InMuSIC: an Interactive Multimodal System for Electroacoustic Improvisation

Giacomo Lepri
STEIM - Institute of Sonology, Royal Conservatoire in The Hague, The Netherlands
leprotto.giacomo@gmail.com

ABSTRACT

InMuSIC is an Interactive Musical System (IMS) designed for electroacoustic improvisation (clarinet and live electronics). The system relies on a set of musical interactions based on the multimodal analysis of the instrumentalist's behaviour: observation of embodied motion qualities (upper-body motion tracking) and sonic parameters (audio features analysis). Expressive cues are computed at various levels of abstraction by comparing the multimodal data. The analysed musical information organises and shapes the sonic output of the system influencing various decision-making processes. The procedures outlined for the real-time organisation of the electroacoustic materials intend to facilitate the shared development of both long-term musical structures and immediate sonic interactions. The aim is to investigate compositional and performative strategies for the establishment of a musical collaboration between the improviser and the system.

1. INTRODUCTION

The design of IMS for real-time improvisation poses significant research questions related to human computer interaction (e.g. [1]), music cognition (e.g. [2]), social and cultural studies (e.g. [3]). An early important work is George Lewis' *Voyager* [4]. In *Voyager*, the author's compositional approach plays a crucial role: specific cultural and aesthetic notions are reflected in the sonic interactions developed by the system. More recently, systems able to generate improvisations in the style of a particular performer (e.g. Pachet's *Continuator* [5] and *OMax* from IRCAM [6]) were developed. In these systems, the implementation of a particular type of finite-state machine, highly refined for the modelling of cognitive processes, allows for the simulation of humanised behaviours such as imitation, learning, memory and anticipation.
In this field of research, the chosen framework for the composition of sonic interactions reflects particular cultural and musical models, performative intuitions, as well as specific cognitive paradigms and technological notions. Music improvisation is here conceived as a wide-ranging creative practice: a synthesis of intricate processes involving physicality, movement, cognition, emotions and sound. The design approach of InMuSIC derived from an embodied cognition of music practice [7]. The majority of the interactive system for improvisation developed during the last years are not based on an embodied cognition of music practice and they focus on the sonic aspects of the performance. Nevertheless, a multimodal approach for the design of improvising IMS was adopted within various research. For example, Ciufo [8], Kapur [9] and Spasov [10] developed IMS able to extract in real-time both gestural and sonic qualities of the performer interacting with the machine. However, these applications are concerned with the recognition of specific body parts and particular gestures (e.g. hands movements). One of the main goal of the presented research is related to the definition of strategies for a qualitative analysis of upper-body features pertinent to a wide range of gestures, not restricted to specific types of movement. This paper presents the system's overall design approach sketching a strategy for the real-time multimodal analysis and representation of instrumental music practice.

2. THE INTERACTIVE FRAMEWORK

The notion of interaction here investigated is inspired by the spontaneous and dialogical interactions characterising human improvisation. The intention is to provide the system with an autonomous nature, inspired by the human ability to focus, act and react differently in relation to diverse musical conditions. In regards to each specific performance, the close collaboration between the musician and InMuSIC should enable the constitution and emergence of specific musical forms. The generation, modification and temporal organisation of new sonic materials are established negotiating the musical behaviour of the performer and the systems internal procedures. In order to facilitate the development of a spontaneous musical act, the platform should then be able to assess different degrees of musical adaptiveness (e.g. imitation/variation) and independence (e.g. contrast/discontinuity). InMusic has been conceived for real-time concert use within contexts related to electroacoustic improvisation. The compositional research has developed alongside a specific musical aesthetic concerned with the exploration of sonic spectral qualities within flexible fluctuations in time rather than actual melodic/harmonic progressions and metrical tempo [11].
The IMS presented relies on the analysis and comparison of sonic and motion qualities. This is by identifying and processing abstracted expressive musical hints of the performer. The attempt of composing and exploring sonic and gestural interdependences is the foundation of the inquired interactive paradigm. Thus, the framework composed to frame and shape the musical interactions, in addition to the sonic dimension, aims to take into account fundamental performative and expressive aspects complementary to the sound production.

3. THE COMPOSITIONAL MODEL

In this section, the InMuSIC's conceptual model is presented. Figure 1 illustrates a layered model based on the work of Leman and Camurri [12]. It is composed of five modules located on three different levels of abstraction, ranging from the representation of physical energy to the more compositional extent related to performative intuitions. Consequently, it is possible to conceive a *continuum* linking the physical world to its musical interpretation. The lowest level is associated to those units that perform tasks related to the physical domain (i.e. detection of sound and movements). The highest level is related to the more abstract components of the system, responsible for compositional choices that govern the real-time sonic interactions. This representation defines an interactive loop and it offers the possibility to frame the essential functions associated to the musical behaviour of the system.
In addition, the conceptual model presented is inspired by the work of von Bertalanffy [13]. The design approach of the relations between the various system's units is influenced by specific criteria: (i) any change in a single unit causes a change in all the units, (ii) the system's behaviour reacts to the incoming data and modifies them in order to either cause change, or to maintain the stationary state (positive and negative feedback) and (iii) the same results may have different origins (i.e. the same causes do not produce the same effects, and *vice versa*). The individual modules will be now briefly introduced.

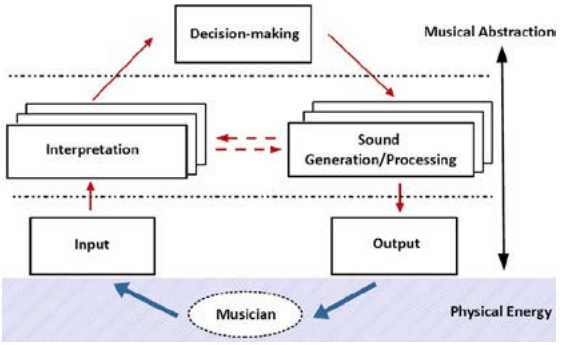

Figure 1. The conceptual model of InMuSIC.

- **Input** - The module executes two main functions: (i) detection of the movements and sounds articulated by the musician and (ii) conversion of this energy (i.e. kinetic and sonic) into digital information. InMuSIC foresees the use of two sensors: the instrument's sound is detected using a condenser microphone and the movement of the performer is captured using the 3D sensor Microsoft Kinect 2.

- **Interpretation** - The information is here interpreted through several parallel processes. Specific sonic and movement features are derived. The comparison of the various analyses provides a second level of interpretation related to the musician's behaviour. In particular conditions, the unit analyses the interventions generated by the system itself. This feedback contributes to the system's self-organisation processes.

- **Decision-making** - The module is located on the highest level of abstraction within the model. Its main function concerns the time-based organisation of the procedures for the generation and manipulation of new sound materials. The decision-making strategies are based on a *negotiation* between the system's internal stochastic processes and the analysed performer's behaviour.

- **Sound generation/processing** - The unit consists of a set of algorithms for sound synthesis and processing: the electronic materials proposed by the system are here actually generated and shaped. In order to establish direct interactions, the system can assign the control of the parameters of the algorithms directly to the data extracted from the modules related to the sound and movement analyses.

- **Output** - The module transfers into the physical domain the information generated by the most abstract units. The processes involved are: (i) the amplification of the generated signal, (ii) the signal's conversion from digital to analogue and (iii) the projection of the sound in the performative space.

4. THE SYSTEM ARCHITECTURE

From a practical point of view, whilst a musician plays a freely improvised session, the system performs five main tasks: movement analysis, sound analysis, sound and movement comparison, decision-making and sound generation. Specific software units compute each of these tasks. The various components are implemented using Max/MSP and EyesWeb. The two platforms communicate through an Open Sound Control (OSC) protocol. A description of the five modules and their functions will now be presented.

4.1 Sound analysis

The unit extracts three low-level audio features: loudness, onset detection and fundamental frequency. The audio signal is analysed by matching and evaluating the outputs of several algorithms [14, 15, 16]. Each of these is tuned for specific dynamic and frequency ranges.
A first level of analysis is associated to the variation in time of the detected data. Initially the features are interpreted through different low-pass filtering and moving average processes. Subsequently the derivative of each feature is computed. By mapping the obtained values using different logistic functions, two thresholds are fixed. In relation to the data previously analysed, the information extracted is defined by three possible states: higher, lower or stable. Consequently, this procedure displays a minimal representation of each audio feature: (i) high, low or stable dynamics (*crescendo vs. diminuendo*); (ii) high, low or stable onset detection (increase *vs.* decrease of the events density); (iii) high, low or stable pitch deviation (expansion *vs.* reduction of the used frequency range). The algorithms implemented interpret the incoming values by means of an *inertial behaviour*. In order to detect any positive or negative change, a certain amount of variation is required. This conduct, simulating the function of a short-term memory, is specifically calibrated for each feature. This is crucial to the fine-tuning of the system's sensitivity.

The understanding of the performer's sonic behaviour is therefore associated to the variation in time of the extracted features. The methodology adopted is influenced by psychological research on human communication [17]. The main assumption is that we can only perceive the relationships or models of relationships that substantiate our own experience. Our perceptions are affected by processes of variation, change or motion. Any phenomenon is perceived only in relation to a reference: in this case the music previously played.

4.2 Movement analysis

Based on the research by Glowinski et al. [18] for the analysis of affective nonverbal behaviour using a reduced amount of visual information, the module extracts expressive gestural features. This interpretation implies the analysis of behavioural features pertinent to a wide range of gestures and not restricted to specific types of movement. The challenge consists of detecting information representative of an open sphere of possible expressive motions: the chosen strategy focuses on a minimal representation of affective movements. A qualitative approach to the analysis of upper-body movements and affect recognition, is hereby adopted [19]. Considering a reduced amount of visual information (i.e. 3D position, velocity, and acceleration of the musicians head, hands and elbows - see 2), three expressive features are extracted: smoothness (degree fluidity associated to the head movement), contraction index (degree of posture openness) and quantity of motion (QOM) (overall kinetic energy).
Applying the same procedure, illustrated in the sound analysis section, the features are further interpreted. Each analysis is reduced to three possible states: (i) high, low or stable smoothness (detection of fluidity and continuity *vs.* jerky or stillness in regards to the head movements); (ii) high, low or stable QOM (overall QOM variation - presence of motion *vs.* stillness or isolated movements); (iii) high, low or stable contraction index (variations in the degree of posture - open *vs.* close).

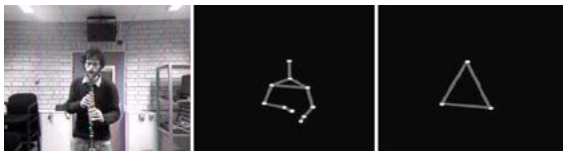

Figure 2. The detected skeleton of a musician playing the clarinet. The motion analysis is based on a minimal representation of affective gestures.

4.3 Sound and movement comparison

The module is designed to combine and compare the data coming from the movement and sound analyses. The various *stable* states are ignored: the detection of a *stable* state does not produce any change to the internal conditions of the system (i.e. maintenance of the current stationary state). Figure 3 illustrates the available combinations in regard to each *high-low* state. Through a Graphical User Interface (GUI) it is possible to manually select which combinations the module will consider during the performance. Figure 3 presents a possible selection of the states combinations often used by the author performing with InMuSIC. Once a specific combination is chosen (e.g. low QOM and low loudness), the unit constantly verifies if the two states are simultaneously detected: to each selected combination, a simple boolean condition is applied. In addition, the unit tracks how long each condition is verified. In short, during the performance, the data sent to the decision-making module defines (i) which condition selected is currently true and (ii) the time associated to the persistence of each verified condition.
The computation of the various *high-low* states allows for the gathering of information related to the variation in time of the extracted features (continuous inertial interpretation). For instance, in regards to the past trends, the QOM is now increasing or decreasing. The combination and comparison of the *high-low* states associated to the various features is conceived as a further level of abstraction within the expressive analysis of the performer. The organisation of the processes for the generation of new electronics interventions is therefore related to the detection of specific *high-low* conditions (finite-state machine like behaviour). The strategy implemented aims to achieve a minimal and qualitative interpretation of instrumental music practice: the focus is oriented to analyse *how* the musician plays instead of *what* the musician plays.

		Loudness		Events Density		Pitch Deviation	
		high	low	high	low	high	low
QOM	low		✓				
	high			✓			
Contraction Index	low			✓			
	high	✓					✓
Smoothness	low						
	high		✓				✓

Figure 3. The possible comparisons of sound and movement analyses. The ticked boxes are the combination often used by the author while performing with the system.

4.4 Decision-making

The function of the unit mainly concerns the time-based organisation of new musical information (e.g. activation, duration, cross fade and muting of the various system's *voices*). Here the main focus is oriented towards the composition of decision-making processes allowing for the development of both long-term musical structures and immediate sound interventions. The unit establishes sonic interactions that develops inside a *continuum* ranging between two different temporal durations: from short-term immediate re-actions (maximum duration of 4 seconds), to long-term re-actions (maximum duration of 4 minutes). The reference paradigm refers to studies on human auditory memory [20] (short-term and long-term). An *awareness* of different *real-times* is here sought. The overall timing of the unit (i.e. the actual clock that triggers the various sonic processes) is controlled by an irregular *tactus* generated by a stochastic process. The rate of this clock is constantly modified by the variation in time of the onset analysis: the system's *heart beat* increases when the performer articulates a music dense of sonic events and *vice versa*.
The generation and organisation of both short-term and long-term interventions is associated to the detection of the *high-low* conditions occurring during the performance (e.g. simultaneous detection of low QOM and low loudness). To each condition a set of sound processes is applied, a particular type of synthesis can be associated to

more then one condition. The more a condition is detected, the higher the probability is to trigger the related sound processes. Furthermore, stochastic procedures influence the relative weight of each probability with a specific set. The duration of an active sonic process is affected by the persistence in time of the associated *high-low* condition.
Simultaneously, the unit regulates two further parallel procedures. Once a particular sound process is activated, timbral adjustments can occur. The unit can establish a direct link between the performers sonic and gestural behaviours and the processes for the sound synthesis. This relates to the modification of current electronic materials (i.e. manipulation of the control-rate data associated to the triggered sound) using the information coming from the sound and movement analyses. During the performance, the unit can also send the produced electronic materials to the sound analysis module. Thus, a feedback process is activated: instead of evaluating the sonorities produced by the musician, InMuSIC analyses its own output. This situation mainly takes place when the performer is not playing. The possibility of 'listening to itself' is conceived as a further degree of autonomy within the system's agencies.
The described procedures enables the potential generation of a wide range of *musical narratives*, emerging and evolving with regards to each specific performance.

4.5 Sound generation

The sound generation module is conceived to produce heterogeneous sound materials. The sonic interactions generated entail a multiplicity of possible changes concerning diverse musical circumstances. In relation to the different performative and expressive contexts, the variety of timbral and sonic articulation appears to be an important requirement for the development of an engaging interactions. The algorithms implemented for the generation of the electronic materials can be organised into three categories: (i) synthesis (FM, additive, subtractive and physical models [21]), (ii) sampling (real-time processing of pre-recorded sounds) and (iii) live processing (live sampling, live granulation, Fast Fourier transform analysis and re-synthesis and reverberation).
The individual techniques used can be conceived as system's *voices*. Each *voice* is characterised by specific qualities, that are spectro-morphological (i.e. related to the distribution of energy inside the sonic spectrum) and gestural (i.e. associated to the articulation and transformation of sound material over time). In relation to the generated sonorities, each algorithm has been designed to guarantee a certain degree of indeterminacy. The goal is to define processes able to develop extensive variations and manipulations of the electronic materials within predefined physical scopes (e.g. frequency, dynamic and temporal ranges). In other words, every single voice is conceived to explore diverse *sound spaces*. The musician is invited to *navigate* these timbre spaces [22] in collaboration with the system. Once a voice is active, timbre variations may occur: these changes are shaped by the external interventions inferred by the performer's musical behaviour. The intention is to develop a close dialogue/collaboration between acoustic and electronic materials (e.g. fusion, separation, imitation, variation and contrast). This approach allows to partially solve a dichotomy that emerges when attempting to combine the practices of composition and improvisation. Through the real-time interactions with the performer, InMuSIC organises and shapes pre-composed musical materials. The challenge relies on balancing the processes that leads to the development of musical forms within a *performative time* and the musical choices previously made over a *compositional time*.

5. THE PERFORMANCE

InMuSIC has been extensively used by the author in live concerts and it has been presented in several musical events and research contexts.The performance was often evaluated as engaging and successful. The sonic variety generated and the system responsiveness appear to be the most valued traits of the IMS here presented.
InMuSIC was also tested by five expert improvisers in informal settings. The aim was to explore the use of InMuSIC with different players and instruments (two clarinettists, one trombonist, one cellist and one pianist). After a short introduction, the musicians were invited to freely play with the system. Open interviews were undertaken to investigate their impressions. The system was essentially perceived as a generative algorithm allowing for a shared exploration of interesting and engaging musical materials. The experience of playing with InMuSIC was compared to a conversation with a little child: "You don't know very well how it will react. Its a little bit shy at first and you have to draw something out of it". The system was also perceived as able to play both in foreground (leading) and background (either following or leaving space for solos), although some musician felt that InMuSIC was leading too often. Some improvisers perceived a not always bidirectional interaction: the machine was "not listening much". Furthermore, they expressed the desire for a IMS that would more frequently retrieve and develop the materials proposed by them.
Some musicians were slightly frustrated by the impossibility of clearly understand and control the functioning of InMuSIC. Others referred to this aspect positively comparing this situation to the real human-human interaction. Interestingly, some musicians observed that, during the performance, a turning point occurred. After a first clear and simple interaction (i.e. direct action-reaction relationship) the musicians changed their attitude. Once recognised that the machine was listening and responding (even if not constantly) they started to better engage with the system being more open to the electronic material proposed.
During the sessions, the algorithms for the sound and movement analysis were not modified: the settings normally used by the author performing with the clarinet were kept. Compared to the author experience with InMuSIC, it was noticed that the system was less reactive and always performing with a reduced amount of sonic possibilities. This might suggest that the system has to be tuned according to each specific player. In addition, all the musicians agreed on the need of rehearsing in order to achieve a more satisfying performance. There were no significant differences in the system outcome while playing with different instruments. This might be related to the qualitative approach adopted for the analysis of musical behaviour (i.e. looking at how do we play instead of what do we play).

6. CONCLUSIONS

InMuSIC is a multimodal interactive system for electroacoustic improvisation (clarinet and live electronics). It can be defined as a system that composes/improvises music through a dialogical modality. The aim of the research is to design a platform able to establish a close collaboration with the performer, in relation to the analysed musical information. Music improvisation is here conceived as a spontaneous expressive act involving cognitive and technical skills conveyed by sonic and physical behaviours. The interactive paradigm developed is therefore based on the combination and comparison of the performers movement and sound analyses. InMuSIC is tuned to be sensitive to a specific *apparatus* of gestural and sonic behaviours, according to both the instrumental practice of the clarinet and the performative attitudes characterising the author's expressiveness. Future developments of the system may include the possibility of expanding this *apparatus* in order to explore diverse audio and gestural features and widen the performer's analysis. It is not the intention of the author to categorise or attribute any specific semantics to the various expressive cues represented. Instead, the interest relies on the exploration and use (or abuse) of these musical indications in the contexts of composition and improvisation. Nevertheless, the author's impression is that, with a more systematic approach, the multimodal analysis presented might allow for the revealing of performative traits pertinent to specific instruments and players. The conceived performance presumes the development of both musical structures and immediate re-action, emerging from the human-computer cooperation.

7. REFERENCES

[1] A. Cont, S. Dubnov, and G. Assayag, "A framework for anticipatory machine improvisation and style imitation," in *Anticipatory Behavior in Adaptive Learning Systems (ABiALS)*. ABIALS, 2006.

[2] A. R. Addessi, "From Econ to the mirror neurons: Founding a systematic perspective on the reflexive interaction paradigm," *ICMPC-ESCOM2012 Proceedings*, pp. 23–28, 2012.

[3] G. E. Lewis, "Interacting with latter-day musical automata," *Contemporary Music Review*, vol. 18, no. 3, pp. 99–112, 1999.

[4] ——, "Too many notes: Computers, complexity and culture in voyager," *Leonardo Music Journal*, vol. 10, pp. 33–39, 2000.

[5] F. Pachet, "The continuator: Musical interaction with style," *Journal of New Music Research*, vol. 32, no. 3, pp. 333–341, 2003.

[6] G. Assayag, G. Bloch, M. Chemillier, A. Cont, and S. Dubnov, "Omax brothers: a dynamic yopology of agents for improvization learning," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. ACM, 2006, pp. 125–132.

[7] M. Leman, *Embodied music cognition and mediation technology*. Mit Press, 2008.

[8] T. Ciufo, "Design concepts and control strategies for interactive improvisational music systems," in *Proceedings of the MAXIS International Festival/Symposium of Sound and Experimental Music*, 2003.

[9] A. Kapur, "Multimodal Techniques for Human/Robot Interaction," in *Musical Robots and Interactive Multimodal Systems*. Springer, 2011, pp. 215–232.

[10] M. Spasov, "Music Composition as an Act of Cognition: ENACTIV–interactive multi-modal composing system," *Organised Sound*, vol. 16, no. 01, pp. 69–86, 2011.

[11] D. Smalley, "Spectromorphology: explaining sound-shapes," *Organised sound*, vol. 2, no. 02, pp. 107–126, 1997.

[12] M. Leman and A. Camurri, "Understanding musical expressiveness using interactive multimedia platforms," *Musicae Scientiae*, vol. 10, no. 1 suppl, pp. 209–233, 2006.

[13] L. V. Bertalanffy, "General system theory: Foundations, development, applications," Braziller. New York, Tech. Rep., 1968.

[14] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[15] M. Malt and E. Jourdan, "Real-Time Uses of Low Level Sound Descriptors as Event Detection Functions Using the Max/MSP Zsa. Descriptors Library," *Proceedings of the 12th Brazilian Smposium on Computer Music*, 2009.

[16] T. Jehan and B. Schoner, "An audio-driven perceptually meaningful timbre synthesizer," *Analysis*, vol. 2, no. 3, p. 4, 2002.

[17] P. Watzlawick, J. B. Bavelas, D. D. Jackson, and B. O'Hanlon, *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company, 2011.

[18] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a minimal representation of affective gestures," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 106–118, 2011.

[19] A. Camurri, I. Lagerlöf, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International journal of human-computer studies*, vol. 59, no. 1, pp. 213–225, 2003.

[20] B. Snyder, *Music and memory: An introduction*. MIT press, 2000.

[21] D. Trueman and R. DuBois, "PeRColate: A Collection of Synthesis," *Signal Processing, and Video Objects for MAX/MSP/Nato*, vol. 1, p. b3, 2009.

[22] D. L. Wessel, "Timbre space as a musical control structure," *Computer music journal*, pp. 45–52, 1979.