Towards a Cognitive Architecture for Music Perception

Antonio Chella

Department of Chemical, Management, Computer, Mechanical Engineering
University of Palermo, Viale delle Scienze, building 6
90128 Palermo, Italy, antonio.chella@unipa.it

Abstract. The framework of a cognitive architecture for music perception is presented. The architecture extends and completes a similar architecture for computer vision developed during the years. The extended architecture takes into account many relationships between vision and music perception. The focus of the architecture resides in the intermediate area between the subsymbolic and the linguistic areas, based on conceptual spaces. A conceptual space for the perception of notes and chords is discussed along with its generalization for the perception of music phrases. A focus of attention mechanism scanning the conceptual space is also outlined. The focus of attention is driven by suitable linguistic and associative expectations on notes, chords and music phrases. Some problems and future works of the proposed approach are also outlined.

1 Introduction

Gärderfors [1], in his paper on "Semantics, Conceptual Spaces and Music" discusses a program for musical spaces analysis directly inspired to the framework of vision proposed by Marr [2]. More in details, the first level that feeds input to all the subsequent levels is related with *pitch identification*. The second level is related with the identification of *musical intervals*; this level takes also into account the cultural background of the listener. The third level is related with *tonality*, where scales are identified and the concepts of chromaticity and modulation arise. The fourth level of analysis is related with the interplay of pitch and time. According to Gärdenfors, time is concurrently processed by means of different levels related with *temporal intervals*, *beats*, *rhythmic patterns*, and at this level the analysis of pitch and the analysis of time merge together.

The correspondences between vision and music perception have been discussed in details by Tanguiane [3]. He considers three different levels of analysis distinguishing between statics and dynamics perception in vision and music. The first visual level in statics perception is the level of pixels, in analogy of the image level of Marr, that corresponds to the perception of *partials* in music. At the second level, the perception of simple patterns in vision corresponds to the perception of single *notes*. Finally at the third level, the perception of structured patterns (as patterns of patterns), corresponds to the perception of *chords*. Concerning dynamic perception, the first level is the same as in the case of static perception, i.e., pixels vs. partials, while at the second level the perception of visual objects corresponds to the perception of musical notes, and at the third final level the perception of visual trajectories corresponds to the perception of music *melodies*.

Several cognitive models of music cognition have been proposed in the literature based on different symbolic or subsymbolic approaches, see Pearce and Wiggins [4] and Temperley [5] for recent reviews. Interesting systems, representative of these approaches are: MUSACT [6][7] based on various kinds of neural networks; the IDyOM project based on probabilistic models of perception [4][8][9]; the Melisma system [10] based on preference rules of symbolic nature; the HARP system, aimed at integrating symbolic and subsymbolic levels [11][12].

Here, we sketch a cognitive architecture for music perception that extends and completes an architecture for computer vision developed during the years. The proposed cognitive architecture integrates the symbolic and the sub symbolic approaches and it has been employed for static scenes analysis [13][14], dynamic scenes analysis [15], reasoning about robot actions [16], robot recognition of self [17] and robot self-consciousness [18]. The extended architecture takes into account many of the above outlined relationships between vision and music perception.

In analogy with Tanguiane, we distinguish between "static" perception related with the perception of chords in analogy with perception of static scenes, and "dynamic" perception related with the perception of musical phrases, in analogy with perception of dynamic scenes.

The considered cognitive architecture for music perception is organized in three computational areas - a term which is reminiscent of the cortical areas in the brain - that follows the Gärdenfors theory of *conceptual spaces* [19] (see Forth et al. [20] for a discussion on conceptual spaces and musical systems).

In the following, Section 2 outlines the cognitive architecture for music perception, while Section 3 describes the adopted music conceptual space for the perception of tones. Section 4 presents the linguistic area of the cognitive architecture and Section 5 presents the related operations of the focus of attention. Section 6 outlines the generalization of the conceptual space for tones perception to the case of perception of music phrases, and finally Section 7 discusses some problems of the proposed approach and future works.

2 The Cognitive Architecture

The proposed cognitive architecture for music perception is sketched in Figure 1. The areas of the architecture are concurrent computational components working together on different commitments. There is no privileged direction in the flow of information among them: some computations are strictly bottom-up, with data flowing from the subconceptual up to the linguistic through the conceptual area; other computations combine top-down with bottom-up processing.

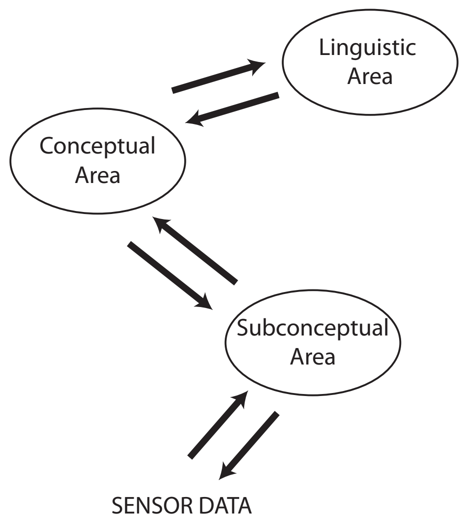

Fig. 1. A sketch of the cognitive architecture.

The *subconceptual* area of the proposed architecture is concerned with the processing of data directly coming from the sensors. Here, information is not yet organized in terms of conceptual structures and categories. In the *linguistic* area, representation and processing are based on a logic-oriented formalism.

The *conceptual* area is an intermediate level of representation between the subconceptual and the linguistic areas and based on conceptual spaces. Here, data is organized in conceptual structures, that are still independent of linguistic description. The symbolic formalism of the linguistic area is then interpreted on aggregation of these structures.

It is to be remarked that the proposed architecture cannot be considered as a model of human perception. No hypotheses concerning its cognitive adequacy from a psychological point of view have been made. However, various cognitive results have been taken as sources of inspiration.

3 Music Conceptual Space

The conceptual area, as previously stated, is the area between the subconceptual and the linguistic area, and it is based on conceptual spaces. We adopt the term *knoxel* (in analogy with the term *pixel*) to denote a point in a conceptual space CS. The choice of this term stresses the fact that a point in CS is the knowledge primitive element at the considered level of analysis.

The conceptual space acts as a workspace in which low-level and high-level processes access and exchange information respectively from bottom to top and from top to bottom. However, the conceptual space has a precise geometric structure of metric space and also the operations in CS are geometric ones: this structure allows us to describe the functionalities of the cognitive architecture in terms of the language of geometry.

In particular, inspired by many empirical investigations on the perception of tones (see Oxenham [21] for a review) we adopt as a knoxel of a *music conceptual space* the set of partials of a perceived tone. A knoxel $\mathbf{k}$ of the music CS is therefore a vector of the main perceived partials of a tone in terms of the Fourier Transform analysis. A similar choice has been carried out by Tanguiane [3] concerning his proposed *correlativity* model of perception.

It should be noticed that the partials of a tone are related both with the pitch and the timbre of the perceived note. Roughly, the fundamental frequency is related with the pitch, while the amplitude of the remaining partials are also related with the timbre of the note. By an analogy with the case of static scenes analysis, a knoxel changes its position in CS when a perceived 3D primitive changes its position in space or its shape [13]; in the case of music perception, the knoxel in the music CS changes its position either when the perceived sound changes its pitch or its timbre changes as well. Moreover, considering the partials of a tone allows us to deal also with microtonal tones, trills, embellished notes, rough notes, and so on.

A *chord* is a set of two or more tones perceived at the same time. The chord is treated as a complex object, in analogy with static scenes analysis where a complex object is an object made up by two or more 3D primitives. A chord is then represented in music CS as the set of the knoxels $[\mathbf{k}_a, \mathbf{k}_b, \dots]$ related with the constituent tones. It should be noticed that the tones of a chord may differ not only in pitch, but also in timbre. Figure 2 is an evocative representation of a chord in the music CS made up by knoxel $\mathbf{k}_a$ corresponding to tone C and knoxel $\mathbf{k}_b$ corresponding to the tone G.

In the case of perception of complex objects in vision, their mutual positions and shapes are important in order to describe the perceived object: e.g., in the case of an hammer, the mutual positions and the mutual shapes of the handle and the head are obviously important to classify the composite object as an hammer. In the same way, the mutual relationships between the pitches (and the timbres) of the perceived tones are important in order to describe the perceived chord. Therefore, spatial relationships in static scenes analysis are in some sense analogous to sounds relationships in music CS.

It is to be noticed that this approach allows us to represent a chord as a set of knoxels in music CS. In this way, the cardinality of the conceptual space does not change with the number of tones forming the chord. In facts, all the tones of the chord are perceived at the same time but they are represented as different points in the same music CS; that is, the music CS is a sort of *snapshot* of the set of the perceived tones of the chord.

In the case of a temporal progression of chords, a scattering occur in the music CS: some knoxels which are related with the same tones between chords will remain in the same position, while other knoxels will change their position in CS, see Figure 3 for an evocative representation of scattering in the music CS. In the figure, the knoxels $\mathbf{k}_a$, corresponding to C, and $\mathbf{k}_b$, corresponding to E,

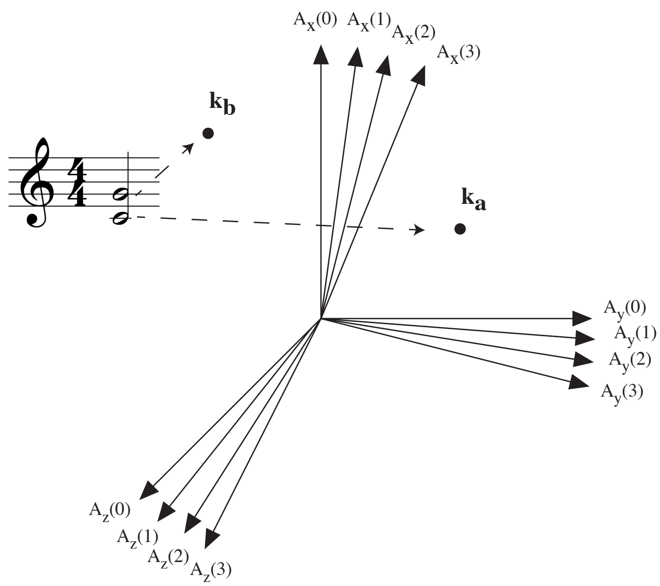

Fig. 2. An evocative representation of a chord in the *music conceptual space*.

change their position in the new chord: they becomes A and D, while knoxel $\mathbf{k}_c$, corresponding to G, maintains its position. The relationships between mutual positions in music CS could then be employed to analyze the chords progression and the relationships between subsequent chords.

A problem may arise at this point. In facts, in order to analyze the progression of chords, the system should be able to find the correct correspondences between subsequent knoxels: i.e., $\mathbf{k}'_a$ should correspond to $\mathbf{k}_a$ and not to, e.g., $\mathbf{k}_b$. This is a problem similar to the *correspondence* problem in stereo and in visual motion analysis: a vision system analyzing subsequent frames of a moving object should be able to find the correct corresponding object tokens among the motion frames; see the seminal book by Ullman [22] or Chap. 11 of the recent book by Szeliski [23] for a review. However, it should be noticed that the expectation generation mechanism described in Section 5 could greatly help facing this difficult problem.

The described representation is well suited for the recognition of chords: for example we may adopt the algorithms proposed by Tanguiane [3]. However, Tanguiane hypothesizes, at the basis of his *correlativity* principle, that all the notes of a chord have the same shifted partials, while we consider the possibility that a chord could be made by tones with different partials.

The proposed representation is also suitable for the analysis of the efficiency in *voice leading*, as described by Tymoczko [24]. Tymoczko describes a geometrical analysis of chords by considering several spaces with different cardinalities,

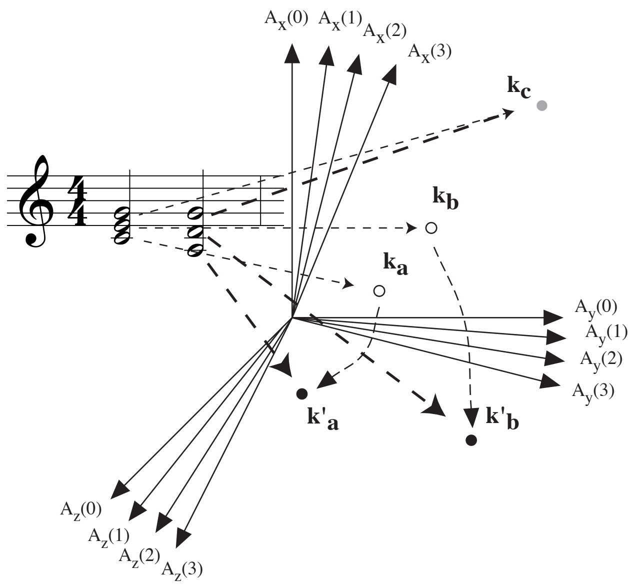

Fig. 3. An evocative representation of a scattering between two chords in the *music conceptual space*.

i.e., a one note circular space, a two note space, a three note space, and so on. Instead, the cardinality of the considered conceptual space does not change, as previously remarked.

4 Linguistic area

In the linguistic area, the representation of perceived tones is based on a high level, logic oriented formalism. The linguistic area acts as a sort of long term memory, in the sense that it is a semantic network of symbols and their relationships related with musical perceptions. The linguistic area also performs inferences of symbolic nature. In preliminary experiments, we adopted a linguistic area based on a hybrid KB in the KL-ONE tradition [25]. A hybrid formalism in this sense is constituted by two different components: a *terminological* component for the description of concepts, and an *assertional* component, that stores information concerning a specific context. A similar formalism has been adopted by Camurri et al. in the HARP system [11][12].

In the domain of perception of tones, the terminological component contains the description of relevant concepts such as chords, tonic, dominant and so on. The assertional component stores the assertions describing specific situations. Figure 4 shows a fragment of the terminological knowledge base along with its mapping into the corresponding entities in the conceptual space.

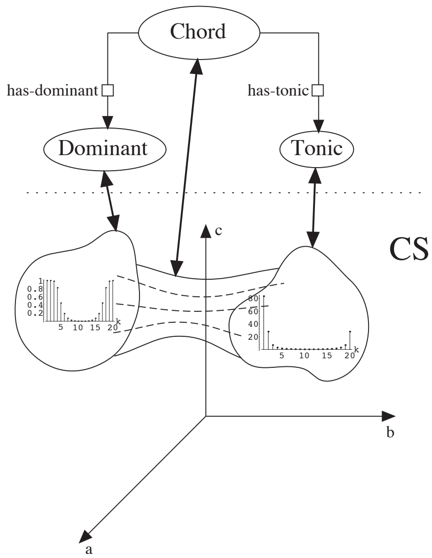

Fig. 4. A fragment of the terminological KB along with its mapping into the conceptual space.

A generic *Chord* is described as composed of at least two knoxels. A *Simple-Chord* is a chord composed by two knoxels; a *Complex-Chord* is a chord composed of more than two knoxels. In the considered case, the concept *Chord* has two roles: a role *has-dominant*, and a role *has-tonic* both filled with specific tones.

In general, we assume that the description of the concepts in the symbolic KB is not exhaustive. We symbolically represent the information necessary to make suitable inferences.

The assertional component contains facts expressed as assertions in a predicative language, in which the concepts of the terminological components correspond to one argument predicates, and the roles (e.g., part_of) correspond to two argument relations. For example, the following predicates describe that the instance `f7#1` of the F7 chord has a dominant which is the constant `ka` corresponding to a knoxel $\mathbf{k}_a$ and a tonic which is the constant `k#b` corresponding to a knoxel $\mathbf{k}_b$ of the current CS:

```
ChordF7(f7#1)
has-dominant(f7#1,ka)
has-tonic(f7#1,kb)
```

By means of the mapping between symbolic KB and conceptual spaces, the linguistic area assigns names (symbols) to perceived entities, describing their structure with a logical-structural language. As a result, all the symbols in the linguistic area find their meaning in the conceptual space which is inside the system itself.

A deeper account of these aspects can be found in Chella et at. [13].

5 Focus of Attention

A cognitive architecture with bounded resources cannot carry out a one-shot, exhaustive, and uniform analysis of the perceived data within reasonable resource constraints. Some of the perceived data (and of the relations among them) are more relevant than others, and it should be a waste of time and of computational resources to detect true but useless details.

In order to avoid the waste of computational resources, the association between symbolic representations and configurations of knoxels in CS is driven by a sequential scanning mechanism that acts as some sort of internal focus of attention, and inspired by the attentive processes in human perception.

In the considered cognitive architecture for music perception, the perception model is based on a focus of attention that selects the relevant aspects of a sound by sequentially scanning the corresponding knoxels in the conceptual space. It is crucial in determining which assertions must be added to the linguistic knowledge base: not all true (and possibly useless) assertions are generated, but only those that are judged to be relevant on the basis of the attentive process.

The recognition of a certain component of a perceived configuration of knoxels in music CS will elicit the *expectation* of other possible components of the same chord in the perceived conceptual space configuration. In this case, the mechanism seeks for the corresponding knoxels in the current CS configuration. We call this type of expectation *synchronic* because it refers to a single configuration in CS.

The recognition of a certain configuration in CS could also elicit the expectation of a scattering in the arrangement of the knoxels in CS; i.e., the mechanism generates the expectations for another set of knoxels in a subsequent CS configuration. We call this expectation *diachronic*, in the sense that it involves subsequent configurations of CS. Diachronic expectations can be related with progression of chords. For example, in the case of jazz music, when the system recognized the $Cmajor$ key (see Rowe [26] for a catalogue of key induction algorithms) and a Dm chord is perceived, then the focus of attention will generate the expectations of G and C chords in order to search for the well known chord progression $ii - V - I$ (see Chap. 10 of Tymoczko [24]).

Actually, we take into account two main sources of expectations. On the one side, expectations could be generated on the basis of the structural information stored in the symbolic knowledge base, as in the previous example of the jazz chord sequence. We call these expectations *linguistic*. Several sources may be taken into account in order to generate linguistic expectations, for example the *ITPRA* theory of expectation proposed by Huron [27], the preference rules systems discussed by Temperley [10] or the rules of harmony and voice leading discussed in Tymoczko [24], just to cite a few. As an example, as soon as a particular configuration of knoxel is recognized as a possible chord filling the role of the first chord of the progression $ii - V - I$, the symbolic KB generates the expectation of the remaining chords of the sequence.

On the other side, expectations could be generated by purely Hebbian, associative mechanisms. Suppose that the system learnt that typically a jazz player adopts the *tritone* substitution when performing the previous described jazz progression. The system could learn to associate this substitution to the progression: in this case, when a compatible chord is recognized, the system will generate also expectations for the sequence $ii - \flat II - I$. We call these expectations *associative*.

Therefore, synchronic expectations refer to the same configuration of knoxels at the same time; diachronic expectations involve subsequent configurations of knoxels. The linguistic and associative mechanisms let the cognitive architecture generate suitable expectations related to the perceived chords progressions.

6 Perception of Music Phrases

So far we adopted a "static" conceptual space where a knoxel represents the partials of a perceived tone. In order to generalize this concept and in analogy with the differences between static and dynamic vision, in order to represent a music *phrase*, we now adopt a "dynamic" conceptual space in which each knoxel represents the whole set of partials of the Short Time Fourier Transform of the corresponding music phrase. In other words, a knoxel in the dynamic CS now represents all the parameters of the *spectrogram* of the perceived phrase.

Therefore, inspired by empirical results (see Deutsch [28] for a review) we hypothesize that a musical phrase is perceived as a whole "Gestaltic" group, in the same way as a movement could be visually perceived as a whole and not as a sequence of single frames. It should be noticed that, similarly to the static case, a knoxel represents the sequence of pitches and durations of the perceived phrase and also its timbre: the same phrase played by two different instruments corresponds to two different knoxels in the dynamic CS.

The operations in the dynamic CS are largely similar to the static CS, with the main difference that now a knoxel is a whole perceived phrase.

A configuration of knoxels in CS occurs when two or more phrases are perceived at the same time. The two phrases may be related with two different sequences of pitches or it may be the same sequence played for example, by two different instruments. This is similar to the situation depicted in Figure 2, where the knoxels $\mathbf{k}_a$ and $\mathbf{k}_b$ are interpreted as music phrases perceived at the same time.

A scattering of knoxels occurs when a change occurs in a perceived phrase. We may represent this scattering in a similar way to the situation depicted in Figure 3, where the knoxels also in this case are interpreted as music phrases: knoxels $\mathbf{k}_a$ and $\mathbf{k}_b$ are interpreted as changed music phrases while knoxels $\mathbf{k}_c$ corresponds to the same perceived phrase.

As an example, let us consider the well known piece *In C* by Terry Riley. The piece is composed by 53 small phrases to be performed sequentially; each player may decide when to start playing, how many times to repeat the same phrase, and when to move to the next phrase (see the performing directions of *In C* [29]).

Let us consider the case in which two players, with two different instruments, start with the first phrase. In this case, two knoxels $\mathbf{k}_a$ and $\mathbf{k}_b$ will be activated in the dynamic CS. We remark that, although the phrase is the same in terms of pitch and duration, it corresponds to two different knoxels because of different timbres of the two instruments. When a player will decide at some time to move to next phrase, a scattering occur in the dynamic CS, analogously with the previous analyzed static CS: the corresponding knoxel, say $\mathbf{k}_a$, will change its position to $\mathbf{k}'_a$.

The focus of attention mechanism will operate in a similar way as in the static case: the synchronous modality of the focus of attention will take care of generation of expectations among phrases occurring at the same time, by taking into account, e.g., the rules of counterpoint. Instead, the asynchronous modality will generate expectations concerning, e.g., the continuation of phrases.

Moreover, the static CS and the dynamic CS could generate mutual expectations: for example, when the focus of attention recognizes a progression of chords in the static CS, this recognized progression will constraint the expectations of phrases in the dynamic CS. As another example, the recognition of a phrase in the dynamic CS could constraint as well the recognition of the corresponding progression of chords in the static CS.

7 Discussion and Conclusions

The paper sketched a cognitive architecture for music perception extending and completing a computer vision cognitive architecture. The architecture integrates symbolic and the sub symbolic approaches by means of *conceptual spaces* and it takes into account many relationships between vision and music perception.

Several problems arise concerning the proposed approach. A first problem, analogously with the case of computer vision, concerns the *segmentation* step. In the case of static CS, the cognitive architecture should be able to segment the Fourier Transform signal coming from the microphone in order to individuate the perceived tones; in the case of dynamic CS the architecture should be able to individuate the perceived phrases. Although many algorithms for music segmentation have been proposed in the computer music literature and some of them are also available as commercial program, as the AudioSculpt program developed by IRCAM[1], this is a main problem in perception. Interestingly, empirical studies concur in indicating that the same Gestalt principles at the basis of visual perception operate in similar ways in music perception, as discussed by Deutsch [28].

The expectation generation process at the basis of the focus of attention mechanism can be employed to help solving the segmentation problem: the linguistic information and the associative mechanism can provide interpretation

[1] http://forumnet.ircam.fr/product/audiosculpt/

contexts and high level hypotheses that help segmenting the audio signal, as e.g., in the IPUS system [30].

Another problem is related with the analysis of *time*. Currently, the proposed architecture does not take into account the metrical structure of the perceived music. Successive development of the described architecture will concern a metrical conceptual space; interesting starting points are the geometric models of metrical-rhythmic structure discussed by Forth et al. [20].

However, we maintain that an intermediate level based on conceptual spaces could be a great help towards the integration between the music cognitive systems based on subsymbolic representations, and the class of systems based on symbolic models of knowledge representation and reasoning. In facts, conceptual spaces could offer a theoretically well founded approach to the integration of symbolic musical knowledge with musical neural networks.

Finally, as stated during the paper, the synergies between music and vision are multiple and multifaceted. Future works will deal with the exploitation of conceptual spaces as a framework towards a sort of *unified* theory of perception able to integrate in a principled way vision and music perception.

References

1. Gärdenfors, P.: Semantics, conceptual spaces andthe dimensions of music. In Rantala, V., Rowell, L., Tarasti, E., eds.: Essays on the Philosophy of Music. Philosophical Society of Finland, Helsinki (1988) 9–27
2. Marr, D.: Vision. W.H. Freeman and Co., New York (1982)
3. Tanguiane, A.: Artificial Perception and Music Recognition. Number 746 in Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin Heidelberg (1993)
4. Wiggins, G., Pearce, M., Müllensiefen: Computational modelling of music cognition and musical creativity. In Dean, R., ed.: The Oxford Handbook of Computer Music. Oxford University Press, Oxford (2009) 387–414
5. Temperley, D.: Computational models of music cognition. In Deutsch, D., ed.: The Psychology of Music. Third edn. Academic Press, Amsterdam, The Netherlands (2012) 327–368
6. Bharucha, J.: Music cognition and perceptual facilitation: A connectionist framework. Music Perception: An Interdisciplinary Journal **5**(1) (1987) 1–30
7. Bharucha, J.: Pitch, harmony and neural nets: A psychological perspective. In Todd, P., Loy, D., eds.: Music and Connectionism. MIT Press, Cambridge, MA (1991) 84–99
8. Pearce, M., Wiggins, G.: Improved methods for statistical modelling of monophonic music. Journal of New Music Research **33**(4) (2004) 367–385
9. Pearce, M., Wiggins, G.: Expectation in melody: The influence of context and learning. Music Perception: An Interdisciplinary Journal **23**(5) (2006) 377–406
10. Temperley, D.: The Cognition of Basic Musical Structures. MIT Press, Cambridge, MA (2001)
11. Camurri, A., Frixione, M., Innocenti, C.: A cognitive model and a knowledge representation system for music and multimedia. Journal of New Music Research **23** (1994) 317–347

12. Camurri, A., Catorcini, A., Innocenti, C., Massari, A.: Music and multimedia knowledge representation and reasoning: the HARP system. Computer Music Journal **19**(2) (1995) 34–58
13. Chella, A., Frixione, M., Gaglio, S.: A cognitive architecture for artificial vision. Artificial Intelligence **89** (1997) 73–111
14. Chella, A., Frixione, M., Gaglio, S.: An architecture for autonomous agents exploiting conceptual representations. Robotics and Autonomous Systems **25**(3-4) (1998) 231–240
15. Chella, A., Frixione, M., Gaglio, S.: Understanding dynamic scenes. Artificial Intelligence **123** (2000) 89–132
16. Chella, A., Gaglio, S., Pirrone, R.: Conceptual representations of actions for autonomous robots. Robotics and Autonomous Systems **34** (2001) 251–263
17. Chella, A., Frixione, M., Gaglio, S.: Anchoring symbols to conceptual spaces: the case of dynamic scenarios. Robotics and Autonomous Systems **43**(2-3) (2003) 175–188
18. Chella, A., Frixione, M., Gaglio, S.: A cognitive architecture for robot self-consciousness. Artificial Intelligence in Medicine **44** (2008) 147–154
19. Gärdenfors, P.: Conceptual Spaces. MIT Press, Bradford Books, Cambridge, MA (2000)
20. Forth, J., Wiggins, G., McLean, A.: Unifying conceptual spaces: Concept formation in musical creative systems. Minds and Machines **20** (2010) 503–532
21. Oxenham, A.: The perception of musical tones. In Deutsch, D., ed.: The Psychology of Music. Third edn. Academic Press, Amsterdam, The Netherlands (2013) 1–33
22. Ullman, S.: The Interpretation of Visual Motion. MIT Press, Cambridge,MA (1979)
23. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer, London (2011)
24. Tymoczko, D.: A Geometry of Music. Harmony and Counterpoint in the Extended Common Practice. Oxford University Press, Oxford (2011)
25. Brachman, R., Schmoltze, J.: An overview of the KL-ONE knowledge representation system. Cognitive Science **9**(2) (1985) 171–216
26. Rowe, R.: Machine Musicianship. MIT Press, Cambridge, MA (2001)
27. Huron, D.: Sweet Anticipation. Music and the Psychology of Expectation. MIT Press, Cambridge, MA (2006)
28. Deutsch, D.: Grouping mechanisms in music. In Deutsch, D., ed.: The Psychology of Music. Third edn. Academic Press, Amsterdam, The Netherlands (2013) 183–248
29. Riley, T.: *In C*: Performing directions. Celestial Harmonies (1964)
30. Lesser, V., Nawab, H., Klassner, F.: IPUS: An architecture for the integrated processing and understanding of signals. Artificial Intelligence **77** (1995) 129–171