THE SMUSE: A SITUATED COGNITION APPROACH TO INTERACTIVE MUSIC COMPOSITION

Sylvain Le Groux
GTCMT
Georgia Tech, Atlanta

Paul F.M.J. Verschure
SPECS, UPF, Barcelona
ICREA, Barcelona

ABSTRACT

The evolution of computer-based music systems has gone from computer-aided composition, which transposed the traditional paradigms of music composition to the digital realm, to complex feedback systems that allow for rich multimodal interactions. Yet, a lot of interactive music systems still rely on outdated principles in the light of modern situated cognitive systems design. Moreover, the role of human emotional feedback, arguably an important feature of musical experience, is rarely taken into account into the interaction loop. We propose to address these limitations by introducing a novel situated synthetic interactive composition framework called the SMuSe (for Situated Music Server). The SMuSe is based on the principles of parallelism, situatedness, emergence and emotional feedback and is built on a cognitively plausible architecture. It addresses questions at the intersection of music perception and cognition while being used as a creative tool for interactive music composition.

1. BACKGROUND

Interactivity has now become a standard feature of many multimedia systems and plays a fundamental role in contemporary art practice. Specifically, real-time human/machine interactive music systems are now omnipresent as both composition and live performance tools. Yet, the term “interactive music system” is often misused. The interaction that takes place between a human and a system is a process that includes both control and feedback, where the real-world actions are interpreted into the virtual domain of the system. The SMuSe is based on the principles of parallelism, situatedness, emergence and emotional feedback and is built on a cognitively plausible architecture. It addresses questions at the intersection of music perception and cognition while being used as a creative tool for interactive music composition.

2. FROM EMBODIED COGNITIVE SCIENCE TO MUSIC SYSTEMS DESIGN

2.1. Classical View

A look at the evolution of our understanding of cognitive systems put in parallel with the evolution of music composition practices, gives a particularly interesting perspective on some limitations of actual interactive music systems.

The classical approach to cognitive science assumes that external behavior is mediated by internal representations and that cognition is basically the manipulation of these mental representations by sets of rules. It mainly relies on the sense-think-act framework, where future actions are planned according to perceptual information.

Interestingly enough, a parallel can be drawn between classical cognitive science and the development of classical music which also heavily relies on the use of formal structures. It puts the emphasis on internal processes (composition theory) to the detriment of the environment or the body, with a centralized control of the performance (the conductor). Disembodiment in classical music composition can be seen at several levels. Firstly, by training, the composer is used...
to compose in his head and translate his mental representations into an abstract musical representation: the score. Secondly, the score is traditionally interpreted live by the orchestra’s conductor who “controls” the main aspects of the musical interpretation, whereas the orchestra musicians themselves are left with a relatively reduced interpretative freedom. Moreover, the role of audience as an active actor of a musical performance is mostly neglected.

2.2. Modern View

An alternative to classical cognitive science is the connectionist approach that tries to build biologically plausible systems using neural networks. Unlike more traditional digital computation models based on serial processing and explicit manipulation of symbols, connectionist networks allow for fast parallel computation. Moreover, it does not rely on explicit rules but on emergent phenomena stemming from the interaction between simple neural units. Another related approach, called embodied cognitive science, put the emphasis on the influence of the environment on internal processes. In some sense it replaced the view of cognition as a representation by the view that cognition is an active process involving an agent acting in the environment. Consequently, the complexity of a generated structure is not the result of the complexity of the underlying system only, but partly due to the complexity of its environment.

A piece that gives a good illustration of situatedness, distributed processing, and emergence principles is In C by Terry Riley. In this piece, musicians are given a set of pitch sequences composed in advance, but each musician is left in charge of choosing when to start playing and repeating these sequences. The piece is formed by the combination of decisions of each independent musician that makes her decision based on the collective musical output that emerges from all the possible variations.

Following recent evolution of our understanding of cognitive systems, we emphasize the crucial role of emergence, distributed processes and situatedness (as opposed to rule-based, serial, central, internal models) in the design of interactive music composition systems.

2.3. Human-In-The-Loop

With the advent of new sensor technologies, the influence of the environment on music systems can be sensed via both explicit and implicit interfaces, which allow access to the behavioral and physiological state of the user. Music is often referred to as the language of emotion, hence human emotions seems to be a natural feedback channel to take into account in the design of a situated music system. We believe that in order to be complete, the design of a situated music system should take into consideration the emotional aspects of music.

2.3.1. Explicit Gestural Interfaces

The advent of new sensing technologies has fostered the development of new kind of interfaces for musical expression. Graphical User Interfaces, tangible interfaces, gesture interfaces have now become omnipresent in the design of live music performance or compositions. Most of these interfaces are gesture-based interfaces that require explicit conscious body movements from the user. They can give access to behavioral or self-reported information, but not to implicit emotional states of the user.

2.3.2. Implicit Biosignal Interface

Thanks to the development of more robust and accurate biosignal technologies, it is now possible to derive emotion-related information from physiological data and use it as an input to interactive music systems. Although the idea is not new, the past few years have witnessed a growing interest from the computer music community in using physiological data such as heart rate, electrodermal activity, electroencephalogram and respiration to generate or transform sound and music. Providing emotion-based physiological interface is highly relevant for a number of applications including music therapy, diagnosis, interactive gaming, and emotion-aware musical instruments.

2.3.3. Emotional Mapping

Music and its effect on the listener has long been a subject of fascination and scientific exploration from the Greeks speculating on the acoustic properties of the voice to Musak researcher designing “soothing” elevator music. It has now become an omnipresent part of our day to day life. Music is well known for affecting human emotional states, and most people enjoy music because of the emotions it evokes. Yet, although emotions seem to be a crucial aspect of music listening and performance, the scientific literature on music and emotion is scarce if compared to music cognition or perception. The relationship between specific musical parameters and time-varying emotional responses is still not clear. Biofeedback interactive music systems appear to be an ideal paradigm to explore the complex relationship between emotion and music.

2.4. Music Perception and Cognition

2.4.1. Hierarchy

What are the most relevant dimensions of music and how should they be represented in the system? Here, we take a cognitive psychology approach, and define a set of parameters that are the most salient perceptually and the most meaningful cognitively. Music is a real-world stimulus that is meant to be appreciated by a human listener. It involves a complex set of perceptive and cognitive processes
The different levels of sequential grouping for musical material: event fusion, melodic and rhythmic grouping and formal sectioning (from [42]).

that take place in the central nervous system. The fast advances in the neuroscience of music over the past twenty years have taught us that these processes are partly interdependent, are integrated in time and involve memory as well as emotional systems [16, 32, 33]. Their study shed light on the structures and features that are involved in music processing and stand out as being perceptually and cognitively relevant. Experimental studies have found that musical perception happens at three different time scale, namely the event fusion level when basic musical events such as pitch, intensity and timbre emerge (∼50 ms); the melodic and rhythmic grouping when pattern of those basic events are perceived (∼5 s), and finally the form level (from 5 s to 1 hour) that deals with large scale sections of music [42] (Figure 2). This hierarchy of three time scale of music processing forms the basis on which we built SMuSe’s music processing chain.

2.4.2. Modularity
Research on the brain substrates underlying music processing has switched in the last twenty years from a classical view emphasizing a dichotomy between language (supposedly processed in left hemisphere) and music (respectively right hemisphere) to a more modular view [11]. There is some evidence that music processing modules are organized into two parallel but largely independent submodules that deal with pitch content (“What?”) and temporal content (“When?”) respectively [33, 18]. This evidence suggests that they can be treated separately in a computational framework. Additionally, studies involving music-related deficits in neurologically impaired individuals (e.g. subjects with amusias who can’t recognize melodies anymore) have shown that music faculty is composed of a set of neurally isolable processing components for pitch, loudness and rhythm [32]. The common view is that pitch, rhythm and loudness are first processed separately by the brain to then later form (around 25-50ms) an impression of unified musical object [27] (see [16] for a review of the neural basis of music perception). This modularity as well as the three different levels and time scales of auditory memory (sound, groups, structure) form a set of basic principles for the design of our computational framework.

3. A COMPUTATIONAL MODEL BASED ON A SOCIETY OF MUSICAL AGENTS

3.1. The SMuSe’s Architecture

The architecture of SMuSe follows a hierarchical and modular structure, and has been implemented as a set of agents using data-flow programming. The musical material is represented at three different hierarchical levels, namely event fusion, event grouping and structure corresponding to different memory constraints. From the generative point of view, SMuSe modules are divided into time modules (“when”) that generate rhythmic pattern of events and content modules (“what”) that for each time event choose musical material such as pitch and dynamics (Figure 3).

At the low event fusion level, SMuSe provides a set of synthesis techniques validated by psychoacoustic tests [26, 23] that give perceptual control over the generation of timbre as well as the use of MIDI information to define basic musical material such as pitch, velocity and duration. Inspired by previous works on musical performance modeling [7], the SMuSe also allows to modulate the expressiveness of music generation by varying parameters such as phrasing, articulation and performance noise [24].

At the medium melodic and rhythmic grouping level, the SMuSe implements various state of the art algorithmic
composition tools (e.g. generation of tonal, Brownian and serial series of pitches and rhythms, Markov chains, ...). The time scale of this mid-level of processing is in the order of 5s. for a single grouping, i.e. the time limit of auditory short-term memory.

The form level concerns large groupings of events over a long period of time (longer than the short-term memory). It deals with entire sequences of music and relates to the structure and limits of long-term memory. Influenced by experiments in synthetic epistemology and situated robotics, this longer term structure is accomplished via the interaction with the environment [45][44].

The modularity of the music processing chain is also reflected in different SMuSe modules that specifically deal with time (“when”) or material (“what”).

3.2. Agent Framework

The agent framework is based on the principle that complex tasks can be accomplished through a society of simple cross-connected self-contained agents [29]. Here, an agent is understood as “anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors [39]. In the context of cognitive science, this paradigm somehow takes a stand against a unified theory of mind where a diversity of phenomena would be explained by a single set of rules. The claim here is that surprising, complex and emergent results can be obtained through the interaction of simple non-linear agents. The agent framework is particularly suited to building flexible real-time interactive musical systems based on the principles of modularity, real-time interaction and situatedness.

3.3. Data-flow Programming

We chose to implement this hierarchy of musical agents in SMuSe in a data-flow programming language called Max/MSP [46]. Data flow programming conceptually models a program as a directed graph of data flowing between operations. This kind of model can easily represent parallel processing which is common in biological systems, and is also convenient to represent an agent-based modular architecture.

3.4. Distributed Control

All the musical agents in SMuSe can be controlled and accessed from anywhere (including over a network) at any time in a distributed fashion (via OSC commands). This gives great flexibility to the system, and allows for shared collaborative compositions where several clients/performers can access and control a common music server/generator. In this collaborative composition paradigm, every performer builds on what the others have done. The action of one performer on the server (e.g. modulation of a set of musical parameters) stimulates the performance of a next action for another performer. The result is a complex sound structure that keeps evolving as long as different performers contribute changes to its current shape. A parallel could be drawn with stigmergic mechanisms of coordination between social insects like ants [41][3][11]. In ant colonies, the pheromonal trace left by one ant at a given time is used as a means to communicate and stimulate the action of the others. Hence they manage to collectively build complex networks of trails towards food sources. Similarly, in a distributed collective music paradigm one performer leaves a musical trace (i.e. the result of a control command to the music server) to the shared composition via independent control, which in turn stimulate the other co-performers to react and build on top of it (via other control commands).

3.5. Concurrent and On-the-fly Control of Musical Processes

We have proposed a biologically inspired memory and process architecture for SMuSe as well as a computational model based on software agents. The OSC communication protocol allows to easily send text-based commands to specific agents in the hierarchy. It allows for flexible and intuitive time-based, concurrent and on-the-fly control of musical processes.

The different musical agents in the SMuSe all have a specific ID/ address at which they receive commands and data. The addresses are divided into /global (affecting the whole hierarchy), /voice[n] (affecting specific voices), and /synth[n] (affecting specific sound generators). The OSC syntax supports regular expressions which allows to address several modules at the same time with a compact syntax.

Patterns of perceptually-grounded musical features are sent to the short-term (STM) and long-term memory (LTM) modules at any moment in time via specific commands.

/* Example: fill up the STM */
/voice1/rhythm/pattern 4n 4n 8n 16n 16n 4n
/voice1/pitch/pattern 0 0 5 7 10 0
/voice1/pitch/register 4 5 4 4 4 5
/voice1/velocity 12 12 16 16 12 32

Musical sequence generation follow different Selection principles, a term inspired by the reflexions of Koenig on serial music and algorithmic composition [19]. It refers to the actions taken by the system to generate musical events using the available short and long term memory content.
Figure 4. SMuSe’s environment: the SMuSe can interact with its environment through different sensors such as biosignals, camera, gazers, lasers, pressure sensitive floor, MIDI, audio, but also via OSC commands sent from client applications (such as console terminal, IQR, Iannix graphical score, Torque game engine, etc.) to the music server over the network.

These actions can be deterministic (e.g. playback of a stored sequence) or based on probability of occurrence of specific events (series, Markov chains, random events). This allows for an hybrid approach to algorithmic composition where complex stochastic processes are mixed with more deterministic repeating patterns (Cf. Table ??). Expressivity parameters such as articulation, tempo and dynamics can be continuously accessed and modulated.

3.6. Human in the loop

We tested the SMuSe within different sensing environments ranging from physiology sensors that can provide implicit emotional user interaction (heart rate, electrodermal activity, electroencephalogram), to virtual and mixed-reality sensors for behavioral interaction (camera, gazers, lasers, pressure sensitive floors) and finally MIDI and audio (microphone) for direct musical interaction (Figure 4). SMuSe integrates sensory data from the environment (that conveys information about the human participants interacting with the system) in real time and send this interpreted data to the music generation processes after appropriate fixed or learned musical mappings. The initial musical material generated by SMuSe is amplified, transformed and nuanced as the interaction between the system and the participant evolves.

3.7. Emotional Mappings

We have built a situated cognitive music system that is sensitive to its environment via musical, behavioral and physiological sensors. Thanks to a flexible architecture, the system is able to memorize, combine and generate complex musical structures in real-time. We have described various sensate environments that provide feedback information from the human interactor and allow to close the interaction loop. As proposed previously, we take an approach that focuses on emotion-related feedback.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>Emotional expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplitude envelope</td>
<td>Musical</td>
<td>Diminuendo, accentuate, fade, fortissimo, piano (Schneider &amp; Ockel- sky, 1977), crescendo, fade, accentuate (Schneider &amp; Ockelsky, 1977), piano (Joshi, 1987)</td>
</tr>
<tr>
<td>Sharp</td>
<td>Pleasantness, unpleasantness, surprise, activity (Schneider &amp; Ockel- sky, 1977), anger (Joshi, 1987)</td>
<td></td>
</tr>
<tr>
<td>Articulation</td>
<td>Staccato</td>
<td>Fast, slow (Joshi, 1987)</td>
</tr>
<tr>
<td>Long</td>
<td>Trancendence, sedate (Joshi, 1987)</td>
<td></td>
</tr>
<tr>
<td>Harmony</td>
<td>Simple / compound</td>
<td>Relaxed, toned-down (Ladmiral, 1997), tension, tenor (Kruimhah, 1996)</td>
</tr>
<tr>
<td>Loudness</td>
<td>Level</td>
<td>Anger (Joshi, 1987)</td>
</tr>
<tr>
<td>Soft</td>
<td>Fear, tenderness, sadness (Joshi, 1987)</td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>Fear (Schneider &amp; Ockelsky, 1977)</td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>Happiness, unpleasantness, activity (Schneider &amp; Ockelsky, 1977)</td>
<td></td>
</tr>
<tr>
<td>Rapid change</td>
<td>Fear (Kruimhah, 1997)</td>
<td></td>
</tr>
<tr>
<td>Melodic range</td>
<td>Wide</td>
<td>Fear (Kruimhah, 1997), joy (Scholl &amp; Thompson, 1999)</td>
</tr>
<tr>
<td>Narrow</td>
<td>Sadness (Scholl &amp; Thompson, 1999)</td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>Major</td>
<td>Harmony (Schneider &amp; Ockelsky, 1977), Kruimhah (1996), diatonic, anger (Schneider &amp; Ockelsky, 1977)</td>
</tr>
<tr>
<td>Minor</td>
<td>Harmony (Kruimhah, 1997), diatonic, anger (Schneider &amp; Ockelsky, 1977)</td>
<td></td>
</tr>
<tr>
<td>Pitch level</td>
<td>High</td>
<td>Harmony, potency, anger, fear, activity (Schneider &amp; Ockelsky, 1977)</td>
</tr>
<tr>
<td>Low</td>
<td>Harmony, unpleasantness, sedate (Schneider &amp; Ockelsky, 1977)</td>
<td></td>
</tr>
<tr>
<td>Tempo</td>
<td>Fast</td>
<td>Activity, surprise, happiness, unpleasantness, potency, fear, anger (Schneider &amp; Ockelsky, 1977), happiness (Joshi, 1987), anger (Joshi, 1987)</td>
</tr>
<tr>
<td>Slow</td>
<td>Sadness, boredom, diatonic (Schneider &amp; Ockelsky, 1977), seda- te, tenderness (Joshi, 1987), sedate (Kruimhah, 1997)</td>
<td></td>
</tr>
<tr>
<td>Timber</td>
<td>Few harmonics</td>
<td>Pleasantness, boredom, happiness, sedate (Schneider &amp; Ockel- sky, 1977)</td>
</tr>
<tr>
<td>Many harmonics</td>
<td>Potency, anger, diatonic, fast, activity, surprise (Schneider &amp; Ockelsky, 1977)</td>
<td></td>
</tr>
<tr>
<td>Soft</td>
<td>Tenderness, sedate (Joshi, 1987)</td>
<td></td>
</tr>
<tr>
<td>Sharp</td>
<td>Anger (Joshi, 1987)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. The relationship between musical parameters and perceived emotion. A summary of the main results. (adapted from [10])

3.7.1. Predefined Mappings

From advanced behavioral and physiological interfaces, we can infer emotion-related information about the interaction with SMuSe. One possible way to take this feedback into account is to design fixed mappings based on previous results from experimental psychology studies that have investigated the emotional responses to specific musical parameters. This a priori knowledge can be used to drive the choice of musical parameters depending on the difference between the goal emotion to be expressed or induced, and the emotional state detected by the system via its sensors. A number of reviews have proposed generic relationships between sound parameters and emotional responses [9, 13, 12]. These results serve as the basis for explicit emotional mappings (Cf. Table 1) and have been confirmed in the specific context of SMuSe’s parameter space. This explicit design approach is detailed in [26, 20].

3.7.2. Adaptive Mappings

In cases where the relationship between the change of musical parameters and the emotional response has to be learned, explicit mappings is not possible. One solution in SMuSe is to use a Reinforcement Learning (RL) agent that learns to adapt the choice of musical parameters based on the interaction of the system with the environment [43]. Reinforcement learning is a biologically plausible learning algorithm particularly suited to an explorative and adap-
Figure 5. The reinforcement-based music system is composed of three main components: the music engine (the SMuSe), the reinforcement learning agent and the listener who provides the reward signal.

Figure 6. Artistic realizations: A) Re(PER)curso (Art Futura and MACBA, Barcelona 2007) B) The Multimodal Brain Orchestra (FET, Prague 2009) C) XIM sonification (2007)

Barcelona in the same year. The performance was composed by several interlaced layers of artistic and technological activities. The music controlled had three components: a predefined soundscape, the percussionist who performed from a score and the interactive composition system synchronized by SMuSe; the physical actors, the percussionist and the dancer were tracked by a video based active tracking system that in turn controlled an array of moving lights that illuminated the scene. The spatial information from the stage obtained by the tracking system was also projected onto the virtual world where it modulated the avatar’s behavior allowing it to adjust body position, posture and gaze to the physical world. In 2009, the Brain Orchestra, a multimodal performance using brain computer interfaces, explored the creative potential of a collection of brains directly interfaced to the world. During the performance, four “brain musicians” were controlling a string quartet generated by the SMuSe using their brain activity alone. The orchestra was conducted by an “emotional conductor”, whose emotional reactions were recorded using biosignal interfaces and fed back to the system. The Brain Orchestra was premiered in Prague for the FET 09 meeting organized by the European Commission. Finally, a live performance of a piece inspired by Terry Riley’s “in C” served as an illustration of the principles of parallelism, situatedness and emergence exhibited by the SMuSe at the Ernst Strungmann Forum on Language, Music and the Brain: a mysterious relationship.

5. CONCLUSIONS

The SMuSe illustrates a novel situated approach to music composition systems. It is built on a cognitively plausible architecture that takes into account the different time frames of music processing, and uses an agent framework to model a society of simple distributed musical processes. It takes advantage of its interaction with the environment to go beyond the classic sense-think-act paradigm.
It combines cognitively relevant representations with perceptually grounded sound synthesis techniques and is based on modern data-flow audio programming practices. This provides an intuitive, flexible and distributed control environment that can easily generate complex musical structure in real-time. SMuSe can sense its environment via a variety of sensors, notably physiology-based sensors. The analysis and extraction of relevant information from sensor data allows to re-inject emotion-based feedback to the system based on the responses of the human participant. The SMuSe proposes a set of "pre-wired" emotional mappings from emotions to musical parameters grounded on the literature on music and emotion, as well as a reinforcement learning agent that performs online adaptive mapping. It provides a well grounded approach towards the development of advanced synthetic aesthetic systems and a further understanding of the fundamental psychological processes on which it relies.

6. REFERENCES


