MULTIMODAL INTERFACES FOR EXPRESSIVE SOUND CONTROL

Antonio Camurri

InfoMus Lab,
DIST - University of Genova, Genova, Italy
http://www.infomus.dist.unige.it antonio.camurri@unige.it

ABSTRACT

This paper introduces research issues on multimodal interaction and interfaces for expressive sound control. We introduce Multisensory Integrated Expressive Environments (MIEEs) as a framework for Mixed Reality applications in the performing arts. Paradigmatic contexts for applications of MIEEs are multimedia concerts, interactive dance / music / video installations, interactive museum exhibitions, distributed cooperative environments for theatre and artistic expression. MIEEs are user-centred systems able to interpret the high-level information conveyed by performers through their expressive gestures and to establish an effective multisensory experience taking into account expressive, emotional, affective content. The lecture discusses some main issues for MIEEs and presents the EyesWeb (www.eyesweb.org) open software platform which has been recently redesigned (version 4) in order to better address MIEE requirements. Short live demonstrations are also presented.

1. INTRODUCTION

A number of interactive systems are currently available to process audio and/or video streams: e.g. PureData, Max/MSP (www.cycling74.com), Isadora (www.troikatronix.com). These and other systems are particularly oriented toward a single modality of interaction, i.e., they might perform well when working with audio only (PureData, Max) or video only (Isadora). Support to sensory fusion and multimodality is very low if any. In the recent years several new requirements emerged for interactive systems (see e.g. [1]). Multimodal interfaces are not only a matter of working with streams of different types or with different sensors, but mainly concerns the ability to work at different abstraction levels and to support integrated processing of different channels [7]. The EU-IST project MEGA (Multisensory Expressive Gesture Applications, www.megaproject.org) defined a conceptual framework for multimodal expressive gesture processing, structured on four layers [17].
This paper gives a personal view on the perspective on next generations of interactive systems, by introducing Multisensory Integrated Expressive Environments (MIEEs), and introduces some latest developments around the EyesWeb research project, which aims at a partial implementation of MIEEs.

2. BEYOND INTERACTIVE SYSTEMS

The real-time multimodal processing and the modelling of expressive gesture in on-stage interactive performances is a challenge for both scientific and artistic research [1,9]. Multimodality and expressiveness [9,17] are intended to contribute to improve state of the art hyper- and virtual musical instruments [2,3,8], interactive dance, and interactive performances in general where technology is not only a tool, but rather it is integrated with art at the level of language and it becomes something intrinsic to the artwork. This leads to the concept of Multisensory Integrated Expressive Environments (MIEEs) [17]. MIEEs can be conceived of as a new generation of musical instruments based on real-time and intelligent human-machine interaction [18]: new musical instruments as a holistic human-machine concept based on an assembly of modular input/output devices and musical software components that are arranged according to essential human musical content processing capabilities. MIEEs aim at providing the extended digital platforms for the exchange of expressiveness through cross-modal interactions at levels that go beyond the classical multimedia approaches in art.
An example of a partial exploitation of the concept of MIEEs in large scale performances is in the work "Cronaca del Luogo" by Luciano Berio (opening of Salzburg Festival, 1999) in which the EyesWeb system (www.eyesweb.org) was used to control in real-time the processing of the voice of the main character depending on an analysis of the performer's gestures. More recent examples include the public performances in the framework of the EU-IST Project MEGA (Multisensory Expressive Gesture Applications, www.megaproject.org), ranging from medium-scale events, e.g., the concert "Allegoria dell'opinione verbale" by Roberto Doati, to large-scale events, e.g., "Medea" by Adriano Guarnieri [5].
The design of MIEEs is challenging and many research issues have still to be faced. For example, systems must be endowed with the capability of interpreting performers' gestures, and in particular expressiveness in the context where and when the gesture is performed. A MIEE should keep into account of spatial, temporal, and content memory. Information contained in performers' gesture may be structured on several layers of complexity. A particular emphasis is on affective, expressive, emotional information. In fact, it is the capability of interpreting expressive information that allows interaction of technology and art at the level of the language art employs to convey content and to provide the audience with an aesthetical experience. In this framework, a central role is assumed by research on *expressive gesture*, i.e., on the high-level emotional, affective content gesture conveys, on how to analyse and process this content, on how to use it in the development of innovative multimodal interactive systems able to provide users with natural expressive interfaces [9]. Related previous research concerns Affective Computing [10], KANSEI Information Processing [11], recent studies on the communication of expressive content or "implicit messages" [12], work by psychologists (e.g.,[13,14,15]) and - from art and humanities - theories from

choreography (e.g., Rudolf Laban's Theory of Effort) and music composition (e.g., Schaeffer's Morphology).
Another key issue is the development of strategies for controlling and/or generating audio in real-time. That is, even if algorithms able to correctly and reliably interpret high-level expressive information from gesture were available, the problem of if and how to use such information in an artistic performance still remains open. In particular, a challenging direction is on the control of sound synthesis techniques using multimodal gesture analysis cues. A model for the mapping of cues (at different levels of detail) from multimodal expressive gesture to the parameters controlling sound synthesis is a very interesting direction in which some steps have already done (see for example [19]). For example, an interesting research direction is to explore models where the natural "physicality" and meaning of cues related to expressive gesture is mapped on parameters of synthesis techniques by physical models or similar techniques where parameters have "physical", non abstract meanings.
A further difficulty is due to artistic choices of the designer of the performance, i.e., how much degrees of freedom the designer wishes to leave to the automatic systems in the control process, therefore abandoning the interaction metaphor of the "musical instrument" and going toward a dialog metaphor: in other words the role of technology in the artwork and, from a certain point of view, the concept of artwork. These aspects have been partially faced with the definition of the concept of *expressive autonomy* [16] and have been further investigated in the framework of the aforementioned EU IST MEGA project with particular reference to the definition of a conceptual architecture for modelling possible control (mapping) strategies.

3. MODELING EXPRESSIVE GESTURE

An important issue concerns the modeling and processing of expressive gesture. To this aim, a first problem concerns the identification of a suitable collection of descriptors (cues) that can be used for describing expressive gesture. Secondly, algorithms have to be defined and implemented to extract measures for such descriptors. Finally, data analysis has to be performed on these measures in order to obtain high-level information. This is only a rough sketch of the analysis process: a review of some our recent research on algorithms for extraction of cues at several level of abstraction, from low-level signal-related cues to high-level analysis of expressive content is available in [9]. Such algorithms provide the input for the strategies for gestural control of multimedia output, and in particular of sound synthesis.
Expressive cues are likely to be structured on several layers of complexity. In analysis of dance fragments using video cameras for example, some cues can be directly measured on the video frames coming from a single video camera observing the dancer. Others may need more elaborate processing or 3D information. For example, it may be needed to identify and separate expressive gestures in a movement sequence in order to compute features that are strictly related to single gestures (e.g., duration, directness, fluency).
In the framework of the EU-IST project MEGA (Multisensory Expressive Gesture Applications, `www.megaproject.org`) a conceptual framework for expressive gesture processing has been defined, structured on four layers [17]. Layer 1 (Physical Signals) includes algorithms for gathering data captured by sensors such as video cameras, microphones, on-body sensors (e.g., accelerometers), sensors of a robotic system, environmental sensors. Layer 2 (Low-level features) extracts from the sensors data a collection of low-level cues describing the gesture being performed. In case of dance, for example, cues include kinematical measures (speed, acceleration of body parts), detected amount of motion, amount of body contraction/expansion. Layer 3 (Mid-level features and maps) deals with two main issues: segmentation of the input stream (movement, music) in its composing gestures, and representation of such gestures in suitable spaces. Thus, the first problem here is to identify relevant segments in the input stream and associate to them the cues deemed important for expressive communication. For example, in dance analysis a fragment of a performance might be segmented into a sequence of gestures where gesture's boundaries are detected by studying velocity and direction variations. Measurements performed on a gesture are translated to a vector that identifies it in a semantic space representing categories of semantic features related to emotion and expression. Sequences of gestures in space and time are therefore transformed in trajectories in such a semantic space. Trajectories can then be analysed e.g., in order to find similarities among them and to group them in clusters. Layer 4 (Concepts and structures) is directly involved in data analysis and in extraction of high-level expressive information. In principle, it can be conceived as a conceptual network mapping the extracted features and gestures into (verbal) conceptual structures. For example, a dance performance can be analysed in term of the performer's conveyed emotional intentions, e.g., the basic emotions anger, fear grief, and joy. However, other outputs are also possible: for example, a structure can be envisaged describing the Laban's conceptual framework of gesture Effort, i.e., Laban's types of Effort such as "pushing", "gliding", etc. Experiments can also be carried out aiming at modelling spectators' engagement. Machine learning techniques can be employed ranging from statistical techniques (e.g., multiple regression and generalized linear techniques), to fuzzy logics or probabilistic reasoning systems (e.g., Bayesian networks), to various kinds of neural networks (e.g., classical back-propagation networks, Kohonen networks), support vector machines, decision trees. In a recent experiment described in [4] we tried to classify expressive gesture in dance performance in term of the four basic emotions anger, fear, grief, and joy. Results showed a rate of correct classification for the automatic system (five decision tree models) in between chance level and spectators' rate of correct classification. In another experiment, discussed in the same paper, we measured the engagement of listeners of a music performance (a Skriabin's Etude) and analysed correlations with extracted audio cues and with cues obtained from the movement of the performer (a pianist).

4. THE EYESWEB 4 OPEN PLATFORM

The EyesWeb open platform (`www.eyesweb.org`) has been designed with a special focus on the multimodal analysis and processing of non-verbal expressive gesture in human movement and music signals. It was developed at InfoMus Lab at DIST - University of Genova and recently it has been enhanced and reengineered to support features of MIEEs (version 4). EyesWeb consists of a number of integrated hardware and software modules that can be easily interconnected and extended in a visual environment. The EyesWeb software includes a development environment and a set of libraries of reusable software components that can be assembled by the user in a visual language to build

patches as in common computer music languages inspired to analog synthesizers. EyesWeb is open so it is easily possible to extend it by third parties with libraries and plugins.
Besides its wide use in artistic projects, EyesWeb is used to support experiments on computational models of non-verbal expressive communication, on mapping, at different levels, gestures from different modalities (e.g., human full-body movement, music) onto real-time generation of multimedia output (e.g., sound, music, visual media, mobile scenery). It allows fast development and experiment cycles of interactive performance setups.
Recent improvements concern better support to cross-media and integrated real-time processing of different streams (e.g. audio and video), new libraries for real-time processing of expressive gesture, support to XML and many other features in the direction of support to MIEEs. Modules and patches collocated at different layers in the conceptual framework (see previous Section) are available: modules to extract low-level parameters from audio and motion data (e.g., the coordinates of the baricenter of a dancer's silhouette and its bounding rectangle, the loudness and roughness of an audio excerpt), modules to extract mid-level features and expressive cues (e.g., body contraction/expansion, amount of detected motion, music tempo, articulation), high-level mappers (e.g., neural networks, Bayesian networks, Support Vector Machines). Many of such modules are included in the EyesWeb Expressive Gesture Processing Library [9], which now also includes 3D cues.
EyesWeb also supports distributed applications, e.g., patches running on several PCs, multi-user patches. In order to help programmers in developing blocks and extend the system, the EyesWeb Wizard software tool has been developed. Users can develop autonomously (i.e., possibly independently from EyesWeb) the algorithms and the basic software skeletons of their own modules. Then, the Wizard supports them in the process of transforming algorithms in integrated EyesWeb modules.
Multiple versions of modules (versioning mechanism) are supported by the system, e.g., allowing the use in patches of different versions of the same data-type or module. EyesWeb has been the basic platform of the MEGA EU IST project. In the EU V Framework Program it has also been adopted in the IST CARE HERE and IST MEDIATE projects on therapy and rehabilitation and by the MOSART network for training of young researchers. In the EU VI Framework Program it has been adopted by the TAI-CHI project (Tangible Acoustic Interfaces for Computer-Human Interaction) and by the Networks of Excellence ENACTIVE and HUMAINE. EyesWeb is fully available at its website (`www.eyesweb.org`). Public newsgroups also exist and are daily managed to support the growing EyesWeb community (several thousands of users), including individuals, universities, research institutes, and industries.

5. CONCLUSIONS

This paper introduces MIEEs and presents a partial implementation: EyesWeb 4 and related developments. This is only a first step in the directions sketched in the paper: much work remains to be done. We are in particular working at control strategies and on the so-called META-EyesWeb [9]. In particular, META-EyesWeb is a layer above EyesWeb able to supervise and to schedule execution of patches and subpatches according to adaptive interactive narrative structures (therefore beyond the metaphor of musical instrument). For example, the META-EyesWeb layer can support simple timelines of activation of patches in live electronics performances as, e.g., in Max. But, more interestingly, it supports a dynamic graph of execution (i.e., an interactive narrative structure) where each node is a (sub)patch and each link defines the semantics on how to pass from its input patch to its output patch. For example, it can be defined a "fading" behaviour between two patches, whose parameters can depend on previous history and concurrent active patches.

6. ACKNOWLEDGEMENTS

I am grateful to the colleagues of the Staff of the InfoMus Lab, with whom I had and have the pleasure to work in the research projects presented in this paper. Research is partially supported by the EU IST TAI-CHI project.

7. REFERENCES

[1] A. Camurri, "Interactive Dance/Music Systems," in *Proc. Intl. Computer Music Conference ICMC-95*, The Banff Centre for the Arts, Canada, Sept.3-7, 1995, pp. 245–252, Intl. Comp. Music Association (ICMA).

[2] D. Arfib, J. M. Couturier, and L. Kessous, "Design and use of some new digital musical instruments," in A. Camurri, G. Volpe (Eds.), *Gesture-based Communication in Human-Computer Interaction*, LNAI 2915, Springer Verlag, 2004.

[3] T. Machover and J. Chung, "Hyperinstruments: Musically intelligent and interactive performance and creativity systems," in *Proc. International Computer Music Conference (ICMC89)*, 1989, pp. 186–190.

[4] A. Camurri and P. Ferrentino, "Interactive environments for music and multimedia," *Multimedia Systems*, vol. 7, pp. 32–47, Springer-Verlag, 1999.

[5] A. de Götzen (2004), "Enhancing Engagement in Multimodality Environments by Sound Movements in a Virtual Space," *IEEE Multimedia Magazine*, pp. 4–7, April-June 2004.

[6] A. Camurri and G. Volpe, Eds., *Gesture-Based Communication in Human-Computer Interaction*, Springer, 2004.

[7] A. Camurri and T. Rikakis, Eds., *IEEE Multimedia*, Special Issue on Multisensory Communication and Experience Through Multimedia, vol. 11, no. 3, IEEE CS Press, Jul-Sept. 2004.

[8] R. Rowe, *Interactive Music Systems*. MIT Press, 1993.

[9] A. Camurri, B. Mazzarino, and G. Volpe, "Expressive gestural control of sound and visual output in multimodal interactive systems," in *Proc. Intl. Conf. Sound and Music Computing*, Ircam, Paris, October 2004.

[10] R. Picard, *Affective Computing*. Cambridge, MA, MIT Press, 1997.

[11] S. Hashimoto, "KANSEI as the Third Target of Information Processing and Related Topics in Japan," in *Proceedings of the International Workshop on KANSEI: The technology of emotion,* Camurri A. (Ed.)*,* AIMI (Italian Computer Music Association) and DIST-University of Genova, pp. 101-104, Genova, Italy, 1997.

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine,* no. 1, 2001.

[13] M. Argyle, *Bodily Communication*. Methuen & Co Ltd, London, UK, 1980.

[14] H. G. Wallbott, "The measurement of Human Expressions," in *Aspects of communications*, Walbunga von Rallfer-Engel (Ed.), pp. 203–228, 1980.

[15] K. R. Scherer, "Why music does not produce basic emotions: pleading for a new approach to measuring the emotional effects of music," in *Proceedings Stockholm Music Acoustics Conference (SMAC-03)*, KTH, Stockholm, Sweden, 2003, pp. 25–28.

[16] A. Camurri, P. Coletta, M. Ricchetti, and G. Volpe, "Expressiveness and Physicality in Interaction," *Journal of New Music Research*, vol. 29, no. 3, pp. 187–198, Swets & Zeitlinger, Lisse, The Netherlands, 2000.

[17] A. Camurri, G. De Poli, M. Leman, and G. Volpe, "Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications," *IEEE Multimedia Magazine*, in print.

[18] R. Rowe, *Machine musicianship*, MIT Press, Cambrige MA, 2001.

[19] M. Battier and M. Wanderley (Eds.), *Trends in Gestural Control of Music*, Ircam Publ.

Antonio Camurri (born in Genova, Italy, in 1959; '84 Master Degree in Electrical Engineering in 1984; Ph.D. in Computer Engineering in 1991) is Associate Professor at DIST-University of Genova (Faculty of Engineering), where he teaches "Software engineering" and "Multimedia Systems". He is founder and scientific director of the InfoMus Lab at DIST-University of Genova (www.infomus.dist.unige.it). InfoMus Lab participated to the exploitation of research results in several international artistic productions (e.g. Salzburg Festival 1999, Teatro La Scala, Milano, 1996), in museum and science centre interactive exhibits, in entertainment multimedia, in multimedia systems for therapy and rehabilitation. His research interests include sound and music computing, multimodal intelligent interfaces, computational models of non-verbal expressive gesture, interactive multimodal-multimedia systems for museum, theatre, music, entertainment, therapy and rehabilitation. Project Coordinator of the EU IST-E3 Project MEGA (Multisensory Expressive Gesture Applications, 2000-2003, www.megaproject.org), currently he is local project manager of several EU Projects in the VI Framework Programme (e.g. NoEs ENACTIVE, HUMAINE; IST TAI-CHI; CA 5252) and of industry contracts. He is member of the ExCom of the IEEE CS TC on Computer Generated Music and Associate Editor of the international "Journal of New Music Research".