Emotional control and visual representation using advanced audiovisual interaction

Vasillis Psarras

Camberwell College of Arts,
University of the Arts London,
Peckham Road,
London SE5 8UF, UK
E-mail: billhaze85@gmail.com

Andreas Floros*, Konstantinos Drosos and Marianne Strapatsakis

Department of Audiovisual Arts,
Ionian University,
Plateia Tsirigoti 7,
Corfu 49100, Greece
Fax: +30 26610 48491
E-mail: floros@ionio.gr
E-mail: kdrosos@ionio.gr
E-mail: maristra@ionio.gr
*Corresponding author

Abstract: Modern interactive means combined with new digital media processing and representation technologies can provide a robust framework for enhancing user experience in multimedia entertainment systems and audiovisual artistic installations with non-traditional interaction/feedback paths based on user affective state. In this work, the 'Elevator' interactive audiovisual platform prototype is presented, which aims to provide a framework for signalling and expressing human behaviour related to emotions (such as anger) and finally produce a visual outcome of this behaviour, defined here as the emotional 'thumbnail' of the user. Optimised, real-time audio signal processing techniques are employed for monitoring the achieved anger-like behaviour, while the emotional elevation is attempted using appropriately selected combined audio/visual content reproduced using state-of-the-art audiovisual playback technologies that allow the creation of a realistic immersive audiovisual environment. The demonstration of the proposed prototype has shown that affective interaction is possible, allowing the further development of relative artistic and technological applications.

Keywords: affective interaction; voice anger detection; audio and emotions; emotional control; audiovisual arts.

Reference to this paper should be made as follows: Psarras, V., Floros, A., Drosos, K. and Strapatsakis, M. (2011) 'Emotional control and visual representation using advanced audiovisual interaction', *Int. J. Arts and Technology*, Vol. 4, No. 4, pp.480–498.

Biographical notes: Vasillis Psarras graduated from the Department of Audiovisual Arts at Ionian University in 2009. Currently, he is a Post-Graduate Student (MA in Visual and Digital Arts) from Camberwell College of Arts at the University of the Arts London. His main fields of interest include video installations, video art, the visualisation of emotion, conceptual art through the prism of digital arts, as well as music composition and soundscapes. He has presented artworks in various video art and audiovisual festivals such as Athens Video Art Festival 2007 and 2008, Crash Fest 08, International Panorama of Film and Video, Festival Miden 2007 and 2008, etc. He was a Member of the Creative Team in an Interdisciplinary Video-Game Project at Ionian University, where he composed video game soundtracks and sound effects. His last artwork developed during his diploma thesis was an interactive audiovisual installation that converts viewer's anger to visual output.

Andreas Floros received his Engineering and PhD degrees from the Department of Electrical and Computer Engineering at the University of Patras, in 1996 and 2001, respectively. His research was mainly focused on digital audio signal processing and conversion techniques. In 2001, he joined the semiconductor industry, working in projects related with digital audio delivery over wireless networks and lately with audio encoding and compression implementations in embedded processors. In 2005, he was appointed as a Visiting Assistant Professor in the Department of Audiovisual Arts at Ionian University. Since January 2009, he is an Assistant Professor at the same department, with research interests including binaural and three-dimensional audio reproduction, interactive-intelligent digital audio signal processing and emotionally driven sound synthesis. He is the Secretary of Audio Engineering Society Greek Section, a Member of the Hellenic Institute of Acoustics and the Technical Chamber of Greece.

Konstantinos Drosos received his BEng equivalent degree from the Department of Technology of Sound and Musical Instruments, Technological Educational Institute of Ionian Islands, Greece, and his MSc in Sound and Vibration Studies from the Institute of Sound and Vibration Research, University of Southampton, UK, in 2005 and 2007, respectively. Since 2008, he is an Acoustics and Electroacoustics Consultant in various projects and Laboratory Coordinator from the Department of Technology of Sound and Musical Instruments at the Technological Educational Institute of Ionian Islands. In 2009, he started his PhD Dissertation from the Department of Audiovisual Arts, Ionian University, in the area of psychoacoustics. He is a Student Member of the Hellenic Institute of Acoustics.

Marianne Strapatsakis first degree was from the Athens College of Technology in Interior Architecture. She continued her studies in Paris, where she graduated with the following degrees from: the Ecole Nationale Superieure des Beaux Arts, a Diploma in Painting and a Certificate in Sketch, the Ecole du Louvre, a Certificate in History of Art and the University of Paris I, Sorbonne a Diploma, Licence of Arts Plastiques. Since 2004, she has served as a Substitute Professor in the Department of Audiovisual Arts at the Ionian University, where she is currently an Associate Professor. In 2005, she was appointed as a Visiting Professor at the Athens School of Fine Arts. She has realised 31 individual exhibitions and has participated in 52 group exhibitions worldwide.

1 Introduction

Emotional expression represents one of the most important aspects of human everyday life and an important factor in communication. As the concept of interaction continuously evolves by invoking new techniques and ideas in the area of interactive multimedia environments, it is expected that in the very near future the term 'user interface' will expand to include new human life contexts, like emotions. Obviously, this procedure requires the development of mechanisms that accurately recognise emotions and associate them in order to provide the user affective mood. However, prior to that, two more fundamental issues should be addressed: the definition of the exact emotions that will be considered, taking into account the requirements of the specific application, as well as their accurate modelling that will further allow their quantitative description and estimation.

In typical interactive audiovisual environments, the previously described role of emotions as the triggering event for interaction can be inverted. This means that alternative means of interaction can be targeted to raise single or complex emotional states. A typical example is music: starting from a single piano tone (Baraldi et al., 2006), up to complex sound sequences organised in the well-known form of music (Gabrielsson and Lindström, 2001), specific emotions can be expressed and invoked to any human listener.

Invoking emotions through music have recently attracted the research interest from both the analysis and synthesis sides (Friberg, 2008) for algorithmically extracting emotions from music signals and for realising parametric music synthesis by taking into account the desired emotional target, respectively. Moreover, a number of published works have investigated the retrieval and definition of the emotional information contained e.g. in human speech (Kienast and Sendlmeier, 2000), aiming to analyse and determine the fundamental characteristics of speech that also carry the affective voice content.

This work goes a step further by combining both the above emotional analysis and synthesis perspectives: we focus on identifying, controlling and visually representing the expressed intensity of human behaviour related to emotions through combined audiovisual means and technologies. More specifically, in this work we define three different and distinct processes:

1 intensity *identification* (or tracking) i.e. performed in real-time through audio (human voice) recordings and appropriate audio signal processing and information retrieval techniques

2 intensity *control*, a process that aims to 'elevate' the affective intensity to a specific value (hence the term 'Elevator' in the title of the work)

3 intensity *representation* that attempts to convert the elevated human emotion-related behaviour into an audio/visual snapshot (a kind of emotional 'thumbnail') that uniquely and efficiently describes the current human emotional conditions.

For the purposes of this work, the above three processes were realised and integrated within an interactive audiovisual installation prototype termed 'Elevator'. It is nowadays well known that interactive audiovisual installations represent a new form for realising complex human-oriented experiments that most commonly involve human–machine interactions. Moreover, they are also employed by modern artists as a new artistic

expression approach (Birchfield et al., 2006). Hence, new terms and ideas originating from the general concept of interaction are nowadays frequently used to provide novel means of audio and visual production, where the human audience is actively participating in the production process (Birchfield et al., 2005; Boxer, 2005).

'Elevator' represents an interactive audiovisual installation designed and developed during this work. The main purpose of 'Elevator' is to signal and control in real-time human behaviour related with emotions using appropriately combined audiovisual content, capture the elevated emotional state and finally create a visual output that corresponds to its intensity. For the purposes of this work, we focus on a very common emotion that extensively characterises every being in nature: anger. Hence, in the context presented here, the proposed 'Elevator' platform provides the necessary framework for recognising/signalling, expressing and visualising anger-like human behaviour.

The rest of this paper is organised as follows: in Section 2, a brief summary of the relationship between audio signals in general and the emotions is presented, focusing mainly on existing methods that are employed for extracting the affective audio content. Next, the analytic description of the 'Elevator' interactive installation is provided in Section 3, followed by a brief analysis of the functional and behavioural observations made during the installation prototype demonstration in Section 4. Finally, Section 5 concludes this work and accents further interaction and audio/visual enhancements that may be integrated in the 'Elevator' prototype in the future.

2 Audio and emotions

Studying the impact of music signals on human emotions has been a challenging research prospective since late 1930s (Hevner, 1936). However, recently many researchers are trying to assess the answer to the question 'why music is so closely related to emotions?' by developing analytic models that aim to represent the relation between music and raised emotions (Wallis et al., 2008).

The initial approach on identifying the relation between music and corresponding emotions was to define (and further quantify) the emotional variations that are induced by specific features and parameters of music signals, such as tempo and rhythm (Gundlach, 1935; Husain et al., 2002). Additional experiments have considered systematically varied compositions that were heard by subjects that rated the perceived emotional feelings (Juslin and Laukka, 2003).

In general, in order to analytically extract a general relationship between music and emotions, efficient and accurate modelling of emotions is required. Towards this aim, there are a number of theories and interdisciplinary approaches for emotion research (Scherer, 2004). For example, a simplified approach is to use a limited set of 'basic' perceived emotions, such as happiness, anger, sadness, fear and tenderness, which can be clearly distinguished by the human subjects. Although such an approach oversimplifies the overall concept, it has been shown that it provides meaningful result for perceived (but not induced) emotions (Juslin, 1997). An alternative affective modelling approach is to express emotions in multidimensional spaces as vectors of basic emotional parameters. For example, a previously published work (Russell, 1980) considered the activity-valence space for analytically expressing discrete emotions as points in the two-dimension space.

Provided that musical emotions are modelled based on one of the above theories, an analytic mapping of the modelled emotions to audio/music features must be established. Recent works have investigated the above issue by experimentally derived weighted mappings between emotions and audio features using regressive models (Oliveira and Cardoso, 2008).

In the context of the current work, the aim of the employed model that relates audio signals and raised emotional behaviour is restricted to a single emotion: the human anger. Hence, significant simplifications can be performed, mainly in terms of one-dimensional affective models. As it will be presented in Section 3.3, such a simplified model was developed and employed here for both signalling and estimating the human affective behaviour intensity, i.e.

1 for creating an audiovisual environment that aims to control and elevate the human anger-like behaviour

2 for detecting and measuring the achieved affective elevation through voice recorded signals, respectively.

Focusing on the audio-driven anger control and elevation process, the employed audio signals may not constrainedly be simple music content. Previous examples borrowed from experiences in cinema industry have shown that film sound design is based on distinct audio components layers, namely dialogs, music, sound of human actions, ambient sounds and sound effects (Alves and Roque, 2009). The combination of the above sound design components enables human multisensory and emotional perception in additional interactive application fields, such as video games. Hence, the audio signals employed here for anger control can be the outcome of an advanced audio-events' generation process that aims to synthesise the desired soundscape by taking into account the exact time and spatial position of specific sound samples in the three dimensional space, as well as the concurrent integration of audio and visual content. This renders the parameterised emotional elevation and control a very flexible and efficient process, but on the other hand it represents a very challenging task, since most of the previously published studies have exclusively considered the impact of music signals to emotions, while the relationship of the human audible events in general and the emotions they induced has not been yet exploited (Juslin and Laukka, 2003).

As mentioned previously, the simplified emotion detection model was employed for detecting and measuring the achieved anger-like behaviour from voice recorded signals. Emotion detection from speech represents a research topic targeted to a wide range of applications, such as virtual reality and avatar realistic emotional representation (Feng et al., 2001), as well as emotionally driven speech synthesis (Oudeyer, 2003). Such emotion recognition and speech synthesis techniques typically employ specific statistical moments and features of speech (i.e. pitch or fundamental frequency variation, signal energy distribution in specific frequency bands, the speech rate, etc.) in order to detect the affective mood of the speaker or to provide a realistic framework for affective speech synthesis. Following the simplifications allowed by the requirement of detecting a simple emotion (anger), in this work a simple, optimised algorithm for measuring the anger-like behaviour intensity level in real-time was developed and employed, which combines the estimation of three instantaneous speech feature values (pitch and energy). More details on this algorithm will be provided in Section 3.3.

3 Installation prototype design and development

Figure 1 illustrates the general architecture of the 'Elevator' installation prototype comprised of five discrete submodules:

1 *The content selection module*, which is responsible for the instantaneous choice of the multimedia audiovisual content to be reproduced, according to the installation operational mode. As it will be explained in detail in Section 4, the discrimination of the operational mode was realised for evaluating the performance of the prototype, through two different test cases.

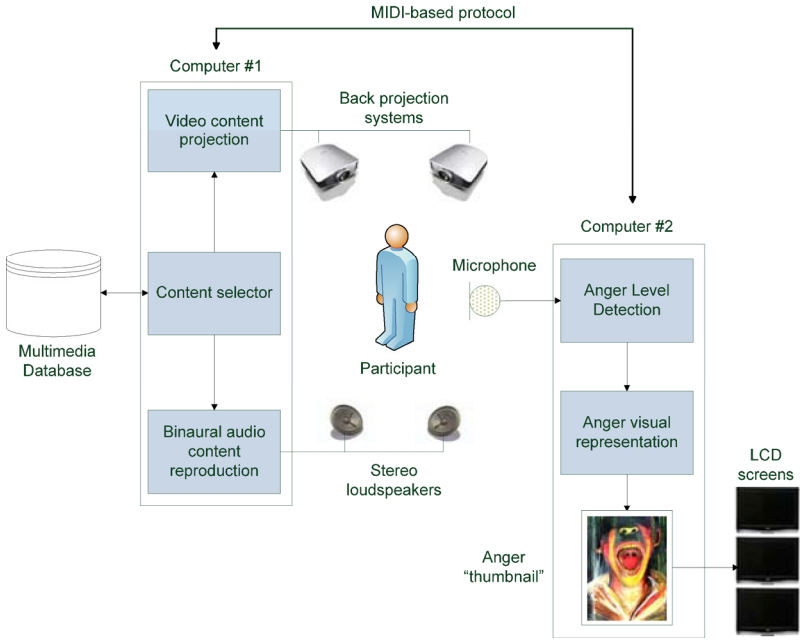
2 *The video content projection subsystem*, which employed two synchronised high-resolution video projectors for back-projecting the visual content onto two panels placed to the left and right sides of the installation.

3 *The audio content playback subsystem*, which was responsible for creating the desired three-dimensional (3D) sound field.

4 *The anger level detection module*, which realises the algorithm for detecting the instantaneous anger-like behaviour intensity through the voice signal recorded during an installation session.

5 *The anger visual representation subsystem*, which is responsible for producing the visual output that corresponds to the achieved affective elevation (i.e. the anger visual thumbnail).

Figure 1 The 'Elevator' installation prototype architecture (see online version for colours)



All the above modules were implemented in real-time software, using the processing open source platform,[1] which represents a powerful software sketchbook and professional production tool used in many fields of audio and image signal processing, science/technology and arts. The selection of processing induced significant reduction on the final real-time implementation complexity and allowed the concurrent processing and control of both audio and visual playback processes. The software developed was installed and executed on two computer systems. The first one hosted the multimedia database and realised the content selection and audiovisual playback module, while the second was responsible for the remaining software functionality.

The necessary communication between these two computer systems was realised using the MIDI serial standard. Typical application areas for MIDI include music, lighting control, show and machine control (audio and video), robotics, etc. For the purposes of the current work, a number of specific MIDI messages were

1 mapped to distinct events (such as the detection of the human presence or absence into the installation space)

2 selected for information exchange (e.g. delivering the intensity values derived by the anger level detection system to the content selector module).

Due to the above simplified communication requirements and the small dynamic range of the transmitted information, MIDI was selected as a low-cost and very-low installation and functional complexity *ad hoc* protocol compared to other alternatives (i.e. Bluetooth or the common Ethernet-based wired or wireless networking protocols), the employment of which would require the development of an appropriate application layer networking module for producing and parsing the necessary signalling/instructions between the two remote computers.

The 'Elevator' interactive prototype environment was designed for and installed in a close room with dimensions $2 \times 3 \times 3$ m (see Figure 2). A cardioid condenser microphone installed at the roof of the room was responsible for recording and providing the audio input and feedback to the anger level detection subsystem. Moreover, three LCD screens were placed in parallel with the front wall, for

1 reproducing a video loop representing the concept of logic, passion and sensation (more details will be provided in Section 3.1)

2 reproducing the anger visual thumbnail.

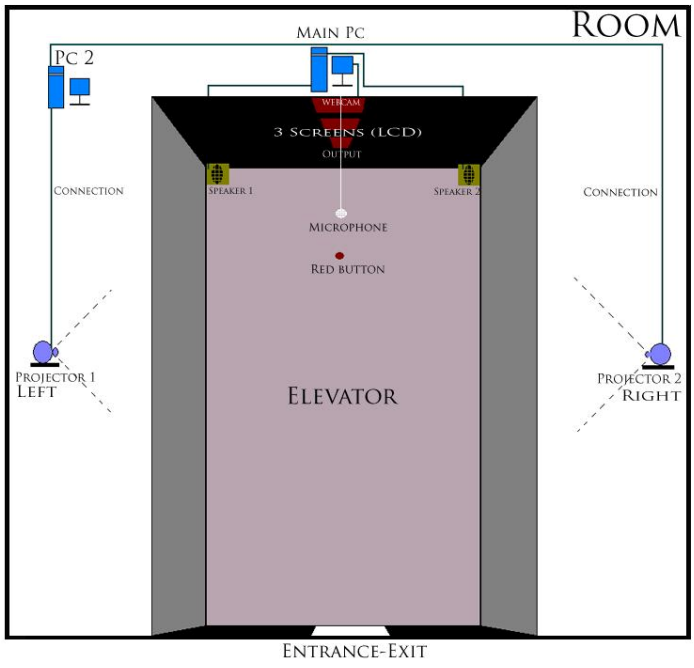
The typical narrative structure of an 'Elevator' session is the following: by the moment the user is entering the installation space, the anger level detection and the content selector subsystems are both initiated. The first one continuously records the voice signals produced by the participant. However, due to the concurrent audio content reproduction, the detected anger level feedback is calculated only during pre-defined silent reproduction intervals. Each 'Elevator' participant is informed about this limitation by reading the general usage guidelines of the installation prior to entering it. It should be also noted here that, since the presence of these silent intervals is critical for interfering with the emotional experience of the participants (i.e. large silence duration may result into a perceivable discontinuity of the emotional triggering conditions), their duration definition was based on the minimum time interval required by the anger level detection algorithm to successfully operate. After a period of simulations employed for calibrating this time interval value, it was found that a period of 3 sec was adequate for producing nearly time-invariant anger level elevation results. The final silence duration was finally

set to 5 sec, considering an additional 2 sec interval for allowing the participant to perceive the presence of the audio reproduction gap.

The content selector module finally selects and retrieves the audiovisual content to be reproduced from a multimedia database. During the first moment of operation for one participant, the content is randomly selected, but after this period of time, the audiovisual content selection is adaptively performed according to the monitored intensity levels: low intensity levels result into more stressful audiovisual content and vice versa.

To realise the above dynamic selection process, the audiovisual content stored in the multimedia database was appropriately tagged with the anger-like behaviour elevation levels it is aimed to achieve. This information was derived by linearly modelling the subjective influence of the visual components and parameters that were used to synthesise the audiovisual content (see Sections 3.1 and 3.2) on the 'target' intensity value. More specifically, in a typical tagging session, each human subject was exposed to a sequence of visual content that was produced by linearly varying specific video (and audio) synthesis components and parameters (as those defined in Section 3.1). The subject then had to define the anger intensity value (in the scale 25%, 70%, 75% and 100%) that he/she believed that the specific content represented. This approach resulted into a linear parametric model that allowed the analytic correlation of the basic audio/visual components and synthesis parameters that were used for the final audiovisual content with the target anger intensity values. This model was in succession used for tagging the complete multimedia database content.

Figure 2 The 'Elevator' installation layout (see online version for colours)



The overall duration of an 'Elevator' session is defined by the exact time-length of the audiovisual content selected, but it does not exceed the limit of 6 min. Moreover, if

1 no human presence is detected for a period of 30 0sec, the session automatically ends, and, in this case, no visual anger representation is produced

2 the user wants to prematurely terminate the current Elevator session, he/she may press the red button illustrated in Figure 2, a task that immediately produces the instantaneous visual anger thumbnail.

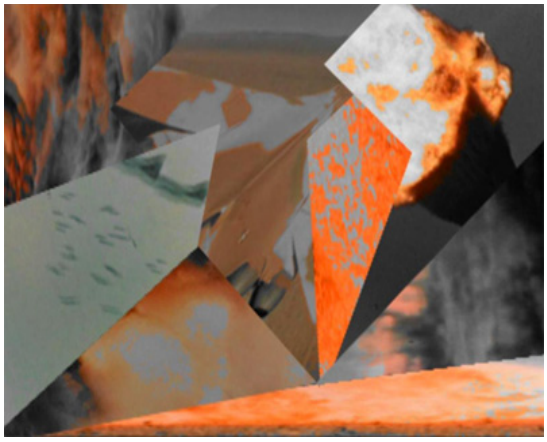
It should be also noted that the participants were informed about the aim of the installation prior to enter the installation environment. For that reason, a computer system with a large monitor was installed in the area next to the installation entrance, providing information about the structure of an interactive session and the process of estimating anger-like behaviour through voice feedback. Additionally, just after an installation session the participant was able to watch an overview of the complete system architecture.

3.1 Visual content and playback

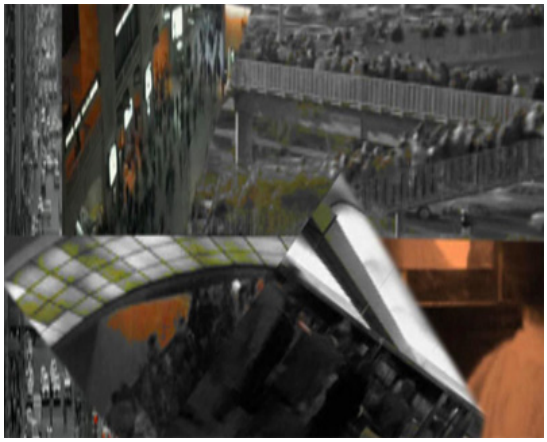
As it is shown in Figure 2, the visual content is back-projected on the left and right sides of the human participant. Video projection is successively realised using the left and right projectors. This renders the observation of the visual content a more active process for the user. Concerning content construction, the final video streams were synthesised by appropriately mixing short and small video parts, forming a video mosaic. These short video components were related to different everyday concepts, such as war, living in alarge city, etc. (see Figure 3). Their exact placement in the final video stream was randomly produced, provided that no-overlapping between them was caused. Moreover, they typically contained geometrical shapes with corners (i.e. triangles) that result into a more aggressive visual effect. The shapes' colours were carefully selected between a number of grey-based combinations (mainly red–grey, yellow–grey and green–grey), depending on the instantaneous palette of the overall video synthesis. The final mosaic mixing parameter values (such as transparency and brightness) were linearly varied in units of 10% increments for each parameter.

The above process resulted into a large number of different video stream versions. The complete library of the video content produced through the above process was finally tagged through the subjective tests mentioned in Section 3 and stored in the multimedia database.

Figure 3 Typical video stills related to (a) war and (b) large city concepts (see online version for colours)



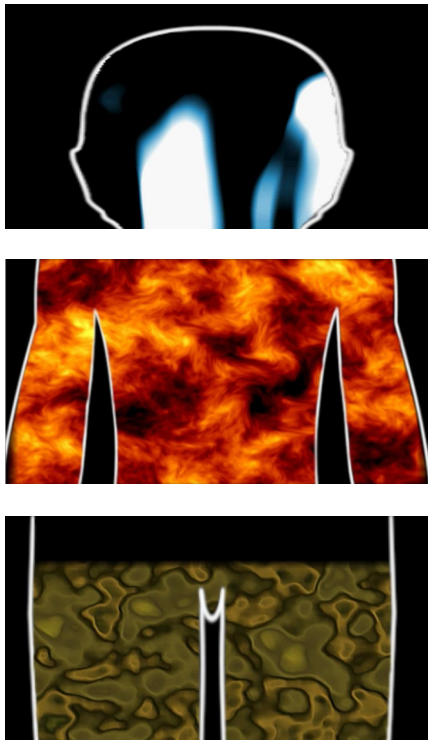
(a)



(b)

As mentioned in Section 3, during an 'Elevator' session, three LCD screens were employed, which continuously played three video streams that represented the concept of logic, sensation and passion (see Figure 4) as three significant and interacting parameters that affect the human life and affective conditions. The reproduction of these video streams was necessary for helping the participant immerse into the installation environment during his entrance, while they also contribute as a conjectural component, aiming to be considered as the mirror-projected image of the participant. The LCD screens were also employed for displaying the anger visual thumbnail produced by the anger visual representation subsystem.

Figure 4 Visual representation of human logic–sensation–passion (see online version for colours)



3.2 Audio content and playback

As mentioned in Section 2, anger-related behaviour signalling was also attempted by reproducing selected audio material. Apart from the audio content itself, this material was organised in the context of pre-defined and complete spatial audio-events in the three-dimensional (3D) installation environment. The selection of this approach for audio data representation was based on the requirement for creating an accurate and authentic immersive sound environment that fully attracts the attention and the senses of the human user.

Although there are a number of different analogue and digital technologies for 3D sound reproduction (Berkhout et al., 1992; Brandenburg and Bosi, 1997; Davis, 1993), the final selection of binaural technology was based on the degree of the sound field representation authenticity achieved under the functional conditions of the installation environment, as well as the loudspeaker installation simplicity. Binaural source localisation (Viste and Evangelista, 2004) represents a highly accurate technique for achieving 3D audio environment recreation by synthesising a two-channel audio signal using the well-known head-related transfer functions (HRTFs) (Moller et al., 1995) between the sound source and each listener's human ear. Hence, a simple set of stereo headphones is required for binaural audio playback. The simple set-up of a binaural reproduction system renders it convenient for a number of state-of-the-art applications, including mobile applications and communications. Alternatively, binaural left and right

signals can be reproduced directly using a pair of conventional loudspeakers. In this case, the additional undesired cross-talk paths that transit the head from each speaker to the opposite ear must be cancelled using cross-talk cancellation techniques (Ward and Elko, 1999). The above binaural synthesis process can be also combined with existing sound field models producing binaural room simulations and modelling. Hence, the above sound field models can output the exact spatial-temporal characteristics of the reflections in a space.

In this work, binaural audio synthesis was performed using a real-time binaural processing engine introduced by Tsakostas and Floros (2007a), that provides a set of significant features (such as optimised HRTF equalisation, efficient room acoustics modelling and cross-talk cancellation for transaural reproduction through stereo loudspeakers) that result into high 3D audio playback quality. This engine was combined with a time-varying binaural convolution/filtering algorithm that takes into accounts both physical and psychoacoustic criteria for efficiently representing moving sound sources (Tsakostas and Floros, 2007b).

Binaural playback was performed using stereo loudspeakers, appropriately placed in space. Cross-talk cancellation was also employed. As explained previously, each audio-event was the combination of an audio track and the corresponding spatial information expressed as the instantaneous Cartesian coordinates. Multiple audio-events were tagged (following the same methodology applied for the visual content) and stored in the multimedia database mentioned previously, and apart from random selection algorithms, a number of pre-determined audio-events selection paths were defined. The original audio material synthesised and used consists of music short tracks and audio samples. Music data were mainly used for creating the ambient sound environment, although in many cases (i.e. during the video reproduction), music is assigned to virtual moving sound sources linearly following the video reproduction switch between the two projection systems. A typical example of music data employed is the initial music track reproduced at the beginning of a session, which is the recording of an old analogue movie playback system. This sound strongly symbolises the beginning of the session, of what the human user is going to experience.

On the other hand, apart from music, real-world and artificially synthesised sounds were also employed. Typical examples include heart-beats, very-low frequency artefacts and urban noises, closely related to the video content reproduced. Various digital audio effects were also applied (such as level control, low or high-pass filtering, echoes, reverberation, delay, etc.). All these sounds, parameters and effects were the fundamental components for algorithmically synthesising audio streams stored in the multimedia database. These audio streams were also tagged with a target anger value, following the tagging process described in Section 3. Additionally, the moving sound source effects were enhanced by a number of signal processing techniques (i.e. linear pitch variation – increment or decrement), which provided a mechanism for fading between opposite moving sound sources.

3.3 Anger level detection

The anger level detection subsystem was responsible for tracking and estimating the intensity of the anger-like behaviour elevation level. In particular, this subsystem was fed in real-time by the human voice recorded by the cardioid microphone, producing an

output intensity level during pre-defined silent audio reproduction intervals, using two features of the voice signal:

1 the variation of the voice signal fundamental frequency (i.e. pitch)

2 the instantaneous values of the recorded voice energy.

The selection of these two voice features was based on previous research works (Razak et al., 2003; Vogt et al., 2008) which have shown that voice fundamental frequency (f_0) can be considered as a very important parameter to identify the emotional state of the human speaker. Additionally, other research efforts (Borchert and Dusterhoft, 2005; Nakatsu et al., 1999) have successfully combined the above voice features in emotion recognition tasks, showing that the combination of increased voice fundamental frequency along with the concurrent increased voice energy can be evaluated as an increase in anger. Recent research results on the field of emotion's modelling and recognition from voice (Scherer, 2003) have also shown that anger is a twofold emotion, i.e. anger can be either hot – rage or cold – controlled. Thus, in order to evaluate anger's level, one should focus on the anger's quality ('what anger') in advance of anger's quantity ('how much anger') calculation. Previously published works have shown that the variation of f_0 and of the acoustic energy in the voice signal can be regarded as an increase in anger's level in both cases of anger (Peter and Herbon, 2006).

For the purpose of this work, the previous approach that combines the voice signal fundamental frequency and acoustic energy variation was used for determining the anger-related elevation level expressed as the percentage of the initial and final anger state of the participant. Four discrete anger elevation differences were defined: 25%, 50%, 75% and 100%. These anger elevation scales were mapped to voice pitch and energy values, using the following procedure: four actors (two male and two female) were invited to perform one word with increasing level of anger defined by the previous scales. This resulted into a speech corpus of 16 audio files which were appropriately processed using time windowing, Fast Fourier and Mel frequency transformations (Korba et al., 2008; Potamianos and Potamianos, 1999), and the average fundamental frequency and voice energy variation was measured for all voice recordings. This allowed the direct map of these variations to the above defined scales of anger elevation.

To measure the achieved anger-like behaviour intensity elevation value, the partial results from the above two criteria (i.e. the variation in f_0 and the change in the acoustic energy of the voice signal) had to be combined. Assuming that $\Delta L_{A,f}$ and $\Delta L_{A,E}$ denotes the (%) anger elevation scale value obtained from the above criteria, respectively, the overall anger elevation value was calculated as the equal-weighted sum of the above values, i.e.

$$\Delta L_A = \frac{\Delta L_{A,f} + \Delta L_{A,E}}{2} \tag{1}$$

3.4 Anger visual representation

The fundamental idea for deriving the anger visual thumbnail is to detect the user face and to algorithmically process it, according to the estimated behavioural elevation value provided by the anger level detection subsystem described previously. The first face detection process was implemented using the open-source Open Computer Vision

(OpenCV) framework,[2] which incorporates a number of real-time image processing libraries for face/gesture recognition, motion tracking, etc., suitable for developing advanced human–computer interfaces and computer vision applications. More specifically, the signal of a video camera oriented towards the user was used for deriving an image containing only the human face. Due to the OpenCV optimised performance, the above detection was found to be very accurate in all test cases considered during the design phase. Moreover, the subareas position of the human face major components (i.e. the eyes, nose and mouths) were also determined in 2D by taking into account their average position measured in a typical set of ten face images, after being appropriately transformed to the captured image size length ratios. This approach was found to exhibit adequate efficiency for the purposes of the present work, while it introduced low computational load.

Following the previous face delimitation process, the anger level visual representation was performed by adding specific visual components (such as clouds, variably sized fonts, screws, acanthuses, etc.), using variable colour masks, layer transparencies, contrast and brightness values. The number of the visual components and the absolute values of the above visual control parameters were linearly mapped to the anger elevation level detected. Figure 5 shows four typical examples of the anger visual representation subsystem output that correspond to the four anger elevation scales considered in this work.

Figure 5 Derived visual thumbnails for anger scales 25%, 50%, 75% and 100%, respectively (see online version for colours)

4 Demonstration and results

The ‘Elevator’ installation prototype was demonstrated during the 3rd annual audiovisual festival organised by the Department of Audiovisual Arts, Ionian University on May 2009 (Figure 6). During this demonstration, a number of subjective tests have been performed for verifying the accuracy of the control, signalling/elevation and visual representation processes. For realising this sequence of tests, a number of participants used and interacted with the Elevator installation prototype. Only one actor was allowed

to enter the installation and interact with it, in order to avoid any impact on the raised emotions imposed by the presence of a group of people. Two cases of tests were considered. In the first case, for each participant, the target anger elevation value was pre-defined (i.e. the selection of the sequence of the audiovisual content reproduced was pre-determined, in order to achieve a specific anger elevation), assuming an initial anger-related affective condition equal to zero (i.e. zero anger intensity value). For that reason, only initial nearly zero anger confirmed participants' cases were further considered for obtaining the following results. The aim of this sequence of tests was to determine the consistency and the statistics of the anger elevation values obtained under specific anger triggering conditions. Indeed, from these tests it was found that the measured anger-like behaviour was always in the range of the targeted anger scale (25%, 50%, 75% and 100%). Table 1 also summarises the statistics of the collected results in terms of the measured mean and standard deviation values. These results also validate the accuracy of the affective elevation process.

The second test case represents the normal 'Elevator' operational mode: the reproduced audiovisual content was randomly selected from the multimedia database, taking into account the initially recognised anger-like state of the participant and a (randomly selected) target anger intensity value which was always greater than the initial one. These tests allowed to obtain an overview of the overall impression and experience of the participants. After each 'Elevator' session, the participants were asked to fill in a questionnaire. By analysing and aggregating the results obtained by these questionnaires, the following conclusions were drawn:

1 Nearly 71% of the total number of the participants felt anger. Moreover, 19% of them described that they felt stressed, but they could not clearly define that they felt anger.

2 About 10% of the participants were not able to accomplish a complete 'Elevator' session, as they were over-stressed by the reproduced audiovisual content and the overall dynamic installation environment.

3 During the first demonstration day, 30% of the participants (in the majority students) answered that they were mainly focusing on the affective interaction path and the way the visual elevated anger thumbnail was produced. Hence, no significant attention was paid on the concept of 'Elevator' itself. However, during the second and third day, the majority (90%) of the same participants decided to interact again with 'Elevator', focusing this time on the audiovisual content and allowing the spontaneous participation and reaction.

An additional aim of the second test case was to verify the accuracy of the final visual anger representation. More specifically, the visual emotional thumbnails produced by the Elevator installation in all test cases described previously were given to all participated human subjects for grading the anger-like behaviour intensity they believed that each particular thumbnail represented. After summarising the results obtained through these tests, it was found that nearly the 90% of the human subjects had graded the emotional thumbnails in fine-accordance with the derived anger elevation level.

Figure 6 'Elevator' demonstration snapshot (see online version for colours)

Table 1 Measured anger elevation mean and standard deviation values

Targeted anger scale	*0–25%*	*25–50%*	*50–75%*	*75–100%*
Measured mean	17.9	42.5	68.0	86.0
SD	4.3	5.6	7.1	6.2

5 Conclusions

In this work, an interactive audiovisual installation prototype called 'Elevator' is presented which aims to signal, express and monitor and finally measure human behaviour related to emotions (focusing particularly on anger, a very common everyday life emotion). The signalling and control of anger-like behaviour is based on a novel approach which employs the reproduction of appropriate visual and audio content, forming a highly immersive and authentic audiovisual environment. The above elevation process is adaptive, with the appropriate feedback provided by the anger level detection subsystem which employs signal processing techniques in the frequency domain for deriving the instantaneous anger-related behaviour intensity values.

It should be noted here that the above signal processing techniques were successfully used in the past by a number of researchers for emotional (anger) detection. However, in this work, provided the constraints imposed by the realisation methodology followed (i.e. the fact that the participants were pre-informed about the aim of the installation, rendering the subjective outcomes related to emotional detection biased), these techniques were targeted to provide a display of anger-like behaviour and not of anger itself. The achieved measured levels of anger-like behaviour are finally mapped to adaptive visual content that represent the thumbnail of the emotional state of each participant. After a sequence of subjective tests, it was shown that the accuracy of the

achieved anger-like behaviour elevation value and the corresponding emotional thumbnail mapping is significantly high. This conclusion also verifies the efficiency of the overall recognition and signalling process.

Taking into account the above results, the demonstration of the 'Elevator' installation prototype has shown that the development of a new generation of multimedia interactive environments that employs novel and innovative interaction concepts (such as affective feedback) is feasible. It is the authors' near future intention to further exploit the derived results and conclusions of this work for developing an adaptive version of interaction that takes into account wide-spread and commonly met audio and visual-based emotion-triggering conditions (i.e. visual content and sounds in noisy urban environments), under strict experimental conditions that will additionally allow the recognition and detection of the human anger.

References

Alves, V. and Roque, L. (2009) 'A proposal of soundscape design guidelines for user experience enrichment', *Proceedings of the AudioMostly 2009 Conference on Interaction with Sound*, Glasgow, UK.

Baraldi, F.B., Poli, G.D. and Roda, A. (2006) 'Communicating expressive intentions with a single piano note', *Journal of New Music Research*, Vol. 35, No. 3, pp.197–210.

Berkhout, A.J., Vogel, P. and Vries, D. (1992) 'Use of wave field synthesis for natural reinforced sound', *Proceedings of the Audio Engineering Society 92nd Convention*, preprint 3299.

Birchfield, D., Lorig, D. and Phillips, K. (2005) 'Network Dynamics in Sustainable: a robotic sound installation', *Organised Sound*, Vol. 10, pp.267–274.

Birchfield, D., Phillips, K., Kidané, A. and Lorig, D. (2006) 'Interactive Public Sound Art: a case study', *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06)*, Paris, France.

Borchert, M. and Dusterhoft, A. (2005) 'Emotions in speech – experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments', *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp.147–151.

Boxer, S. (2005) 'Art that puts you in the picture, like it or not', *New York Times*.

Brandenburg, K. and Bosi, M. (1997) 'ISO/IEC MPEG-2 advanced audio coding: overview and applications', *Proceedings of the Audio Engineering Society 103rd Convention*, New York, preprint 4641.

Davis, M. (1993) 'The AC-3 multichannel coder', *Proceedings of the Audio Engineering Society 95th Convention*, New York, preprint 3774.

Feng, Y., Chang, E., Xu, Y. and Shum, H.Y. (2001) 'Emotion detection from speech to enrich multimedia content', *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, pp.550–557.

Friberg, A. (2008) 'Digital audio emotions: an overview of computer analysis and synthesis of emotional expression in music', *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland.

Gabrielsson, A. and Lindström, E. (2001) 'The influence of musical structure on emotional expression', in P.N. Juslin and J.A. Sloboda (Eds.), *Music and Emotion: Theory and Research*. New York: Oxford University Press, pp.223–248.

Gundlach, R. (1935) 'Factors determining the characterization of musical phrases', *American Journal of Psychology*, Vol. 47, No. 4, pp.624–643.

Hevner, K. (1936) 'Experimental studies of the elements of expression in music', *American Journal of Psychology*, Vol. 48, pp.246–286.

Housain, G., Thompson, W.F. and Schellenberg E.G. (2002) 'Effects of musical tempo and mode on arousal, mood, and spatial abilities', *Music Perception*, Vol. 20, No. 2, pp.151–171.

Juslin, P.N. (1997) 'Perceived emotional expression in synthesized performances of a short melody: capturing the listener's judgment policy', *Musicae Scientiae*, Vol. 1, No. 2, pp.225–256.

Juslin, P.N. and Laukka, J. (2003) 'Communication of emotions in vocal expression and music performance: different channels, same code?' *Psychological Bulletin*, Vol. 129, No. 5, pp.770–814.

Kienast, M. and Sendlmeier, W.F. (2000) 'Acoustical analysis of spectral and temporal changes in emotional speech', *Proceedings of the ISCA ITRW on Speech and Emotion*, Newcastle, pp.92–97.

Korba, C.A.M., Messadeg, D., Djemili, R. and Bourouba, H. (2008) 'Robust speech recognition using perceptual wavelet denoising and Mel-frequency product spectrum cepstral coefficients features', *Informatica*, Vol. 32, No. 3, pp.283–288.

Moller, H., Sorensen, M., Hammershoi, D. and Jensen, C. (1995) 'Head-related transfer functions of human subjects', *Journal of the Audio Engineering Society*, Vol. 43, No. 5, pp.300–321.

Nakatsu, R., Nicholson, J. and Tosa, N. (1999) 'Emotion recognition and its application to computer agents with spontaneous interactive capabilities', *Proceedings of the 3rd Conference on Creativity and Cognition*, Loughborough, pp.135–143.

Oliveira, A.P. and Cardoso, A. (2008) 'Emotionally-controlled music synthesis', *Proceedings of the 10th Regional Conference of AES Portugal*, Lisboa.

Oudeyer, P.Y. (2003) 'The production and recognition of emotions in speech: features and algorithms', *Int. J. Human–Computer Studies*, Vol. 59, pp.157–183.

Peter, C. and Herbon, A. (2006) 'Emotion representation and physiology assignments in digital systems', *Interacting with Computers*, Vol. 18, No. 2, pp.139–170.

Potamianos, G. and Potamianos, A. (1999) 'Speaker adaptation for audio-visual speech recognition', *Proceedings of Eurospeech*, Vol. 3, pp.1291–1294.

Razak, A., Yusof, M.H. and Komiya, R. (2003) 'Towards automatic recognition of emotion in speech', *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology*, pp.548–551.

Russell, J. (1980) 'A circumplex model of affect', *Journal of Personality and Social Psychology*, Vol. 39, pp.1161–1178.

Scherer, K.R. (2004) 'Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them?' *Journal of New Music Research*, Vol. 33, No. 3, pp.239–251.

Tsakostas, C. and Floros, A. (2007a) 'Optimized binaural modelling for immersive audio applications', *Proceedings of the Audio Engineering Society 122th Convention*, Vienna, preprint 7100.

Tsakostas, C. and Floros, A. (2007b) 'Real-time spatial representation of moving sound sources', *Proceedings of the Audio Engineering Society 123rd Convention*, New York, preprint 7279.

Viste, H. and Evangelista, G. (2004) 'Binaural source localization', *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04)*, pp.145–150.

Vogt, T., Andre, E. and Nikolaus, B. (2008) 'EmoVoice – A framework for online recognition of emotions from voice', *Proceedings of 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Kloster Irsee, Germany.

Wallis, I., Ingalls, T. and Campana, E. (2008) 'Computer generating emotional music: the design of an affective music algorithm', *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland.

Ward, A.B. and Elko, G.W. (1999) 'Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation', *IEEE Signal Processing Letters*, Vol. 6, No. 5, pp.106–108.

Notes

[1] www.processing.org.

[2] http://sourceforge.net/projects/opencvlibrary/.