10 Visualization of Tonal Content in the Symbolic and Audio Domains

Petri Toiviainen
Department of Music
PO Box 35 (M)
40014 University of Jyväskylä
Finland
ptoiviai@campus.jyu.fi

Abstract

Various computational models have been presented for the analysis and visualization of tonality. Some of these models require a symbolic input, such as MIDI, while other models operate with an audio input. The advantage of using a MIDI representation in tonality induction is the explicit representation of pitch it provides. The advantage of the audio representation, on the other hand, is wider availability of musical material and closer correspondence to perception. To obtain a better understanding of tonality perception and computational modeling thereof, it would be crucial to compare analyses of tonality obtained from computational models operating in these two representational domains. This article presents a dynamic model of tonality perception based on a short-term memory model and a self-organizing map (SOM) that operates in both MIDI and audio domains. The model can be used for dynamic visualization of perceived tonal content, making it possible to examine the clarity and locus of tonality at any given point of time. This article also presents a method for the visualization of tonal structure using self-similarity matrices. Two case studies are presented, in which visualizations obtained in the MIDI and audio domains are compared.

Tonal Theory for the Digital Age (Computing in Musicology 15, 2007), 187–199.

10.1 Introduction

Music in many styles is organized around one or more stable reference tones (the tonic, in Western tonal music). This is reflected in Western music theory by the key of the music. Krumhansl and Shepard (1979) introduced the probe-tone technique to investigate one aspect of how a tonal context influences the perception of pitch, in particular the perceived stability of each pitch within a tonal context. The results of these studies were in line with music-theoretic predictions, with the tonic highest in the hierarchy, followed by the third and fifth scale tones, followed by the remaining scale tones, and finally the non-diatonic tones. Pitch-class distributions of various Western musical styles have been found to bear a great similarity to the tonal hierarchies. It has been suggested that listeners acquire the tonal hierarchies by learning these statistical distributions while listening to music (for an opposing view, see Leman 2000). The key-finding algorithm by Krumhansl and Schmuckler (see Krumhansl 1990) is based on the comparison between the pitch-class distribution of the piece under examination and the tonal hierarchies. More specifically, it correlates the pitch-class distribution of the piece with the tone profiles of each of the 24 keys. The key with the highest correlation with the pitch-class distribution is considered to be the key of the piece.

As music unfolds in time, the tonality percept often changes. In particular, the tonality can be clearer at one point than at some other point. Furthermore, a particular piece of music may contain modulations from one key to another. These changes in perceived tonality may be important in the creation of expectancies and tension.

Toiviainen and Krumhansl (2003) introduced a method for quantifying the temporal evolution of tonality percept. In this method, referred to as the continuous probe-tone method, listeners were presented with a piece of music to one ear and a continuously sounding probe tone to the other ear. The listeners' task was to rate the degree to which the probe tone fitted the music at each point in time. The process was repeated using as probe tones each tone of the chromatic scale. This yielded a dynamically changing 12-dimensional stability profile. This dynamic process was modeled with a system consisting of a model of short-term memory and a self-organizing map (SOM; Kohonen 1997). The output of the model was found to correlate significantly with the subjects' ratings obtained by the continuous probe-tone method.

A number of computational models of tonality induction have been presented (for an overview, see Krumhansl 2004). A fundamental distinction can be made within the models based on the kind of representation of music they assume. More specifically, some of these models require a symbolic input, such as a MIDI file, while other models operate with an audio input. The advantage of using a MIDI representation in tonality induction is the explicit representation of pitch it provides. The advantage of the audio representation, on the other hand, is wider availability of musical material and closer correspondence to perception. To obtain a better understanding of tonality perception and the computational modeling thereof, it would be crucial to compare analyses of tonality obtained from computational models operating in these two representational domains.

The model presented in this article can accept both MIDI and audio input, therefore allowing the comparison of tonality visualizations obtained from these two representational domains. In what follows, the model is first described. Subsequently, it is applied to the MIDI and audio representations of F. Chopin's *Prelude in A♭ Major, Op. 28, No. 17*, and O. Messiaen's *Vingt regards sur l'enfant Jésus: Regard IV*. Visualizations of tonal structure of these compositions, obtained from MIDI and audio representations, are compared.

10.2 Self-Organizing Map

The SOM is an artificial neural network that simulates the formation of ordered feature maps. It consists of a two-dimensional grid of units, each of which is associated with a reference vector. Through repeated exposure to a set of input vectors, the SOM settles into a configuration in which the reference vectors approximate the set of input vectors according to some similarity measure; the most commonly used similarity measures are the Euclidean distance and the direction cosine. The direction cosine between an input vector $\mathbf{x}$ and a reference vector $\mathbf{m}$ is defined by

$$\cos\theta = \frac{\sum_i x_i m_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i m_i^2}} = \frac{\mathbf{x}\cdot\mathbf{m}}{\|\mathbf{x}\|\|\mathbf{m}\|}. \qquad (1)$$

Another important feature of the SOM is that its configuration is organized in the sense that neighboring units have similar reference vectors. For a trained SOM, a mapping from the input space onto the two-dimensional grid of units can be defined by associating any given input vector with the unit whose reference vector is most similar to it. Because of the organization of the reference vectors, this mapping is smooth in the sense that similar vectors are mapped onto adjacent regions. Conceptually, the mapping can be thought of as a projection onto a non-linear surface determined by the reference vectors.

10.3 Dynamic Model of Tonality

10.3.1 Representation of Pitch-Class Content

The pitch-class content of a given analysis window can be easily computed from a MIDI representation by applying a mod 12 operator to the note number values and summing the total duration of notes belonging to each modulo class. This leads to a 12-component vector indicating the prevalence of each pitch-class within the window; this vector is subsequently referred to as the pitch-class distribution. If the input consists of audio, the chromagram provides a similar kind of representation (e.g., Gómez and Bonada 2005). The chromagram can be calculated, for instance, by estimating the amplitude spectrum of the windowed signal with the FFT transform, and summing for each pitch-class the amplitude of the bins of the spectrum whose frequencies correspond to that particular pitch-class. Alternatively, it can be calculated using a constant-Q filterbank with semitone spacing between adjacent filters, and summing the power of the outputs of the filters whose center frequencies correspond to the same pitch-class. It must be noted that, because of the contribution of the overtones, the chromagram is not an exact representation of the pitch-class content of the signal.

With both MIDI and audio input, the pitch-class content analysis is carried out using a short sliding window; the exact length of the window is not crucial as long as it is sufficiently small (i.e., of the order of 100 ms).

10.3.2 Short-Term Memory Model

Regardless of the representational domain, the short-term memory is implemented as a bank of twelve leaky integrators, each representing one pitch-class, and at each given point of time contains information about recent pitch-class content in the music. The length of the memory is determined by the time constant of the leaky integrators. For details about the short-term memory model, see Toiviainen & Krumhansl (2003).

10.3.3 Long-Term Memory Model

To create a long-term memory model, a SOM of 36 by 24 units was first trained. For MIDI input, the training set consisted of the 24 K-K profiles. For audio input, the contribution of overtones in the chromagram was modeled assuming a simple exponential relationship between the amplitudes of overtones, a_i=0.8^{i-1}, where a_i, i = 1,...,6, denotes the amplitude of overtone i, and performing a cyclic convolution of each of the K-K profiles with the chromagram of a modeled single tone. Regardless of the set of vectors used in training, the final configuration of the map is similar in terms of key relationships. The SOM is specified in advance to have a toroidal configuration, that is, the left and right edges of the map are connected to each other, as are the top and bottom edges. This choice is based on the fact that octave equivalence implies circularity of pitch. The resulting map is displayed in Figure 10.1. The map shows the units with reference vectors that correspond to the K-K profiles.

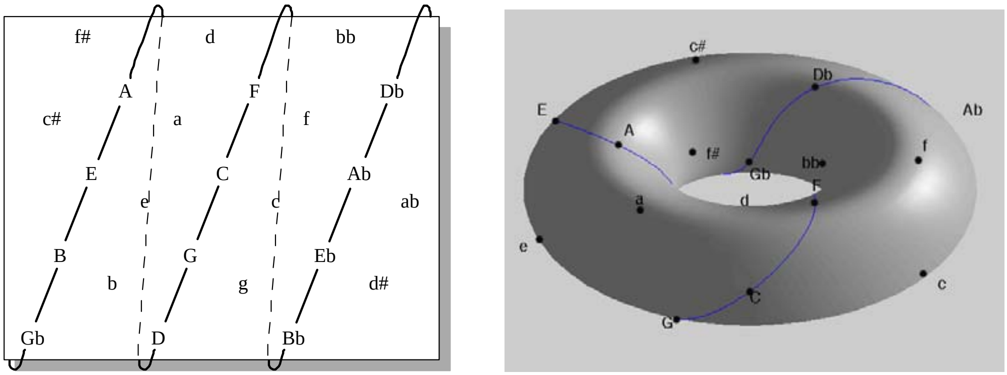

Figure 10.1. Structure of a self-organizing map trained with the tonal hierarchies (original or modified) of the 24 keys (12 major and 12 minor). The subfigure on the left depicts the map in two dimensions (opposite edges are considered to be joined to each other); the subfigure on the right depicts the map in three dimensions.

As can be seen, the configuration of the map corresponds to music-theoretic notions. For instance, keys that are a perfect fifth apart (e.g., C and G) are proximally located, as are relative (e.g., C and a) as well as parallel (e.g., C and c) keys.

10.3.4 Activation Pattern on the SOM

In the trained SOM, a distributed mapping of tonality is defined by associating each unit with an activation value. For each unit, this activation value depends on the similarity between the input vector and the reference vector of the unit. Specifically, the units whose reference vectors are highly similar to the input vector have a high activation, and vice versa. The activation value of each unit can be calculated, for instance, using the direction cosine of Equation 1. The location and spread of this activation pattern provides information about the perceived key and its strength. More specifically, a focused activation pattern implies a strong sense of key and vice versa. Figure 10.2 displays examples of activation patterns on the SOM.

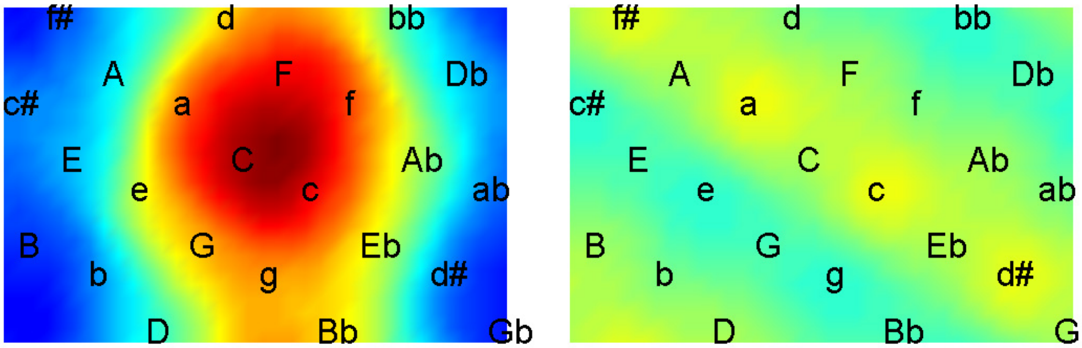

Figure 10.2. Two activation patterns of a SOM evoked by short-term pitch-class memory. Left: clear tonality at the vicinity of C major. Right: unclear tonality.

As time goes by, the contents of the short-term memory constantly change as new notes are being played. As a consequence, the activation pattern of the SOM also changes.

10.2.5 Visualizing Tonal Self-Similarity

Structural features within a piece of music have been visualized with a self-similarity matrix (e.g., Foote, Cooper, and Nam 2002), a matrix that shows the degree of similarity between different parts of a musical piece. Let $\mathbf{v}_i$ denote a vector representing any musical feature at instant *i*. The self-similarity matrix $\mathbf{M} = (m_{ij})$ is defined as

$$m_{ij} = s(\mathbf{v}_i, \mathbf{v}_j), \qquad (2)$$

where *s* denotes any similarity measure. By definition, the matrix is symmetrical across its diagonal. Figure 10.3 illustrates schematically the calculation of a self-similarity matrix.

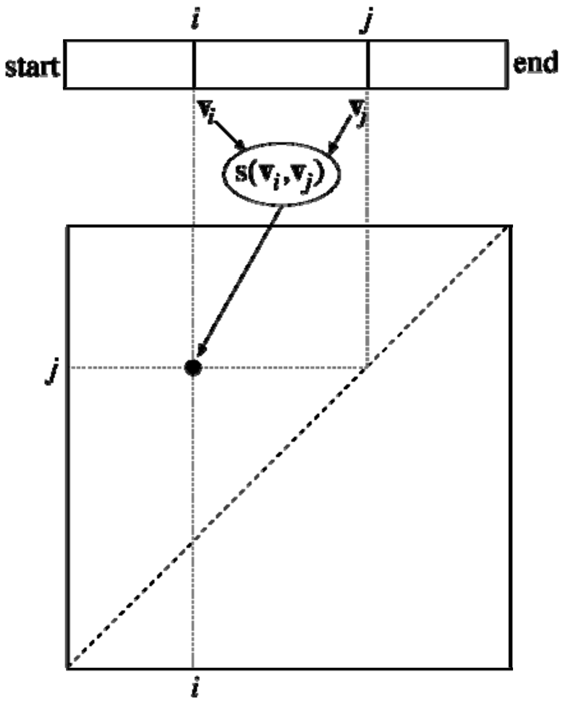

Figure 10.3. *Calculation of a self-similarity matrix.*

To visualize tonal structure, the similarity matrix was in the subsequent analyses derived from the activation patterns of the SOM. The similarity measure used was the negative of the city-block distance,

$$s(\mathbf{v}_i, \mathbf{v}_j) = -\sum_k \left| v_{ik} - v_{jk} \right| \qquad (3)$$

where v_{ik} denotes the activation value of unit *k* in the activation pattern calculated at instant *i*. The contents of a self-similarity matrix can be visualized as a square using different colors to indicate different degrees of similarity. In the present paper, the matrices are visualized so that bright shades of gray stand for high degrees of similarity and dark shades for low degrees of similarity.

10.4 Case Studies

In what follows, the dynamic model of tonality is applied to two pieces of music. These are *Prélude No. 17 in A♭ Major* by F. Chopin and *Vingt regards sur l'enfant Jésus: Regard IV* by Olivier Messiaen. In both cases, three kinds of input are used: (1) MIDI file, (2) audio input rendered from the MIDI file, and (3) audio recording of a musical performance. The output of the SOM and the obtained self-similarity matrices are compared among these input types. In all simulations the time constant of the short-term memory was set to 3 seconds, because this value has been found to yield the best match with behavioral results (see Toiviainen & Krumhansl 2003).

10.4.1 Chopin: Prelude No. 17 in A♭ Major

Figure 10.4 gives some idea of the tonal vocabulary of the Chopin *Prélude*.

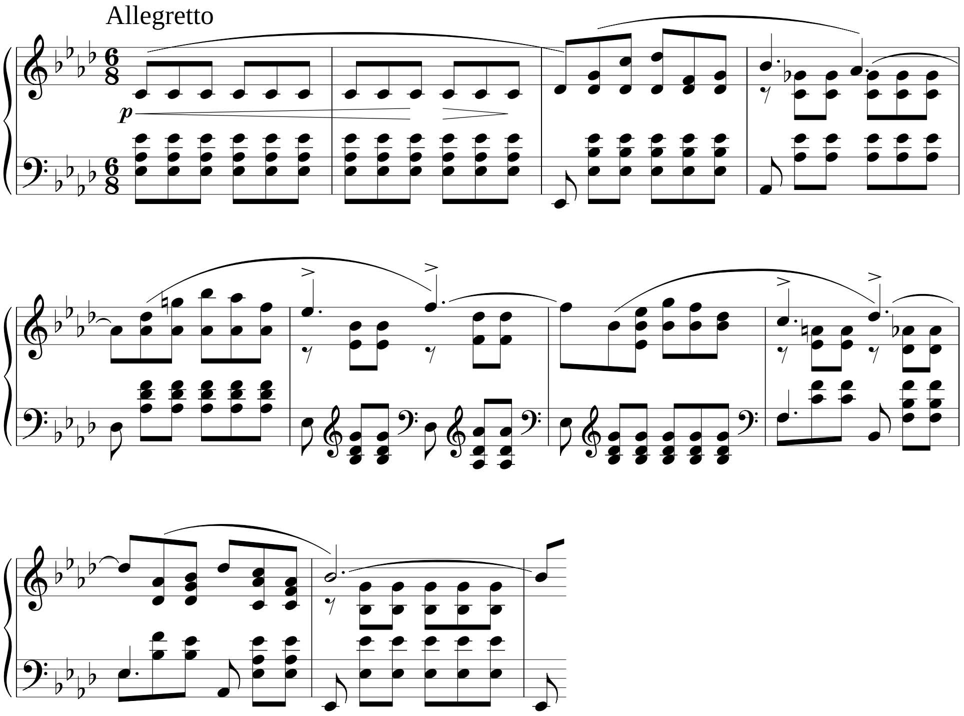

Figure 10.4. *First ten bars of Chopin's Prélude in A♭ Major, Op. 28, No. 17.*

Figure 10.5 shows the activation patterns on the SOM using three different input types and four different sections of the piece as input. The activation patterns obtained from the MIDI file, the audio file rendered from the MIDI file, and the audio recording[1] are displayed in the top, middle and bottom rows, respectively. The four columns in the figure correspond, from left to right, to sections at 0–7, 33–40, 75–83, and 83–90 seconds from the beginning of the recording, and the respective sections in the MIDI file and the rendered audio. These particular sections were chosen because they represent a wide range of tonalities within the composition.

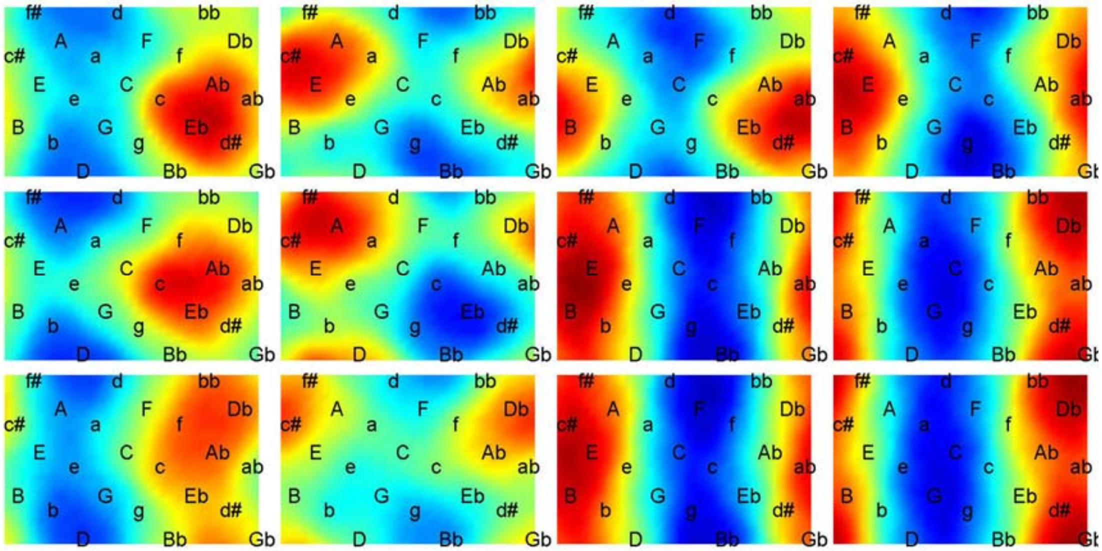

Figure 10.5. *Activation patterns of a SOM of keys evoked by F. Chopin's Prélude No. 17 in A♭ Major, and obtained from a MIDI representation (top row), an audio representation rendered from a MIDI file (middle row), and an audio recording of the composition (bottom row). The four columns correspond to different sections in the piece (see text). Bright shades of gray correspond to a high degree of activation on the SOM.*

Overall, the activation patterns derived from the three different representations appear as similar, suggesting that analyses of tonality from an audio representation yield results similar to those obtained from a MIDI representation and thus correspond to a certain degree with results obtained from listening tests (see Toiviainen and Krumhansl 2003). On a more detailed level, the activation patterns obtained from the two audio representations of the piece of music seem to be more similar to each other than to the one obtained from the MIDI representation.

A global view of the tonal structure can be obtained by calculating self-similarity matrices from the activation patterns. These are displayed in Figure 10.6, using a window length of 3 seconds in the analyses. As can be seen, the self-similarity matrices bear a great degree of similarity to each other, suggesting that the particular representation of music (i.e., MIDI vs. audio) used in such structural analysis of tonality may not be critical.

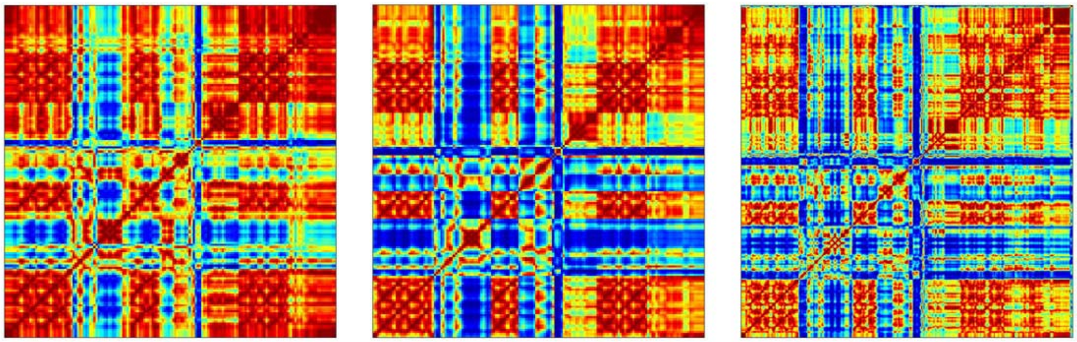

Figure 10.6. *Self-similarity matrices calculated from the activation patterns of the SOM for F. Chopin's Prélude No. 17 in A ♭ Major using different music representations. Left: MIDI input. Middle: audio rendered from MIDI. Right: audio recording of a performance. Bright shades of gray denote a high degree of similarity.*

10.4.2 Messiaen: Vingt regards sur l'enfant Jésus: Regard IV

Compared to the Chopin *Prélude*, the rate of harmonic change is much more rapid in Messiaen's *Regard IV*, the first five bars of which are shown in Figure 10.7.

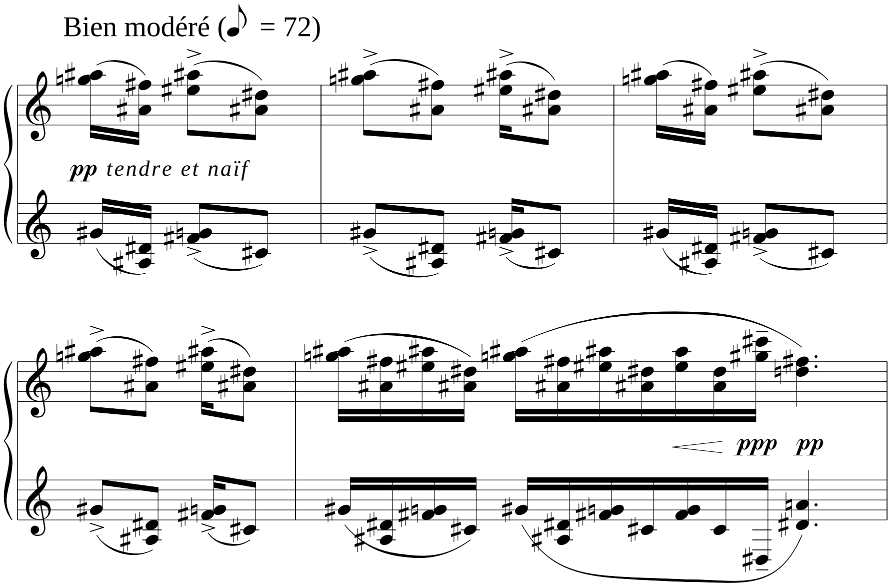

Figure 10.7. *The opening bars of Regard IV from Messiaen's Vingt regards sur l'enfant Jésus.*

Figure 10.8 displays the activation patterns on the SOM using three different input types and four different sections of the piece as input. The activation patterns obtained from the MIDI file, the audio file rendered from the MIDI file, and the audio recording[2] are displayed in the top, middle and bottom rows, respectively. The four columns in the figure correspond, from left to right, to sections at 0–5, 15–17, 55–60, and 60–66 seconds from the beginning of the recording, and the respective sections in the MIDI file and the rendered audio. Again, these particular sections were chosen because they represent a wide range of tonalities within the composition.

As can be seen, there is more difference in terms of the activation patterns between the forms of music representation than in the previous composition by Chopin. This might be due to the fact that *Vingt regards sur l'enfant Jésus: Regard IV* has, overall, a less clear tonality than *Prélude No. 17*, and in such cases the resulting activation pattern might be more dependent on the particular representation of music used.

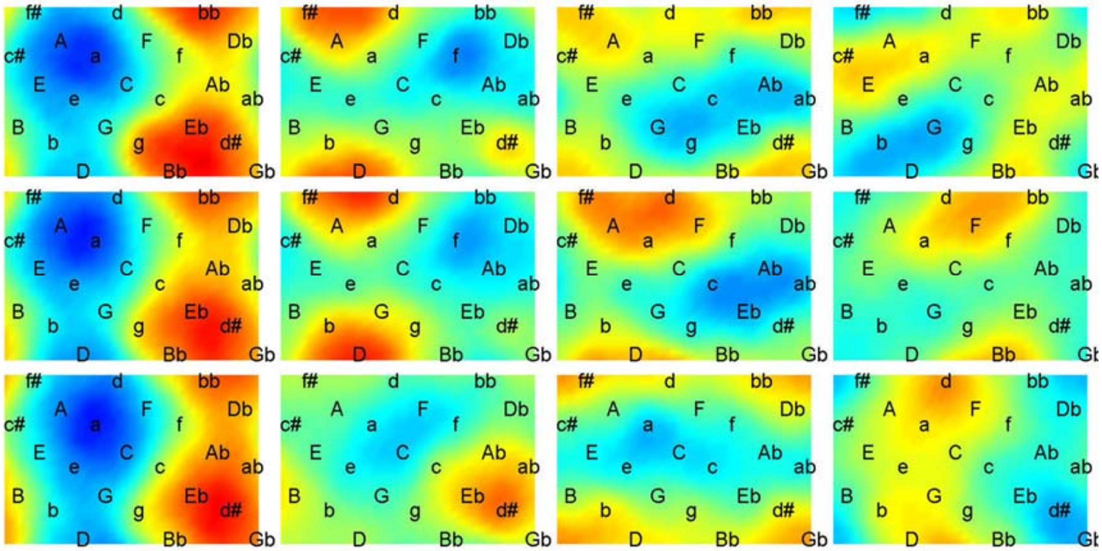

Figure 10.8. *Activation patterns of a SOM of keys evoked by Messiaen's Vingt regards sur l'enfant Jésus: Regard IV, and obtained from a MIDI representation (top row), an audio representation rendered from a MIDI file (middle row), and an audio recording of the composition (bottom row). The four columns correspond to different sections in the piece (see text). Bright shades of gray correspond to a high degree of activation on the SOM.*

Again, a global view of the tonal development can be obtained with the self-similarity matrices (see Figure 10.7). As in the previous example, the length of the analysis window is 4 seconds. Although the activation patterns depicted in Figure 10.9 vary across different music representations, the self-similarity matrices of Figure 10.10 display a strikingly similar structure. This may suggest that, although the visualization of instantaneous tonal content with the method described here may depend on the particular music representation used, the representation of tonal structure by means of self-similarity matrices is more robust in this respect.

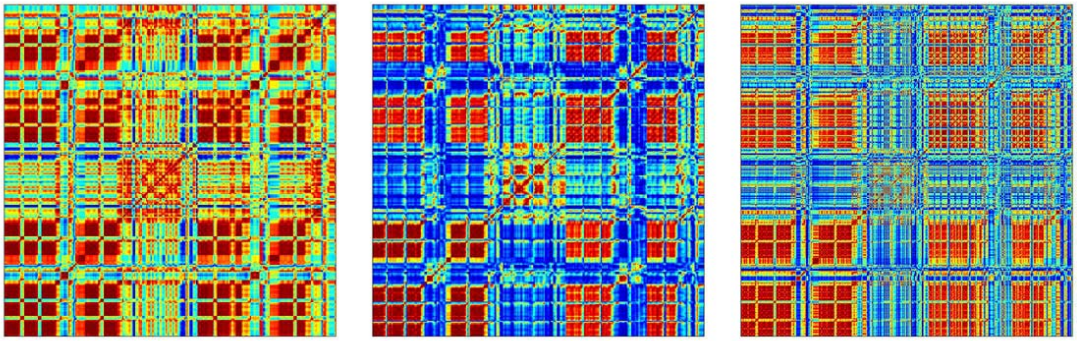

Figure 10.9. *Self-similarity matrices calculated from the activation patterns of the SOM for Messiaen’s* Vingt regards pour l'enfant Jésus: Regard IV *using different music representations. Left: MIDI input. Middle: audio rendered from MIDI. Right: audio recording of a performance. Bright shades of gray denote a high degree of similarity.*

10.5 Tonality Visualization Software

The visualizations above were created with the MIDI Toolbox (Eerola and Toiviainen 2004) and the MIR Toolbox (Lartillot and Toiviainen 2007), which are collections of MATLAB functions for the analysis, visualization, and manipulation of MIDI and audio files, respectively. The author has also implemented an application of the model, called *AudioKeySOM*, which allows real-time visualization of tonal content from various kinds of audio input, such as microphone, line-in, or an audio file. Currently, this software supports only Mac OS. Figure 10.10 displays a screenshot of the *AudioKeySOM* application. MIDI Toolbox, MIR Toolbox and *AudioKeySOM* are freely downloadable at *http://www.jyu.fi/music/coe/materials*.

10.6 Conclusion

This article has presented a model for the visualization of tonality and investigated outputs produced by it using two kinds of music representation, MIDI and audio. The examples presented above suggest that the two representations yield relatively similar visualizations of instantaneous tonality as activation patterns on the SOM. With tonally less clear material, however, greater differences in the activation patterns were observed. When the tonal structure is visualized using a self-similarity matrix calculated from the activation patterns of the SOM, the presented examples suggest a relatively minor dependence on the particular music representation used, suggesting that this visualization method is robust with respect to the representational domain. As these observations are based on only a few examples, it is evident that more research is needed to corroborate them.

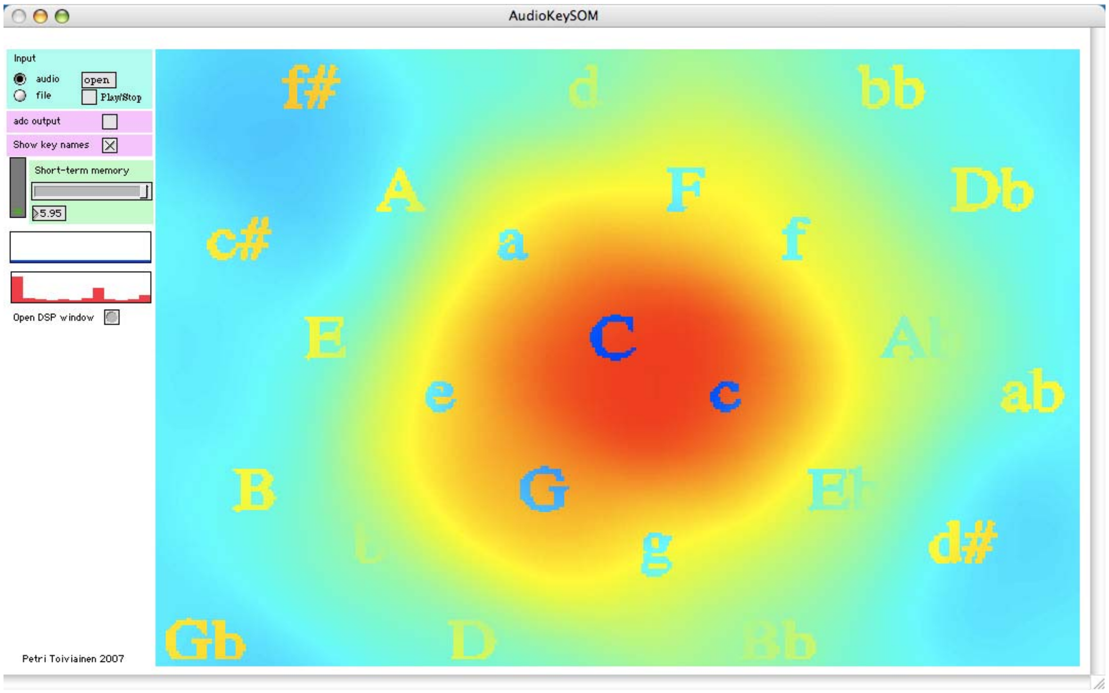

Figure 10.10. *A screenshot of the AudioKeySOM application.*

The *AudioKeySOM* application has a number of possible uses. For instance, it could be used in education as a tool for teaching concepts of tonality. Further, it could be used for artistic purposes as a means for adding a visual element to musical performances that is controlled by the tonal structure of music.

Notes

1. The recording was played by Philippe Giusiano, from the CD *Chopin: Préludes op. 28 et Sonate en Si Mineur op. 58*, published by Alphée.

2. The recording was played by Pierre-Laurent Aimard on the CD *Messiaen: Vingt regards sur l'enfant Jésus*, published by Teldec Classics.

References

Eerola, Tuomas, and Petri Toiviainen (2004). *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä: Kopijyvä, Jyväskylä, Finland. Available at *http://www.jyu.fi/music/coe/materials*.

Foote, Jonathan, Matt Cooper, and Unjung Nam (2002). "Audio Retrieval by Rhythmic Similarity" in *Proceedings of the [Third] International Conference on Music Information Retrieval*, Paris. Available at *http://www.ismir.net/proceedings/*.

Gómez, Emilia, and Jordi Bonada (2005). "Tonality Visualization of Polyphonic Audio," *Proceedings of International Computer Music Conference* 2005. Available at *http://www.iua.upf.edu/mtg/publications/9d0455-ICMC2005-GomezBonada.pdf*.

Kohonen, T. (1997). *Self-Organizing Maps*. Berlin: Springer-Verlag.

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

Krumhansl, Carol L., and Roger N. Shepard (1979). "Quantification of the Hierarchy of Tonal Functions within a Diatonic Context," *Journal of Experimental Psychology: Human Perception and Performance* 5, 579–94.

Krumhansl, Carol. L. (2004). "The Cognition of Tonality—As We Know It Today," *Journal of New Music Research* 33/3, 253–68.

Lartillot, Olivier, and Petri Toiviainen. (2007). "MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio," in *Proceedings of the International Conference on Music Information Retrieval*, Vienna. Available at *http://www.ismir.net/proceedings/*.

Leman, Marc (2000). "An Auditory Model of the Role of Short-Term Memory in Probe-Tone Ratings," *Music Perception* 7/4, 435–63.

Toiviainen, Petri, and Carol L. Krumhansl (2003). "Measuring and Modeling Real-Time Responses to Music: The Dynamics of Tonality Induction," *Perception* 2/6, 741–66.

Submitted: 15 August 2007. Final revisions: 30 November 2007.