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)

Temporal multi-scale considerations in the modeling of tonal cognition from continuous rating experiments

Agustín Martorell[*1], Petri Toiviainen[#2], Emilia Gómez[*3]

[*] *Music Technology Group, Universitat Pompeu Fabra, Spain*
[#] *Department of Music, University of Jyväskylä, Finland*
[1]agustin.martorell@upf.edu, [2]petri.toiviainen@jyu.fi, [3]emilia.gomez@upf.edu

ABSTRACT

Modeling tonal induction dynamics from naturalistic music stimuli usually involves slide-windowing the stimuli in analysis frames or leaky memory processing. In both cases, the appropriate selection of the time-scale or decay constant is critical, although rarely discussed in a systematic way. This study shows the qualitative and quantitative impact that time-scale has in the evaluation of a simple tonal induction model, when the concurrent probe-tone method is used to capture continuous ratings of perceived relative stability of pitch-classes. Music stimulus is slide-windowed using many time-scales, ranging from fractions of second to the whole musical piece. Each frame is analysed to obtain a pitch-class profile and, for each temporal scale, the time series is compared with the empirical annotations. Two commonly used frame-to-frame metrics are tested: a) Correlation between the 12-D vectors from ratings and model. b) Correlation between the 24 key activation strengths, obtained by correlation of the 12-D vectors with the Krumhansl and Kessler's key profiles. We discuss the metric artifacts introduced by the second representation, and we show that the best performing time-scale, minimizing the root mean-square of the frame-to-frame distances along time, is far longer than short-time memory conventions. We propose a temporal multi-scale analysis method as an interactive tool for exploring the effect of time-scale and different multidimensional representations in tonal cognition modeling.

I. INTRODUCTION

Tonal context induction refers to the development of a sense of key in listeners exposed to music stimuli. Cognitive and computational modeling of such process is challenged by the elusive description of tonality and by the relatively large and undefined temporal spans required for capturing that sense of context. For complex polyphonic stimuli, the time-scales involved in the process depend, in general and among other factors, on the tonal material temporal *delivery*, which can vary notably as music unfolds in time. The concurrent probe-tone method allows capturing real-time responses from subjects exposed to music of any complexity, providing quantitative ratings of the perceived relative stability of pitch-classes over time. This information is related to the concept of key strength, and it can be used to evaluate computational models of tonality induction. Many of the current models, however, apply rigid temporal analysis assumptions, obscuring the interpretation of the tonality phenomenon respecting time.

II. BACKGROUND

Most empirical methodologies aiming to model tonal context cognition, usually under the terms *sense of key* or *key induction,* have relied upon stop-and-rate retrospective judgement tasks (Krumhansl & Kessler, 1982). In these settings, listeners are exposed to some music stimulus, after which they are asked to rate a certain subjective perceptual magnitude. These methods provide a notable control of the experimental variables, but they present interpretative concerns about the nature of the captured features, given that sense of key is derived from indirect measurements -e.g. the relative stability between pitch-classes using the probe-tone technique-, and because ratings are produced *after* the stimulus (Krumhansl, 1990). Modeling tonal cognition dynamics -the evolution of the sense of key as music unfolds in time- from these approaches is even more problematic and time consuming (Vos & Leman, 2000).

Real-time response experimental tasks have been proposed to deal with some of these problems, such as the concurrent probe-tone method (Toiviainen & Krumhansl, 2003). Under this approach, a realistic complex music stimulus is used, and the probe tones are played concurrently along music listening. Subjects rate continuously the goodness-of-fit between the probe tones and the music stimulus by dragging a slider in a non-stop setting. Two important aspects make this approach differ substantially from the classical probe-tone tasks. First, the probe tones sound simultaneously with the music instead of being presented after the stimuli. While this might be seen as an advantage respecting the retrospective judgements, it represents a very different musical reality. A quasi-continuous tone sounding along an *independent* tonal discourse might have a variety of psychological effects on listeners, and it is not clear whether listeners would keep a constant criterion in their attention and judgements. The second problem comes with the calibration of the rating responses, in terms of the consistent use of continuous scales and the unpredictable motoric mediation delays (Koulis et al., 2008). This has an enormous impact on the statistical significance of the intra-subject and inter-subject analysis.

Several problems arise for evaluating computational models of tonal induction, when continuous ratings are taken as reference for comparison. Temporal scale is a fundamental parameter for describing contextual musical features, as it is critical for time series analysis in general. Many window-based key-finding models from symbolic notation (Krumhansl, 1990; Temperley, 2001) analyse the stimuli in beats or bars, in order to produce 12-D profiles comparable to the probe-tone ratings. In the audio domain, where metric segmentation is not always reliable, this is often implemented by an overlapping sliding window of constant duration (Gómez, 2006). Similar time-scale decisions apply to most models inspired on auditory processing (Leman, 2000), by using a decay constant to simulate leaky memory processing. One recurring argument about the time-scale selection (Toiviainen & Krumhansl, 2003) is that it should fall within the short-time memory constraints, agreed around 3-8 seconds, and being tuned according to rhythmic or metric assumptions, or as a compromise between smoothness and discontinuity of the resulting signals.

There have been few systematic approaches to understanding the role of temporal scale in modeling tonal induction from probe-tone methods, most of them around the discussion about short-term and long-term memory implications. In (Leman, 2000), the echo constant of the proposed global image model is manipulated in steps of 0.2 seconds, spanning a range of 0.2 to 5 seconds, in order to find the optimal value fitting empirical correlation trends among several context inducing sequences (Krumhansl & Kessler, 1982).

Manipulation of the leaky memory decay constant has been proposed as interactive parameter in a real-time visualization model of tonal induction (Toiviainen, 2008), but no attempts of empirical validation were done. Several window sizes have been explored using Chew's spiral array for modeling tonal boundaries (Chew, 2006). In her work, evaluation was based on coarse structural -offline- annotations by experts or key indications in the score. Multiple time-scale analysis methods have been used, in both MIDI (Sapp, 2005) and MIDI/audio domains (Martorell & Gómez, 2011), for hierarchical visual analysis of tonality, key profile comparison, and pitch-space interactive inspection, but no validation has been approached respecting empirical ratings over time.

Aside from time-scale, evaluation issues arise from the multidimensional (12-D) nature of the involved time series. Tonal induction models are often evaluated through an indirect space of key strengths. This mapping is achieved by correlating the 12-D vectors with the ring-shifted key profiles proposed by Krumhansl and Kessler -henceforth *K-K's profiles*-. Vectors from both ratings and model are projected into this space in order to be compared, and model evaluation is discussed in these terms -after mapping- (Krumhansl, 1990; Toiviainen & Krumhansl, 2003). This mapping has been proposed in a variety of dimensional reduction solutions, most of them for visual comparison purposes, including multidimensional unfolding into toroidal inter-key spaces (Krumhansl & Kessler, 1982; Krumhansl & Toiviainen, 2003) and self-organized maps (Toiviainen, 2008; Janata, 2007). Despite their visual informativeness, such frame-based representations do not provide a proper quantitative comparison between models, given that they do not represent the actual listener's ratings, but their relationships with predefined -assumed- categories. Very few direct quantitative careful annotations of *key induction* dynamics along a complete complex music piece have been approached (Krumhansl, 1990). As a case study of her key finding algorithm, a Bach's prelude was analysed offline by two experts, which provided measure-based estimations of key according to a variety of musicological criteria, and rating up to four possible weighted key candidates. In most studies, however, subjects rated their perceptual relative stability of pitch-classes instead of direct key strengths -for obvious methodological reasons-. Despite general warnings are posed about this issue (Krumhansl, 1990), no study has covered the quantitative stress or evaluation artifacts introduced by the different dimensional projections. Such concerns include the mapping through correlation, which is non-metric in nature since it does not hold triangular inequality. Thus, comparison of two input vectors -the actual ratings and the prediction- is not equivalent to their comparison after mapping, and this has a quantitative impact on the evaluation.

In experimental settings capturing continuou*s* ratings, the autocorrelation of the involved signals is a very relevant factor as well, introducing notable statistical significance problems. However, there is no consensus about its treatment in literature (Schubert, 2001), obscuring the comparison between published models of continuous tonal induction.

III. AIMS

In this study, we show the impact that time-scale and multidimensional mappings have in the evaluation of a simple tonal induction model against ratings collected by the concurrent probe-tone technique. We reanalyse empirical data from a previous experiment (Toiviainen & Krumhansl, 2003), provided by the second author of the present work, to unveil additional limitations of working with preexisting -often preprocessed- data. From the tonal perception modeling standpoint, we question about temporal conventions and timing in tonal cognition. From the evaluation perspective, we discuss some requirements that published studies might consider to make tonal modeling comparison more informative. We propose an interactive temporal multi-scale tonal analysis tool for exploring the implications of time-scale and dimensional projections in tonal modeling.

IV. METHOD

A. Empirical data gathering

In (Toiviainen & Krumhansl, 2003), concurrent probe-tone tasks were carried out for capturing real-time responses to Bach's organ duet BWV 805 in A minor. Full experimental details can be consulted in the original article, the main ones follow. Eight highly trained musicians participated in the rating task. Stimuli consisted on 12 versions of the duet, each of them including a quasi-continuous probe tone from the chromatic scale. A church organ timbre was used for rendering the resulting files. In order to prevent blending and peripheral sensory dissonance, probe tones were slightly interrupted at the end of each measure, and there were presented to the opposite ear than the duet through headphones. Tempo was fixed to 75 bpm. After a training session with the interface using similar stimuli, subjects were exposed to the 12 versions in random order, adjusting a horizontal on-screen slider with the mouse according to the perceived degree of fit between the music and the probe tones. Slider's position was recorded each 200 ms. Raw data were smoothed by averaging over each 800 ms. and then averaged across subjects. This resulted in a 12-dimensional time series of 216 samples, representing the relative perceived stability of each pitch-class over time, which is the empirical data used in what follows -henceforth, *ratings*-

B. Model of tonal induction

In order to show the evaluation impact of time-scale and dimensional mapping, a very simple model of tonal induction is implemented, avoiding sophistications that might obscure the interpretation. Most of the discussion, however, would apply similarly for more refined models, as long as sliding windows or decay constants were used for analysing the stimuli.

The stimulus (free of probe tones) is first converted into a chroma representation. In the audio domain, Harmonic Pitch-

Class Profiles (HPCP) (Gómez, 2006) are computed from the signal every 50 ms. This results on a 12-dimensional time series, representing an estimation of the pitch-class relative energies. In the symbolic domain, the preprocessing just removes the octave information from the MIDI score. Both versions of the stimulus, as used in the original experiment, were provided by the second author of this work.

Then, a multi-scale temporal segmentation is applied to the preprocessed signals. The minimum time-scale is fixed to 800 ms., matching the sampling period of the empirical ratings, and the maximum window size fits the whole musical piece. This range is covered following a logarithmic ratio between adjacent scales, as a compromise between number of resolutions and computation cost. This logarithmic approach also contributes to a pleasant visual inspection of the resulting time vs. time-scale representation. A hop-size of 800 ms. is used for all time-scales, to provide temporal alignment with the ratings time series. Placement of the corresponding analysis and rating frames is done by matching their endings, that is, tonal context is defined from past to present, as most models do. To avoid artifacts in the estimation as time-scale increases, only full-sized segments of music are analysed, so we discard the beginning of the piece accordingly.

A pitch-class profile is then computed for each segment. In the audio domain, HPCP vectors within the frame are averaged and normalized to estimate the relative pitch-class energies of the segment. In the symbolic domain, we adapt the method implemented in the MIDI Toolbox (Eerola & Toiviainen, 2004) to get the relative duration of each pitch-class within the frame. Parncutt's durational accent (Parncutt, 1994) is not used here for two reasons. First, we are interested in applying the closest criteria for both audio and MIDI versions, but the onset information reliability is challenged in the audio domain. Second, the psychological effect of sustained church organ sounds is not likely to be well represented by the predominance of onsets assumed in Parncutt's model -e.g. the perceptual presence of a suspended note is substantially different when played by a harpsichord compared with a pipe organ-. The resulting 12-D time series are the output of our simple tonal induction model, and the information to be evaluated against the empirical ratings.

We will discuss two different representations for the tonal-related time series, as commonly found in literature. For both the model and the ratings, we will consider: a) the 12-D time series described above; b) the 24-D time series obtained after correlation of the 12-D vectors with the ring-shifted K-K's key profiles. Frame-to-frame distance between the model and the ratings is computed as one minus the correlation of the corresponding vectors. The final comparison between both multidimensional time series is considered as the root mean-square of the frame-to-frame distances along time.

C. Time vs. time-scale representation

Quantification of the overall similarity between both time series, however, is incomplete for representing the *quality* of a continuous tonal induction model. Aside from the numeric result, it is of interest to evaluate which sections of the music stimulus are being poorly represented by the model and for which ones the algorithm performs the best. An interactive visual inspection method is proposed as a complementary qualitative validation of the model, respecting both the impact of time-scale and the alignment of similar frames between the ratings and the model time series. This visualization is used as an index for exploring the different dimensional mappings over time and across time-scales.

The algorithm, based on Sapp's *keyscapes* (Sapp, 2005), is fed with the output of our tonal induction model -a 12-D time series for each time-scale of analysis-. Each frame is projected as a 4-D tonal centroid into K-K's space of inter-key distances -coordinates of key centres are available in (Krumhansl, 1990)-, using the multidimensional unfolding technique described in (Krumhansl & Toiviainen, 2003). This method can handle some degree of uncertainty, allowing centroids to be located *between* key centres, anywhere in the continuous pitch-space, according to their relative key strengths.

These 4-D points are then mapped into a single perceptual dimension (colour), using the method described in (Martorell & Gómez, 2011). In this approach, depicted in Figure 1, a 3-D projection of the K-K's 4-D space is geometrically inscribed into the CIE 1976 L*a*b* (known as CIELAB) colourspace (CIE, 2004). This solution approaches three desirable properties of the 4-D inter-key space: a) it provides a unique colour for each point in the continuous space; b) it approximates *perceptual uniformity*, that is, perceptual similarity between any pair of colours is correlated with the Euclidian distance between their spatial locations; c) it keeps both properties throughout the double circularity of the toroidal space, providing smooth colour transitions in any direction.

After this process, the resulting coloured centroids are organized in a 2-D image, representing the tonal estimates over time (x-axis) at many time-scales (y-axis). Figure 2 shows the keyscape computed from the MIDI version of the stimulus. The ratings data are processed in the same way as the model's output, and the resulting coloured time series is shown aligned below the keyscape.

V. RESULTS

A. Qualitative and quantitative analysis

A qualitative visual analysis of Figure 2 shows an overall coarse alignment between the estimates and the ratings -similar colours account for close distances in pitch-space-, and the effect of time-scale in the model's output. In this representation, a notable quantitative stress is introduced by the estimates (12-D) mapping into the 4-D inter-key space, and additional distortion appears as a consequence of the geometrical colouring through a 3-D projection of the pitch-space. However, it provides a visual intuition about the matching *quality* between the involved time-series, as well as a useful index for interactive exploration along time and across time-scales.

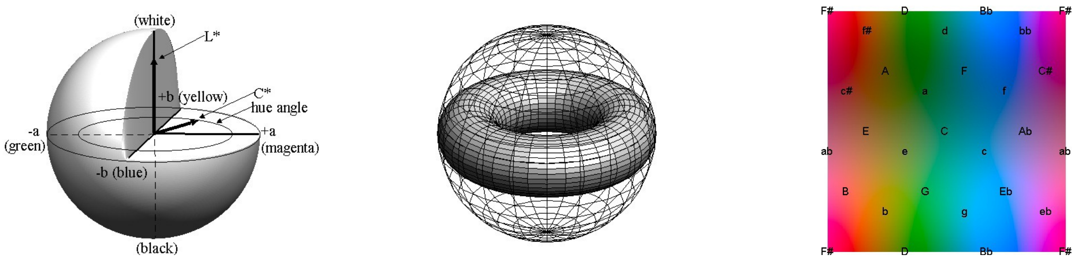

Figure 1. Colouring process. Left: CIELAB colourspace. Centre: geometrical inscription of pitch-space in CIELAB. Right: unfolded coloured pitch-space (colour legend for the keyscape).

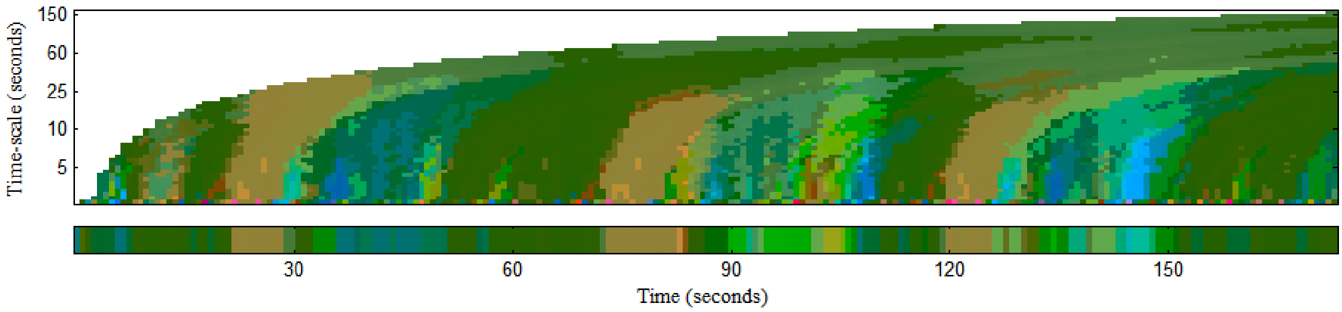

Figure 2. Keyscape (above) aligned with perceptual ratings (below).

Regarding quantitative analysis, we considered time-scales ranging from 800 ms. -sampling period of the perceptual ratings- to 60 secs. Larger time-scales were discarded for two reasons. First, the frequent modulations of the music stimuli make very unlikely that subjects were considering such long segments in their ratings. Additionally, we want to compare a representative amount of music over time and the impact of time-scale over the same stimuli, but larger time-scales require discarding quite a notable segment at the beginning. Consequently, we remove the data for the first 60 seconds from both the model and the ratings, and the time-scales above that value from the model. As side effects of this decision, we evaluate just the most interesting and tonally rich excerpt of the music stimulus, and we avoid potential response artifacts during the subject's ratings at the beginning of the piece.

Figure 3 shows the root mean-square deviation (considering error as the frame-to-frame distances) vs. time-scale, computed for both the 12-D and 24-D representations discussed above.

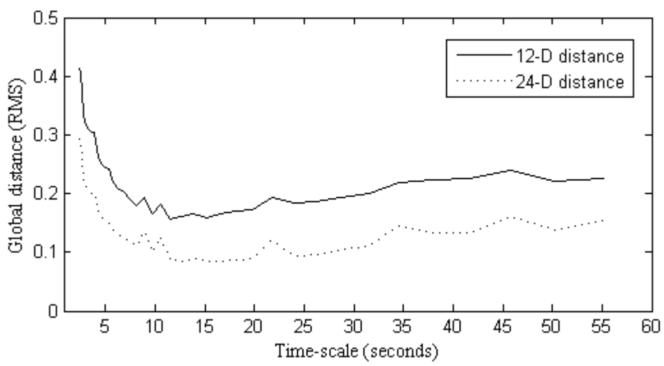

Figure 3. Global time-series distance vs. time-scale

Two aspects become evident from the figure. First, the best matching is achieved for time-scales of 11.5 s. (12-D) and 15 s. (24-D). Both curves show a clear trend with a minimum around that value, and they behave quite similarly across time-scales. Second, the 24-D representation introduces a clear metric artifact in the evaluation, seeming to produce a notable better matching.

B. Discussion

The best performing time-scales, around 13 seconds, are significantly larger than the agreed conventions for short-term memory, usually in the range of 3-8 secs. Several explanations could account for this result.

First, we could argue that such large time-scales can be actually involved in the processing of contextual tonal information, as several hierarchical tonal theories support (Lerdahl, 2001). This specific music stimulus is far more complex than simple chord progressions, but it is quite tonal, so main tonic references are not difficult to be *sustained* for highly trained musicians -as were the subjects in the experiment-, bridging short tonicizations and uncertainties even in the presence of *foreign* probe tones. Two aspects of the composition might probably contribute to this: the main theme lasts 8 bars, which corresponds to 12.8 secs. at 75 bpm. The piece is structured around such duration, not only for the thematic expositions but for the progressions as well. Moreover, the writing style might impose a strong thematic listening from the very beginning, when the main theme is presented in isolation by a single voice.

We should also consider some signal processing issues. The available empirical ratings data were the result of, at least, two averaging processes: one intended to minimize the erratic fluctuations of the subject's motoric action during the task, and an additional inter-subject averaging. The resulting signals, although downsampled to reduce its autocorrelation, keep a considerable degree of smoothness. In our temporal multi-scale model, smoothing is inherent to the use of larger time-scales, and this might contribute to better matching with the ratings time series.

Regarding metric artifacts, we should notice that the 24-D space used in the evaluation by key strengths is heavily coupled. In fact, the 24-D vectors resulting from the correlation with K-K's profiles are just *extended* versions of the input 12-D vectors, covering only a reduced 24-D subspace. However, correlation between vectors after mapping assumes a true space with 24 degrees of freedom, which results in a better fitting for essentially the same information than the original 12-D vectors. Figure 4 shows the evolution over time of the frame-to-frame correlations computed for the best performing time-scale (according to the 12-D representation). Here, only the first 15 seconds of music were removed, to show a more significant portion of the music stimulus.

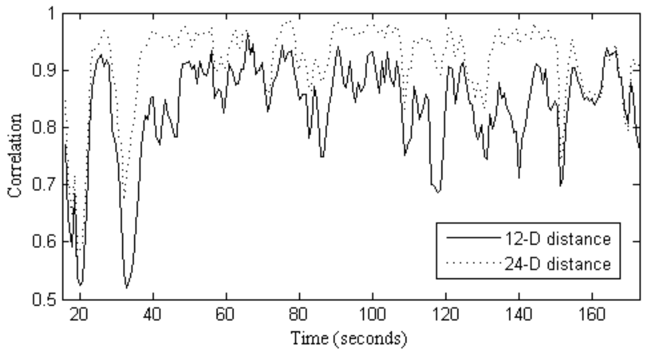

Figure 4. Frame-to-frame correlation for best time-scale.

The figure clearly shows that the 24-D mapping *outperforms* the plain 12-D representation for virtually all frames, but this is actually a mathematical artifact. Additionally, it is evident that both metrics are not equivalent, as can be observed from their temporal evolution. This is not surprising, since correlation does not hold triangular inequality, so we cannot expect it to behave as a metric in the strict sense.

In any case, the use of K-K's profiles is not justified for the stimulus used in the experiment. These profiles are assumed to be the categorical references respecting the perceived relative pitch-class stability, but they were derived from subjects' exposure to clearly unambiguous tonal contexts, which is not the case of the stimulus used here. It's evident that by using other key profiles, such as Temperley's (Temperley, 2001), this kind of evaluation would result in a different quantification of similarity. We thus claim that the most appropriate information to be used in the quantitative evaluation of tonal induction models, against ratings from probe-tone methods, are the plain 12-D output vectors from the model.

Figure 4 also shows the music sections for which the model performance drops significantly, which claims for including evaluation over time in any comparison between models to complement the whole piece statistics. It might happen, for instance, that a given model performs well for the complex music sections, but behaves poorly for the basic tonal material.

C. About fixed time-scales in modeling. Future scenarios.

The variable performance of the model along time leads to questioning the motivations of using fixed time-scales in the modeling of tonal context cognition. Aside computational convenience, there are no supporting evidence in literature for time-scale to be fixed in general. Most discussion about time in tonality focuses on the short-term vs. long-term debate, but fixed time-scales are implemented in virtually all cases. However, such temporal constraint seems counterintuitive to music experience in general. Even assuming that tonal context is induced in listeners just from the stimuli -which is quite a strong claim in a general sense-, we could argue that the time-scales involved in the listeners' processing of tonal information would depend on the tonal material itself, which can be -and usually is- quite variable as music unfolds in time. This is not in conflict with some parsimony arguments, which claim for minimizing the temporal memory resources. Such dynamic mechanism might actually optimize the use of cognitive buffers, adjusting the required size to the complexities of the information to be processed.

In our multi-scale evaluation scenario, it is evident that the model can outperform respecting any single resolution by using a dynamic time-scale over time. This would be actually consistent with the intuitions of musical listening. Tonal cognition is unlikely a passive process respecting time, as attention can be driven towards short-term or long-term activity dynamically, depending on both the musical content and the listener's background and intentions. Under this hypothetical *dynamic listening mode*, the time-scale would be adjusted along time to find an optimal path across the keyscape maximizing the fitting between the model and the perceptual ratings.

VI. CONCLUSION

Time-scale is a critical factor for describing tonality at contextual level, and a main parameter to care about in any model of tonal cognition. However, very few studies have addressed the problem of temporal resolution in a systematic way for tonality. Experimental methods capturing real-time continuous ratings related to tonal induction provide a potential framework for such research.

This work addressed some of the evaluation issues of approaching the time-scale problem, and it calls for more adequate quantification methods beyond global statistics. Standard methods and metrics borrowed from information retrieval research, such as precision and recall analysis, together with a more transparent use of statistics, could contribute to richer discussions and more informative comparison between models. We also argue that the availability of shared datasets of raw perceptual ratings would benefit the research on tonal cognition, contributing in terms of modeling comparison, multidimensional analysis standardization, and statistical adequateness.

ACKNOWLEDGMENT

This work was partially supported by the following projects: Classical Planet: TSI-070100-2009-407 (MITYC), DRIMS: TIN2009-14247-C02-01 (MICINN) and MIReS: EC-FP7 ICT-2011.1.5 Networked Media and Search Systems, grant agreement No. 287711.

REFERENCES

Chew, E. (2006). Slicing it all ways: mathematical models for tonal induction, approximation and segmentation using the spiral array. *INFORMS Journal on Computing*, 18(3), 305-320.

CIE 015 (2004). *Colorimetry*, 3rd ed. Commission Internationale de L'Éclairage.

Eerola, T. & Toiviainen, P. (2004). MIDI Toolbox: Matlab tools for music research. Finland: University of Jyväskylä. Electronic version: http://www.jyu.fi/musica/miditoolbox/

Gómez, E. (2006). Tonal description of music audio signals. PhD diss. Barcelona: Universitat Pompeu Fabra.

Janata, P. (2007). Navigating tonal space. In W. B. Hewlett & E. Selfridge-Field (Eds.), *Tonal Theory for the Digital Age (Computing in Musicology 15)*. Cambridge, Ma: MIT Press.

Koulis, T., Ramsay, J. O. & Levitin, D. J. (2007). From zero to sixty: calibrating real-time responses. *Psychometrika*, 73(2), 321-339.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. New York: Oxford University Press.

Krumhansl, C. L. & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89, 334-368.

Krumhansl, C. L. & Toiviainen, P. (2003). Tonal cognition. In I. Peretz & R. Zatorre (Eds.), *The Cognitive Neuroscience of Music* (pp. 95-108). New York: Oxford University Press.

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception*, 17(4), 481-510.

Lerdahl, F. (2001). *Tonal Pitch Space*. New York: Oxford University Press.

Martorell, A. & Gómez, E. (2011). Two-dimensional visual inspection of pitch-space, many time-scales and tonal uncertainty over time. In C. Agon, E. Amiot, M. Andreatta, G. Assayag, J. Bresson & J. Mandereau (Eds.), *Mathematics and Computation in Music* (pp. 140-150), LNCS/LNAI 6726. Berlin: Springer-Verlag.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11(4), 409-464.

Sapp, C. S. (2005) Visual hierarchical key analysis. Computers in entertainment, 3(4), 1-19.

Schubert, E. (2001). Continuous measurement of self-report emotional responses to music. In P. N. Juslin & J. A. Sloboda (Eds*.), Music and Emotion* (pp. 393-414). Oxford: Oxford University Press.

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

Toiviainen, P. (2008). Visualization of tonal content in the symbolic and audio domains. In W. B. Hewlett & E. Selfridge-Field (Eds.), *Tonal Theory for the Digital Age (Computing in Musicology 15)*. Cambridge, Ma: MIT Press.

Toiviainen, P. & Krumhansl, C. L. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, 32, 741-766.

Vos, P. G. & Leman, M. (Eds.) (2000). Tonality induction. *Music Perception*, 17(4), 401-548.