

Sound And Music for Everyone Everyday Everywhere
Every way

CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

SAME

D4.2 CONCEPTUAL MODEL

<i>Version</i>	<i>Edited by</i>	<i>Changes</i>
V0.2	F.Bevilacqua, N. Rasamimanana, A.Cont	First draft
V0.3	F.Bevilacqua, N. Rasamimanana, A.Cont	Reformulated after Sept5 meeting
V0.4	A. Camurri, G. Volpe	Integrated text from IRCAM and UGDIST
V0.5	R. Bresin	Added text from KTH
V0.6	F. Bevilacqua	Corrections/additions from Ircam
V0.7	A. Camurri, G. Volpe	Pre-final version

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

TABLE OF CONTENTS

Table of contents	2
1. Introduction	3
2. Background.....	4
2.1 Layered analysis of expressive gesture	4
2.2 Gesture follower, recognition and interaction	7
2.3 Emotion in music performance: analysis and synthesis	8
3. Concepts.....	9
3.1 Enaction and ecological knowledge	9
3.2 Multimodality/cross-modality/a-modality	10
3.3 Collaborative and social interaction.....	11
3.4 Context awareness	12
3.5 Time modeling and synchronization	13
3.5.1 Time scale and prediction	13
3.5.2 Sound and gesture descriptors	14
3.5.3 Time modeling and social descriptors	15
4. The SAME model.....	15
4.1 Model description	16
4.1.1 Interaction with the physical space.....	17
4.1.2 Interaction with physical objects.....	18
4.1.3 Feature spaces.....	19
4.1.4 Affective/emotional spaces.....	19
4.1.5 Context-awareness	20
4.1.6 Social interaction	21
4.2 Techniques	21
4.3 Types of interaction	23
5. References.....	25

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

1. INTRODUCTION

In this document we describe a conceptual model for multimodal processing of gesture and embodied music content. The document is organized as follows. First, we recall important previous works since this conceptual model can be seen as a continuation of previous researches by SAME partners. Second, we describe concepts and key elements for the conceptual model coming from the analysis of the Use Cases and Frameworks described in D2.2 and D.3.1. Third, we explicit the conceptual model by describing its components, the different types of gesture and social descriptors used in it, and their relationships to the various concepts. Descriptors are technically illustrated in D3.1 and D4.1. We also discuss the interaction models and in particular how this conceptual model extend previous established models.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

2. BACKGROUND

2.1 Layered analysis of expressive gesture

A relevant source for the work in the SAME Project is the research carried out in previous FP5 and FP6 projects dealing with topics strictly related to SAME issues. One of such projects, the FP5 EU-IST MEGA Project (Multisensory Expressive Gesture Applications, November 2000 – October 2003), worked out a multilayered conceptual model for multimodal analysis of expressive content in human full-body movement and gesture and for synthesis of expressive music as well as visual output (Camurri et al., 2005).

Whereas, from the one hand, such a model mainly refers to full-body movement and gesture and more limited interaction with physical objects (e.g., a mobile device moved with a hand) is taken into account only implicitly (in the sense that the model does not prevent to be applied also in such a case), on the other hand, it represents a quite consolidated model, exploited in several real world applications, that can be extended and adapted in order to encompass the aspects that are peculiar to the SAME Project, i.e., physical interaction with objects (mobiles), context-awareness, and social interaction.

A major purpose of the MEGA conceptual model is modeling of expressive gesture. Expressive gesture is understood as a general concept including musical, human movement, visual (e.g. computer animated) gesture. An attempt of defining the concept of expressive gesture can be found in (Camurri et al., 2004). Such definition finds its basis on Kurtenbach and Hulteen's (1990) definition of gesture stating that gesture is "a movement of the body that contains information". Especially in performing arts (the application field mainly addressed by the MEGA project), gesture is not only intended to denote things or to support speech as in the traditional framework of natural gesture, but the information it contains and conveys is often related to the affective, emotional domain. From this point of view, gesture can be considered "expressive" since it carries what Cowie and colleagues (2001) call "implicit messages", and what Hashimoto (1997) calls KANSEI. That is, expressive gesture is the responsible of the communication of information called "expressive content". Expressive content is different and in most cases independent from, even if often superimposed to, possible denotative meaning. Expressive content concerns aspects related to feelings, moods, affect, intensity of emotional experience.

The multi-layered model for analysis of expressive gesture developed in MEGA builds on four different layers/steps following a bottom-up approach (Camurri et al., 2004). Figure 1 provides a sketch of the MEGA model.

Sound And Music for Everyone Everyday Everywhere Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

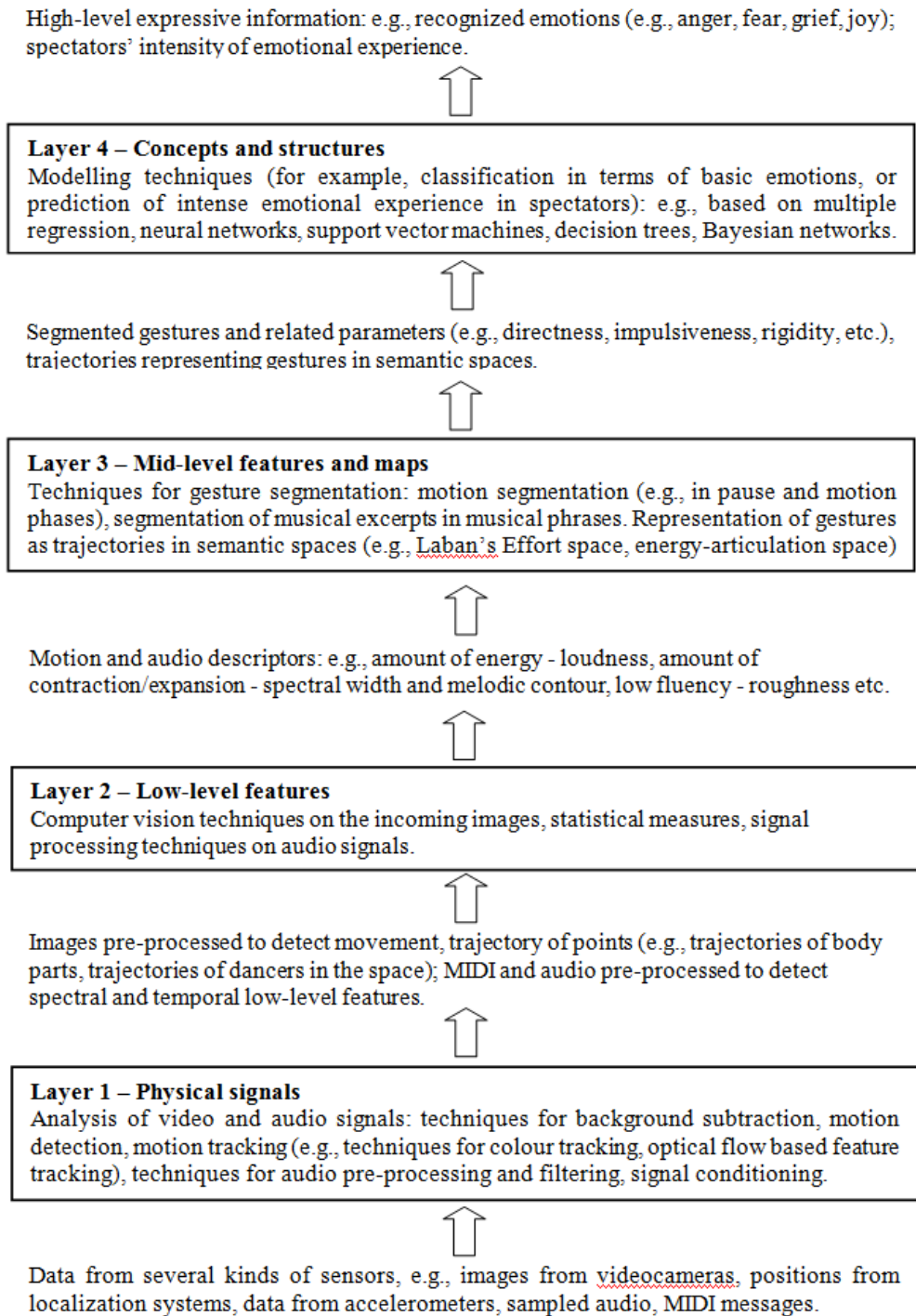


Figure 1. A snapshot of the multilayered conceptual model for analysis of expressive gesture worked out in the framework of the FP5 EU-IST Project MEGA (Multisensory Expressive Gesture Applications).

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

Layer 1 (Physical Signals) includes techniques for gathering data captured by sensors such as videocameras, on-body sensors (e.g., accelerometers), sensors of a robotic system, environmental sensors. Data are collected from different sources in order to obtain a vector of measurements of the same event (i.e., a movement) as rich as possible. Sources at this level may be either physical sensors (tactile, haptic, infrared, or ultrasound sensors) or low-level measures performed on single video frames (e.g., employing techniques like optical flow, motion templates, etc.). Layer 1 produces as output vectors of measures and processed video frames.

Layer 2 (Low-level features) extracts from the sensors data a collection of motion features describing the movement being performed. The MEGA Project focused on specific kinds of cues derived from research by psychologists (e.g., Wallbott, 1998; Boone and Cunningham, 1998) and artists: an important set of features are those inspired to the Effort dimensions described by the choreographer Rudolf Laban in his Theory of Effort (Laban, 1947, 1963). Features developed in MEGA included for example kinematical measures (speed, acceleration of body parts), detected amount of motion, amount of body contraction/expansion. However, mainly because of technological constraints (e.g., low space and time resolution videocameras at that time), research in MEGA could not address subtler features such as impulsiveness, rigidity, heaviness, fluidity, etc., nor the project directly focused on cues measured from manipulation of objects. Also, MEGA did not implement the whole Laban's Effort framework and limited research to some of the Laban's dimensions (e.g., the space dimension through the directness index).

Layer 3 (Mid-level features and maps) deals with two major issues: segmentation of movement in its composing gestures, and representation of such gestures in suitable spaces. Thus, the first problem here is to identify relevant segments in the movement stream and associate to them the qualities/features 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 analyzed e.g., in order to find similarities among them and to group them in clusters. The intermediate feature space proposed in the SAME model can be considered as an extension of this layer.

Layer 4 (Concepts and structures) is conceived as a conceptual network mapping the extracted features and gestures into (possibly verbal) conceptual structures. For example, a dance performance can be analyzed in term of the performer's conveyed emotional intentions: in the MEGA Project an experiment was carried out aiming at distinguishing among the four basic emotions (anger, fear, grief, and joy) in dance performances. Another research direction, started in MEGA and which is still open, is modeling spectators' engagement. Machine learning techniques can be employed in this layer ranging from

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

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.

The MEGA conceptual model already included some mechanisms for taking context into account and for enabling dynamic adaptation of the model to different contexts. The aim was to adapt the behavior of the analysis and synthesis Layers, while preserving at the same time the modularity of the conceptual framework. This was obtained by including some intermediate modules in between the Layers. Suppose for example to have a module at Layer 3 extracting the “scenic presence” of a dancer. The “scenic presence” would be a mid-level feature that could be employed by Layer 4 for classifying the dancer’s current expressive intention. Modules in Layer 1 and 2 provide Layer 3 with the information needed to perform this task. On the basis of this information Layer 3 calculates an index of scenic presence. Suppose now that such an analysis is done in a situation in which lights are particularly relevant. In this case, lighting on stage can strongly affect (in a non-linear way) the scenic presence index. For example, if the dancer were standing in a lighted area in front of the stage, his/her scenic presence would sensibly grow. Scenic presence therefore needs to be emphasized in a non-linear way. This can be done by means of an intermediate (between Layer 3 and Layer 4) module taking as inputs the calculated index of scenic presence, the outputs from the physical layer (stage coordinates, lighting position and intensity), and from the low-level features layer (e.g., amount of detected motion to understand whether the dancer is or is not moving), and generating as output a modified (enhanced) index of scenic presence. The mid/high-level feature “index of scenic presence” already implicitly depends on the actual values of the low-level features from Layers 1 and 2: the mechanism implemented by the intermediate module allows to adapt and tune the index according to the desired focus of attention.

Such dynamic adaptation to changing contexts, even if already considered at a conceptual level, was never fully implemented in the framework of MEGA. Instead, it represents a primary goal for SAME.

2.2 Gesture follower, recognition and interaction

Another relevant source for SAME, especially with respect to analysis of gesture in its temporal development, is the research carried out at IRCAM concerning the development of interactive systems for performing arts: digital sound and visuals are mediated in real-time by the gesture and motion of instrumentalists/dancers. For this, IRCAM have established recently a general framework of analysis based on machine learning techniques. This is informed by IRCAM fundamental research studies on music and dance performances. This research focused on the analysis of multimodal movement and audio data, and particularly investigated the different constraints instrumentalists face when playing acoustic instruments. Such studies bring forth basic components of embodied

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

interaction and communication concepts that are exploited for the development of electronic interactive systems. For example, the analyses of bowed string movements revealed some of the constraints, in particular acoustical or biomechanical (Rasamimanana, 2006; 2008a; 2008b; 2008c) from which instrumentalists create expressive music.

This research lead to a first important model taking into account fundamentally the temporal behavior of gesture, instead of just considering posture. Machine learning approaches provide a strong base for temporal modeling. For example, technologies such as score following (Cont, 2008) and gesture following (Bevilacqua et al., 2007) are based on temporal modeling of musical objects. Importantly, the modeling of temporal processes allows for the consideration of past events as well as upcoming events through anticipation.

The research findings are currently applied to the creation of novel types of interactions for artistic performance (Bevilacqua et al., 2006; Cont, 2008).

2.3 Emotion in music performance: analysis and synthesis

Results from previous research which will contribute to the realization of the SAME conceptual model include those obtained at KTH in the field of analysis and synthesis of emotion in expressive music performance.

Research in automatic expressive music performance was started at KTH in the 70's and it is summarized in a recent overview paper (Friberg et al, 2006). Main achievements previous to the start of the SAME projects have been:

- the design of the Director Musices (DM) system for expressive music performance implementing about 30 performance rules, which modify acoustic parameters such as duration, articulation, sound level, vibrato, attack time, envelope for each note (Bresin et al., 2002)
- the identification and definition of macro-rules for automatic emotionally expressive music performance (Bresin & Friberg, 2000; Juslin et al., 2002)
- the design of pDM real-time system for expressive music performance (Friberg, 2006)
- methods and tools for the control of music performance using gesture controllers, sensors, and body movements (Bresin et al., 2003; Friberg, 2005a, 2005b)

In summary the tools produced from KTH and made available in the SAME project allow to modify the emotional expression of a music score in real-time by manipulating a set of acoustic parameters. This modification can be done using different tools, both software and hardware.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

3. CONCEPTS

The SAME conceptual model should keep into account results from theories of perception. Listening to sound and seeing a movement are intimately linked from the point of view of perceptual and cognitive processes. Sounds and music evoke motor imagery. How perception is linked to action? Perception is an active and explorative process. Consider for example SAME Framework 3 (Mobile Orchestra Explorer): users in the Orchestra Explorer perform free movements that may be explained as explorative behavior of perceivers attending and acting on a sonic event. Explorative movement is based on motor imagery, rooted in sensorimotor skills. Similarly, in the SAME Framework 2, the way that users spontaneously move on sound and music are exploited as initial inputs for iterative and adaptive models. Here follows some links to important theories and concepts of relevance to the SAME conceptual model.

3.1 Enaction and ecological knowledge

The SAME Project focuses on gesture data produced as mutual interaction with sound data. In this framework, gestures are considered as a result of a sensorimotor perception/action coupling, which can be referred as an *enactive* approach. This approach questions the classical cognitive science paradigm, based on a linear schema of the perceptual system where sensory inputs are interpreted with respect to internal symbolic representations and where actions are considered as the result from this cognitive process. On the contrary, enaction builds on the following principles:

- Perception is an activity of sensorimotor coupling with the environment (or a destined result);
- Features are presented as available rather than as represented;
- Action and perception are inextricably bound.

In this view, gestural/corporeal activity can act as a mode of exploration of all given possibilities drawing on implicit understanding of sensorimotor regularities. The perception of emotion and the sense of interaction in music can be strongly linked to our expectation on dependencies between some gestural activities and sound environments. For example, the sound emanating from an approaching object tends to become louder and more intense: this example refers to a sort of pre-understanding, referring to the role of "*ecological knowledge*" (Gibson) in the mapping of sonic and movement signals.

This view is supported by recent literature in vision and music cognition (Noe, 2004; Leman, 2007; London, 2004).

This cognition paradigm has direct modeling implications that have proved to be very fruitful with real-world data. While traditional processing approaches to recognition and

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

decision making consider analysis and action as separate steps, this cognitive inspired approach considers the two as part of the same process.

3.2 Multimodality/cross-modality/a-modality

Multimodal analysis enables the integrated analysis of information from different sensorial modalities (auditory, visual). It allows integration of features and use of complementary information, e.g., use of information in a given modality for supplementing lack of information in another modality or for reinforcing analysis in another modality.

Cross-modal analysis enables exploiting potential similarities in the approach for analyzing information in different modalities: so, for example techniques developed for analysis in a given modality (e.g., audio) can also be used for analysis in another modality (e.g., video); further, commonalities at mid- and high-level in representations of different sensory channels are an important perspective for developing models based on *a-modal*, converging representations.

In multisensory research there is a reference to a-modal perception, related to the observation that something we hear is similar to something we see, or that we do with our body. We perceive something "similar" regardless to sensory modality.

Intensity, rhythm, shape, texture, spatial extent, spatial location, duration, temporal rate are attributes that are transferred from one modality to another. The ability to perceive a-modally seems to be a feature innate in the perceptual system or learned at an early developmental stage. It is deemed important also in the development of multisensory perceptual skills (Lewkowicz and Kraebel, 2004; Lickliter and Bahrick, 2004).

Also the theory developed by Stern (2000) puts into evidence a-modal perception as a key component of caregiver/infant relation: e.g., a sound may be heard as being similar to a seen movement. He proposes a theory based on the concept of "vitality affects" to denote the almost infinite variations of an emotion, e.g., joy, or of an action (the close of a door - see e.g., Pollick). These elusive qualities are better captured by dynamic, kinetic terms, such as "surging", "fading away", "fleeting", "explosive", "crescendo", "diminuendo", "bursting", "drawn out", etc. Each of these variations can be described in terms of an "activation contour", i.e., a dynamic profile: the way the dynamics, or the intensity, or the emotion or the action evolve in time. Such activation contours are deemed to be a-modal. A-modal perception seems to provide a source to our sensitivity to create relations (automatic, reactive) between sonic content and movement. Complementary points on time modeling are further discussed in section 3.5.

Further relevant concepts in Stern's theory are (i) *cross-modal correspondence*, intended as imitation and affect attunement; (ii) *imitation* as reproduction of overt behavior; and (iii) *affect attunement* (in mother-infant relation) referred as features in mother's execution of actions, i.e., subtleties in the mother's performance that not only reflect or mimic the overt behavior of the child, but that are rooted in the mother's intention of making known to the

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

child that she understands the inner-feeling state of the child. Attunement involves an element of imitation, but it is distinguished from imitation on the basis of the kind of intuition or empathy that adults exhibit when with small children.

Stern proposes three (a-modal) features as important for affect attunement: intensity, timing, shape. He then lists a set of features that in more detail explains how these overall features may be involved in affect attunement:

- “1. absolute intensity: the level of the intensity of the mother's behaviour is the same as that of the infant's, irrespective of the mode or form of the behaviour[...]
2. Intensity contour
3. Temporal beat. Regular pulsation in time is matched
4. Rhythm. A pattern of pulsations of unequal stress is matched
5. Duration. The time span of the behaviour
6. Shape. Some spatial feature of a behaviour that can be abstracted and rendered in a different act is matched" (Stern, 2000: p.146)

Activation contour is closely related to the experiencing and expression of vitality affects. How “good” is the match required to be accepted as an imitation? Certain salient or accentuated moments should be precise, while before and after these moments precision may be lower (Wohlschlager, Gattis, Bekkering, 2003). Therefore, an action that is intended to imitate another action does not have to be a perfect reproduction to be regarded as an imitation.

3.3 Collaborative and social interaction

Social interaction and collaboration, intended as cooperation between subjects in order to fulfill a specific goal, is a relevant aspect in the framework of SAME.

In particular, music making and listening are a clear example of a human activity that is above all interactive and social. However, nowadays mediated music making and listening is usually still a passive, non-interactive, and non-context sensitive experience. The current electronic technologies, with all their potential for interactivity and communication, have not yet been able to support and promote this essential aspect of music making and listening. This can be considered a degradation of traditional listening experience, in which the public can interact in many ways with performers to modify the expressive features of a piece. The need of recovering such active attitude with respect to music is strongly emerging and a major objective of SAME is to develop research and technologies enabling such a recovery.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

Relevance of collaborative and social interaction is further witnessed by the Roadmap of future research in Sound and Music Computing (SMC, www.smcnetwork.org). The roadmap indicates “Address social concerns” among the future research challenges in the SMC field, and “Expand existing SMC methodologies emphasising user-centred and group experience-centred research” as a strategy for addressing such challenge. In particular the roadmap says: “In practice, however, music is most of the time a social activity in which musical engagement is influenced by the behaviour of other participants. Existing empirical and experimental methodologies should be expanded towards understanding aspects of social music cognition. These involve the study of the social context in which musicians and listeners influence each other during musical activities. Also the tools for collaboration, information and communication exchange are now developed in the context of e-science and e-learning and there are no collaborative tools that incorporate all the music specific information, such as audio files, scores, or extracted audio features. Such tools should take into account the profile and experience of users.”

Since long time collaborative models have been used in a lot of application contexts, e.g., in AI and in HCI in the field of conversational agents (see for example, Guinn and Biermann, 1993; Pérez-Quinones and Sibert, 1996). In the literature the term “collaborative” is often used with reference to Collaborative Virtual Environments (CVEs), intended as systems that “use VR technology to visualise a space inhabited by multiple users, usually geographically remote in the real world” (Benford et al., 1997), and provide a framework for enhancing cooperation among users finalized to a given group work (Benford et al., 1996, 1997).

Taking inspiration from Benford’s definition, in the context of SAME “collaborative” means that users cooperate in the common group “work” consisting in generating a music performance (Framework 1 – Collaborative music performance), in exploring and manipulating a music piece (Framework 2 - Learning and context adaptive gestural music exploration and Framework 3 – Mobile orchestra explorer), in playing a sound story or a sound game (Framework 4 – Sound Stories). In other words, music content can evolve and be molded on the basis of joint and coordinated actions by the users who can directly collaborate in generating and transforming the content. The word “collaborative” is therefore used mainly with reference to its social meaning (i.e., bringing together people cooperating in the fulfillment of a goal), rather than in its technological implications (such as requirements of CVEs).

3.4 Context awareness

See Annex of D2.2 (context descriptors feasibility study), Sections 3 and 4.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

3.5 Time modeling and synchronization

Time plays a fundamental role in SAME with respect to three different aspects: (i) time scale for descriptors and relation to prediction, (ii) time modeling of gesture, i.e., the analysis of temporal dynamics of gesture, and (iii) time modeling in the framework of social interaction, e.g., synchronization among members of a group during interaction. These three aspects, that need to be taken into account in the SAME conceptual framework, are briefly summarized below.

3.5.1 Time scale and prediction

Most motion analysis systems are based on the analysis of data frames. A single frame can be considered as a “posture”, which is an “instant” view of the movements (precisely the shortest possible considering the motion capture system). Generally, this small time scale is on the order of few milliseconds to tens of millisecond. However, if this time granularity is generally enough to analyze postures/events, it is not sufficient to grasp a complete view of movements, gestures, and interaction mechanisms. An ensemble of frames over a given time interval, as schematically illustrated in Figure 2, must therefore be considered. Depending on this time window, different motion characteristics can be highlighted, typically from milliseconds to several minutes. Importantly, the time window should also be compared to particular sound and music characteristics (Levitin et al., 2002).

For example, the following time scales can be considered:

- 1-10ms: typical time resolution necessary for percussive gestures, and time accuracy required for rhythmic control of sound. This also corresponds to sound transients.
- 10-500ms: typical time interval when considering fast gestures and their transitions. In music, this typically corresponds to “notes”.
- 500-5000ms: typical time scale when considering a group of gesture. In music, this corresponds to a phrase or group of notes.
- 5s-minutes: time scale for slow movements. In music, this is generally related to music/composition structure

Using probabilistic models, the prediction of the gesture unfolding in the future can be estimated from the analysis of past data. This fact is further developed in the section Techniques (4.2). An important point resides in the fact that there is a relationship between observation time window and the prediction time window.

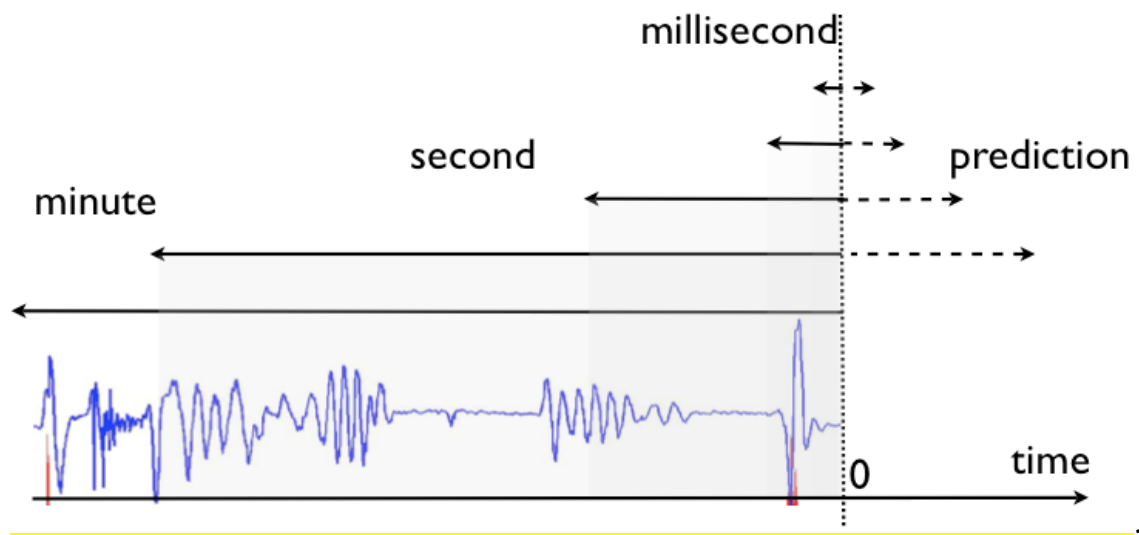


Figure 2. Illustration of various time scales when analyzing gestures. The blue curve represents an example of gesture data (here acceleration from bowing). Depending on the time window of observation, prediction over similar time scale can be estimated using a probabilistic model.

3.5.2 Sound and gesture descriptors

As described previously, music playing, or more generally interacting with sound, is narrowly tied to certain acquired sensorimotor couplings of a physical action with a given musical instrument/interface (whether real or virtual) leading up to sonic results. These couplings can be studied by observing temporal dynamic profiles of both sound and gesture descriptors. These different temporal features actually contribute to a great deal of what we often referred to expressivity (Rasamimanana, 2008a).

These profiles can be related to particular types of articulation, i.e. the manner successive gestures/sounds are merged together. Moreover, the actual temporal relationships between gesture and sounds descriptors reveal also important information about playing mechanisms. Such temporal relationships between sound and gesture descriptors are complex, since gesture and sound cannot be simply considered as two separate phenomena in a causal relationship, but as mutually interacting.

While there are many algorithms in the literature designed for sub-problems encountered in pattern recognition domains, they often lack consideration for temporal dynamics: most methods are geared towards stationary or quasi-stationary data-structures with little or no concern for interactive learning. Specific methods for time modeling do exist, such as Hidden Markov Models, but they are often used considering relatively coarse time resolution. The work at IRCAM on gesture follower and score follower partially fills this gap, by proposing approaches for the modeling of fine temporal behaviors, based on Hidden Markov Models and Semi-Markov Models. Nevertheless, appropriate modeling should also take into account the interactive coupling between gesture and sounds

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

properties. For this, an approach based on concurrent interacting agents is under consideration at IRCAM.

3.5.3 Time modeling and social descriptors

In the framework of social interaction, synchronization takes place at several levels, from synchronization of single gestures and types of gestures (e.g., all members in the group, mimicking each other, perform a guitar-like gesture), up to emotional synchronization, emotional contagion, empathy.

As for synchronization at the level of single gestures and/or types of gesture, the focus is on the measurement of a set of features describing gestures performed by each participant and on finding correlations and similarities for such features between participants. Thus, it is possible to obtain an index of the degree of coherence between participants, i.e., whether they perform or not similar gestures (e.g., according to a given similarity measure). As a further result, it is also possible to isolate rare, unexpected, or incoherent behavior by a single participant in comparison with the average behavior of the group, when a high degree of coherence is detected. If gestures are mapped onto trajectories in abstract feature spaces, coherent behavior will be represented by trajectories (associated to single participants) occupying about the same area in the feature space or moving together along the space. A group showing incoherent or non-homogeneous behavior will instead be represented by trajectories moving chaotically and in an uncorrelated way along the whole space.

Analysis of synchronization can also be carried out at the emotional level in order to investigate subtle phenomena such as emotional entrainment and empathy. This encompasses the analysis of high-level emotional, expressive descriptors for individuating the establishment of synchronization in the emotional state or expression between participants. This phenomenon is particularly emerging in the case of music ensembles, where performers achieve such a level of emotional synchronization that the group can be considered as a single organism (i.e., the ensemble), rather than a group of performers. Preliminary studies in this direction, showing promising results, have been carried out at UGDIST on a duo of violin players and on string quartets (Camurri et al., 2008b).

4. THE SAME MODEL

The model proposed by UGDIST for research and applications in the SAME project grounds its bases on the concept of simultaneous navigation/exploration and manipulation, by

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

multiple users, of multiple maps at different levels of abstraction. The metaphor of navigation/exploration has been chosen because of its familiarity and easiness to understand and handle for the users. Many existing systems for several application domains are based on the concept of navigation/exploration (e.g., navigation in virtual worlds, navigation and retrieval of multimedia content, Google Earth and GPS applications, games, etc.).

Peculiar aspects of the SAME model are:

- *Physicality and embodiment*: this is obtained in two complementary ways: (i) associating the navigation/exploration concept to a real, physical space and to user's expressive movement and gesture, and/or (ii) exploiting interaction with physical objects endowed with suitable sensors. Both navigation and manipulation exploit enaction (see Section 3.1), i.e., the knowledge the user gains by moving and acting in the space.
- *Context-awareness*, i.e., the maps, at all levels, change and adapt according to the context where music is experienced; moreover, the model adapts to what the user is doing (e.g., whether she is at home, walking, running, driving, etc.) and to the input/output devices that are available.
- *Social interaction*, i.e., navigation/exploration/manipulation is performed or better exploited by a group of (collaborative) users. This is an important contribution in a shared social as well as active experience of music.

4.1 Model description

The multilayered model proposed for SAME is an evolution of previous research (Camurri et al., 1994) and of the multilayered model for expressive gesture processing presented in (Camurri et al., 2005; see also Section 2.1). With respect to the original multilayered framework, the model presented here (i) explicitly takes into account the role of context and context-awareness, (ii) explicitly takes into account the social interactions between multiple users, (iii) better defines the content of each layer through a dimensional approach (the maps). Figure 3 provides a sketch of the SAME model.

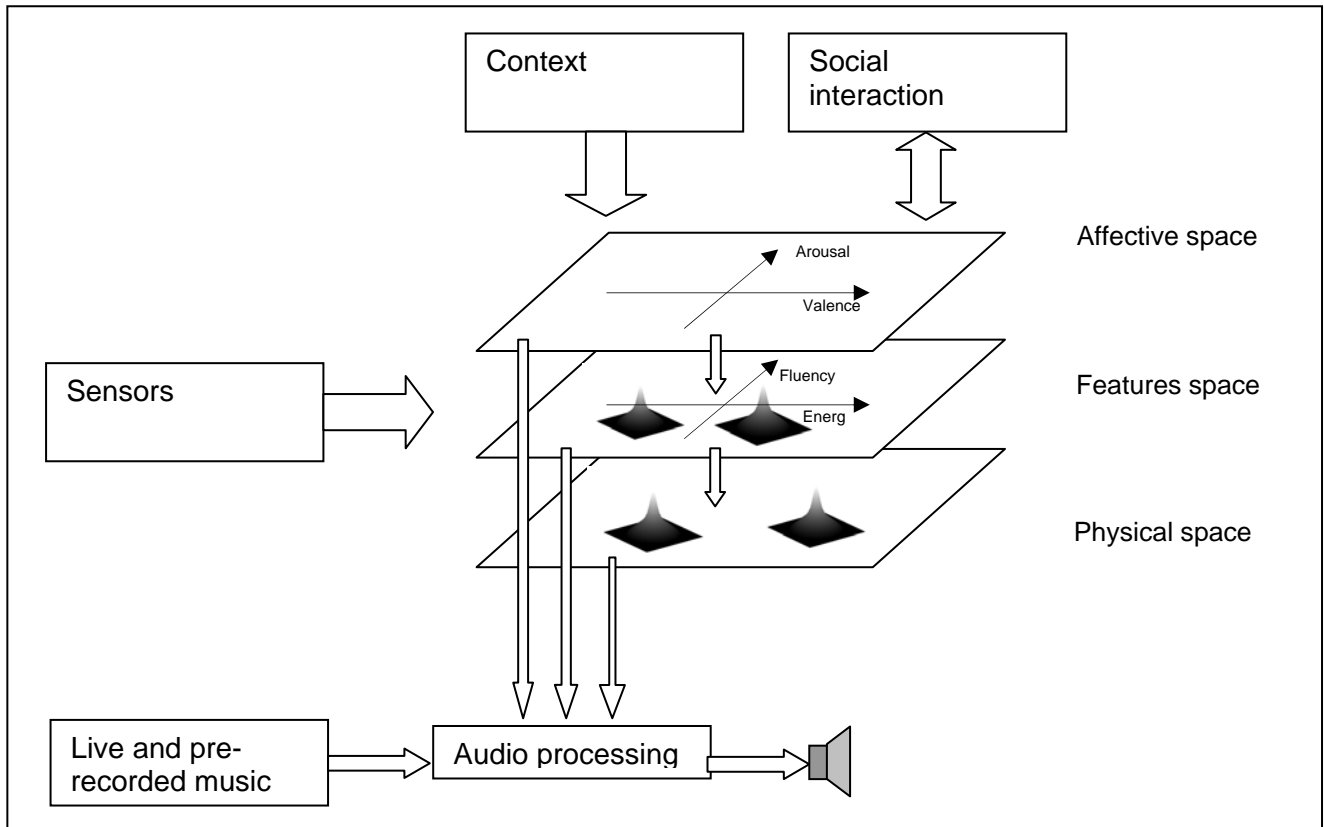


Figure 3. A snapshot of the SAME model, based on the concept of simultaneous navigation/exploration and manipulation, by multiple users, of multiple maps at different levels of abstraction.

4.1.1 Interaction with the physical space

Interaction with the physical space is obtained by superimposing onto the map of the real space a collection of 2D potential functions. Each potential function is associated to a specific audio content, e.g., a single instrument in a polyphonic pre-recorded music piece, a sound in an interactive sound story, a live input from the user, e.g., a virtual instrument. While each listener navigates the physical space, the values of such potential functions in correspondence of her (x, y) position are computed. These values are directly used to control audio processing in several ways, ranging from simple applications such as real-time mixing, where values are weights for the sound level of the corresponding channel, up to complex real-time control of audio effects and processing modules.

This approach can be applied to both when the user moves in a physical space (with her mobile device) and her position is tracked along time and when it is the mobile device which the user moves (e.g., with her hand) in a more limited space (e.g., upon a table). An example of the first context is the “Mobile Orchestra Explorer” (Framework 3) where users explore a physical space populated by a virtual orchestra. An example of the second

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

context is the “Sound Stories” framework (Framework 4), where a possible application is a game board upon which users can move their mobiles.

In early prototypes, potential functions given by the weighted sum of two components, an exponential one and a logarithmic one have been used (see for example Camurri et al., 2007). However, other kinds of potential functions can be used as well. For example, in case the occupation rate of an area in the space has to be calculated (e.g., to check whether the mobile device is in a given location of a game board or to compute how much time has been spent in that location), such a potential function can be used.

The type and the shape of potential function can change according to the context and to input coming from the upper levels. Potential function can also be time-varying. For example a given sound may be activated by a user (or a mobile device) entering in a specific area; then the same sound may gradually fade out (according to a time-varying potential function) when the user leaves that area as long as she does not occupy it again.

Note that these potential functions can also be defined as probability density functions, and thus can be directly related to probabilistic approaches derived from machine learning techniques. For example, the likelihood that a user is following a particular path in the space can be expressed as probability function associated at each (x, y) position. Parameter of such probability functions (e.g. maximum, variance) can then be mapped to parameters of various sound processing.

4.1.2 Interaction with physical objects

In case users interact with physical objects (e.g., mobile devices) different approaches are possible according to the application and the context. For example, if it makes sense to transform manipulation of the object into a navigation in a space (e.g., a repetitive movement of the mobile toward the left causes a displacement in the map toward the left), a similar approach as the one depicted above can be used. Otherwise, the features measured by the on-board sensors (e.g., acceleration along the three axes) will directly contribute to navigation in the feature spaces at the second level, that is, the first level of physical map is skipped and audio processing is directly controlled through navigation in the feature and affective maps. This may be the case of, e.g., Framework 1 (collaborative music performance) where mobile devices are used to control virtual music instruments. Features extracted from the movement of the mobile may be used to compute higher level features (e.g., repetitiveness, frequency of repetitions, energy, impulsiveness, etc.) that are then mapped onto audio processing parameters (e.g., see Framework 2).

Note that interaction with physical objects and navigation in the physical space can take place simultaneously. That is, users can manipulate a mobile device while being tracked during their exploration of the physical space.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

4.1.3 Feature spaces

While users move and manipulate objects, features are extracted describing user's movement and gestures. In the context of the SAME Project, inputs of the feature extraction modules are data from sensors (e.g., accelerometers on mobile devices), even if images from one or more videocameras may also be processed if available (e.g., if interaction takes place in a space that can be monitored with videocameras; such information is part of the contextual information addressed in the project). Examples of features are those described in Deliverables 3.1 and 4.1. Features are mapped on mid-level feature spaces such as for example the Energy-Fluency space described in (Camurri et al., 2005) and used for the ExpressiveHiFi application at the International Broadcasting Conference 2001 (IBC2001). Another example is the 3D space choreographer Rudolf Laban (1947; 1963) introduces in his Theory of Effort.

Thus, movement and gesture can be represented as a trajectory in such a feature space. Again, (N-dimensional) potential functions (or probability functions) can be superimposed on these spaces so that feature extraction results in the navigation/exploration of the feature space through user's movement and gesture. The values of the potential function corresponding to the trajectories followed by the user's gesture can be used in two different ways: (i) to directly control audio processing, i.e., enabling direct manipulation of sound and music content through movement features, and (ii) to dynamically modify the shape of the 2D potential functions at the physical level (if the physical map is used).

Thus, different styles of movement or different gestures can modify the way music is experienced. So, for example, repetitive guitar-like gestures will result in the occupation of an area of the feature space associated to that kind of gesture. A potential function associated to the guitar and located in that area of the feature space will make the guitar sound gradually emerge as long as the user insist with such gestures. Here also, the potential function can be related to probability function from recognizing particular gestures. As another example, repetitions of high-energy gestures may result in making specific kinds of instruments emerging and other hiding: highly repetitive, strong, "percussive" gestures may make percussions emerging over the other sections (e.g., either directly or by increasing the amplitude of the 2D potential functions associated with percussions at the physical level). A change to a gentler, calmer, "string-like" movement may make the string section emerge, while the amplitude of the potential functions associated to the percussions gradually decreases as they fade out. At this level, sonic interaction design issues (Rocchesso and Serafin, 2008) are one of the most important components in order to create effective active listening environments.

4.1.4 Affective/emotional spaces

At the higher level, the user can intervene on the expressive features of the music performance. This is done through the navigation of an emotional, affective space. Starting from the extracted features, a further analysis is carried out for classifying the expressive

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

intention the user conveys with her expressive movement and gesture. Expressive intentions are translated in a position (or a trajectory) in an affective, emotional space. Such space can also be divided in several areas, each one corresponding for example to a different performance of the same instrument with a different expressive intention. Several examples of such affective spaces are available in the literature, for example the spaces used in dimensional theories of emotion, e.g., (Russell, 1980; Tellegen et al., 1999), or those especially developed for analysis and synthesis of expressive music performance, e.g., (Canazza et al., 2000; Juslin, 2000; Vines et al., 2005a). Classification of expressive intentions from movement and gesture features can be performed starting from the results obtained in psychological research (e.g., Boone and Cunningham, 1998; De Meijer, 1989; Wallbott, 1998).

The higher level has a twofold influence on the lower ones:

- (i) it can directly intervene on the audio processing chain, e.g., by triggering different intentions either by suitably sequencing audio fragments from pre-recorded performances of the same piece with different expressive intentions, or by applying techniques such as the KTH real-time rule-based system (Friberg, 2006);
- (ii) it can operate on the parameters of the potential functions at the feature and physical levels so that their profiles are dynamically adapted to the expressive behavior of the user.

In an early prototype (Camurri et al., 2008a), four different expressive intentions have been classified and the results used to control expressive music performance with dance.

In music recommendation system such as those described in Framework 5 (Music for my life) music pieces can be placed in the feature and affective spaces according to their content (both at the level of features and at the level of the emotional content the pieces convey). As the user navigates such spaces through her movement and her expressive intentions, the music pieces that are located in the area the user occupies may be provided to her as suggested music to listening to.

4.1.5 Context-awareness

Context information is considered at all levels. This is information related for example to where active music listening takes place (e.g., at home, in the car, in a pub, in a discotheque) and to what the user is doing while listening to music (e.g., walking, running, driving, seated on the sofa). Context information can intervene on all the parameters of the model: the classification technique and its parameters at the higher-level, the parameters of the feature extraction techniques, the parameters of the potential or probability functions.

For example, classification of gestures or of specific expressive intentions has deep cultural and contextual roots. At level two (features maps) and three (affective maps) context information can be used to dynamically adapt the model to such changing contexts. In

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

other words, the multidimensional maps where gestures are represented according to their type of the expressive intention they convey can be dynamically changed and/or adapted (e.g., by adapting the potential function located on them) according to contextual information. The classification techniques have to adapt accordingly (e.g., in case a supervised machine learning technique is used the training set against which classification is performed can change according to contextual information).

4.1.6 Social interaction

The model, as sketched up to now, refers to a single user. However, a fundamental aspect of the work in SAME is to enable active experience of music by multiple users and groups. Thus, the model encompasses a further level where synchronization among the elements in a group is evaluated. Synchronization is considered at the emotional level, that is it concerns the degree of common feeling, engagement, empathy which is reached within the group. Preliminary studies in this direction have been carried out leading to promising results (Camurri et al., 2008b). Research on emotional synchronization makes it possible to (i) individuate a common behavior/feeling in the group and (ii) evaluating synchronization and coherence among its elements (see also Section 3.5.3). When high emotional synchronization and coherence is detected, the group becomes like a single person experiencing the same feelings, and thus it can be represented as a single individual. Similarly, several groups can be represented as several single individuals. This allows to replicate the model, with suitable adaptations, also in case of active experience by single and multiple groups.

4.2 Techniques

Several different techniques are applied at the different levels of the conceptual model, in order to compute suitable gesture, motion, social descriptors and to map them onto sound and music parameters. Such techniques can be grouped in two major categories, depending on how they are calculated: descriptors can result from a direct, explicit calculation, or they can result from an indirect calculation, using machine-learning techniques.

Direct calculation is mainly used for low and mid-level descriptors. Thus, descriptors obtained with direct calculation are usually found at level one (physical level) and level two (features level). Input to direct calculation techniques are the data coming from mobile on-board sensors (e.g., data from accelerometers, audio from the on-board microphone, images from the on-board camera) and data coming from possible other sensors in the environment, if available (e.g., images from video-cameras placed in the environment). Direct descriptors are usually computed in a single step either on each input sample (e.g., on each input image), or on a frame of input samples, i.e., in a time window (e.g., a frame

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

of N accelerometer samples). Typical direct descriptors are statistical descriptors of a given input signal (e.g., average and peak velocity) or other simple computations (e.g., the volume of an audio stream, the area of a blob extracted from an image, etc.). Expressive descriptors (see for example Camurri et al., 2004, 2005) can also results from a (quite) direct calculation, e.g., the amount of detected motion (motion index) computed on a small number of video frames, the contraction index of a blob computed on a single frame, a simple impulsiveness index computed on a frame of accelerometer samples.

Higher-level descriptors (located at levels two and three in the proposed model) are obtained using machine learning techniques. Contrary to direct descriptors, which are computed from low-level data in a single step process, these descriptors are computed in two steps, after a learning phase. Precisely, they can typically be seen as “distance” or “divergence” from a performed movement to one or multiple reference movement. This approach inherently draws on probabilistic modeling. For example, Hidden Markov Models (Rabiner, 1990) have been used to model movement data (Bevilacqua, 2007). Several descriptors can hence be deduced from the likelihoods output from these models. Examples of such descriptors are the likelihood index, giving the probabilistic similarity between the performed gesture and the reference (it therefore reflects the probability that the currently performed movement is one reference movement); temporal deviation, indicating how temporally different the performed movement is with respect to the reference movement; shape distances, used to compute the statistical difference/similarity between a performed and a reference movement (Euclidian distance or Kullback Leibler divergence and other relevant measures as proposed by Basseville, 1989).

More generally, approaches building statistics on movements can be used to model part or whole of movements. Decomposition approaches and functional data analyses can provide explicit temporal and shape descriptions of movement data. An approach consists of decomposing movement data on one or several bases where each basis component is linked to an elementary significant movement feature, determined statistically (Lee and Seung, 1999; Hoyer, 2004) or empirically. This approach permits to sum up a movement as a contribution of elementary components, and besides it also addresses segmentation issues for movement data by factorizing a movement as a temporal concatenation of basic elements. At the same time, techniques can be used that explicitly model temporal shapes. Different approaches can be considered, either based on a physical modeling explaining movements (Nelson, 1983; Hogan and Sternad, 2007) or based on a functional and/or differential calculus regression (Ramsay and Silverman, 2002; Vines et al., 2005b). One main benefit of such techniques is that they provide a physics-based framework for interpreting movement data.

Further machine learning techniques can be used for classification of gestures e.g., according to expressive intentions. These include “classical” neural networks, self-organizing maps, support vector machines, decision trees, etc. Clustering algorithms can also be employed e.g., for grouping feature vectors according to a specific gestural quality.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

4.3 Types of interaction

The SAME model takes into account different types of interaction with the sound and music content that are implied in the SAME Frameworks and Use Cases.

Various categories can be considered:

- “Interactions based on mimetic skills, or rehearsed action scenarios, such as playing a musical instrument” (Leman, 2007). This type of interaction encompasses conscious control of the user, as playing an instrument. It is also one of the simplest forms of embodied listening (the user mimics the instrument she wants to listen to). In the proposed conceptual model this type of interaction can be obtained at level two (features maps). At that level, features extracted from user’s gestures are classified in terms of the type of gesture the user is performing (e.g., a percussive movement, a movement mimicking the bow of a string instrument, a guitar-like movement, etc.). Such categories are placed in a map at level two. The current location of the user in such a map is used for making the corresponding instruments (or sections) emerging and other hiding, or to trigger sounds associated to the kind of instrument being mimicked. In case interaction takes place in a physical space, the type of gesture can directly influence the shape of the potential functions at level one.
- “Interactions based on goal-directed gestures that do not require highly developed skills but nevertheless may be highly culture-dependent such as symbolic gestures” (Leman, 2007). This type of interaction encompasses context dependant, person dependant, culture dependant gestures. Semi/conscious control may also appear in this type of interaction (e.g., gestures that have become automatic). In the proposed conceptual model this type of interaction is dealt with at level two. Again, gesture classification techniques are used for distinguishing between different types of gestures (e.g., different symbolic gestures) and for placing such different types on a map. The current position of the user in such a map controls musical output. In this type of interaction, however, context information plays a fundamental role because classification may be context-dependent. Thus context information is used for switching between different context-dependent classifications and different maps. Thus, several instances of level two exist and the instance currently in use depends on the context.
- “Interactions based on direct episodic action sequence, involving responses based on our emotive, affective, and expressive capabilities” (Leman, 2007). This type of interaction encompasses expressive gestures and it is often characterized by unconscious control. In the conceptual framework this type of interaction is addresses at level three (affective maps), where expressive gestures are analyzed, e.g., in terms of trajectories in a space associated to a dimensional theory of emotion. Emotional responses are treated as processes rather than as static states. The current position in such affective, emotional space influences musical output

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

either directly (e.g., intervening on expressive features of music performance) or through modifications (e.g., change of shape of potential functions) at the features and physical levels.

Another fundamental aspect in SAME is social/collaborative interaction. If from the one hand the types of interaction introduced above usually refer to how a single user interact with the sound and music content, on the other hand they can be generalized to the (social) interaction among members of a group of users and to the interaction of the whole group with sound and music content.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

5. REFERENCES

- Basseville M., 1989. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18:349–369.
- Benford, S., Brown, C., Reynard, G., Greenhalgh, C., 1996. Shared Spaces: Transportation, Artificiality, and Spatiality, in *Proc. CSCW'96*, 77-85, Boston, Massachusetts, ACM Press, 1996.
- Benford, S., Snowdon, D., Colebourne, A., O'Brien, J., Rodden, T., 1997. Informing the design of collaborative virtual environments, in S. C. Hayne and W. Prinz (eds.): *GROUP'97*, in *Proc. of the ACM SIGGROUP Conference on Supporting Group Work*, 71-80, Phoenix, Arizona.
- Bevilacqua, F., Guedy, F., Fléty, E., Leroy, N., Schnell N., 2007. Wireless sensor interface and gesture-follower for music pedagogy, in *Proc. 2007 Intl. Conference on New Interfaces for Musical Expression (NIME07)*, New York, USA.
- Bevilacqua, F., Rasamimanana, N., Fléty, E., Lemouton, S., Baschet, F., 2006. The augmented violin project: research, composition and performance report, in *Proc. 2006 Intl. Conference on New Interfaces for Musical Expression (NIME06)*, Paris, France.
- Boone, R. T., Cunningham, J. G., 1998. Children's decoding of emotion in expressive body movement: The development of cue attunement, *Developmental Psychology*, 34, 1007-1016.
- Bresin, R., & Friberg, A. (2000). Emotional coloring of computer controlled music performance. *Computer Music J*, 24(4), 44-63.
- Bresin, R., Friberg, A., & Sundberg, J. (2002). Director musices: The KTH performance rules system. In *Proceedings of SIGMUS-46* (pp. 43-48). Kyoto.
- Bresin, R., Hansen, K. F., & Dahl, S. (2003). The Radio Baton as configurable musical instrument and controller. In Bresin, R. (Ed.), *Proceedings of SMAC 2003, Stockholm Music Acoustics Conference* (pp. 689-691).
- Camurri, A., Frixione, M., Innocenti, C., 1994. A Cognitive Model and a Knowledge Representation System for Music and Multimedia, *Journal of New Music Research*, 23, 317-347.
- Camurri, A., Mazarino, B., Ricchetti, M., Timmers, R., Volpe, G., 2004. Multimodal analysis of expressive gesture in music and dance performances, in A. Camurri, G. Volpe (Eds.), *Gesture-based Communication in Human-Computer Interaction*, LNAI 2915, Springer Verlag.
- Camurri, A., De Poli, G., Leman, M., Volpe, G., 2005. Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications. *IEEE Multimedia*. 12(1), 43-53.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

Camurri, A., Canepa, C., Volpe, G., 2007. Active listening to a virtual orchestra through an expressive gestural interface: The Orchestra Explorer, in: Proceedings of the 7th Intl. Conference on New Interfaces for Musical Expression (NIME2007), New York, USA.

Camurri, A., Canepa, C., Coletta, P., Mazzarino, B., Volpe, G., 2008a. Mapped Affetti Erranti: a Multimodal System for Social Active Listening and Expressive Performance, in: Proceedings of the 8th Intl. Conference on New Interfaces for Musical Expression (NIME08), Genova, Italy, 134-139.

Camurri, A., Varni, G., Volpe G., 2008b, Emotional Entrainment in Music Performance, in: Proceedings 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG2008), Amsterdam, The Netherlands.

Canazza, S., De Poli, G., Drioli, C., Rodà, A., Vidolin, A., 2000. Audio Morphing Different Expressive Intentions for Multimedia Systems. IEEE Multimedia, 7(3), 79 – 83.

Cont, A., 2008. Antescofo: Anticipatory synchronization and control of interactive parameters in computer music, in Proc. of the International Computer Music Conference (ICMC2008), Belfast, UK.

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

De Meijer, M., 1989. The contribution of general features of body movement to the attribution of emotions, Journal of Nonverbal Behavior, 13, 247-268.

Friberg, A. (2005a). A fuzzy analyzer of emotional expression in music performance and body motion. In Brunson, W., & Sundberg, J. (Eds.), Proceedings of Music and Music Science, Stockholm 2004.

Friberg, A. (2005b). Home conducting: Control the overall musical expression with gestures. In Proceedings of the 2005 International Computer Music Conference (pp. 479-482). San Francisco: International Computer Music Association.

Friberg, A., 2006. pDM: an expressive sequencer with real-time control of the KTH music performance rules. Computer Music Journal, 30(1), 37-48.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. Advances in Cognitive Psychology, Special Issue on Music Performance, 2(2-3), 145-161.

Guinn, I.C., Biermann, A., 1993. Conflict Resolution in Collaborative Discourse, in Computational Models of Conflict Management in Cooperative Problem Solving, in Proceedings 13th International Joint Conference on Artificial Intelligence (IJCAI), Chambéry, France.

Hashimoto, S., 1997. KANSEI as the Third Target of Information Processing and Related Topics in Japan, in Camurri A. (Ed.), Proceedings of the International Workshop on KANSEI:

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

The technology of emotion, AIMI (Italian Computer Music Association) and DIST-University of Genova,101-104.

Hogan, N., Sternad, D., 2007. On rhythmic and discrete movements: reflections, definitions and implications for motor control. *Experimental Brain Research*, 181:13–30.

Hoyer P., 2004. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469.

Juslin, P. N., 2000. Cue utilization in communication of emotion in music performance: relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(6), 1797-1813.

Juslin, P. N., Friberg, A., & Bresin, R. (2002). Toward a computational model of expression in performance: The GERM model. *Musicae Scientiae*, Special issue 2001-2002, 63-122.

Kurtenbach, G., Hulteen, E., 1990. Gestures in Human Computer Communication, in Brenda Laurel (Ed.), *The Art and Science of Interface Design*, Addison-Wesley, 309-317.

Laban, R., Lawrence, F.C., 1947. *Effort*. Macdonald & Evans Ltd., London.

Laban, R., 1963. *Modern Educational Dance*. Macdonald & Evans Ltd., London.

Lee D., Seung H., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 410:788–791.

Leman, M., 2007. *Embodied music cognition and mediation technology*. The MIT Press.

Levitin, D., McAdams, S., Adams, R., 2002. Control parameters for musical instruments: a foundation for new mappings of gesture to sound. *Organised Sound*, 7(2):171–189.

Lewkowicz, D.J., Kraebel, K.S., 2004. The value of Multisensory Redundancy in the Development of Intersensory Perception, in G. Calvert, C. Spence, B.E. Stein (Eds), *The handbook of multisensory processes*, 665-678, Cambridge, MA London: A Bradford Book, The MIT Press.

Lickliter, R., Bahrick, L.E., 2004. Perceptual Development and the origins of multisensory responsiveness, in G. Calvert, C. Spence, B.E. Stein (Eds), *The handbook of multisensory processes*, 665-678, Cambridge, MA London: A Bradford Book, The MIT Press.

London, J., 2004. *Hearing in Time: Psychological Aspects of Musical Meter*. Oxford University Press.

Nelson, W.L., 1983. Physical principles for economies of skilled movements. *Journal Biological Cybernetics*, 46(2):135–147.

Noe, A., 2004. *Action in Perception*. MIT Press , Cambridge.

Sound And Music for Everyone Everyday Everywhere
Every way CONFIDENTIAL

A. Camurri, F. Bevilacqua, and G. Volpe (Eds.)

29.09.2008

- Pérez-Quinones, M., Sibert, J. L., 1996, A Collaborative Model of Feedback in Human-Computer Interaction, in Proc. Conference on Human Factors in Computing Systems (CHI'96).
- Rabiner, L. R., 1989. A tutorial on hidden markov models. In Proceedings of the IEEE, pages 257-286..
- Ramsay, J., Silverman, B., 2002. Applied functional data analysis: Methods and case studies. New York: Springer-Verlag.
- Rasamimanana, N., Fléty, E., Bevilacqua F., 2006. Gesture analysis of violin bow strokes, in Gesture in Human-Computer Interaction and Simulation, volume 3881 of Lecture Notes in Computer Science (LNCS), Springer Verlag, 145–155.
- Rasamimanana, N., 2008a, Geste instrumental du violoniste en situation de jeu: analyse et modélisation, PhD thesis, Université Paris 6 - IRCAM UMR STMS.
- Rasamimanana, N., Bernardin, D., Wanderley, M., Bevilacqua, F., 2008b, String bowing gestures at varying bow stroke frequencies: A case study, in Gesture in Human-Computer Interaction and Simulation, Lecture Notes in Computer Science. Springer Verlag (accepted).
- Rasamimanana, N., Bevilacqua, F., 2008c. Effort-based analysis of bowing movements: evidence of anticipation effects, Journal of New Music Research (accepted for publication).
- Rocchesso, D., Serafin S., 2008. Sonic interaction design: sound, information and experience, in: Proceedings CHI 2008 conference, Florence, Italy.
- Russell, J.A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161-1178.
- Stern, 2000. The interpersonal world of the infant, Basic books, New York.
- Tellegen, A., Watson, D., Clark, L. A., 1999. On the dimensional and hierarchical structure of affect. Psychological Science, 10(4), 297-303.
- Vines, B. W., Krumhansl, C.L., Wanderley, M.M., Ioana, M. D., Levitin, D.J., 2005a. Dimensions of Emotion in Expressive Musical Performance. Ann. N.Y. Acad. Sci., 1060, 462-466.
- Vines, B., Nuzzo, R., Levitin, D. 2005b. Analyzing temporal dynamics in music: differential calculus, physics, and functional data analysis techniques. Music Perception, 23(2):137–151.
- Wallbott, H. G., 1998. Bodily expression of emotion, European Journal of Social Psychology, 28, 879-896.
- Wohlschlager, Gattis, Bekkering, 2003. Action generation and action perception in imitation: an instance of the ideomotor principle. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 358, 501-515.