Biologically Inspired Cognitive Architectures 25 (2018) 135–140

Contents lists available at ScienceDirect

Biologically Inspired Cognitive Architectures

journal homepage: www.elsevier.com/locate/bica

Research article

Simulating music with associative self-organizing maps

Miriam Buonamente[a], Haris Dindo[a], Antonio Chella[a], Magnus Johnsson[b,c,*]

[a] RoboticsLab, DICGIM, Polytechnic School, University of Palermo, Viale delle Scienze, Ed. 6, 90128 Palermo, Italy
[b] Magnus Johnsson AI Research AB, Kamrersgatan 7, Höör, Sweden
[c] Department of Intelligent Cybernetic Systems, NRNU MEPhI, Moscow, Russia

ARTICLE INFO

Keywords:
Neural network
Associative self-organizing map
Music
Internal simulation

ABSTRACT

We present an architecture able to recognise pitches and to internally simulate likely continuations of partially heard melodies. Our architecture consists of a novel version of the Associative Self-Organizing Map (A-SOM) with generalized ancillary connections. We tested the performance of our architecture with melodies from a publicly available database containing 370 Bach chorale melodies. The results showed that the architecture could learn to represent and perfectly simulate the remaining 20% of three different interrupted melodies when using a context length of 8 centres of activity in the A-SOM. These promising and encouraging results show that our architecture offers something more than what has previously been proposed in the literature. Thanks to the inherent properties of the A-SOM, our architecture does not predict the most likely next pitch only, but rather continues to elicit activity patterns corresponding to the remaining parts of interrupted melodies by internal simulation.

Introduction

Visual and auditory stimuli are important sensory modalities in the human brain. As with the recognition of actions performed by other conspecifics, listening to music is likely to fire two kinds of processes in the brain: first, the recognition and then the internal simulation of the heard melody (Hesslow, 2002). We are in facts able to recognise a musical piece by hearing a few notes only and by internally playing the rest of the song.

Current understanding of music cognition, including music perception and internal music simulation, seems to be less advanced than other areas of human cognition, such as – for example – visual perception. Music is a complex phenomenon, and for this reason many researchers have developed a variety of cognitive models trying to capture it. Though some societies live without writings, and some even without visual arts, there seems to be no human culture without musical arts. Unlike other human activities, such as vision and language, which primarily seem to be based on localised parts of the brain, music seems to involve almost all of the brain (Ball, 2010). Moreover, musical activities involve the whole body and the real-time body coordination with other people, while also taking into account significant perceptual and cognitive demands (Leman, 2007).

Our aim is to develop a cognitive model able to endow an agent with the ability to *recognise* and *simulate* perceived musical stimuli. The architecture we propose here is an extension of our previous efforts in developing a self-organizing system able to anticipate and to understand what is happening in its environment, by providing it with the necessary tools to interact and to efficiently and safely cooperate with mate humans (Buonamente, Dindo, & Johnsson, 2015). To that aim we adopted a novel variant of the Self Organising Map, called the Associative Self-Organising Map (A-SOM), in order to develop an architecture able both to recognise and to simulate observed behaviour. In addition to the advantages offered by the SOM, the A-SOM can remember perceptual sequences by associating the current network activity with its own earlier activity. Due to this ability, the A-SOM can receive initial input and then continue to elicit the likely continuation, i.e. to carry out sequence completion of perceptual activity over time. The developed solution has proven to be able to parsimoniously represent and to efficiently discriminate human actions in real-time (Buonamente, Dindo, & Johnsson, 2013a), as well as to simulate the likely continuation of the recognised actions (Buonamente, Dindo, & Johnsson, 2013b).

The present research is thus aimed at generalising the previous research by using the A-SOM to recognise, and to internally simulate, perceived music. Music presents several and interesting challenges, in particular in terms of data and knowledge representation. According to Conklin's multiple viewpoints representation (Conklin & Witten, 1995), notes or events in a melody (musical sequence) have multiple attributes

* Corresponding author at: Magnus Johnsson AI Research AB, Kamrersgatan 7, Höör, Sweden.
E-mail addresses: antonio.chella@unipa.it (A. Chella), magnus@magnusjohnsson.se (M. Johnsson).

https://doi.org/10.1016/j.bica.2018.07.006
Received 18 June 2018; Accepted 6 July 2018

such as pitch, duration and onset time. These attributes are either observable such as pitch and duration or abstract derived from them such as inter-onset interval and pitch contour. A given piece of music is decomposed into parallel streams of features, known as viewpoint types. We chose to work with an event-based representation extracted from symbolic music data named Kern. The Kern representation is a text-based description used to represent the underlying syntactic information conveyed by a musical score. The scheme allows the encoding of all the attributes that describe a score, such as pitch, duration, accidentals, ties, slurs and many others. Even individual musical parts or voices, typical of musical scores, are represented in the Kern format using different columns; each column of data represents a single part or voice. The Kern format is one of several in the humdrum syntax (Huron, 1997) that provides ease in readability and capabilities of search, comparison and editing.

To implement music prediction, we exploited an analogy with natural language. A statistical language model can be represented by the conditional probability of the next word given all the previous ones. In the note-by-note technique (Todd, 1989), notes are produced sequentially and linearly, from the start of a piece to the end, each note depending on the preceding context. We extended the note-by note technique in order to handle also the case in which no more input is provided. In this case, even if we stopped the execution of a piece, the new generalised A-SOM can predict the future depending on the earlier pitch and the related context.

One of the problems we faced was to find the proper length of the preceding context, i.e. to find the right context that enables the A-SOM to make right predictions, without ambiguity. We initially tried to accomplish the task by using a simplified implementation of the A-SOM. The idea was to use the self-organizing ability of the A-SOM to infer grammatical and semantic structure of the sequence of pitches, and the A-SOM's recurrently connected ancillary connections to simulate the context. Through the ancillary connections, the A-SOM would associate its activity with its own delayed activity enabling the network to remember and complete musical sequences. However, we found that the simplified A-SOM implementation was unable to internally simulate the learnt melodies in a satisfactory way, and hence a stronger mechanism was needed to represent the ambiguity present in music. To solve this problem we employed a novel and more capable version of the A-SOM with generalised ancillary connections, thus improving its capacity to associate ancillary input and activity, together with a context module that represents the sequence of the k last activations in the A-SOM.

The architecture was tested on melodies taken from a corpus of 370 Bach chorale melodies coded in the Kern format publicly available at the website " http://kern.ccarh.org/". The implementation of all code for the experiments presented in this paper was done in C++ using the neural modelling framework "Ikaros" (Balkenius, Morén, Johansson, & Johnsson, 2010).

The remainder of this paper is structured as follows. Section "Related works" presents related work. Section "Proposed architecture" presents the novel version of the A-SOM with generalised ancillary connections and the proposed architecture. Section "Experiment" presents the experiments for evaluating the architecture, and finally the conclusions are outlined in Section "Conclusion".

Related works

The research presented in this paper is concerned with modeling cognitive processes in the perception and simulation of melodies. In particular, the problem studied here is the recognition of listened melodies and the prediction of their likely continuation when no more input is provided to the system.

Music is a complex human phenomenon and many researchers have invested time and effort to develop a variety of models for its understanding. Several computational models of music cognition have been proposed in the literature, see (Wiggins, Pearce, & Müllensiefen, 2009; Temperley, 2012) for reviews. Representative systems are, among others: MUSACT (Bharucha, 1987; Bharucha, 1991) based on various kinds of neural networks; the IDyOM system based on a probabilistic model of music perception (Pearce & Wiggins, 2004; Pearce & Wiggins, 2006; Wiggins et al., 2009); the Melisma system developed by (Temperley, 2001) and based on preference rules of symbolic nature; the HARP system, aimed at integrating symbolic and subsymbolic levels (Camurri, Frixione, & Innocenti, 1994). Kohonen (1989) proposes a generative grammar based on SOMs that derives its grammatical production automatically from examples and optimizes the length of context for each individual production rule on the basis of conflicts occurring in the given examples.

Generative models have been successfully applied for analysis, prediction and generation of new melodies. Conklin developed the Multiple Viewpoint System (Conklin & Witten, 1995) for music prediction. This approach is based on the observation that music is structured at multiple different levels: rhythmic, melodic, and harmonic, and each of these dimensions can be described by different attributes. The multiple viewpoint system thus represents each attribute with a separate statistical model and then it combines all the predictions of all the models. We took cue from the multiple viewpoint idea in order to represent music data for our experiment. The idea of discrete attributes allow us to focus our attention only on the pitch attribute at this stage of our research, leaving for future experiments all the other attributes. A similar approach has been followed by the IDyOM system by employing Markov models (Pearce & Wiggins, 2004; Pearce & Wiggins, 2006).

A connectionist approach allows for flexibility and ability to predict music, which makes it a suitable method for our experiments. Moreover, neural networks overcome the problem related to data sparsity that affects Markov models. Our approach is based on the idea to use a self-recurrent neural network to predict notes (pitches) at times $>t$ using notes until time t as inputs. Similar to our experiment, Todd in Todd (1989) proposes a parallel distributed processing network as solution for the note-by-note prediction task. Todd uses a Jordan recurrent network (Jordan, 1986) to reproduce classical songs and then to produce new songs. The outputs are recurrently fed back as inputs. In addition, self-recurrence on the inputs provides a decaying history of these inputs. Mozer and Soukup (1990) developed CONCERT, a back-propagation-through-time (BPTT) recurrent network, which uses a set of melodies written in a given style to compose new melodies in that style. CONCERT is an extension of a traditional algorithmic composition technique in which transition tables specify the probability of the next note as a function of previous context.

Proposed architecture

We propose a novel approach based on a new version of the A-SOM for the problem of perception and simulation of music. Our proposed system consists of this variant of the A-SOM with generalised ancillary connections and a Context Module, see Fig. 1.

The role of the Context Module is to represent the sequence of the k last activations (each representing a pitch) in the A-SOM. The A-SOM uses the so created context as ancillary input to elicit an activity pattern likely to follow the previous sequence of pitches without actually receiving any input. This internally simulated activity, representing the likely next pitch, is then used to create a new context, which will again elicit a simulated activity representing another pitch in the A-SOM and so on.

The new version of the A-SOM substitutes the original ancillary connections with a feed forward neural network which uses the generalised delta rule to adapt the weights. As the original A-SOM, the new version develops an ordered representation of the input space (in this case a pitch representation) by unsupervised learning, while simultaneously self-supervising the adaptation of its ancillary connections (which associates the context sequences with the activity in the A-SOM that represents the next pitch of the melody). In this way, the A-SOM based system is able to simulate the likely continuation of a melody.