scale degree and are not separated by silent frames.

The obtained transcription, as shown in Figure 1, seems qualitatively quite close to the desired transcription, except for low-pitch notes, where, due to the acoustical properties of guitar, the fundamental is missing. In Figure 1, the artifacts produced by the wrong handling of missing fundamentals, corresponding to the parts indicated with *ottava bassa* notations ("8vb"). We plan to take into consideration the treatment of missing fundamental in future works.

B. Motivic and Rhythmic Analysis

Rhythm cannot be simply described as temporal positions and duration expressed in seconds, because such purely technical representation of the signal does not relate to the way rhythm is actually processed by listeners. Rhythmic values are more classically expressed with respect to a hierarchical metrical grid, defined by a pulsation that is further organized into several layers of beats. The duration of a given note is usually related to its distance with a successive note, expressed proportionally to the range of durations contextually defined by the metrical grid, which needs therefore to be induced in the same time as the transcription itself, and not after it. The induction of pulsations requires a detection of repetitions, and even more generally a detection of repeated motifs (Lartillot, 2010). As such, metrical analysis is dependent on a motivic analysis, which needs to be carried out in parallel. Even in music not founded on a strict metrical pulse – such as the improvisation considered in this study –, implicit pulsations that develop locally play an important role in the understanding of the music. For instance, in bar 7 of Figure 2, we can notice successive repetitions of a same simple motif that creates an implicit but clearly recognizable pulsation. In the actual piece, the pulsation accelerates significantly during this successive motivic repetition. As such, there is no simple relation between rhythmic values (as proposed in the transcription) and actual duration in seconds; on the contrary, the rhythm is discretized, based on the motivic pattern that forms a constantly evolving metrical grid.

C. Modal Analysis

Similarly, pitch cannot be simply expressed as frequency in Hertz, as this signal description does not correspond to the symbolic representation of pitch actually understood by listeners. Pitch is more classically expressed with respect to a modal or tonal scale, which is defined not only by a series of pitch intervals, but also by a tuning of the scale – or even of separate degrees in the scale –, which adapts to the progressive fluctuation of the frequency related to each separate scale degree as time goes by. This requires therefore a discretization of the frequencies actually observed in the signal, based on a modal and/or tonal analysis, which needs to infer, in parallel to the transcription itself, modulations from one scale to another. We propose to model modal/tonal analysis as a progressive induction of symbolic representations based on a clustering of the recent pitch values that have been observed into sets that are mapped to theoretical systems that are culturally predefined (Lartillot and Ayari, 2011). It can be noticed here that this modeling is partly founded on motivic and rhythmical analysis as well, which indicates once again how complex is the interdependencies between the different areas of music analysis.

In the previous discussion, it was taken from granted that individual notes can be extracted in a first step, and then further processed for rhythm quantization and pitch spelling that were shown to be dependent on a rather complex musicological context. Yet, even the detection of notes in the continuous sound signal is generally dependent on that same context: the segmentation of note events from a continuous pitch curve, for instance, as discussed in the next section, is based on a detection of significant pitch gaps, which, once again, is generally dependent on the given modal or tonal context. Example of the importance of the higher-level musicological context for note onset detection is given in (Lartillot, 2012).

D. Structural Analysis

Structural analysis of the piece can also improve the overall quality of the transcription. This is even more important when there is no clear metrical structure, such as in this free unmetered improvisation. A detailed motivic analysis can improve the initial transcription (Figure 1), such that main themes are indicated in separate bars, and short patterns are shown with slurs (Figure 2). The modeling of such detailed structural analysis is planned for future works, and will extend the principles of the modeling of motivic analysis, discussed in previous paragraph B.

IV. CONCLUSIONS

This study challenges the traditional dichotomy between transcription and structural analysis and suggests instead a multi-layer structuring of events of various scales (notes, gestures, motifs, chords, phrases, etc.), where higher-level structures contextually guide the progressive discovery of lower-level elements. The model will be further compared with listeners' annotation of the piece collected experimentally.

The results of the model will be further confronted with the musician's and musicological analyses. Various versions of the models will be tested, and intrinsic parameters will be varied as well, in order to enrich the specifications of the model and suggest improvements. Dialogs between the musician, cognitive science and musicology will allow a critique of the methodology and of the model, suggesting new directions for further research.

The purely bottom-up processes of pitch-curve based note onset extraction have been integrated into the new version 1.4 of *MIRtoolbox* (Lartillot and Toiviainen, 2007). The integration of top-down factors related to motivic, metrical, harmonic, modal, structural and stylistic analyses are being integrated in our new computational framework under development, called *The MiningSuite* (Lartillot, 2011). The impact of culture on the establishment of those higher-level musicological analyses and even on the pitch transcription itself, has been studied with the analysis of music from different cultures, such as Arabic maqam improvisations (Lartillot, 2012).

ACKNOWLEDGMENT

We thank Teemu Viinikainen for very kindly offering us a very nice jazz improvisation, and for accepting to be interviewed afterwards and to give a very detailed musical analysis of his piece, as well as performing an accompaniment of his improvisation. The recordings were carried out with the technical assistance of Mikko Leimu, at the Musica studio of the University of Jyväskylä. Martin Hartmann, PhD student at the Finnish Centre of Excellence in Interdisciplinary Music Research, is helping us to perform the musicological analysis of the improvisation, and is collaborating in the project.

REFERENCES

Ayari, M. (2008). Performance and musical perception analysis. *Intellectica*, 48-49.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proc*. 17, 97–110.

Lartillot, O. (2010). Reflexions towards a generative theory of musical parallelism. *Musicae Scientiae*, *Discussion Forum 5*, 195-229.

Lartillot, O. (2011). A comprehensive and modular framework for audio content extraction, aimed at research, pedagogy, and digital library management. *130th AES Convention Proc*.

Lartillot, O. (2012). 'Computational analysis of Maqam music: from audio transcription to musicological analysis, everything is tightly intertwined. *Acoustics Hong Kong*.

Lartillot, O., & Ayari, M. (2011). Cultural impact in listeners' structural understanding of a Tunisian traditional modal improvisation, studied with the help of computational models. *J. Interdisciplinary Music Studies, 5*, 85–100.

Lartillot, O., & Toiviainen, P. (2007). MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio. *International Conference on Music Information Retrieval.*

Peeters, G. (2006). Music pitch representation by periodicity measures based on combined temporal and spectral representations. *ICASSP 2006 Proc*.

Tolonen, M., & Karjalainen, M. (2000). A computationally efficient multipitch analysis model. *IEEE Trans. Speech and Audio Proc. 8*, 708–716.

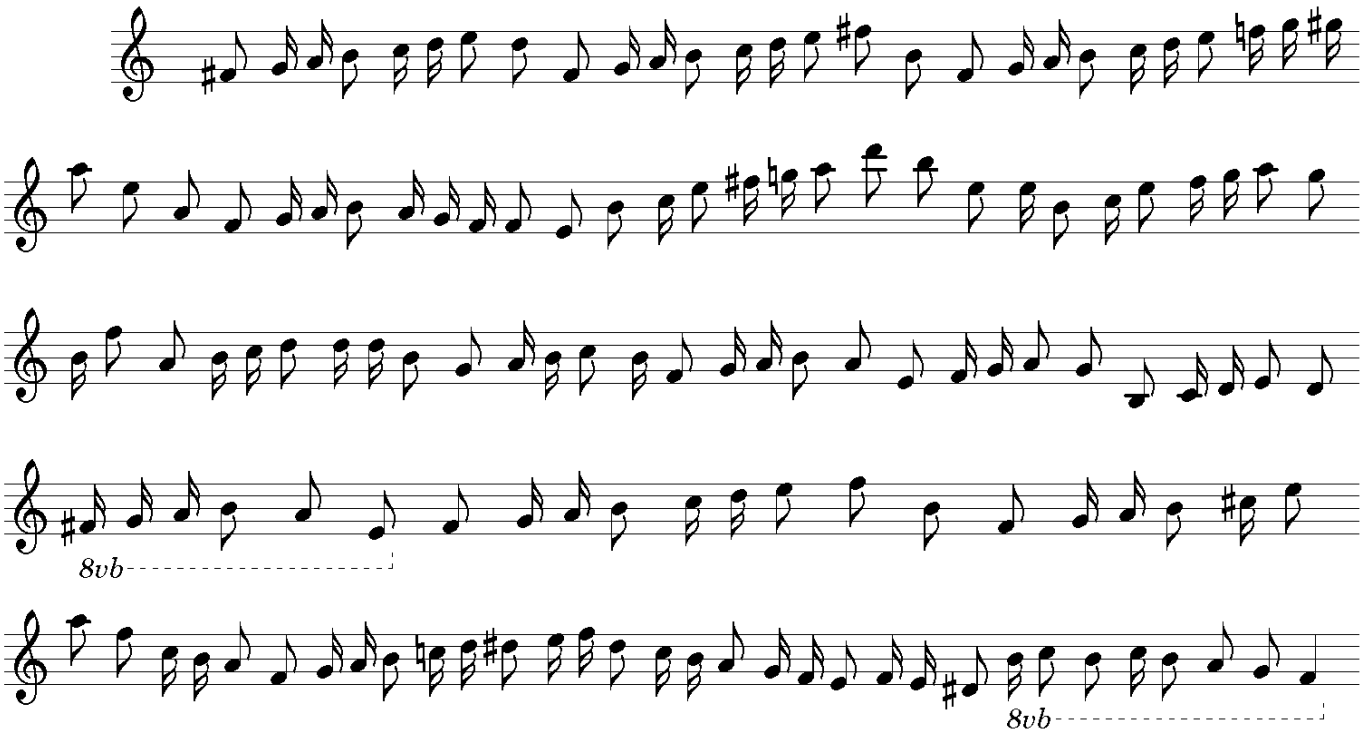

Figure 1. Transcription of the first part of Teemu Viinikainen's improvisation.

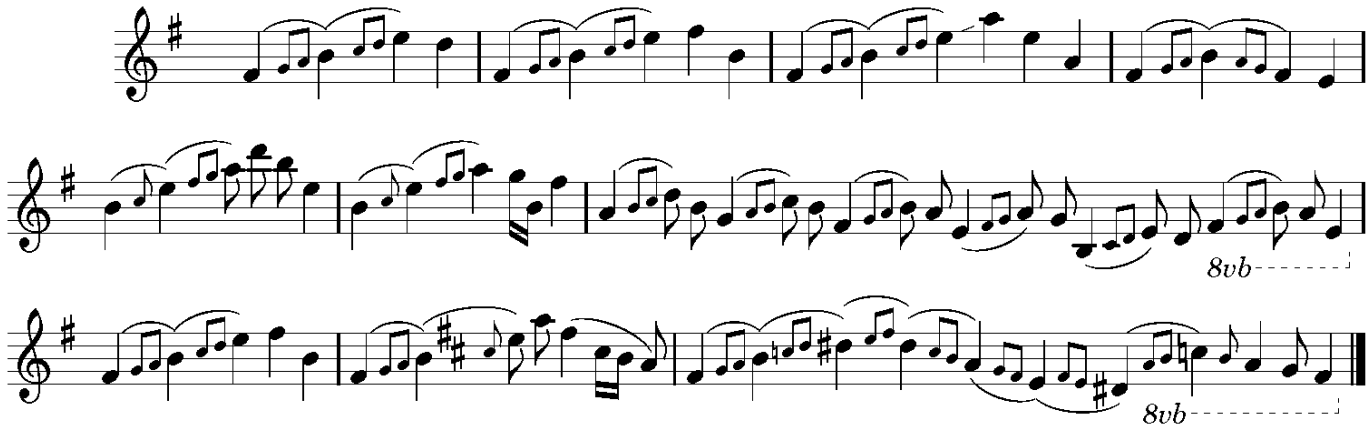

Figure 2. Improvement of the transcription based on structural analysis.