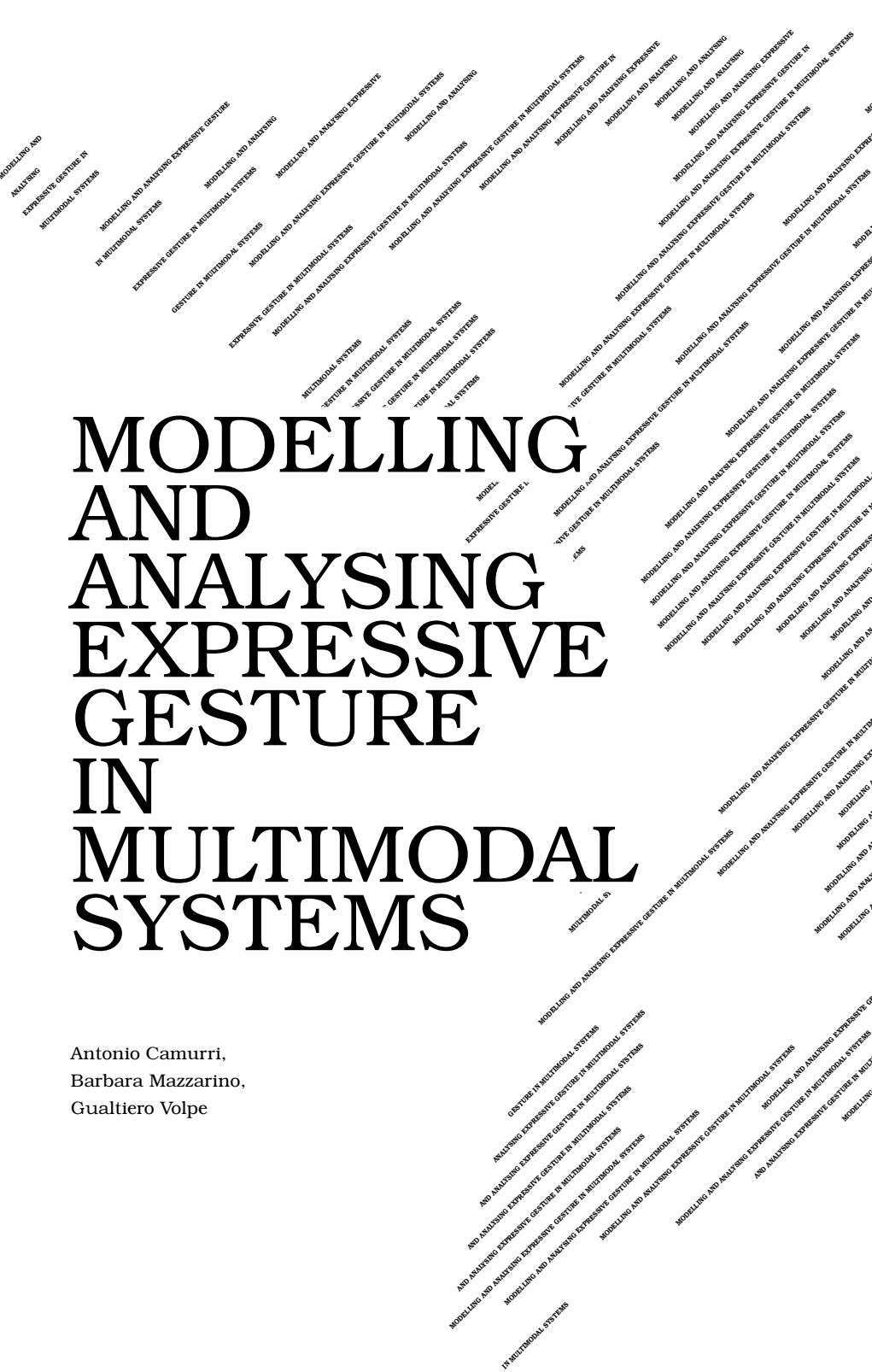

MODELLING AND ANALYSING EXPRESSIVE GESTURE IN MULTIMODAL SYSTEMS

Antonio Camurri,
Barbara Mazzarino,
Gualtiero Volpe

1. Introduction

This chapter presents research on the modelling of expressive gesture in multimodal interaction and on the development of multimodal interactive systems explicitly taking into account the role of expressive gesture in the communication process. In this perspective, a particular focus is on dance and music performances as first-class conveyors of expressive and emotional content.

Expressive gesture is a key concept in our research.[1] This paper tries to deal with it, and introduces two experiments aiming at understanding the non-verbal mechanisms of expressive/emotional communication.

Several definitions of gesture exist in the literature. The most common use of the term is with respect to natural gesture, which is defined as a support to verbal communication. For Cassel and colleagues (1990) "[a] natural gesture means the types of gestures spontaneously generated by a person telling a story, speaking in public, or holding a conversation." McNeill (1992) in his well-known taxonomy divides the natural gestures generated during a discourse into four different categories: iconic, metaphoric, deictic, and beats. In a wider perspective, Kurtenbach and Hulteen (1990) define gesture as "a movement of the body that contains information."[2]

In artistic contexts, and in particular in the field of performing arts, gestures are often not intended to denote things or to support speech as in the traditional framework of natural gesture, but the information they contain and convey is related to the affective/emotional domain. From this point of view, gestures can be considered "expressive" depending on the kind of information they convey: expressive gestures carry what Cowie et al. (2001) call "implicit messages"[3], and what Hashimoto (1997) calls KANSEI. That is, they are responsible of the communication of a kind of information (what we call *expressive content*) which is different and in most cases independent from, even if often superimposed on, a possible denotative meaning, and which concerns aspects related to feelings, moods, affect, and emotional intentions.

For example, the same action can be performed in several ways, by stressing different qualities of movement: it is possible to recognise a person from the way she or he walks, but it is also possible to obtain information about the emotional state of a person by looking at her or his gait, e.g., if she or he is angry, sad, happy. In the case of gait analysis, we can therefore distinguish among several objectives and layers of analysis: a first one aiming at describ-

1 Camurri et al. 2005

2 A survey and a discussion of existing definition of gesture can be found in Cadoz and Wanderley 2000.

3 Cowie et al. 2001

ing the physical features of the movement, for example in order to classify it;[4] a second one aiming at extracting the expressive content gait conveys, e.g., in terms of information about the emotional state that the walker communicates through her or his way of walking. From this point of view, walking can be considered as an expressive gesture: even if no denotative meaning is associated with it, it still communicates information about the emotional state of the walker, i.e., it conveys a specific expressive content. In fact, in this perspective the walking action fully satisfies the conditions stated in the definition of gesture by Kurtenbach and Hulteen (1990): walking is "a movement of the body that contains information." Some studies can be found aiming at analysing the expressive intentions conveyed through everyday actions: for example, Pollick (2001) investigated the expressive content of actions like knocking or drinking.

If on the one hand expressive gestures partially include natural gestures, that is, natural gestures can also be expressive gestures, we face on the other hand a more general concept of expressive gesture that includes not only natural gestures but also musical, human movement, and visual (e.g. computer-animated) gestures. Our concept of expressive gesture is therefore somewhat broader than the concept of gesture as defined by Kurtenbach and Hulteen, since it also considers cases in which, with the aid of technology, communication of expressive content takes place even without an explicit movement of the body, or, at least, the movement of the body is only indirectly involved in the communication process. This can happen, for example, when using visual media. The expressive content is conveyed through a continuum of possible ways ranging from realistic to abstract images and effects: cinematography, cartoons, virtual environments with computer-animated characters and avatars, and expressive control of lights in the context of a theatre (e.g. related to actor's physical gestures). Consider, for example, a theatre performance: the director, choreographer, composer can ask actors, dancers, musicians, to communicate content through a number of expressive gestures (e.g., dance and/or music phrases). At the same time, technology allows the director to extend the language available to him. He can map motion or music features onto particular configurations of lights, in movements of virtual characters, in automatically generated computer music and live electronics. In this way, he can create an "extended" expressive gesture that, while still having the purpose of communicating an expressive content, is only partially related to explicit body movements: in a way, such "extended expressive gesture" is the result of a juxtaposition of several dance, music, and visual gestures, but it is not just the sum of them, since it also includes the artistic point of view

4 Quite a lot of research work can be found in the computer vision literature about gait analysis, see for example Liu et al. 2002.

of the director who created it, and it is perceived as multimodal stimuli by human spectators.

Our research on expressive gesture aims at the development of interactive multimedia systems enabling novel interaction paradigms and allowing a deeper engagement of the user by explicitly observing and processing his/her expressive gestures. Since artistic performances use non-verbal communication mechanisms to convey expressive content elaborately, we focused on performing arts, and in particular on dance and music performances, as a test-bed where computational models of expressive gesture and algorithms for expressive gesture processing can be developed, studied, and tested.

In particular, our attention has been focused on two aspects:

- Expressive gesture as a way to convey a particular emotion to the audience;
- Expressive gesture as a way to emotionally engage the audience.

Each of these has recently been the subject of experiments at our lab aiming at understanding which features in an expressive gesture are responsible for the communication of the expressive content, and how the dynamics of these features correlates with a specific expressive content.

In this paper, we concretely illustrate our approach by presenting two experiments focused on these two aspects.

The first one aims at (i) individuating which motion cues are mostly involved in conveying the dancer's expressive intentions (in term of basic emotions) to the audience during a dance performance and (ii) testing the models and algorithms developed by comparing their performances with spectators' ratings of the same dance fragments.

The second one investigates the mechanisms responsible for the audience's engagement in a musical performance. The aim of this experiment is again twofold: (i) individuating which auditory and visual cues are most involved in conveying the performer's expressive intentions and (ii) testing the developed model by comparing their performances with spectators' ratings of the same musical performances.

For the analysis of expressive gesture in these experiments a unifying conceptual framework was adopted.

2. The Layered Conceptual Framework

The experiments presented in this paper address expressive gesture in music and dance performance.

While gesture in dance performance mainly concerns the visual/physical modality (even if the auditory components can be relevant as well), gesture in music performance uses both the auditory and the visual channels to convey expressive information, and, thus, it is multimodal in its essence. Gestures in music performance are not only the expressive and functional gestures that a performer physically makes, but also include expressive gestures present in the sound produced. When we define gestures as structural units that have internal consistency and are distinguished in time and quality from neighbouring units, it is possible to analyse gestures in both modalities. Multimodality is therefore a key issue. In order to deal with multimodal input a unifying conceptual framework has been adopted.[5] It is based on a layered approach ranging from low-level physical measures (e.g., position, speed, acceleration of body parts for dance gestures, sampled audio signals or MIDI messages for music gesture) toward descriptors of overall gesture features (e.g., motion fluency, directness, impulsiveness for dance gestures, analysis of melodic and harmonic qualities of a music phrase for music gestures).

This layered approach is sketched in Figure 1. Each layer is depicted with its inputs, its outputs, and the kind of processing it is responsible for. In the following sections, each layer will be discussed with respect to its role in the two experiments.

Our conceptual framework, here presented for analysis, can also be applied for synthesis of expressive gesture: for example for the generation and control of the movement of avatars, virtual characters, or robots in Mixed Reality scenarios, as well as for the synthesis and interpretation of music. Examples of synthesis of expressive movement and expressive audio content are well documented in literature.[6]

Finally, it should be noticed that in the perspective of developing novel interactive multimedia systems for artistic applications, such a framework should be considered in the context of a broader Mixed Reality scenario in which virtual subjects (e.g., virtual characters) who behave both as observers and as agents perform the four layers of processing in the analysis of observed expressive gestures and in the synthesis of expressive gestures to communicate (directly or remotely) with other real and virtual subjects.

5 Camurri et al. 2005

6 See e.g. the EMOTE system (Chi et al. 2000) for generation of movement of avatars and virtual characters based on high level motion qualities, and the systems for synthesis of expressive music performances developed at KTH (Friberg et al. 2000) and by the DEI-CSC group at the University of Padova (Canazza et al. 2000).

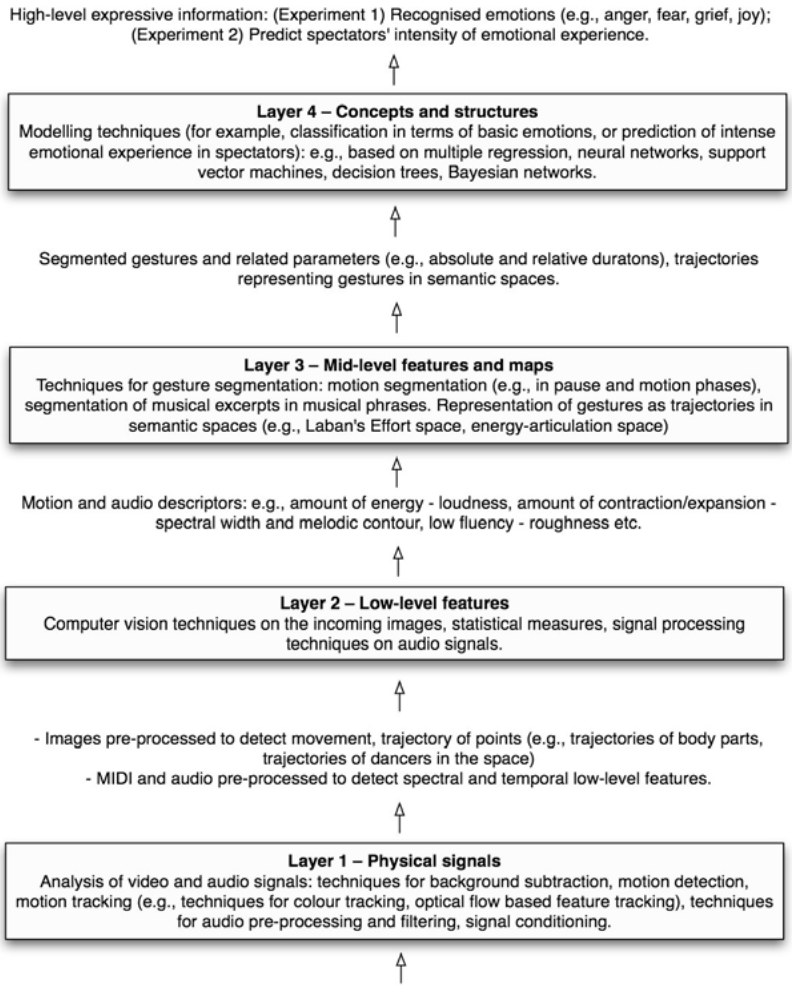

Fig. 1. The layered conceptual framework and its instantiation in the two experiments

3. Modeling expressive gesture in dancers

As an example of analysis of expressive gesture in dance performance, we discuss an experiment carried out in collaboration with the Department of Psychology of the University of Uppsala (Sweden) in the EU-IST MEGA project.

The aim of the experiment was twofold: (i) individuating which motion cues are mostly involved in conveying the dancer's expressive intentions to the audience during a dance performance and (ii) testing the developed models and algorithms by comparing their performances with spectators' ratings of the same dance fragments.

In the case of this experiment, expressive gesture was analysed with respect to its ability to convey emotions to the audience. The study focused on the communication through dance gesture and recognition by spectators of the four basic emotions: anger, fear, grief, and joy.

The research hypotheses are grounded in the role of the Laban's dimensions in dance gesture, as described in Laban's Theory of Effort:[7]

- The time dimension in terms of overall duration of time and tempo changes also elaborated as the underlying structure of rhythm and flow of the movement;
- The space dimension in its aspects related to Laban's "personal space" e.g., to what extent limbs are contracted or expanded in relation to the body centre;
- The flow dimension in terms of analysis of shapes of speed and energy curves, and frequency/rhythm of motion and pause phases;
- The weight dimension in terms of amount of tension and dynamics in movement, e.g., vertical component of acceleration.

These cues were predicted to be associated in different combinations to each emotion category.[8]

3.1 The experiment

An experienced choreographer was asked to design a choreography such that it excluded any propositional gesture or posture and it avoided stereotyped emotions.

7 Laban 1963; Laban/Lawrence 1947

8 Details can be found in Camurri/Lagerlöf/Volpe 2003.

In Uppsala, five dancers performed this same dance with four different emotional expressions: anger, fear, grief and joy. Each dancer performed all four emotions. The dance performances were video-recorded by two digital videocameras (DV recording format) standing fixed in the same frontal view of the dance (a spectator view). One camera obtained recordings to be used as stimuli for spectators' ratings. The second video camera was placed in the same position but with specific recording conditions and hardware settings to simplify and optimise automated recognition of movement cues (e.g., high speed shutter). Dancers' clothes were similar (dark), contrasting with the white background, in an empty performance space without any scenery. Digitised fading eliminated facial information and the dancers appeared as dark and distant figures against a white background.

The psychologists in Uppsala then proceeded in collecting spectators' ratings: the dances were judged with regard to perceived emotion by 32 observers, divided in two groups. In one group ratings were collected by 'forced choice' (choose one emotion category and rate its intensity) for each performance, while the other group was instructed to use a multiple choice schema, i.e., to rate the intensity of each emotion on all four emotion scales for each performance.

At the same time, at the InfoMus Lab we proceeded in extracting motion cues from the video recordings and in developing models for automatic classification of dance gestures in terms of the basic emotion conveyed.

3.2 Automated Extraction of Motion Cues

Extraction of motion cues followed the conceptual framework described in Section 2.

3.2.1 Layer 1

In the case of analysis of dance performance from video, layer 1 is responsible for the processing of the incoming video frames in order to detect and obtain information about the motion that is actually occurring. It receives as input images from one or more videocameras and, if available, information from other sensors (e.g., accelerometers). Two types of output are generated: processed images and trajectories of body parts. Layer 1 accomplishes its task by means of consolidated computer vision techniques usually employed for real-time analysis and recognition of human motion and activity.[9] It

9 See for example the temporal templates technique for representation and recognition of human movement described in Bobick/Davis 2001.

should be noted that in contrast to the research of Bobick and J. Davis, we do not aim at detecting or recognising a specific kind of motion or activity. The techniques we use include feature-tracking based on the Lucas-Kanade algorithm[10], skin colour tracking to extract positions and trajectories of hands and head, an algorithm to divide a body silhouette into sub-regions, and Silhouette Motion Images (SMIs). A SMI is an image carrying information about variations of the silhouette shape and position in the last few frames. SMIs are inspired by motion-energy images (MEI) and motion-history images (MHI).[11] They differ from MEIs in the fact that the silhouette in the last (more recent) frame is removed from the output image: in such a way only motion is considered, while the current posture is skipped. Thus, SMIs can be considered as carrying information about the "amount of motion" which has occurred in the last frames. Information about time is implicit in SMI and is not explicitly recorded. We also use an extension of SMIs, which takes into account the internal motion in silhouettes. In such a way we are able to distinguish between global movements of the whole body in the General Space and internal movements of body limbs inside the Kinesphere.

3.2.2 Layer 2

Layer 2 is responsible for the extraction of a set of motion cues from the data coming from low-level motion tracking. Its inputs are the processed images and the trajectories of points (motion trajectories) coming from Layer 1. Its output is a collection of motion cues describing movement and its qualities. According to the research hypotheses described above, the cues extracted for this experiment include:

- Cues related to the amount of movement (energy) and in particular what we call Quantity of Motion (QoM). QoM is computed as the area (i.e., number of pixels) of an SMI.[12] It can be considered as an overall measure of the amount of detected motion, involving velocity and force;
- Cues related to body contraction/expansion, and in particular the Contraction Index (CI). CI is a measure, ranging from 0 to 1, of how the dancer's body uses the space surrounding it. The algorithm to compute the CI[13] combines two different techniques: the individuation of an ellipse approximating the body silhouette and computations based

10 Lucas/Kanade 1981
11 Bradsky/Davis 2002; Bobick/Davis 2001
12 Camurri/Lagerlöf/Volpe 2003
13 Camurri/Lagerlöf/Volpe 2003

on the bounding region. The former is based on an analogy between the image moments and mechanical moments:[14] the eccentricity of the approximating ellipse is related to body contraction/expansion. The latter compares the area covered by the minimum rectangle surrounding the dancer with the area currently covered by the silhouette;
- Cues derived from psychological studies[15] such as amount of upward movement, dynamics of the Contraction Index (i.e., how much CI was over a given threshold along a time unit);
- Cues related to the use of space: length and overall direction of motion trajectories;
- Kinematical cues (e.g., velocity and acceleration) calculated on motion trajectories.

For those cues depending on motion trajectories a Lucas-Kanade feature tracker was employed in Layer 1. A redundant set of 40 points randomly distributed on the whole body was tracked. Points were reassigned each time dancers stopped their motion (i.e., a pause was detected) so that a small and non-significant amount of points was lost during tracking. Overall motion cues were calculated by averaging the values obtained for each trajectory.

3.2.3 Layer 3

Layer 3 is in charge of segmenting motion in order to individuate motion and non-motion (pause) phases. QoM was used to perform such segmentation. QoM is related to the overall amount of motion and its evolution in time can be seen as a sequence of bell-shaped curves (*motion bells*). In order to segment motion, a list of these motion bells was extracted and their features (e.g., peak value and duration) computed. An empirical threshold was defined for these experiments: the dancer is considered to be moving if its current QoM is above 2.5% of the average value of the QoM computed along each whole dance fragment.

Segmentation allows further higher-level cues to be extracted, e.g., cues related to the time duration of motion and pause phases. A concrete example is the Directness Index (DI), calculated as the ratio between the length of the straight trajectory connecting the first and the last point of a motion trajectory and the sum of the lengths of each segment constituting the trajectory. Moreover, segmentation can be considered as a first step toward the analysis of the rhythmic aspects of the dance. Analysis of the sequence of pause and motion phases and their relative time durations can lead to a first evaluation

14 Kilian 2001

15 See for example Boone/Cunningham 1998

of dance tempo and its evolution in time, i.e., tempo changes and articulation (the analogue to music legato/staccato). Parameters from pause phases can also be extracted to differentiate real standing-still positions from active pauses involving low-motion (hesitation, subtle swaying or tremble, e.g., due to instable equilibrium or fatigue).

Furthermore, motion fluency and impulsiveness can be evaluated. They are related to Laban's Flow and Time axes. Fluency can be estimated starting from an analysis of the temporal sequence of motion bells. A dance fragment performed with frequent stops and restarts (i.e., characterised by a high number of short pause and motion phases) will gain the result of being less fluent than the same movement performed in a continuous, "harmonic" way (i.e., with a few long motion phases). The hesitating, bounded performance will be characterised by a higher percentage of acceleration and deceleration in the time unit (due to the frequent stops and restarts), a parameter that has been demonstrated to be of relevant importance in motion flow evaluation.[16]

A first measure of impulsiveness can be obtained from the shape of a motion bell. In fact, since QoM is directly related to the amount of movement detected, a short motion bell having a high peak value will be the result of an impulsive movement (i.e., a movement in which speed rapidly moves from a value near or equal to zero, to a peak and back to zero). On the other hand, a sustained, continuous movement will show a motion bell characterised by a relatively long time period in which the QoM values have little fluctuations around the average value (i.e., the speed is more or less constant during the movement).

Fluency and impulsiveness are also related to the spectral content of the QoM: a movement having significant energy at high frequencies is a candidate to be characterised by low fluency.

3.2.4 Layer 4

In this experiment, Layer 4 collects inputs from Layers 2 and 3 (18 variables have been calculated on each detected motion phase) and tries to classify a motion phase in terms of the four basic emotions anger, fear, grief and joy.

As a first step, statistical techniques have been used for a preliminary analysis: descriptive statistics and a one-way ANOVA have been computed for each motion cue.[17]

16 See for example Zhao 2001, where a neural network is used to evaluate Laban's flow dimension.

17 Results of such preliminary analysis can be found in Mazzarino 2002; Camurri/Lagerlöf/Volpe 2003; Volpe 2003.

Decision tree models were then built for classification. Five training sets (85% of the available data) and five test sets (15% of the available data) were extracted from the data set. The samples for the test sets were uniformly distributed along the four classes and the five dancers. Five decision trees were built on the five training sets and evaluated on the five test sets. The Gini's index of heterogeneity was used for building the decision trees. Decision trees were selected for this study since they produce rules that can be used to interpret the results. Comparison with other classification techniques (e.g., Neural Networks, Support Vector Machines) remains a task for possible future work.

The above-described techniques in the four layers were implemented in our EyesWeb open software platform.[18] The Expressive Gesture Processing Library[19] includes these and other processing modules.

3.3 Results

Results from spectators' ratings are described in Camurri/Lagerlöf/Volpe 2003. The results obtained on the five decision trees can be summarised as follows (results for the best model are reported in Tables 1 and 2 (see p. 232) showing the confusion matrices for the training set and for the test set respectively).

Two models (3 and 5) fit the data set quite well; the rates of correct classification on the training set for these two models averaged over the four classes are 78.5% and 61.6%, respectively. Three models (1, 2, and 4) have difficulties in classifying fear. The rates of correct classification on the training set for these three models averaged over the four classes are 41.9%, 38.7%, and 36.0%, respectively. Models 2 and 4 also have problems with joy, which means that they distinguish correctly only between anger and grief.

A similar situation can be observed in the evaluation carried out on the test set: only models 3 and 5 are able to classify all four emotions correctly. Model 1 cannot classify fear, while models 2 and 4 cannot classify fear and joy.

The rates of correct classification on the test set for the five models averaged on the four classes are respectively: 40%, 36%, 36%, 26%, and 40%. Thus the average rate of correct classification on the five models is 35.6%. Except for model 4, they are all above chance level (25%). Model 5 can be considered as the best model, since it has a rate of correct classification of 40% and is able to classify all four emotions.

18 Camurri et al. 2000; Free download of technical documentation and full software environment are available at <http://www.eyesweb.org>.

19 Camurri/Mazzarino/Volpe 2003

Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	64	71.9	28.1	46	10	2	6
Fear	60	61.7	38.3	15	37	1	7
Grief	86	47.7	52.3	10	19	41	16
Joy	74	64.9	35.1	13	8	5	48

Table 1. Confusion matrix for the training set for the best decision tree

Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	12	41.7	58.3	5	3	0	4
Fear	13	30.8	69.2	6	4	2	1
Grief	12	41.7	58.3	2	0	5	5
Joy	13	46.1	53.8	4	0	3	6

Table 2. Confusion matrix for the test set for the best decision tree

These rates of correct classification, which at first glance seem to be quite low (40% being the best model), should however be considered with respect to the rates of correct classification by spectators who were asked to classify the same dances. In fact, spectators' ratings collected by psychologists in Uppsala show a rate of correct classification (averaged over the 20 dances) of 56%.

The rate of correct automatic classification (35.6%) is thus in between chance level (25%) and the rate of correct classification for human observers (56%).

Furthermore, if the rate of correct classification for human observers is considered as a reference, and percentages are recalculated taking it as 100% (i.e., relative instead of absolute rates are computed), the average rate of correct automatic classification with respect to spectators is 63.6%, and the best model (i.e., model 5) obtains a rate of correct classification of 71.4%.

By observing the confusion matrix of the best model (both for the test set and for the training set) it can be noticed that fear is often classified as anger. This particularly holds for the test set, where fear is the basic emotion which receives the lowest rate of correct classification, since 6 of the 13 motion phases extracted from fear performances are classified as anger. Something similar can be observed in spectators' ratings.[20]

20 Camurri/Lagerlöf/Volpe 2003

A deeper comparison with spectator's ratings shows that while anger is generally well classified both by spectators and by the automatic system (60% for automatic recognition vs. 60.6% for spectators), quite bad results are obtained for fear (below chance level for the automatic classification). The biggest overall difference between spectators and automatic classification was observed for joy (70.4% for spectators vs. 27.7%, just above chance level, for automatic classification). In the case of grief instead, automatic classification performs better than human observers (48.3% for automatic classification vs. 39.8% for spectators): this happens in five cases and mainly for grief. In seven cases, the rate of correct classification for the automatic system is below chance level (and this always happens for fear). In one case, automatic classification did not succeed in finding the correct emotion (Fear – Dancer 4), but spectators obtained 67% of correct classification. In another case, spectators' ratings are below chance level (Grief – Dancer 5), but automatic classification could obtain a rate of correct classification up to 50%.

Dancer 1 obtained the lowest rates of correct classification both from spectators and from the models. Dancer 5 obtains similar rates from both. Dancer 2 is the best classified by spectators and also obtains a quite high rate (with respect to the other dancers) in automatic classification.

4. Analysis of expressive gesture in music performance

The second experiment investigates the mechanisms responsible for the audience's engagement in a musical performance.[21] The aim of this experiment is again twofold: (i) individuating which auditory and visual cues are most involved in conveying the performer's expressive intentions and (ii) testing the model developed by comparing its performance to spectators' ratings of the same musical performances.

In this experiment, expressive gesture was analysed with respect to its ability to convey the intensity of emotion to the audience. The study focused on communication through visual and auditory performance gestures of emotional intensity and the effect of it on spectators' emotional engagement.

The research hypotheses combine hypotheses from Laban's Theory of Effort[22] with hypotheses stemming from performance research[23] and research on the intensity of emotion and tension in music and dance:[24]

21 More detailed description of such an experiment is available from Timmers et al. 2006.

22 Laban 1947, Laban/Lawrence 1963

23 Palmer 1997; Timmers 2002

24 Krumhansl 1996; Krumhansl/Schenck 1997

1. Emotional intensity is reflected in the degree of openness (release) or contraction (tension) of the torso of the performer.
2. Emotional intensity is communicated by the main expressive means for a pianist: tempo and dynamics.
3. Intensity increases and decreases with energy level (speed of movements, loudness, tempo).
4. Intensity is related to the performer's phrasing: it increases towards the end of the phrase and decreases at the phrase boundary with the introduction of new material.

4.1 Method

4.1.1 Musical performance

A professional pianist was asked to perform an emotionally engaging piece of his choice at a concert that was organised for the experiment's purpose. He performed the piece first without public in a normal manner and an exaggerated manner and then with public in a normal, concert manner. Exaggerated meant with enhanced expressivity, which was, according to the pianist, consistent with the style of performance of the early 20th Century.

He performed on a Yamaha Disklavier, which made it possible to register MIDI information of the performance. In addition, audio recordings were made, and video recordings from four sides (Fig. 2). The video recordings from the left were presented to the participants of the experiment.

The pianist chose to perform Etude Op. 8 no. 11 by Alexander Skriabin, which is a slow and lyrical piece (Andante cantabile) in a late Romantic style that has a considerable amount of modulations. According to the pianist, the piece can be played with a lot of freedom. Theoretically, the piece has a simple A B A with coda structure (A A' B A" A"' C, to be more precise), but the pianist interpreted the line of the music differently: the first main target of the music is a release of tension halfway through the B section. Everything preceding this target point is a preparation for this tension release. The A section is anyway preparatory; it leads towards the start of the B section, which is the real beginning of the piece. After this release of tension, the music builds up towards the dramatic return of the theme of the A section. This prepares for the second possible point of tension release halfway through the coda at a general pause. The release, however, is not continued, and the piece ends most sadly.

Fig. 2. Video recordings of the piano performances (right, top, left, and front views)

4.1.2 Participants

12 people participated in the experiment; among them were four musicians. The participants varied greatly in musical experience. Some of them never had had music lessons and hardly listened to classical music, while others had basically performed classical music for their entire lives.

4.1.3 Procedure

The participants saw the performances on a computer screen and heard them on loudspeakers. They saw and heard the performances two times. During the first hearing, they indicated the phrase boundaries in the music by pressing the button of the joystick. During the second hearing, they indicated to what extent they were emotionally engaged with the music by moving a MIDI-slider up and down. The order of the repeated performances was randomised over participants. The whole procedure was explained to them by a written instruction and a practice trial.

4.2 Analyses

4.2.1 Auditory Performance Data

The key velocity and onset times of notes were extracted from the MIDI files (layer 1). From this, the average key velocity for each quarter note was calculated as well as inter-onset intervals (IOI's) between successive quarter notes (layer 2). The quarter note IOI is an accurate measure of local duration, while key velocity corresponds well to local loudness. These measures were interpreted as a direct expression of emotional intensity and as an expression of musical phrasing.[25]

4.2.2 Visual Performance Data

For the analysis of the movement of the pianist, we concentrated on the movement of the head, which shows both backward-forward movement (y-direction) and left-right movement (x-direction). The position of the head was measured, using the Lucas and Kanade feature-tracking algorithm[26] that assigns and tracks a specified number (in our case 40) of randomly assigned moving points within a region (layer 1). Velocity and acceleration were calculated for each trajectory using the symmetric backward technique for the numeric derivative (layer 2). Average values of position and velocity among the forty trajectories were calculated for both the x and y component. In addition, the velocity values were integrated for the x and y movement to get a general measure of amount of movement over time. Redundancy in the number of points (i.e., forty points instead, for example, of just the barycentre of the blob) allowed us to get more robust and reliable values for velocity. A low-pass filter was applied to smooth the data obtained. Measures were summarised per quarter-note in order to be comparable to the other measures.

4.2.3 Spectators' ratings

For each quarter note in the performance, the number of people who indicated a phrase boundary was calculated by summing the number of boundary indications per quarter note over participants. This sum per quarter note was expressed as a multiple of chance-level, where chance-level corresponded to an even distribution of the total of segment-indications over quarter notes. This segmentation measure will be abbreviated as SM.

25 See 4.3 Results.

26 Lucas/Kanade 1981

The indication of emotional engagement was measured at a sampling rate of 10 Hz using a MIDI-slider that had a range from 0 to 127. The average level of the MIDI-slider (emotion measure, abbreviation EM) per quarter note was calculated for each participant separately.

An EyesWeb patch application was developed to store and process participants' data in real-time.

4.3 Results

4.3.1 Auditory Performance Data (layer 3)

The resulting profiles of quarter note key velocity and quarter note IOI were highly similar for the three performances: they all started at a slow tempo and with soft dynamics, and had considerable crescendi and accelerandi in the A section, a diminuendo and crescendo in the B section accompanied by first a highly variable tempo and thereafter an accelerando, a fast and loud return of the A section with limited variation in tempo and dynamics, a soft and slower repeat of the theme, and a coda that fades away in dynamics and tempo (Fig. 3). This global pattern is indicated by arrows at the bottom of Figure 3.

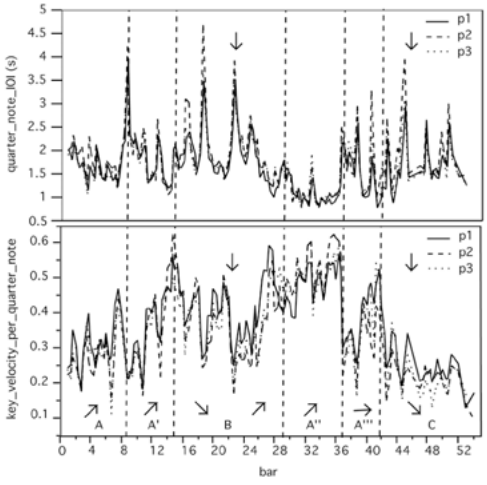

Fig. 3. The duration per quarter note and the key velocity per quarter note as it varies throughout the Skriabin Etude. Separate plots for the three performances of the piece. Vertical lines indicate section boundaries. Arrows are explained in the text.

In addition to this global pattern, the IOI profile shows the characteristic peaks of phrase-final lengthenings. It shows this at a fairly high density and large magnitude, except in the forte return of the A section (A''). The key velocity profile shows drops in velocity at most phrase boundaries, but these are compensated by strong crescendi in most sections.

4.3.2 Visual Performance Data (layer 3)

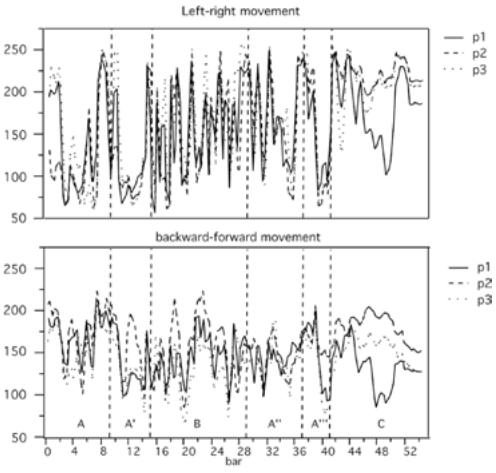

Fig. 4. The position of the head plotted per quarter note. Upper panel shows left-right position (x) and bottom panel the backward-forward position (y). Separate plots for the three performances of the piece. Vertical lines indicate section boundaries.

The position of the head is plotted in Figure 4 for two dimensions: left-right (upper panel) and backward-forward (bottom panel). The movement of the head was especially pronounced and especially consistent over performances in the left-right direction (correlation between p1 and p2 and between p2 and p3 was 0.79; it was 0.83 between p1 and p3). The backward-forward movement becomes more pronounced for the later performances (p2 and p3). The periodic movement is relatively fast in the middle parts of the piece (B and A'''') and slower in the outer parts. This suggests an intensification towards the middle followed by a relaxation towards the end.

4.3.3 Relation between performance data

Correlations between performance measures were calculated to check the coherence between measures. Key velocity and IOI are negatively correlated (r = -0.51 on average). Velocity of head movement is positively correlated with key velocity (r = 0.45 on average) and negatively with IOI (r = -0.25 on average). The low correlation between values is partly due to the asynchrony between the periodicity of the measures. If peak values (maximum for key and movement velocity and minimum for IOI) per two bars are correlated, agreement between movement and sound measures becomes higher. Especially the two velocity measures turn out to be highly correlated (r = 0.79 on average for key and movement velocity, versus r = -0.38 on average for movement velocity and IOI).

All performance measures show a periodic increase and decrease. To check the relation between these periodicities and the musical structure, the location of mimima in key velocity, and maxima in IOI, x-position and y-position were compared to the location of phrase boundaries. Generally, the Skriabin Etude has a local structure of two-bar phrases. The forward and the left position of the performer were taken as start/end point for periodicity. IOI was most systematically related to the two-bar phrasing of the Skriabin piece, followed by key velocity. 55% of the phrase-boundaries were joined by a slowing-down in tempo. The other phrase boundaries were directly followed by a slowing down in tempo (a delay of 1 quarter note). For the key velocity, 42% of the phrase-boundaries coincided with a minimum in key velocity, 15% were anticipated by a minimum and 28% followed by a minimum. The period boundaries of the movement of the pianist hardly synchronised with the score-phrasing. The location of these boundaries varied greatly with respect to the two-bar score-phrasing.

4.3.4 Relation between Performance and Listeners Data (layer 4)

In this study, we had four hypotheses concerning the communication of intensity of emotion in musical performances.

Hypothesis 1 predicts that intensity of emotion is positively correlated with backward-forward movement (y). This hypothesis is easily tested and contradicted: the correlation between listeners' indication of intensity of emotion and backward-forward position is negative (r is -0.23, -0.50, -0.29 for p1, p2 and p3, respectively). It is also contradicted with respect to the other performance data: y-position is negatively correlated with velocity and positively correlated with IOI, which means that the performer moves forward in soft and slow passages and backwards in louder and faster passages.

Hypothesis 2 predicts that tempo and dynamics cooperate to communicate intensity of emotion. This is made problematic by the fairly low correlation between IOI and key velocity and by their different relation with the score-phrasing. Instead the performance data suggests a differentiation in function between the two expressive means, and tempo strongly communicates phrasing.

Hypothesis 3 predicts high movement to correspond with intense dynamics and fast tempi. As we have seen in the previous section, dynamics and movement velocity agree more strongly than movement velocity and tempo. Especially the variation in velocity peaks corresponds.

Hypothesis 4 relates phrasing to intensity of emotion. A clear phrase ending is predicted to coincide with a release in emotional intensity.

A series of multiple regression analyses was carried out. In the first analysis, quarter note IOI, key velocity, and movement velocity were used to predict EM. In the second analysis, the same variables were used to predict SM. In the third analysis, peak values per hyper-bar were used to predict average emotion measure per hyper-bar. All analyses were done directly and with a time-delay of one, two and three quarter notes of the performance data with respect to the listeners' data. The best R^2 obtained will be reported. These were obtained with a delay of either zero or one for SM, and either two or three for the EM.

SM was rather well-predicted by the model, given the Rs^2 of 0.34, 0.33, 0.30 for p1, p2 and p3, respectively. From this model, IOI was the only significant variable. In other words, duration was a fairly good predictor of the variation in number of participants indicating a section-boundary. More participants indicated a phrase-boundary for longer durations.

EM was well-predicted by the quarter note model, but even better by the second model that took the peak values per hyper-bar to predict the average EM per hyper-bar. The quarter note regression analysis had an R^2 of 0.45, 0.68, 0.50 for p1, p2, and p3, respectively, while the hyper-bar peak value regression had an R^2 of 0.53, 0.82, and 0.56. Velocity was always the most significant variable, and was the only significant variable for the hyper-bar peak value regression. For the quarter note regression movement velocity also reached significance for p2 and p3, and IOI for p2. All Rs^2 are relatively high for p2, which suggests that the exaggerated expression of this performance increased communication.

As a comparison, the analyses were repeated with x-position and y-position as independent movement variables instead of the more general movement velocity variable. The results did not improve or change from this alteration, instead x and y-position did not contribute significantly to any of the regressions.

These results confirm a differentiation between expressive means: tempo primarily communicates segmentation, while dynamics communicates emotional intensity. Velocity of the movement is correlated with dynamics and may therefore also reflect emotional intensity, but the sounding parameters are the main communicative factors.

The results are suggestive counter-evidence for hypothesis 4. The failure of tempo to explain variations in emotional intensity contradicts the theory that phrase-final lengthenings cause a release in emotional intensity. There is, however, another way in which phrasing and the dynamics of tension and release do relate, which is at a higher and more global level. Phrase-final lengthenings occur at a high rate and a local scale. At this local scale the relation is weak. Major phrase boundaries that are indicated by a drop in tempo and dynamics are, however followed by a clear release in intensity. Moreover

the global variation of dynamics to which the variation in emotional intensity is so strongly related is the performer's way of communication of the overall form: the first part is an introduction and builds up to the B section, which he considers as the real beginning of the piece. This beginning is again a preparation for the first target of the piece: the release of tension in the middle of the B section.[27] Hereafter, tension builds up towards the dramatic return of the theme, which leads via a repeat of the theme in contrasting dynamics to the second important target of the theme: the second possible release of tension at the general pause. After the general pause, the release is not given and all hope is lost. The piece ends most sadly. The pianist most skilfully expresses this interpretation in the patterning of dynamics.[28] The resulting phrasing is over the entire piece, with subdivisions at measures 22 and 36. The return of the theme is the culminating point of the piece, whereafter tension can release. According to the pianist, this tension cannot, however, be fully resolved.

4.4 Summary

This study had two aims: (i) individuating which auditory and visual cues are most involved in conveying the performer's expressive intentions and (ii) testing the model developed by comparing their performances with spectators' rating of the same musical performances. The auditory and visual cues most involved in conveying the performer's expressive intentions were hypothesised to be key velocity, IOI, movement velocity, and the openness or contraction of the performer's posture. In addition, a relation between phrasing and emotional tension-release was expected.

The analyses of performance data suggested the opposite relation between emotional intensity and the performer's posture. The pianist leaned forward for softer passages and backward for intensive passages. In addition it suggested a differentiation in expressive means, with tempo on one side, and key velocity and movement velocity on the other side.

When relating the performer's data to the listeners' data, this differentiation in expressive means was confirmed. Tempo communicates phrase boundaries, while dynamics is highly predictive for the intensity of emotion felt. Emotional engagement correlated strongly with key velocity, which means that emotional engagement tended to increase with increase of dynamics and decrease at points of softer dynamics. This does not mean that soft passages were without emotional tension, but they were points of relative emotional relaxation. Hardly any evidence was found for movement cues influencing

27 See downward pointing arrows in Fig. 3.

28 See arrows in the key velocity panel of Fig. 3.

listeners' ratings. The sound seemed to be the primary focus of the participants, and vision seemed subsidiary. The local phrase boundaries indicated by tempo did not lead to a release of emotional intensity. The modulation of dynamics over a larger time-span communicates the overall form of the piece and, at that level, intensity did increase and decrease within phrases.

5. Applications of Multimodal Expressive Systems: the Case Study of Active Music Listening

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 the traditional listening experience, in which the public can interact in many ways with performers to modify the expressive features of a piece.

The need to recover such an active attitude with respect to music is emerging strongly, and novel paradigms of *active experience* will be developed. By active experience and *active listening* we mean that listeners are enabled to interactively operate on music content, by modifying and molding it in real-time while listening. Active listening is the basic concept for a novel generation of interactive music systems, which are particularly addressed to a public of beginners, naïve and inexperienced users, rather than to professional musicians and composers.

Active listening is also a major focus for the new EU-ICT Project SAME (Sound and Music for Everyone, Everyday, Everywhere, Every Way).[29] SAME aims at: (i) defining and developing an innovative networked end-to-end research platform for novel mobile music applications, allowing new forms of participative, experience-centric, context-aware, social, shared, active listening of music; (ii) investigating and implementing novel paradigms for natural, expressive/emotional multimodal interfaces, empowering the user to influence, interact, mold, and shape the music content, by intervening actively and physically into the experience; and (iii) developing new mobile context-aware music applications, starting from the active listening paradigm, which will bring back the social and interactive aspects of music to our information technology age.

29 <http://www.sameproject.eu>

In the direction of defining novel active listening paradigms, we recently developed a system, the Orchestra Explorer,[30] allowing users to physically navigate inside a virtual orchestra, to actively explore the music piece the orchestra is playing, and to modify and mold in real-time the musical performance through expressive full-body movement and gesture. By walking and moving on the surface, the user discovers each single instrument and can operate through her expressive gestures on the musical piece which the instrument is playing. The interaction paradigm developed in the Orchestra Explorer is strongly based on the concept of navigation in a physical space where the orchestra instruments are placed. The Orchestra Explorer is intended for use by a single user.

Our novel multimodal system for social active listening, *Mappe per Affetti Erranti*, starts from the Orchestra Explorer and the lessons learned in over one year of permanent installation of the Orchestra Explorer at our site at Casa Paganini, and several installations of the Orchestra Explorer at science exhibitions and public events.

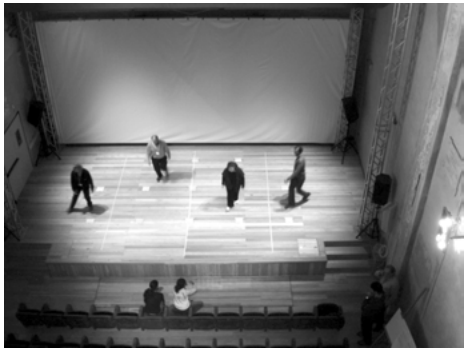

Fig. 5. A group of users interacting with the installation "Mappe per Affetti Erranti" at the auditorium of Casa Paganini.

Mappe per Affetti Erranti extends and enhances the Orchestra Explorer in two major directions. On the one hand it reworks and extends the concept of navigation by introducing multiple levels: from the navigation in a physical space populated by virtual objects or subjects (as it is in the Orchestra Explorer) up to the navigation in virtual affective, emotional spaces populated by different expressive performances of the same music piece. Users can navigate in such affective spaces by their expressive movement and gesture. On the other hand, *Mappe per Affetti Erranti* is explicitly designed for use by multiple users, and encourages collaborative behaviour: only social collaboration allows a correct reconstruction of the music piece. In other words, while users explore the physical space, the (expressive) way in which they move and the degree of collaboration between them allow them to explore at the same time an affective, emotional space.

30 Camurri/Canepa/Volpe 2007

The basic concept of *Mappe per Affetti Erranti* is the collaborative active listening of a music piece through the navigation of maps at multiple levels, from the physical level to the emotional level.

At the physical level the space is divided in several areas. The voice of a polyphonic music piece is associated to each area. The presence of a user (even a single user) triggers the reproduction of the music piece. By exploring the space, the user walks through several areas and listens to the single voices separately. If the users stays in a single area, she listens to the voice associated to that area only. If the user does not move for a given time interval, the music fades out and turns off.

The user can mold the voice she is listening to in several ways. At a low level, she can intervene on parameters such as loudness, density, amount of reverberation. For example, by opening her arms, the user can increase the density of the voice (she listens to two or more voices in unison). If she moves toward the back of the stage the amount of reverberation increases, whereas toward the front of the stage the voice becomes drier.

At a 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. The system analyses the expressive intention which the user conveys with her expressive movement and gesture, and translates it in a position (or a trajectory) in an affective, emotional space. Like the physical space, such affective, emotional space is divided in several areas, each one corresponding to a different performance of the same voice with a different expressive intention. Several examples of such affective, emotional spaces are available in the literature, for example the spaces used in dimensional theories of emotion[31] or those especially developed for the analysis and synthesis of expressive music performance.[32]

Users can thus explore the musical piece in a twofold perspective: navigating the physical space they explore the polyphonic musical structure; navigating the affective, emotional space they explore music performance. A single user, however, can only listen to and intervene on a single voice at a time: she cannot listen to the whole polyphonic piece with all the voices.

Only a group of users can fully experience *Mappe per Affetti Erranti*. In particular, the musical piece can be listened to in its whole polyphony only if a number of users at least equal to the number of voices is interacting with the installation. Moreover, since each user controls the performance of the voice associated to the area she occupies, the whole piece is performed with the same expressive intention only if all the users are moving with the same expressive intention. Thus, the more users move with different, conflicting

31 See for example Russel 1980; Tellegen et al. 1999.

32 See for example Juslin 2000; Canazza et al. 2000; Vines et al. 2005.

expressive intentions, the more the musical output is incoherent and chaotic. But the more users move with similar expressive intentions and in a collaborative way, the more the musical output is coherent and the musical piece is listened to in one of its different expressive performances.

Mappe per Affetti Erranti can therefore be experienced at several levels: by a single user who has a limited but still powerful set of possibilities of interaction, by a group of users who can fully experience the installation, and by multiple groups of users. In fact, each physical area can be occupied by a group of users. In this case each single group is analysed and each participant in a group contributes towards intervening on the voice associated to the area the group is occupying. Therefore, at this level a collaborative behaviour is encouraged among the participants in each single group and among the groups participating in the installation.

The possibility of observing a group or multiple groups of users during their interaction with *Mappe per Affetti Erranti* makes this installation an ideal test-bed for investigating and experimenting group dynamics and social network scenarios.

6. Discussion

The modelling of expressive gesture is being accorded growing importance from both research and industry communities, even if we can consider it as being in its infancy. The main outputs of our research are the definition of a unified multimodal conceptual framework for expressive gesture processing, the experimental results obtained from the two described experiments, and a collection of software modules for cue extraction and processing. The conceptual framework proved to be useful and effective in two different scenarios, well represented by the two experiments described in the paper.

In the first experiment, we focused on the communication of basic emotions from a dancer to the audience, while in the second experiment we focused on the mechanisms that possibly cause emotional engagement in the audience.

The dance experiment can be considered as a first step and a starting point toward understanding the mechanisms of expressive gesture communication in dance. A collection of cues that have some influence in such a communication process was individuated, measured, and studied. A first attempt of automatic classification of motion phases was carried out and some results were obtained (e.g., an average rate of correct classification which was not particularly high, but well above chance level). Some directions for future research have also emerged. For example, other classification techniques could be employed and their performances compared with what

we obtained with decision trees. Some aspects in dance performance that were only marginally considered should be taken into account. In particular, aspects related to rhythm should be further investigated. Expressive cues such as impulsiveness and fluency should be further worked out. Moreover, perceptual experiments would be needed to empirically validate the expressive cues extracted.

The music experiment can be considered as a first step towards the understanding of the relation between movement and sound parameters of a performance, their expressive forms and functions, and their communicative function for spectators. A next step should involve a larger variety of performances and a larger collection of calculated cues, and cues should be fitted to the responses of individual spectators in order to get a deeper as well as broader understanding.

Expressive multimodal systems open a novel range of applications. At the end of this chapter we focused on an application in the field of active music listening.

References

Bobick, Aaron F./Davis, James W. (2001): »The Recognition of Human Movement Using Temporal Templates«. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 257-267.

Boone R. Thomas/Cunningham Joseph G. (1998): »Children's Decoding of Emotion in Expressive Body Movement: The Development of Cue Attunement«. Developmental Psychology 34, 1007-1016.

Bradsky, Gary R./Davis, James W. (2002): »Motion Segmentation and Pose Recognition with Motion History Gradients«. Machine Vision and Applications 13, 174-184.

Camurri, Antonio et al. (2000): »EyesWeb – Toward Gesture and Affect Recognition in Interactive Dance and Music Systems«. Computer Music Journal 24(1), 57-69.

Camurri, Antonio/Lagerlöf, Ingrid/Volpe, Gualtiero (2003): »Emotions and Cue Extraction from Dance Movements«. International Journal of Human Computer Studies, 59(1-2), 213-225.

Camurri Antonio/Mazzarino, Barbara/Volpe, Gualtiero (2004): »Analysis of Expressive Gesture: The Eyesweb Expressive Gesture Processing Library«. In: Antonio. Camurri/Gualtiero Volpe (Eds.), Gesture-based Communication in Human-Computer Interaction, GW'08, LNAI 2915, Berlin: Springer, 2004, 460-467.

Camurri, Antonio et al. (2005): »Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications«. IEEE Multimedia 12(1), 43-53.

Camurri, Antonio/Canepa, Corrado/Volpe, Gualtiero (2007): »Active Listening to a Virtual Orchestra through an Expressive Gestural Interface: The Orchestra Explorer«. In: Proceedings of the 2007 Conference on New Interfaces for Musical Expression (NIME07), 56-61.

Canazza, Sergio et al. (2000): »Audio Morphing Different Expressive Intentions for Multimedia Systems«. IEEE Multimedia, 7(3), 79-83.

Chi, Diane/Costa, Monica/Zhao, Liwei/Badler, Norman (2000): »The EMOTE Model for Effort and Shape«. In: John S. Brown/Kurt Akeley (Eds.) Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New York: ACM Press/Addison-Wesley Publishing Co., 173-182.

Cowie, Roddy et al. (2001): »Emotion Recognition in Human-Computer Interaction«. IEEE Signal Processing Magazine 1, 33-80.

Friberg, Anders et al. (2000), »Generating Musical Performances with Director Musices«. Computer Music Journal 24(3), 23-29.

Hashimoto, Shuji (1997): »KANSEI as the Third Target of Information Processing and Related Topics in Japan«. In: Antonio Camurri (Ed.), Proceedings of the International Workshop on KANSEI: The Technology of Emotion, Genova: AIMI (Italian Computer Music Association) and DIST-University of Genova, 101-104.

Juslin, Patrik 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.

Kilian, Johannes (2001): »Simple Image Analysis By Moments«, OpenCV library documentation. Online available in the repository of the OpenCV Yahoo group (last access: July 2008).

Krumhansl, Carol L. (1996): »A Perceptual Analysis of Mozart's Piano Sonata K. 282: Segmentation, Tension and Musical Ideas«. Music Perception 13(3), 401-432.

Krumhansl, Carol L./Schenck, Diane L. (1997): »Can Dance Reflect the Structural and Expressive Qualities of Music? A Perceptual Experiment on Balanchine's Choreography of Mozart's Divertimento No. 15«. Musicae Scientiae 1, 63-85.

Kurtenbach, Gordon/Hulteen, Eric A. (1990): »Gesture in Human-Computer Interaction«. In: Benda Laurel (Ed.), The Art of Human-Computer Interface Design, Reading: Addison-Wesley, 309-317.

Laban, Rudolf (1963): Modern Educational Dance, London: Macdonald & Evans.

Laban, Rudolf/Lawrence, F. C. (1947): Effort, London: Macdonald & Evans.

Lagerlöf, Ingrid/Djerf, Marie (2001): On Cue Utilization for Emotion Expression in Dance Movements, Manuscript in preparation, Department of Psychology, University of Uppsala.

Liu, Yanxi/Collins, Robert T./Tsin, Yanghai (2002): »Gait Sequence Analysis using Frieze Patterns«. In: Anders Heyden/Gunnar Sparr/Mads Nielsen/Peter Johansen (Eds.), Computer Vision – ECCV 2002, 7th European Conference on Computer Vision, New York: Springer, 657-671.

Lucas, Bruce D./Kanade, Takeo (1981): »An iterative image registration technique with an application to stereo vision«. In: Patrick J. Hayes (Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81), Los Altos: William Kaufmann, 674-679.

McNeill, David (1992): Hand and Mind: What Gestures Reveal About Thought, Chicago/London: University of Chicago Press.

Palmer, Caroline (1997): »Music Performance«. Annual Review of Psychology 48, 115-138.

Russell, James A. (1980): »A Circumplex Model of Affect«. Journal of Personality and Social Psychology 39, 1161-1178.

Pollick, Frank E. et al. (2001): »Perceiving Affect from Arm Movement«. Cognition 82(2), B51-B61.

Tellegen, Auke (1999): »Watson David, and Clark Lee Anna. On the Dimensional and Hierarchical Structure of Affect«. Psychological Science 10(4), 297-303.

Timmers, Renee (2002): Freedom and Constraints in Timing and Ornamentation: Investigations of Music Performance, Maastricht: Shaker Publishing.

Timmers, Renee et al. (2006): »Listeners' Emotional Engagement with Performances of a Skriabin Etude: An Explorative Case Study«. Psychology of Music 34(4), 481-510.

Vines Bradley W. et al. (2005): »Dimensions of Emotion in Expressive Musical Performance«. Ann. N.Y. Acad. Sci., 1060, 462-466.

Volpe, Gualtiero (2003): Computational Models of Expressive Gesture in Multimedia Systems. PhD Thesis, Faculty of Engineering, University of Genova, April 2003.

Wanderley, Marcelo M./Battier, Marc (Eds.) (2000): Trends in Gestural Control of Music, Paris: IRCAM.

Wilson, Andrew. D./Bobick, Aaron. F./Cassell, Justine (1996): Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996, 14-16 Oct, 66-71.

Zhao, Liwei (2001): Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures, PhD Thesis, University of Pennsylvania.

Acknowledgements

We thank Ingrid Lagerlöf for the joint work carried out in the experiment on expressive gesture conveying emotions in dance performances, Renee Timmers and Matja Marolt for their contribution to the experiment on music performance, our colleagues at the InfoMus Lab, and the pianist Massimiliano Damerini for his artistic contributions in providing the material for the piano studies.

The novel applications on active music listening are the main objective of the EU-ICT Project SAME33, focusing on new forms of participative, context-aware, socially active listening to music.

33 <http://www.sameproject.eu>