Towards a multi-layer architecture for multi-modal rendering of expressive actions

G. De Poli[1] F. Avanzini[1] A. Rodà[1] L. Mion[1] G. D'Incà[1] C. Trestino[1] D. Pirrò[1]
A. Luciani[2] N. Castagne[2]

(1) Dep. of Information Engineering DEI/CSC, Padova, Italy
(2) ACROE-ICA, INPG, Grenoble, France
http://www.dei.unipd.it/ricerca/csc http://www.acroe.imag.fr

Abstract

Expressive content has multiple facets that can be conveyed by music, gesture, actions. Different application scenarios can require different metaphors for expressiveness control. In order to meet the requirements for flexible representation, we propose a multi-layer architecture structured into three main levels of abstraction. At the top (user level) there is a semantic description, which is adapted to specific user requirements and conceptualization. At the other end are low-level features that describe parameters strictly related to the rendering model. In between these two extremes, we propose an intermediate layer that provides a description shared by the various high-level representations on one side, and that can be instantiated to the various low-level rendering models on the other side. In order to provide a common representation of different expressive semantics and different modalities, we propose a physically-inspired description specifically suited for expressive actions.

1. Introduction

The concept of expression is common to different modalities: one can speak of expression in speech, in music, in movement, in dance, in touch, and for each of these context the word expression can assume different meanings; this is the reason why expression is an ill-defined concept. In some contexts expression refers to gestures that sound natural (human-like), as opposed to mechanical gestures. As an example, see [11], [9], [10], [3], [4] for musical gestures and [1], [12] for movements.

In other contexts, expression refers to different qualities of natural actions, meaning with this that gestures can be performed following different expressive intentions which can be related to sensorial or affective characteristics. As an example, see [18], [13] for musical gesture, and [15], [16] for movements. These works have shown that this level of expression has a strong impact on non verbal communication, and have led to interesting multi-media applications and to the development of new types of human-computer interfaces.

In this paper we will stick to this latter meaning of expression, therefore when speaking of expression we refer to the deviations from a natural performance of a gesture or action. In section 2 and 3 we will discuss the expressive content in actions from a multimodal perspective. In a rendering system, different application scenarios require different metaphors for expressiveness control. On the other hand, achieving coherence in a multimodal rendering context requires an integrated representation. In order to meet the requirements for flexible and unified representation, we propose in section 4 a multi-layer architecture which comprises three main levels of abstraction. In order to provide a shared representation of different expressive semantics and different modalities, we propose a physically-inspired description which is well suited to represent expressive actions. Some examples and applications are presented in section 5.

2. Multimodal perception and rendering

Looking at how multi-modal information is combined, two general strategies can be identified: the first is to maximize information delivered from the different sensory modalities (sensory combination),

Proceedings of ENACTIVE05
2nd International Conference on Enactive Interfaces
Genoa, Italy, November 17th-18th, 2005

while the second is to reduce the variance in the sensory estimate in order to increase its reliability (sensory integration). Sensory combination describes interactions between sensory signals that are not redundant: they may be in different units, coordinate systems, or about complementary aspects of the same property. By contrast sensory integration describes interactions between redundant signals.

Disambiguation and cooperation are examples for these two interactions: if a single modality is not enough to come up with a robust estimate, information from several modalities can be combined. For example, for object recognition different modalities complement each other with the effect of increasing the information content.

The amount of cross-modal integration depends on the features to be evaluated or the tasks to be accomplished. The modality precision (or appropriateness) hypothesis is often cited when trying to explain which modality dominates under what circumstances. The hypothesis states that discrepancies are always resolved in favor of the more precise or more appropriate modality. In spatial tasks, for example, the visual modality usually dominates, because it is the most precise at determining spatial information. For temporal judgments, however, the situation is reversed and audition, being the more appropriate modality, usually dominates over vision. In texture perception haptics dominates on other modalities, and so on.

When we deal with multimodal rendering of expression in actions, we are interested not only in a fusion at the perceptual level, but also in the modeling and representation level. The architecture of the system should be specifically designed for this purpose, taking into account this problem. Normally a combination of different models, one for each modality, is used. These models map directly intended expression on low level parameters of the rendering system. We believe that a proper definition of a common metaphoric level is a fundamental step for the development of effective multimodal expression rendering strategies.

3. Expression in different modalities

A second point to be addressed when looking for a better definition of expression is the wide range of expressive gestures that are studied in the literature. Roughly, we can identify studies on three level of gestures: single gestures (see [6], [7]), simple pattern-based gestures (see e.g. [5], [8], [1]), and structured gestures (see [13], [18] for musical gesture, and [15], [17] for movement). We can think about analogies between music and movement with reference to these three levels of structural complexity. By single gestures we intend single tones for music or simple movements like arm rotation. These single gestures represent the smallest non structured actions, which combined together form simple patterns. Single patterns in music can be represented by scales or repetition of single tones, while example of basic patterns in movement are a subject walking or turning. Highly structured gestures in music are performances of scores, while in movement we can think about a choreography. This classification yields interesting analogies between the different structures of gestures in music and dance, and provides a path to a common representation of different expressive semantics.

The literature on expressiveness analysis and rendering exhibits an evident lack of research on the haptic modality with respect to the visual and audio modalities. This circumstance can be explained by observing that the haptic modality does not present a range of structured messages as wide as for audio and video (e.g., music or speech, and dance, respectively). In fact, due to the very nature of haptic perception, haptic displays are strictly personal and are not suitable for communicating information to an audience. This is why just a very few kinds of structured haptic languages have been developed along the history.

The haptic modality is indeed hugely important in instrument playing for controlling the expressive content conveyed by other modalities, as shown for example by the haptic interaction between a player and a violin, which quality affects deeply the expressive content in the sound. On the contrary tactile-kinesthetic perception, despite its importance in the whole multisensory system, does not seem to convey expressivity back to the player [31].

4. An architecture for multi-modal expressive rendering

In order to meet the requirements for flexible representation of expressiveness in different application scenarios, we propose a multi-layer architecture which comprises three main levels of abstraction. At the top there is a semantic description, which stays at the user level and is adapted to a specific representation: for example, it should be possible to use a categorical approach (with affective or sensorial labels) or a dimensional approach (i.e. the valence-arousal space) [36].

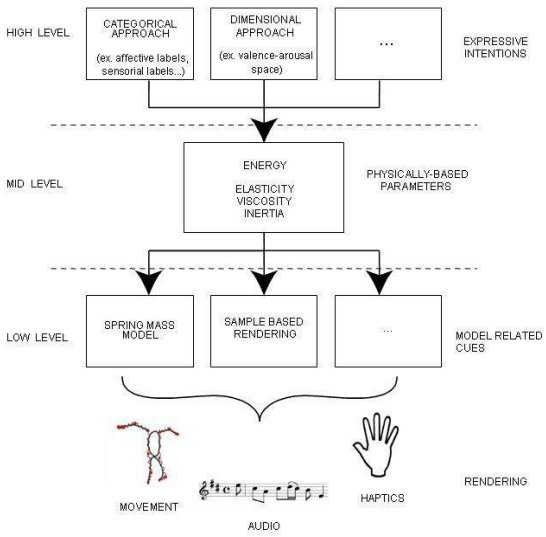

Figure 1: Multi-layer architecture

At the other end are low-level features that describe parameters strictly related to the rendering models. Various categories of models can be used to implement this last level. Sticking to the musical example at this level, signal-based sound synthesis models are adapted to represent note onset, duration, intensity, decay, etc. As depicted by Cadoz *et al.* in [32], physical models can be adapted to render timbre characteristics, interaction properties (collision, friction), dynamic properties as transients (attacks), evolution of decays (laws of damping), memory effects (hysteretic effects), energetic consistency between multisensory phenomena, etc.

Physical-modeling techniques have been investigated for years and have proven their effectiveness in rendering rich and organic sounds [2]. Among such rendering techniques, one of the models that are best suited for controlling expressiveness is made of a network of masses and interactions [32]. Basic physical parameters of the masses and interactions (damping, inertia, stiffness, etc.) determine the behavior of the model. A change in parameters affects the audio rendering, and especially its expressive content.

In between these two extremes, an intermediate layer provides a description that can be shared by the various high-level representations on one side, and can be instantiated to the various low-level rendering models on the other side. In order to provide a common representation of different expressive semantics and different modalities, we propose a physically-based description. For the definition of the intermediate level we need the different modalities to converge towards a common description. In this case, we want this description of the actions (movements, objects and so on) to be based on a physical metaphor. This choice arises from the fact that expressive contents are conveyed by gestures, which are essentially physical events. Therefore, direct or indirect reference to human physical behavior can be a common denominator to all the multi-modal expressive actions and yield a suitable representation.

Using a single model for generating the various categories of phenomena allows to enhance energetic coherency among phenomena [30]. Furthermore, such a physically-based mid-level description is shifted towards the "source side", which is better suited for multi-modal rendering. This amounts to making a shift from existing rendering techniques which are derived from perceptual criteria (at the 'receiver side') and are therefore referred to a specific modality or medium (e.g., music).

The main effort needed at this point is to define a suitable space for this physical metaphor-based description. We have a set of dimensions which describe actions by metaphors. This space must be described by mid-level features, which provide the overall characteristics of the action. As an example, consider a pianist or a dancer who wants to communicate, during a performance, an intention labeled as "soft" (in a categorical approach). Each performer will translate this intention into modifications of his action in order to render it softer, e.g. by taking into account the attack time of single events (such as notes or steps). The actions will therefore be more "elastic" or "weightless". These and other overall properties (like "inertia" or "viscosity"), together with "energy" (used as a scale factor), will be taken into account to define the mid-level description. Citing Castagné, "though users are not commonly confronted in an intellectual manner with the notions of inertia, damping, physical interaction etc., all these notions can be intuitively apprehended through our body and our every-day life" [34].

This kind of multi-layered approach is exemplified in figure 1.

5. Experiments on expression mapping

Previous experiments conducted at CSC/DEI in Padova led to interesting results on automatic detection of expression for different types of gestures.

These studies showed that the expressive content of a performance can be changed, both at the symbolic and signal levels. Psychophysical studies were also conducted in order to construct mappings between acoustic features of sound events and the characteristics of the physical event that has originated the sound in order to achieve an expressive control of everyday sound synthesis.

5.1. Mid-level feature extraction from simple musical gestures

Several experiments on analysis of expression on simple pattern-based musical gestures have been previously carried out. In [5] short sequences of repeated notes recorded with a MIDI piano were investigated, while [19] reports upon an experiment on expression detection on audio data from professional recordings of violin and flute (single repeated notes and short scales). In both works, the choice of the adjectives describing the expressive intention has been considered as an important step for the success of the experiments. In [5], the choice of adjectives has been based on theories of Imberty [20] and Laban [21]. Laban believed that the expressive content of every physical movement is mainly related to the way of performing it, and it is due to the variation of four basic factors: time, space, weight and flow. The authors defined as *basic efforts* the eight combinations of two values (quick/sustained, flexible/direct and strong/light) associated with the first three factors. Each combination gives rise to a specific expressive gesture to which is associated an adjective, as an example a slashing movement is characterized by a strong weight, quick time and flexible space (i.e., a curved line).

It was supposed that sensorial adjectives could be more adequate for an experiment on musical improvisations, since they suggest a more direct relation between the expressive intention and the musical gestures. Starting from Laban theory of expressive movement, the set of adjectives for our experiments was derived by analyzing each of the eight combinations of the values *high* and *low* assigned to articulation, intensity and tempo (velocity). Both value series [quick/sustained, flexible/direct and strong/light] and [articulation, intensity and tempo] have a physical base and can be related to the concepts of energy, inertia, elasticity and viscosity.

Factor analysis on the results of a perceptual test indicated that the sonological parameters tempo and intensity are very important in perceiving the expression of this pattern-based musical gestures. Also, results of a perceptual test showed that listeners can recognize performer's expressions even when very few musical means are used.

Results of analysis were used to tune machine learning algorithms, to verify their suitability for automatic detection of expression. As an example, we used Bayesian networks and a set of HMMs able to give as output the probability that the input performance was played according to an expressive intention [22]. High classification ratings confirmed that automatic extraction of expression from simple pattern-based musical gestures can be performed with a mid-level description.

5.2. Mid-level feature extraction from complex musical gestures

In [27] we showed that a musical performance with a defined expressive intention can be automatically generated by modifying a natural performance of the same musical score. This requires a computational model to control parameters such as amplitude envelope, tempo variation (e.g. accelerando, ritardando, rubato), intensity variation (e.g. crescendo, decrescendo), articulation (e.g. legato, staccato), by means of a set of profiles. A family of curves, which presents a given dynamic evolution, is associated to every expressive intention. Fig.2 shows an example of curves for the control of amplitude envelopes. These curves present strict analogies with motor gestures, as already highlighted by various experimental results (see [28], [29], [10] among others) and the concepts of inertia, elasticity, viscosity and energy can be therefore easily related to them.

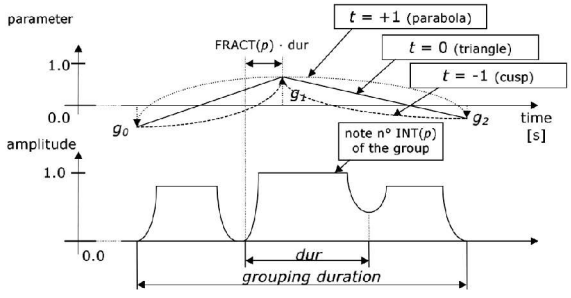

Figure 2: Curves to control the amplitude envelope of a group of notes.

6. Mid- to low-level mappings

As already mentioned, the main open issue for the realization of the multi-layer architecture proposed in this paper is the definition of mappings from the intermediate, shared representation and the low-level features that describe parameters strictly related to the rendering models. In this section we analyze two relevant examples of such mappings.

6.1. Ecological mapping

Many studies in ecological acoustics address the issue of the mapping between acoustic features of sound events and the characteristics of the physical event that has originated the sound [23]. As an example, it has been found that the material of a struck object can be reliably recognized from the corresponding impact sound. In previous studies we have developed a library of physically-based sound models based on modal synthesis [24], that allow simulation of many typologies of such "everyday sounds" and specifically contact sounds (impact, friction, bouncing, breaking, rolling, and so on). Using these models we have conducted a number of psychophysical studies in order to construct mappings between the "ecological level", e.g. object material, or hardness of collision, or viscosity in motion, and the low-level physical parameters of the sound models (see e.g. [25] and [26]). Such an "ecological-to-physical" mapping can be straightforwardly incorporated into the multi-layer architecture that we propose in this paper, where the ecological level corresponds to the mid-level physically-based parameters which maps to the low-level parameters of the modal synthesis models. In this way we realize expressive control of everyday sound synthesis.

6.2. Physically-based expression rendering

In [35] Cadoz demonstrated that physical modeling is suited not only for sound synthesis but also for the synthesis of musical gesture and musical macro-evolution. As explained in that paper, one can obtain a succession of sound events rather than isolated sounds by assembling both high and low frequency mass-interaction physical models into a complex structure. The low frequency structure then stands for a modelling and simulation of instrumental gesture.

In this process, low frequency models are slightly perturbed in a natural manner through feedback from the sound models. Therefore the generated sound events present convincing short-term evolutions, expressiveness and musicality, such as changes in a rhythm or in the timbre of successive musical events – somehow resembling the way a musician would behave.

In motion control and modelling, physically-based particle models can be used to simulate a human body, not as a realistic biomechanical model, but rather as an abstract minimal representation that allows access to the control of the quality of dance motions as they are thought and experienced by dancers during the performance and by teachers [1]: motors of motion, propagation, external resistance, full momentum transfers, etc. This minimal model produces the quality of the motion in a "natural way of performance and thinking" (figure 3 left). In a similar way, Luciani used this type of model in [33] to control the expressive evolution in visual animation as shown in figure 3 right).

Thus, by implementing the middle level of figure 1 through mass-interaction models that stand for musical gesture generators, and by controlling the physical parameters of these models through outputs of the first semantic level, it becomes possible to control the quality of the "instrumental gesture". The instrumental gesture model will then generate accordingly musical events that have some expressive content, and will be mapped onto the last audio rendering level.

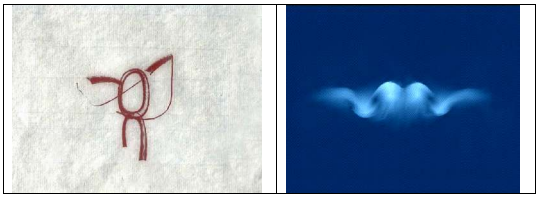

Figure 3. Physically-based particle model for dance and animation

7. Expression rendering systems

In this section we show some concrete examples of instantiation of the proposed architecture, with reference to the models described in previous section.

Our studies on music performances [13] have shown that the expressive content of a performance can be changed, both at the symbolic and signal levels. Models able to apply morphing among performances with different expressive content were investigated, adapting the audio expressive character to the user desires.

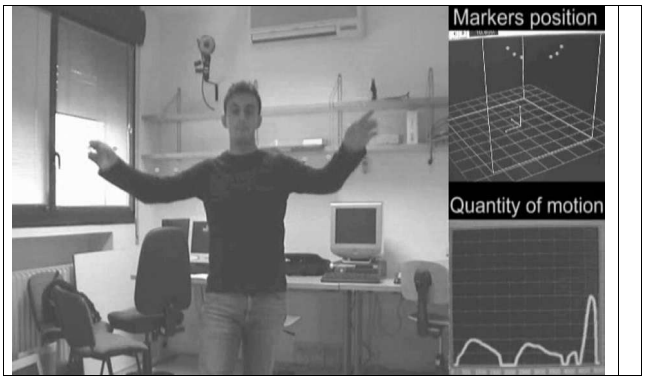

Figure 4. The expressive movements of a dancer control a musical performance

The input of the expressiveness models are composed of a musical score and a description of a neutral musical performance. Depending on the expressive intention desired by the user, the expressiveness model acts on the symbolic level, computing the deviations of all musical cues involved in the transformation. The rendering can be performed by a MIDI synthesizer and/or by driving an audio processing engine. As an example, we can deduce a desired position in the energy–velocity space from analysis and processing of the movement of a dancer in a multimodal setting (fig. 4), and then use this space position as a control input to the expressive content and the interaction between the dancer and the final music performance [15].

On the other side, recent studies at INPG have showed that dynamic models are suitable for the production of natural motions (fig. 3). By designing his own dynamic model, the user has a high level motion control to modify the quality of such dynamically generated movement.

8. Conclusions

In this paper we have proposed a multi-layer architecture which comprises three main levels of abstraction: a semantic description at the top provides the user-level layer and can be adapted to specific user requirements and conceptualization; low-level features at the other end describe parameters strictly related to the rendering model; in between these two extremes, we proposed a physically-inspired description, which is particularly suited to expressive actions and provide a common representation of different expressive semantics and different modalities.

We have proposed direct or indirect reference to human physical behaviour, as a common denominator to multi-modal expressive actions that allows to enhance energetic coherency among phenomena. Furthermore, such a mid-level description is shifted towards the ‘source side’, which makes it suited for multi-modal rendering applications.

Although users are not necessarily familiar with the concepts of inertia, damping, physical interaction etc., all these notions can be intuitively learned through every-day interaction and experience. This amounts to making a shift from existing rendering techniques which are derived from perceptual criteria (at the ‘receiver side’) and are therefore referred to a specific modality/medium (e.g., music).

References

[1] C.M. Hsieh, A. Luciani, ``Physically-based particle modeling for dance verbs", *Proc of the Graphicon Conference 2005*, Novosibirsk, Russia, 2005.

[2] N. Castagné, C. Cadoz, ``GENESIS: A Friendly Musician-Oriented Environment for Mass-Interaction Physical Modeling", *International Computer Music Conference - ICMC 2002* – Goteborg – pp. 330-337, 2002.

[3] B. Repp, ``Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists", *Journal of Acoustical Society of America*, vol. 88, pp. 622-641, 1990.

[4] B. Repp, ``Diversity and commonality in music performance: an analysis of timing microstructure in Schumann's 'Traumerei'", *Journal of Acoustical Society of America*, vol. 92, pp. 2546-2568, 1992.

[5] F. Bonini, A. Rodà, ``Expressive content analysis of musical gesture: an experiment on piano improvisation", *Workshop on Current Research Directions in Computer Music*, Barcelona, 2001.

[6] M. Melucci, N. Orio, N. Gambalunga, ``An Evaluation Study on Music Perception for Content-based Information Retrieval", *Proc. Of International Computer Music Conference*, Berlin, Germany, pp. 162-165, 2000.

[7] E. Cambouropoulos, ``The Local Boundary Detection Model (LBDM) and its Application in the Study of Expressive Timing", *Proceedings of the International Computer Music Conference (ICMC 2001),* 17-22 September, Havana, Cuba, 2001.

[8] L. Mion, ``Application of Bayesian Networks to automatic recognition of expressive content of piano improvisations", in *Proceedings of the SMAC03 Stockholm Music Acoustics Conference*, Stockholm, Sweden, pp. 557-560, 2003.

[9] N. P. Todd, ``Model of expressive timing in tonal music", *Music Perception*, vol. 3, pp. 33-58, 1985.

[10] N. P. Todd, ``The dynamics of dynamics: a model of musical expression", *Journal of the Acoustical Society of America*, 91, pp. 3540-3550.

[11] A. Friberg, L. Frydèn, L. Bodin, J. Sundberg ``Performance Rules for Computer-Controlled Contemporary Keyboard Music", *Computer Music Journal*, 15(2): 49-55, 1991.

[12] D. Chi, M. Costa, L. Zhao, N. Badler, ``The EMOTE Model for Effort and Shape", In *Proceedings of SIGGRAPH00*, pp. 173-182, July 2000.

[13] S. Canazza, G. De Poli, C. Drioli, A. Rodà, A. Vidolin ``Modeling and Control of Expressiveness in Music Performance", *The Proceedings of the IEEE*, vol. 92(4), pp. 286-701, 2004.

[14] R. Bresin, ``Artificial neural networks based models for automatic performance of musical scores", *Journal of New Music Research*, 27(3):239–270, 1998.

[15] A. Camurri, G. De Poli, M. Leman, G. Volpe, ”Communicating Expressiveness and Affect in Multimodal Interactive Systems”, *IEEE Multimedia*, vol. 12, n. 1, pp. 43-53, 2005.

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

[17] K. Suzuki, S. Hashimoto, ``Robotic interface for embodied interaction via dance and musical performance", In G. Johannsen (Guest Editor), *The Proceedings of the IEEE*, Special Issue on Engineering and Music, 92, pp. 656–671, 2004.

[18] R. Bresin, A. Friberg, ``Emotional coloring of computer controlled music performance", *Computer Music Journal*, vol. 24, no. 4, pp. 44–62, 2000.

[19] L. Mion, G. D'Incà, ``An investigation over violin and flute expressive performances in the affective and sensorial domains", *Sound and Music Computing Conference (SMC 05)*, Salerno, Italy, 2005 *(submitted)*.

[20] M. Imberty, *Les ecritures du temps*, Dunod, Paris, 1981.

[21] R. Laban, F.C. Lawrence, *Effort: Economy in Body Movement*, Plays, Inc., Boston, 1974.

[22] D. Cirotteau, G. De Poli, L. Mion, A. Vidolin, and P. Zanon, "Recognition of musical gestures in known pieces and in improvisations", In A. Camurri, G. Volpe (eds.) *Gesture Based Communication in Human-Computer Interaction*, Berlin: Springer Verlag, pp. 497-508, 2004.

[23] W. W. Gaver, ``What in the world do we hear? An ecological approach to auditory event perception", *Ecological Psychology*, 5(1):1 29, 1993.

[24] F. Avanzini, M. Rath, D. Rocchesso, and L. Ottaviani, ``Low-level sound models: resonators, interactions, surface textures", In D. Rocchesso and F. Fontana, editors, *The Sounding Object*, pages 137-172. Mondo Estremo, Firenze, 2003.

[25] L. Ottaviani, D. Rocchesso, F. Fontana, F. Avanzini, ``Size, shape, and material properties of sound models", In D. Rocchesso and F. Fontana, editors, *The Sounding Object*, pages 95-110. Mondo Estremo, Firenze, 2003.

[26] F. Avanzini, D. Rocchesso, S. Serafin, ``Friction sounds for sensory substitution", *Proc. Int. Conf. Auditory Display (ICAD04)*, Sydney, July 2004.

[27] Canazza S., De Poli G., Di Sanzo G., Vidolin A. ``A model to add expressiveness to automatic musical performance", In *Proc. of International Computer Music Conference,* Ann Arbour, pp. 163-169, 1998.

[28] Clynes, M. ``Sentography: dynamic forms of communication of emotion and qualities", *Computers in Biology & Medicine*, Vol, 3: 119-130, 1973.

[29] Sundberg J, Friberg A. ``Stopping locomotion and stopping a piece of music: Comparing locomotion and music performance", *Proceedings of the Nordic Acoustic Meeting Helsinki 1996*, 351-358, 1996.

[30] A. Luciani, “Dynamics as a common criterion to enhance the sense of Presence in Virtual environments”. *Proceedings of “Presence Conference 2004”*. Oct. 2004. Valencia. Spain.

[31] A. Luciani, J.L. Florens, N. Castagné. “From Action to Sound: a Challenging Perspective for Haptics”, *Proceedings of WHC Conference 2005*.

[32] C. Cadoz , A. Luciani, J.L. Florens: "CORDIS-ANIMA: a Modeling and Simulation System for Sound and Image Synthesis- The General Formalism", *Computer Music Journal*, Vol. 17-1, MIT Press, 1993.

[33] A. Luciani, “Mémoires vives”. Artwork. Creation mondiale. *Rencontres Internationales Informatique et Création Artistique*. Grenoble 2000.

[34] N. Castagné, C. Cadoz : "A Goals-Based Review of Physical Modelling" - *Proc. of the International Computer Music Conference ICMC05* - Barcelona, Spain, 2005.

[35] C. Cadoz, "The Physical Model as Metaphor for Musical Creation. pico..TERA, a Piece Entirely Generated by a Physical Model", *Proc. of the International Computer Music Conference ICMC02,* Sweden, 2002.

[36] P. Juslin and J. SLoboda (eds.), *Music and emotion: Theory and research*, Oxford Univ. Press, 2001