



SALERO

Initial Construction of a Limited Emotional Speech Corpus for Analysis

SALERO Deliverable 6.2.1



Initial Construction of a Limited Emotional Speech Corpus for Analysis

SALERO Deliverable D6.2.1

SALERO identifier: SALERO-D6.2.1-DIT-v2.doc

Deliverable number: D6.2.1

Author(s) and company: C. Cullen (DIT)

Work package / task: WP06

Document status: Final

Confidentiality: Public

DOCUMENT HISTORY

Version	Date	Reason of change
1	2008-01-09	document created
2	2008-01-30	Final version to reflect review comments

The work presented in this document was partially supported by the European Community under the Information Society Technologies (IST) priority of the 6th framework programme for R&D.

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

This document contains material, which is the copyright of certain SALERO consortium parties, and may not be reproduced or copied without permission. All SALERO consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the SALERO consortium as a whole, nor a certain party of the SALERO consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

Table of Contents

1	Executive Summary	1
2	Introduction	2
2.1	Purpose of this Document	2
2.2	Scope of this Document	2
2.3	Status of this Document	2
2.4	Related Documents	2
3	Existing Emotional Speech Corpora	3
3.1	Simulated Assets	3
3.1.1	<i>Recording Quality</i>	4
3.2	Broadcast Assets	4
3.2.1	<i>Recording Quality</i>	4
3.3	Natural Assets & MIPs	4
3.3.1	<i>MIP 1</i>	5
3.3.2	<i>MIP 2</i>	5
3.3.3	<i>MIP 3</i>	5
3.3.4	<i>MIP 4</i>	5
3.3.5	<i>MIP 5</i>	5
3.3.6	<i>Rationale for Use of MIP 4</i>	5
3.3.7	<i>Recording Quality</i>	6
4	Experimental Procedures for Emotional Speech Assets	7
4.1	Recording Setup	7
4.2	Experimental Design	8
4.2.1	<i>Experiment 1: Tetris</i>	8
4.2.2	<i>Experiment 2: Console Game Playing</i>	8
4.2.3	<i>Experiment 3: Lego</i>	9
4.3	Case Study	10
4.4	Conclusions	10
5	Corpus Metadata Specification	11
5.1	Metadata Structure	11
5.1.1	<i>Project</i>	11
5.1.2	<i>Session</i>	12
5.1.3	<i>Actor</i>	13
5.1.4	<i>Content</i>	13
5.1.5	<i>Asset</i>	14
6	Implementation of an Emotional Speech Corpus	16
6.1	Requirements	16

6.1.1	<i>Automated Annotation</i>	16
6.2	Architecture	16
6.2.1	<i>Presentation</i>	16
6.2.2	<i>Application</i>	16
6.2.3	<i>Data</i>	16
6.3	Technologies	16
6.3.1	<i>Presentation and Application Tiers</i>	17
6.3.2	<i>Data Tier</i>	17
7	Future Work	18
7.1	Emotional Asset Rating.....	18
7.2	Corpus Workflow	18
7.2.1	<i>Lip-Synching and Character Animation</i>	18
7.2.2	<i>Providing Rules for Emotional Speech Synthesis</i>	18
7.2.3	<i>Defining Metadata for Experimental Production</i>	18
7.3	Corpus Visualisation	19
7.4	Corpora Expansion	19
8	Conclusions	20
9	Appendix 1: Example Corpus Asset Listing	21
10	References	27
11	Glossary	30

1 Executive Summary

This document accompanies the actual SALERO deliverable D6.2.1, a limited Emotional Speech Corpus for analysis. As a supporting document, it describes the design of the corpus and its associated implementation. The document defines:

- The rationale behind the use of Mood Induction Procedures (MIPs) to obtain natural emotional speech assets for analysis.
- The recording conditions used for all experiments.
- The specific Mood Induction Experiments that were used to obtain the assets in the corpus.
- The metadata used to annotate the corpus
- The analysis data produced in accordance with the logical rhythmic tagging framework
- The actual construction of the corpus itself

In conclusion, the document denotes some potential directions for development of the corpus. These directions are a direct result of the work performed during the D6.2.1 deliverable.

2 Introduction

2.1 Purpose of this Document

The purpose of this document is to describe the SALERO deliverable D6.2.1, a limited Emotional Speech Corpus for analysis. Although the actual deliverable is the corpus itself, this document endeavours to explain the method behind the design and implementation of that corpus. An outline of further work is also provided, to indicate how the work carried out in this deliverable may best be continued within the SALERO project.

2.2 Scope of this Document

As an accompaniment document to the deliverable, the scope is limited to description of the methods used. The document will also serve as a reference for future development of this and indeed other corpora, in an effort to produce a consistent method for the archive, retrieval and analysis of speech assets.

2.3 Status of this Document

The status is final.

2.4 Related Documents

Before reading this document it is recommended to be familiar with the following documents:

- D6.1.1
- D6.1.2
- D9.3.2

3 Existing Emotional Speech Corpora

This deliverable relates to the creation of an emotional speech corpus that can subsequently be analysed to determine a possible rule set for the detection of emotional dimensions in speech. In order to construct such a corpus it is important to examine the methods used to create existing speech corpora, particularly in terms of sound quality and the source of the corpus assets. There are three main forms of asset used in existing speech corpora: simulated assets, broadcast assets and induced assets. A few examples of claimed 'natural' emotional speech databases exist [1-3], although the justification for such content is only that it is more natural when compared with simulated content. In the majority of cases, what are termed 'natural' emotional assets are obtained from a broadcast source (mainly television) [3], which calls into question the naturalness of the emotions being expressed.

A high quality natural emotional speech corpus is dependent on obtaining authentic, natural, high quality speech assets. Therefore it is important to investigate the nature of simulated, broadcast, and induced assets in relation to the acquisition of true natural emotional assets. The use of simulated and broadcast assets will be first considered, followed by analysis of the use of Mood Induction Experiments (MIPs) in obtaining induced emotional assets, arguing that this is the ideal method for obtaining natural emotional responses from participants. The use of MIPs also allows a high level of audio quality to be achieved. MIPs take place in a laboratory setting and thus the researcher can ensure that audio is recorded at a high level, ideally at a professional standard of 192 Khz/24bit [4].

3.1 Simulated Assets

Corpora consisting of simulated assets use acted emotional states, read texts and imagined/recalled emotional situations [5-9]. However very little is actually known about how simulated emotion compares to natural emotion [3]. Simulated emotion that involves reading from a text is not a spontaneous expression of emotion with read speech having distinct characteristics from spontaneous speech [10]. Emotional states can be considered to be an important factor in maintaining and negotiating social interaction and relationships; simulated assets are often non-interactive [5-9], consisting of monologues with little or no interaction from other agents. The neglect of the social dimension of emotional speech means that obtained assets may contain only a limited range of emotions. Furthermore, the participants used in simulating emotional assets may have subjectively different interpretations of the emotions that they are required to simulate. Since no objective emotional scale exists, it is difficult to compare two subjective interpretations of an emotional state.

It is arguable that simulated emotion lacks the necessary corresponding physiological states normally associated with real emotions. The underlying physiological state of a subject may also be manifested in spoken communication; as Johnstone [11] argues, emotion can induce changes in speech that the speaker cannot control, with these changes possibly reflecting the underlying physiological changes taking place in the speaker. Actors are able to achieve the change in speech by voluntarily altering it:

"Thus acoustic analyses of actor portrayed speech might not provide an accurate description of spontaneous affective speech modulation, which is likely to differ both qualitatively and quantitatively,...." [11]

It has also been argued that the authenticity of an emotion is dependent on its source and that a simulated emotion:

"can resemble the occurrent [sic] state of a given emotion and yet not be a proper instance of that or, strictly speaking, of any emotion" [12].

It can be argued that the source of acted emotion is therefore not the same as the source of spontaneous or natural (non-simulated) emotion. Similarly the use of emotional recall [9] does not generate natural emotional responses. Recall of an emotional event is not the same as experiencing that event: the recall of a fearful situation is not the same as actually being in danger, the cause of the fear. This distinction is physiologically significant [3, 13] as the context within which the emotion takes place is important: the physiological response during recalled instances of fear and actual instances of fear are not the same [14]. The voluntary nature of simulated emotion and the context within it takes place undermines its authenticity and its suitability as a method of obtaining natural emotional speech assets.

3.1.1 Recording Quality

The greatest single advantage of simulated assets is the potential for control of the recording environment, such that most simulated assets are obtained using studio equipment and conditions. The huge variation of recording quality found in other types of corpus (such as broadcast assets) precludes the definition of cohesive standards, and thus simulated assets are often preferred for this reason.

3.2 Broadcast Assets

Some corpora use assets obtained from broadcast sources, mainly television [1-3], the justification being that they are 'natural' compared to simulated assets [3]. The use of broadcast assets has focused mainly on programmes such as chat shows, interviews and documentaries, these programmes frequently use emotional recall, as is the case with the Reading-Leeds database [15, 16] and the Belfast Naturalistic Database [17]. Using broadcast assets often means that the researcher has to make subjective judgements about the emotional states on display in selecting assets for the corpus. The researcher's judgement of the emotions on display will not necessarily be the same as that of another researcher or even the person originally displaying the emotion. It can be argued that any broadcast is a performance, as the speakers are usually very aware of the recording process taking place. It is recognised in anthropological research that the presence of a camera distorts the reality of the field work situation:

"In using video technology the focal point shifts to the presence of equipment and its effects on rapport. Quiet obviously the presence of video equipment is a major source of field contamination" [18]

The presence of a researcher and equipment may also cause people to act differently or even feel constrained in what can be said and done [18, 19]. It is possible that this distortion and constraint means that televised emotional displays, like simulated emotion, may only be a facsimile of real emotion. The only way to prevent this distortion is to conceal the equipment and covertly record subjects; however this is a highly questionable practice and ethically unsound (Gottdiener 1979). The distorting effect may lessen over time as subjects become used to being recorded [20]. This would suggest that it would be more relevant to use clips taken from the middle or towards the end of a televised program as opposed to clips taken from the start. However, there is an inherent perceptual bias to the recording process [21] and there is no way of knowing if a program was edited in chronological order, or if the footage was taken out of context in order to fulfil a certain editorial agenda. This perceptual bias is inherent in the subjective decisions of the cameraman, the director, the producers and the editor, and it cannot be known how this affects the final outcome of a broadcast piece. While live broadcasts may have less of a perceptual bias, there could still be a bias in the editorial goals of the presenter and the production staff. Even before the researcher has made a subjective decision in selecting clips, a large amount of subjective decisions have already been made in making and broadcasting the program.

3.2.1 Recording Quality

Assets taken from broadcast sources can be of varying audio quality, as 'broadcast quality' is a term rather than a definition; one can not assume that assets obtained from broadcast sources are of uniform quality. Audio quality will also vary depending on the nature of the program, whether it is recorded in a studio or outside in public spaces (as many reality television programs are). Various other factors will affect the audio quality: noise from studio audiences, people talking across each other and environmental noise from outside broadcasts. The equipment used will also affect the sound quality: different broadcast situations may use different recording apparatus (microphones, camera's etc) and methods.

3.3 Natural Assets & MIPs

In order for assets to be considered natural for the purposes of analysis, they should ideally be derived from non-simulated and non-broadcast sources, with audio quality being of paramount importance. The induction of natural emotional responses in a laboratory environment, thus ensuring audio quality can be maintained, is achieved through the use of MIPs. MIPs are procedures that are designed to induce specific emotional states in a test subject within a controlled situation. The emotional states are

temporary and relatively specific as MIPs are designed to induce as single and pure an emotion as possible. Gerrards-Hesse et al [22] carried out an extensive review of MIPs and details five different MIP groups used to elicit real emotion from participants:

3.3.1 MIP 1

Emotion is freely generated using mental techniques such as hypnosis [23] or imagination [24]. With hypnosis the subject enters a trance and is asked to imagine a situation in the past where they experienced a certain emotion. Similarly, using imagination, subjects are encouraged to re-imagine an event or situation in order to invoke within them the intended emotional state.

3.3.2 MIP 2

This MIP relies on external material to induce emotional states, with the subjects also being instructed to get into the mood suggested by the material. The three MIPs in this group are the Velten MIP [25] (using written statements), the Film MIP [26] and the Music MIP [27] where subjects are encouraged, by any means necessary, to get into the mood conveyed by the film or piece of music.

3.3.3 MIP 3

This group is similar to MIP group two, as the Film and Music MIPs also belong to this group. The main difference is that subjects are presented with emotional stimuli, but are not instructed to get into the mood suggested by the material, as it is assumed that the material alone will induce an emotional state. The Gift MIP [28] also belongs in this category and assumes that people will be happy or excited to receive a gift, especially if it is unexpected.

3.3.4 MIP 4

MIPs in this group use the fact that some situations can create emotional states, usually through the satisfaction or frustration of a subjects needs. By placing subjects in a situation where certain needs are activated, such as the need to succeed at a certain task, emotional states can be induced by frustrating or aiding the subject in the attainment of their need. The Success/Failure MIP [29] uses false feedback (positive or negative) concerning a subject's performance in a test that they believe is testing their cognitive ability. The Social Interaction MIP is also in this category and exposes subjects to arranged social situations designed to elicit emotional responses; it is assumed that the behaviours of others will affect the emotional state of participants.

3.3.5 MIP 5

In this group it is presupposed that emotions arise out of an unspecified physiological state, and that a cognitive appraisal of a situation determines the quality of the emotion (see D6.1.1). Therefore by inducing certain physiological states emotions can be elicited from subjects. The Drug MIP [30] uses drugs to induce certain physiological states, while the Facial Expression MIP [31] requires subjects to either smile or frown in order to induce a positive or negative emotional state.

3.3.6 Rationale for Use of MIP 4

Numerous researchers have used MIPs to induce emotional states in participants [32-37]. While emotional induction may be ethically dubious in some cases ([38]), MIPs offer the best solution to obtaining natural emotional assets in a controlled audio environment, specifically the Success/Failure (MIP group 4). The Hypnosis, Imaginary, Velten, Film and Music MIPs have the same problems of authenticity as that of simulated emotion (3.1). Furthermore it can be argued that films and music are subjective artistic fields, while one participant might find a certain film or piece of music to be of a certain emotional quality, another participant might not. Each persons experience of either a film or piece of music is different, resulting from a myriad of factors such as cultural background, socioeconomic situation and personal experiences [39]. The Gift MIP may induce a natural emotional response, but the emotional range is limited to elation [22] while the Drug MIP is ethically dubious and is an impractical method for obtaining assets. The Facial MIP again has problems of authenticity as facial expression is only one aspect of emotional communication: it is not a given that a smile will induce the relevant related emotional state. [31]

While there are varying types of MIP, the Success/Failure MIP would appear to offer distinct advantages. The Success/Failure MIP avoids the problem of demand effects [40] and the numerous issues regarding authenticity relative to other MIPs. The true nature of the experiment is not evident and can be further disguised if needed: participants are engaged in a task and can be led to believe that the completion of the task is the purpose of the experiment. The use of false feedback, either positive or negative, further conceals the true purpose of the experiment. This type of feedback provides the potential means of stimulation of positive and negative emotional states among participants (see D6.1.1). The use of a task based Success/Failure MIP removes the subjective nature associated with some other MIPs, and allows the researcher to control and manipulate the experiment in greater detail. By frustrating or aiding the subjects in their task, without their knowledge, they can be guided towards natural negative or positive emotional states (see D6.1.1 and D6.1.2) without being aware that a certain emotional state is required, thus avoiding the creation of demand effects.

Of particular note is Roland Kehrein [35] who carried out experiments based on the Success/Failure MIP. Kehrein used sound proof rooms and a co-operative task, in this case a Lego construction, with one party giving instructions to what was to be built and the other party following the instructions. By manipulating the Lego available and the time allowed the participants could easily be hindered or aided in the attainment of their goal.

3.3.7 Recording Quality

The use of Mood Induction Procedures to stimulate emotion has the potential for the same recording conditions to be applied as with simulated assets. The difficulties associated with such conditions using MIPs are related to the concealment of recording equipment to avoid revealing the true purpose of the experiment prior to commencement. In Kehrein's experiment, the fact that the participants were seated in separate sound proofed rooms, allowed the conversational interaction to be recorded as two separate high quality audio channels. This allowed both sides of the conversation to be analysed, including overlaps. Similarly, Johnstone [37] used computer games in order to induce real emotional states in test subjects. Johnstone found that computer games were well suited for this purpose as they can be changed and manipulated in order to induce the desired emotions.

A combination of the two experimental designs offers advantages: using computer games as part of a cooperative, task-based MIP offers a high degree of control, either hindering or aiding participants, while the use of separate sound proofed rooms enables high quality audio assets to be obtained. This approach ensures that obtained assets are natural, compared to simulated and broadcast assets, with emotional responses been induced as a result of while the co-operative aspect ensures the social aspect of emotional expression is not neglected. The resulting emotional assets can be claimed to be natural and spontaneous, arising out of the manipulation of the task and the interaction of the participants as opposed to voluntary or knowingly coerced attempts to generate emotional states.

4 Experimental Procedures for Emotional Speech Assets

Three experimental procedures were carried out to obtain the emotional assets used in the corpus detailed in this deliverable. The three procedures are all based on the Success/Failure MIP (3.3.4) and use isolation booths and high quality audio equipment to ensure that natural emotion audio assets are recorded at a very high quality. The equipment used will first be discussed, followed by detailed explanation of three experimental procedures designed to elicit emotional responses from participants. This is followed by a brief discussion of a small case study that was carried out to test the experiments.

4.1 Recording Setup

Researchers have had success using isolation booths in audio-based experiments [35, 41, 42]. For this experiment, two isolation booths are used to ensure that the participants are not distracted and that the recorded audio is free from external and unwanted noise: thus a clean audio signal is maintained and the dialogue between participants can be recorded as two separate audio streams [43]. The basic set-up for the experimental procedures consists of two computers, one running any software as needed (mainly Tetris) and using two flat screen monitors, one in each booth. The second computer is used to record the audio and is connected to a ProTools HD system [44] audio capture device. The Mbox is connected to two Neuman U87 Microphones, one in each booth, situated on two microphone stands with a pop-shield in front of each to minimise plosive utterances. The audio from the microphones is recorded using Pro-Tools audio recording and editing software [44]. The audio is also routed to an external mixer allowing the participants to hear each other via headphones, as well as allowing the external researcher to monitor the audio and to communicate with either one or both of the participants. The audio computer also uses two flat screen monitors with VNC [45] software being used to remotely view the images on the monitors in the booths. This allows the researcher, external to the booths, to see what the participants see and thus manipulate the situation accordingly. This basic set-up is supplemented, when needed, with external games consoles and control pads in each booth (Figure 3).

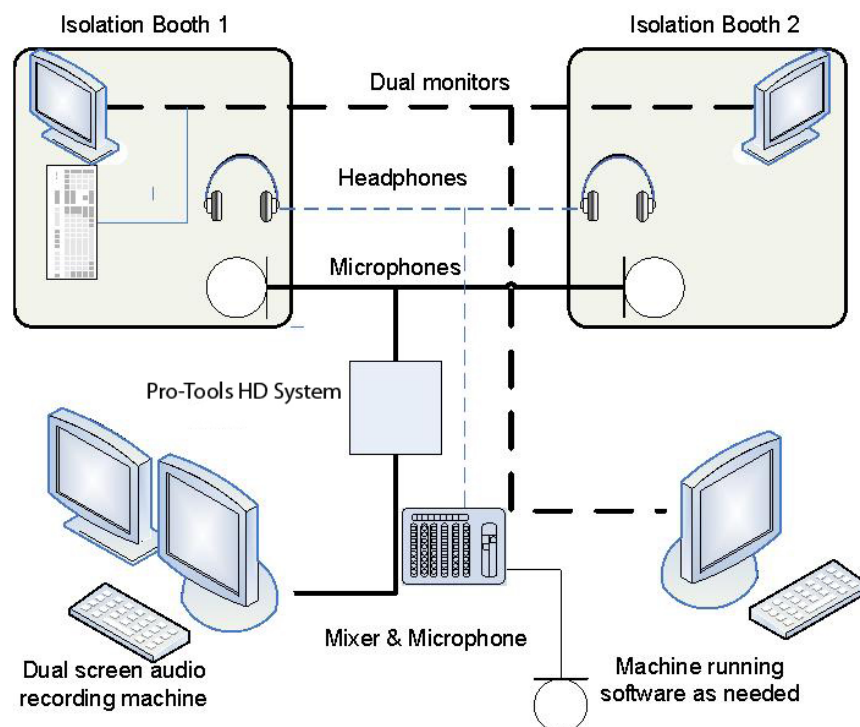


Figure 1: Basic equipment setup. This setup will be considerably augmented within the new Multimodal Interaction Laboratory (MIL) currently being constructed within the DMC

4.2 Experimental Design

In order to elicit emotional responses from participants, three experiments have been designed. Both the Tetris (4.2.1) and the Lego (4.2.3) experiments consist of two participants engaging in a cooperative-based task while the experiment is monitored and recorded externally. The video game experiment consists of the participants competing against each other in a selection of contemporary computer games (4.2.2). All three of the experimental designs are based on the Success/Failure MIP (3.3.4) and Kehreins original experimental design and procedure [35]. The experiments are designed, aside from the Lego task, so that the researcher, external to the isolation booths, will be able to alter various elements of the experiment in order to engender emotional states within the participants. The type and extent of the manipulation depends on which experimental design is being used at the time and varies with circumstance. A small reward will be offered for the successful completion of each cooperative task, providing an added impetus for completing the task within the set boundaries. This reward is offered for successful completion of the set tasks, either beating a certain high-score or completing a task in a certain amount of time; but manipulation of the experiment can ensure that these goals are rarely attained. However the use of a cash reward is a successful method in engendering positive emotional dimensions (elation) in participants (3.3.3) [22], it is a variation of the Gift MIP, and can be used in conjunction with the Success/Failure MIP. Participants taking part in the experiments are not allowed to bring in any devices that display time, thus allowing false information regarding how much time is left or has elapsed to be given.

4.2.1 Experiment 1: Tetris

This experiment uses the hugely popular Tetris puzzle game¹. Tetris consists of seven different shaped blocks that have to be stacked and slotted together, like a vertical jigsaw, in order to complete a horizontal row of blocks; up to four complete rows are possible in any one instance. Each of the seven blocks can be rotated clockwise and counter-clockwise and moved left to right across the screen. Once a player is happy with the position of a block it can be moved straight down into position, otherwise it moves slowly downward until it comes to rest on another block or the bottom of the screen.

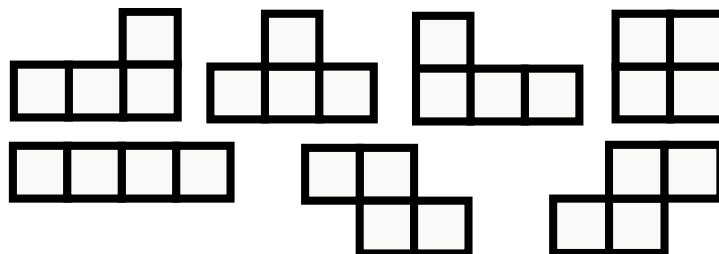


Figure 2: The seven Tetris shapes

For the purposes of this experiment, participant B will manipulate the blocks according to the instructions given by participant A. Participant A will instruct participant B on how to rotate, move and stack the blocks with both working together in order to achieve an agreed score within a certain time frame. Participant B will not be able to see how the blocks are stacking up: only participant A will be able to see the game and the actions carried out by participant B. The co-operative process can be hindered by preventing participant B from hearing the instructions properly (through cutting the audio feed to the booth), by altering the time allowed to achieve the desired goal, or by remotely controlling the falling pieces without either participant knowing. The participants can be aided in the task by giving them more time to complete it, by giving false information regarding how much time is left, or lowering the score that needs to be attained.

4.2.2 Experiment 2: Console Game Playing

This experiment uses modern games consoles and games. The main advent of these console systems is that they have been designed to facilitate multiplayer games and the large amount of games available for it are usually designed with extensive multiplayer options, cooperative and/or competitive in nature.

¹ <http://www.tetris.com/> Home page of Tetris. Details about its history and popularity

Participants taking part in this experiment will mainly have been chosen because of their familiarity with and regular participation in console games. The case study (4.3) indicated that certain types of games were very popular among some of the potential participants and that game play can become highly competitive among competing groups of friends. The case study also indicated that little or no external manipulation was necessary as the game used was designed to be competitive and challenging: the overall game play and style of game is conducive to inducing emotional states in participants. External manipulation is achieved through unplugging a participant's game controller, changing the time limit, giving false information regarding the amount of time left or through offering a cash or material reward.

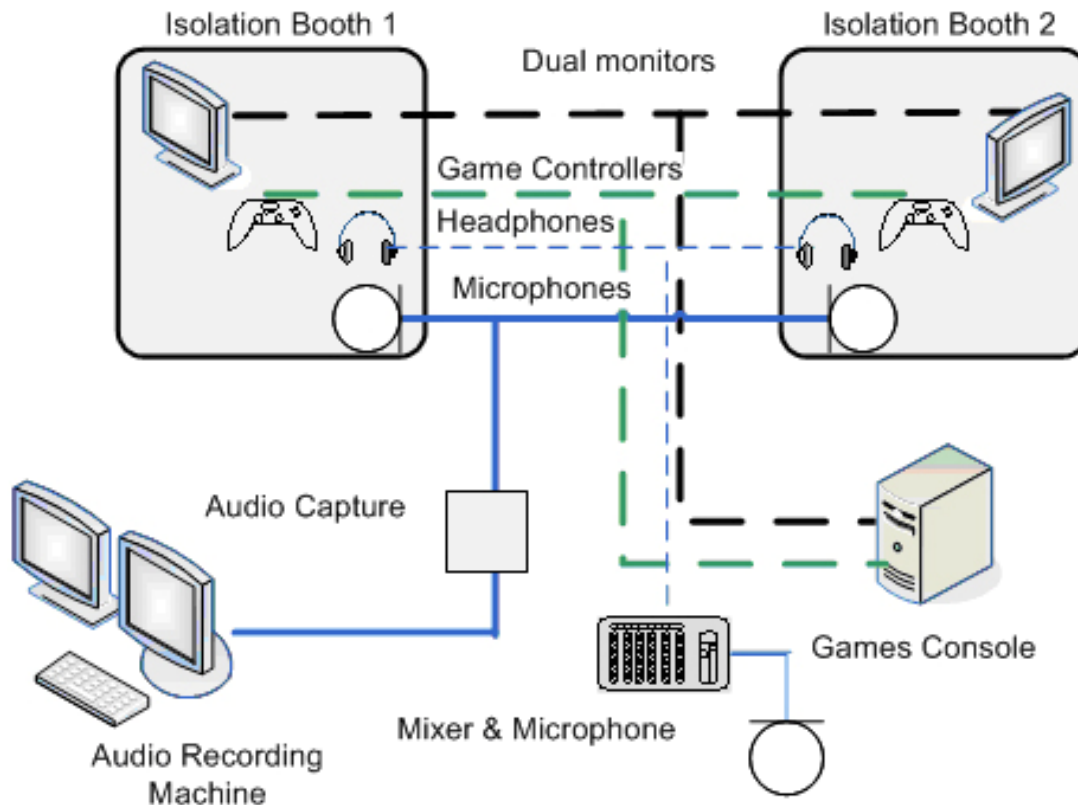


Figure 3: Games console setup

4.2.3 Experiment 3: Lego

This experimental design was first used by Kehrein in 2002 [35]. Two participants in isolation booths must cooperate to build a Lego construction. One participant (participant A) gives instructions on how to build it and the other participant (participant B) follows these instructions. The participant building the construction is the only one who has access to the pieces of Lego needed to complete the structure and must follow the instructions as closely as possible to complete the task. Various aspects of the experiment can be manipulated in order to aid or hinder the participants: necessary pieces of Lego can be removed beforehand, instructions can be prevented from being heard by cutting the audio link, the time allowed for the task can be lengthened or shortened and false sets of illustrated instructions can be provided to the participant giving the instructions.

In this experiment a Lego fire engine is used (Figure 4). This Lego set comprises of a main fire engine, comprising of two separate sections, and a storage trailer. Three separate sets of pictorial instructions are included in the set. This Lego set was chosen because it comprised of three separate constructions that allows the relative difficulty of the task to be varied: the smaller simpler constructions can be used to familiarise the participants with the procedure. The three constructions can be used individually or together in different combinations to hinder or aid the participants. The simpler and smaller constructions can be used with a generous time allowance to aid the participants or all three can be used together, in conjunction with a restricted time allowance, to make it harder, thus hindering the participants, to complete the task.



Figure 4: Image of the Lego Fire truck used in the Lego Experiment

4.3 Case Study

A small scale case study was undertaken to analyse the experimental procedure of each experimental design and to identify any practical considerations that needed to be addressed. One of the main issues raised was that console games were widely played by participants, particularly the multiplayer aspects of the games. Most of the participants in the case study stated that the use of console gaming would attract a lot more participants with very little in the way of external manipulation being needed. Furthermore most participants would be familiar with the particular aspects of console game playing and would feel comfortable playing competitively or co-operatively. Some participants in the case study stated that when playing against each other they got very competitive and immersed in the game being played. This situation is ideal for the elicitation of natural emotional responses with the true nature of the experiment being disguised by the competitive nature of the games being played.

4.4 Conclusions

The three experimental designs are all cased on the Success/Failure MIP, with a variation of the Gift MIP being utilised when needed. Of the three experiments the Tetris and Console game experiments provide the most potential for external manipulation. The Lego experiment can not be manipulated as easily as the removal of Lego pieces or the use of false instruction booklets are performed before the experiment takes place; the only external manipulation available during the experiment is the cutting of a particular audio feed or feed. Examination of the Lego procedure is still ongoing. The Tetris experiment was successful in eliciting emotional responses from participants, however preliminary findings suggest that the console gaming experiment is probably the most suitable for eliciting emotional responses from participants and will be the main experimental procedure used.

5 Corpus Metadata Specification

The definition of specific metadata for use with an emotional speech corpus is crucial, in that poorly (or inaccurately) annotated assets are of little use in analysis. This problem is compounded by the lack of standardisation for speech corpora, particularly in relation to emotion content. The only cohesive attempt at corpus metadata standardisation performed thus far has been by the EAGLE/ISLE consortium [46], which has led to the development of the ISLE Metadata Initiative (IMDI). Although not a comprehensive (or universally adopted) standard, IMDI represents the only current standard for speech corpus metadata available. For this reason, it was decided to implement the IMDI standard within the speech corpus detailed in this deliverable, to maintain as cohesive a standard as possible within current developments. A full listing of the metadata for an example asset is given in Appendix 1.

5.1 Metadata Structure

The metadata used by the emotional speech analysis corpus is organised in the following manner (Figure 5):

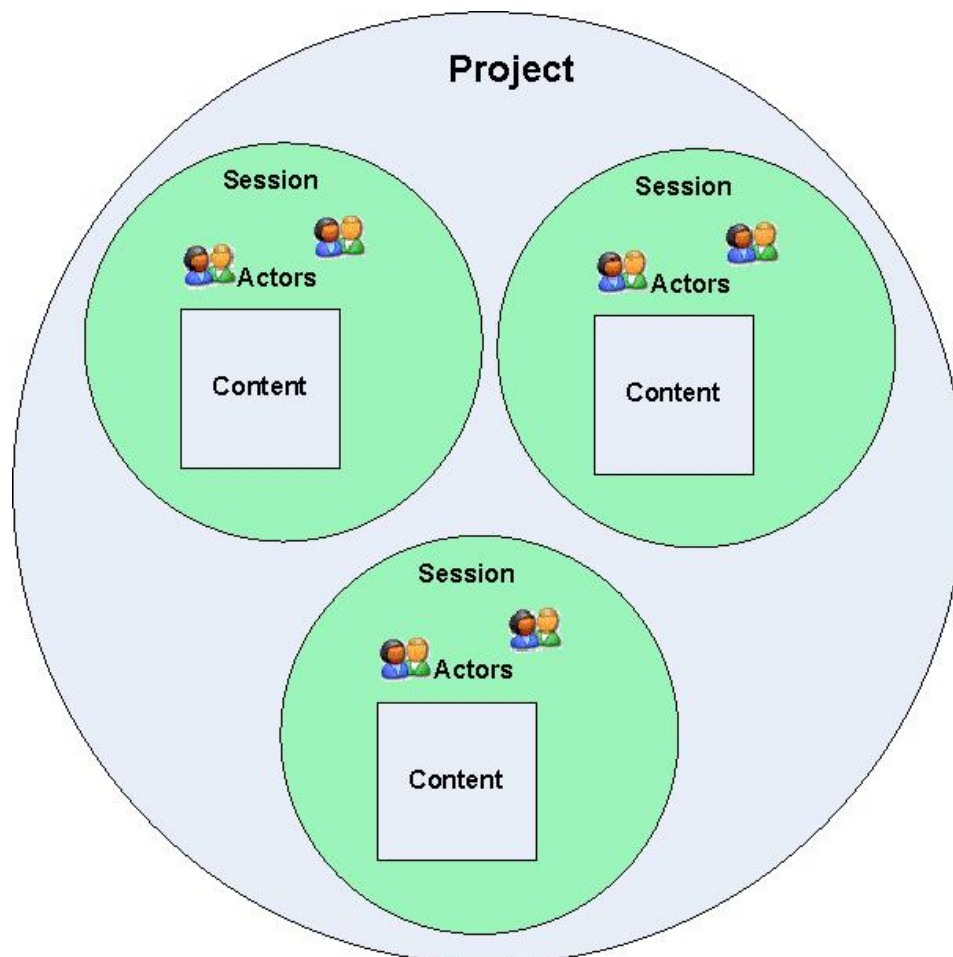


Figure 5: Example block diagram of the IMDI schema organisation. In this example, 3 separate session bundles are grouped logically under a single project. Each session relates to a specific type of content, and involves various actors who deliver the speech acts

5.1.1 Project

The definition of a particular project allows various sessions to be grouped in a logical form. Thus, in the case of the emotional speech corpus described in this deliverable all sessions are organised relative to the SALERO project.

Salero Corpus Browser
New Project

Project Title
SALERO

Project Description
SALERO aims at making cross media-production for games, movies and broadcast faster, better and cheaper by combining computer graphics, language technology, semantic web technologies as well as content based search and

Contact Name
Brian Vaughan

Contact Address
DMC Aungier Street DIT

Contact Email
brian.vaughan@dit.ie

Contact Organisation
Digital Media Center
Dublin Institute of Technology
Aungier Street
Dublin 2

[Edit](#) or [cancel](#)

Figure 6: Example project definition within the corpus database.
This listing gives some brief information about the SALERO project

By grouping sessions logically, it will allow future expansion of the database to include other corpora developed for different purposes (such as language learning or media production).

5.1.2 Session

A session is defined as the common bundle for linguistic events within IMDI metadata, and thus all speech assets are defined relative to a specific session. This allows an audio clip to be taken from a longer recording for specific analysis, while still retaining the same overall metadata as all other files in that session bundle.

Salero Corpus Browser

Navigate

[assets](#) | [projects](#) | [contents](#) | [actors](#) | **[sessions](#)** | [logout](#)

Content Listing

[New session](#)

🕒 Console Gaming Tournament

[Delete Session](#)

Salero Corpus Browser
Edit Session

Title
Console Gaming Tournament

Date
2007 | November | 6

Language
English

Country
Ireland

Address
CSAL, DMC, DIT Aungier Street

Description
Game in a the console gaming tournament that ran from October to December.


[Edit](#) or [cancel](#)

Figure 7: Example gaming tournament session and its associated metadata.
This is an example console gaming tournament bundle, which does not go into specifics of particular games, playing conditions or prize incentives

The session definition provides a convenient way to group assets for analysis, allowing assets taken from different experiments to be assessed either in isolation or within a wider common context.

5.1.3 Actor

The definition of an actor(s) within a session is a very useful aspect of the IMDI standard, as it allows the various participants in a speech recording to be documented for later consideration.



Salero Corpus Browser
Edit Actor

Surname
Tournament Participant

Firstname

Email
NA

Dob
2007 | October | 1

Gender
NA

Address
NA

Organisation
NA

Languages
English

or

Figure 8: Example actor metadata.

This screen shows an anonymised actor, who must still list a valid date of birth as required by the IMDI schema

In many instances, an actors details may be anonymised to ensure that ethical standards are adhered to (this is given as an option for each testing participant). Having said this, it is also very useful to consider database queries based on metadata such as geographical location or language, to allow broader linguistic analysis to be performed. Future work may consider the multi-lingual definition of assets within a corpus for analysis, and thus actor information would be crucial in this regard.

5.1.4 Content

The content metadata defined relates to specific activities for a given session (Figure 9):

Salero Corpus Browser
Edit Content

Profile
Console Gaming

Genre
Stimuli

Subgenre
Stimuli

Interactivity
Interactive

Planningtype
Spontaneous

Involvement
Non-elicited

Socialcontext
Controlled Environment

Eventstructure
Conversation / multi-dialogue

[Edit](#) or [cancel](#)

**Figure 9: Example content definition screen, relating to a session bundle.
In this example, the open vocabularies for genre and sub-genre relate to console gaming
experiments (4.2.2)**

Definitions of genre and sub-genre are open vocabularies, while other terms such as interactivity and planning type are taken from IMDI standard closed vocabularies. By providing more information relating to the type of speech asset being annotated, it is hoped that wider queries can be made in the corpus database as the record set expands over time.

5.1.5 Asset

Each asset in the corpus is defined in terms of its audio quality, which in turn relates to the LinguaTag SMIL analysis data defined in D6.1.1 and D6.1.2 (Figure 10):

Salero Corpus Browser

Navigate

[assets](#) | [projects](#) | [contents](#) | [actors](#) | [sessions](#)

061107_Gaming00269.wav

General Details	
file size	443916 kb
file format	wav
codec	PCM
dit depth	16
sample rate	44100 Hz
sample count	220500
channels	1
duration	5.0 s

Overall Thresholds	
pitch peak	94.1849 Hz
intensity peak	59.8118 db
vowel duration	0.0736445 s

Asset Vowel Details

vowel_1								
clip begin	clip end	jitter	shimmer	voicebreaks	hnr	prominence	syllable text	vowel
0.98729 s	1.06009 s	0.01997	0.07841	0	8.369	2		er

**Figure 10: Corpus listing for asset audio data and LinguaTag vowel analysis data.
Only the first vowel is shown on this screen**

In this manner, an asset could be queried in terms of its emotional dimensions (contained within the LinguaTag SMIL analysis file) or for specific acoustic attributes related to a vowel within the clip. This will allow investigation into the acoustic correlates of emotional speech to be performed using high quality speech assets, subject to emotional clip rating by listener groups (see section 7.1). The SMIL data output by LinguaTag also forms the basis of vowel stress animation methods currently under testing in the PGP experimental production 'My Tiny Planets' (see D9.3.2).

6 Implementation of an Emotional Speech Corpus

6.1 Requirements

There were, from the outset, several considerations that helped to define the technical architecture of the corpus. Firstly, the prototype must provide editors with the ability to insert assets, in the form of wav files, and related linguatag data, in the form of SMIL files. The prototype must parse the SMIL file and populate the corresponding database tables. The corpus, therefore, necessitates a storage layer or database as a persistent back-end. Secondly, editors require remote access to corpus assets. This allows for the addition, deletion and alteration of corpus assets and related metadata. At first, each asset was to be uploaded and annotated individually. However, following initial trials, it was decided to provide the ability for batch uploads, thereby allowing an editor to upload several assets at the any one time. In this case, each asset is annotated with the same metadata.

6.1.1 *Automated Annotation*

A central requirement of the prototype was to reduce the overhead associated with annotating digital assets. Often metadata can be reused for multiple assets, and does not require the editor re-entering this data every time they wish to add a new asset. Therefore the approach involved using Ajax auto-suggest as a way to reduce the annotation overhead when entering new assets. An editor can enter any piece of metadata and re-use that piece of metadata through the autosuggest functionality. Similarly, there are two approaches to inserting metadata; the first allows editors to enter metadata and then upload the asset, annotating the asset with autosuggest functionality as outlined above. The second allows the editor upload the asset and annotate the asset on the fly, effectively creating new metadata.

6.2 Architecture

As with traditional web architecture, the corpus is divided into three separate tiers, presentation, application and data:

6.2.1 *Presentation*

The presentation tier displays the corpus assets and related metadata. As this is a prototype application, a simple style was used to delineate assets from the different forms of metadata.

6.2.2 *Application*

The application tier contains the business logic of the corpus. The prototype, v0.1, provides two approaches to uploading and annotating assets. The first allows the editor to individually upload and annotate a single asset. The second allows an editor to batch upload assets, therefore reducing time when populating the corpus. As mentioned previously, the second approach requires that all assets are annotated with the same metadata.

6.2.3 *Data*

The data tier provides persistent storage for the corpus metadata. The audio assets and related linguatag SMIL files are not stored in the data tier. Conversely, each asset and SMIL file is stored on the web server to reduce overhead when querying and retrieving assets from the data tier.

6.3 Technologies

The technologies used to implement each tier of the corpus are outlined below:

6.3.1 Presentation and Application Tiers

The Presentation and Application tiers were developed using Ruby on Rails². Ruby on Rails is an open source web framework for the rapid application development. The framework is an implementation of the model view controller design pattern and provides an excellent migration mechanism for creating and altering the data tier. Ruby on Rails has support for XML, and consequently SMIL, through the REXML³ ruby-gem, which is packaged with the latest version of Ruby.

6.3.2 Data Tier

The popular open source MySQL⁴ database provides a foundation for the data tier. The creation and manipulation of the database is carried out through the Ruby on Rails framework.

The corpus is available at URL: <http://corpus.dmc.dit.ie/annotate/login>.

² <http://www.rubyonrails.org/>

³ <http://www.germane-software.com/software/rexml/>

⁴ <http://www.mysql.org/>

7 Future Work

7.1 Emotional Asset Rating

Rating of the emotional speech corpus will be performed by online listening tests, which will be delivered to the user using a streaming media server such as Flash Adobe Flash Media Server or Apple Podcast server. Each clip will be delivered to the user at random, with an activation and evaluation slider being provided for rating (see D6.1.1 and D6.1.2). The user will also be given the option not to rate clip, which will also be stored as a rating. This will allow the corpus listening tests to be used as a form of feedback on potentially difficult or unclear clips, allowing the asset groups to be streamlined into more clearly defined emotional dimensions. The intention of the online rating system is to obtain a statistical definition of emotional dimensions for each clip in the corpus, and rate each clip both in terms of its dimensional values and also the confidence rating for that clip. Thus, a clip which has been rated by more listeners will be defined as having a higher confidence level relative to its emotional dimension values, allowing statistical analysis to be performed on groups of assets in as robust a manner as possible. The exact nature of corpus asset rating will be covered in the future deliverable D6.2.2 "Development of an online corpus visualisation and user rating interface".

7.2 Corpus Workflow

There are several potential workflows that are already planned using the corpus detailed in this deliverable. Although all possible avenues of future work cannot be fully defined at this time, it is nevertheless important to indicate where future work (and integration) will focus:

7.2.1 Lip-Synching and Character Animation

Work has already been performed on the use of LinguaTag analysis data as the basis of automated character animation. Initially focussed on lip-synching work, further investigation has led to the consideration of wider application in automated character animation for use in online and mobile content delivery. In current developments, work performed in conjunction with PGP on the 'My Tiny Planets' experimental production (see D9.3.2) considers the automated sequencing of pre-rendered character animations in synchronisation with stress vowels in a speech asset. This work is carried out towards the Milestone M6.5- *Implementation of initial lip-synching animations using refined tagging framework* that is scheduled for T30. Initial results are promising, with future work planned to implement a cohesive plug-in-based workflow in conjunction with PGP and UPF-GTI. In this manner, an asset could eventually be queried within a production speech corpus in terms of its defined metadata, and then its LinguaTag analysis data could be processed by animation plug-ins to be delivered as automated content in an live production.

7.2.2 Providing Rules for Emotional Speech Synthesis

The creation of a limited emotional speech corpus for analysis aims to collate suitable assets for acoustic (and linguistic) analysis, with a view to investigating the acoustic correlates of emotional speech. Although much work has been performed in the area [6, 8, 47, 48], it is still an open research question and thus is worthy of further investigation. This work is carried out with respect to the milestone M6.6- *Initial results of emotional speech corpus analysis into the acoustic correlates of emotional speech* scheduled for T37.

7.2.3 Defining Metadata for Experimental Production

Although the corpus specified in this deliverable is defined for the purposes of speech analysis, links to lip-synching and character animation via experimental productions, alongside the original brief of D6.1.1 to consider interfaces with ontology work in WP03 and search and retrieval work in WP05 leads towards the definition of metadata schemas for production workflows. The use of IMDI in this corpus represents implementation of a standard for speech corpora metadata, but future work will consider means by which the definition of production metadata for speech and audio assets could be implemented. This relates to the current action points 116 and 117 (at time of writing) relating to WP03.

7.3 Corpus Visualisation

Further work will consider how best to develop data visualisations for online analysis of the corpus. Rapid Application Development (RAD) tools such as Adobe Flex Builder will be used to produce interactive online tools for data manipulation that will form the basis of a corpus visualisation. This work will ideally form the basis of a future SALERO deliverable, allowing all SALERO partners quick and easy access to the assets within the corpus and their associated data.

7.4 Corpora Expansion

At time of writing, the corpus contains over 150 fully annotated and tagged assets, and this figure is intended to grow in tandem with further MIP related experiments carried out in the coming months. There is no defined headroom for the size of the corpus, but the experimental criteria, recording conditions and annotation metadata will be upheld in all future work. It is hoped that the results of analysis of this corpus will help in some way to determine the acoustic correlates of emotional speech, and thus standardisation of experiments, audio quality and metadata is essential to providing solid data for this analysis.

8 Conclusions

This document is provided as support for the deliverable D6.2.1: A limited Emotional Speech Corpus for Analysis. As part of this document, the following criteria were covered:

- The rationale behind the use of Mood Induction Experiments
- The recording conditions used
- The Mood Induction Experiments used
- The corpus metadata defined in IMDI format
- The analysis data (output from LinguaTag) defined in SMIL format
- Corpus Construction

Future directions for work were also considered, notably in relation to visualisations of the speech corpus for online analysis. Such visualisations may form part of future SALERO deliverables, and may move toward providing the means of effectively determining the acoustic correlates of emotional speech.

9 Appendix 1: Example Corpus Asset Listing

This appendix contains a full listing of an example asset from the corpus, detailing both IMDI metadata and LinguaTag analysis data:

```
<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE ROOT SYSTEM "corpus.dtd">
<ROOT>
  <actor>
    <id>6</id>
    <firstname></firstname>
    <surname>Tournament Participant</surname>
    <email>NA</email>
    <dob>2007-10-01</dob>
    <gender>NA</gender>
    <address>NA</address>
    <organisation>NA</organisation>
    <languages>English</languages>
  </actor>
  <session>
    <id>2</id>
    <title>Console Gaming Tournament</title>
    <date>2007-11-06</date>
    <language>English</language>
    <country>Ireland</country>
    <address>CSAL, DMC, DIT Aungier Street</address>
    <description>Game in a the console gaming tournament that ran from
October to December.</description>
  </session>
  <session>
    <id>2</id>
    <profile>Console Gaming</profile>
    <genre>Stimuli</genre>
    <subgenre>Stimuli</subgenre>
    <interactivity>Interactive</interactivity>
    <planningtype>Spontaneous</planningtype>
    <involvement>Non-elicited</involvement>
    <socialcontext>Controlled</socialcontext>
    <eventstructure>Conversation</eventstructure>
  </session>
  <project>
    <id>3</id>
    <title>SALERO</title>
    <description>SALERO aims at making cross media-production for games,
movies and broadcast faster, better and cheaper by combining computer
graphics, language technology, semantic web technologies as well as content
based search and retrieval.</description>
    <contactname>Brian Vaughan</contactname>
    <contactaddress>DMC Aungier Street DIT</contactaddress>
    <contactemail>brian.vaughan@dit.ie</contactemail>
    <contactorganisation>Digital Media Center, Dublin Institute of
Technology, Aungier Street, Dublin 2</contactorganisation>
  </project>
</ROOT>

<?xml version="1.0" encoding="UTF-8"?>
<smil xmlns="http://www.w3.org/2001/SMIL21/Language"
xmlns:lg="http://www.dmc.dit.ie/2006/SALERO/linguatag">
  <head>
    <meta name="type" content="Linguatag data"/>
    <meta name="base" content="file:///G:/New Files for John/">
```

```
<paramGroup id="audio_properties">
  <param name="file_name" value="061107_Gaming00269.wav" />
  <param name="file_size" value="443916" />
  <param name="format" value="wav" />
  <param name="codec" value="PCM" />
  <param name="bit_depth" value="16"/>
  <param name="sample_rate" value="44100"/>
  <param name="sample_count" value="220500"/>
  <param name="channels" value="1"/>
  <param name="duration" value="5s"/>
</paramGroup>
</head>

  <smil xmlns="http://www.w3.org/2001/10/synthesis"
xml:base="file:///G:/New Files for John/">
  <audio src="061107_Gaming00269.wav">
    <desc>Transcription</desc>

  </audio>
  <!-- "mark" tags link to smil audio clips -->
</smil>

<body>
  <par>
    <seq id="vowel_stresses">
      <lg:stress_event_thresholds
vowel_duration="0.0736445s"
pitch_peak="94.1849" intensity_peak="59.8118"
/>

      <audio id="vowel_1" src="061107_Gaming00269.wav"
clip-begin="0.98729s" clip-end="1.06009s">

        <param name="jitter" value="0.01997"/>
        <param name="shimmer" value="0.07841"/>
        <param name="voice_breaks" value="0"/>
        <param name="hnr" value="8.369"/>
        <param name="prominence" value="2"/>
        <param name="syllable_text" value=""/>
        <param name="vowel" value="er"/>

        <lg:intensity clip-peak="1.03165s"
begin="60.8944" peak="64.2944"
end="62.4221" mean="63.4945"/>

        <lg:pitch clip-peak="1.01284s"
begin="999" peak="112.862" end="103.928"
mean="108.684"/>
      </audio>
      <audio id="vowel_2" src="061107_Gaming00269.wav"
clip-begin="1.19576s" clip-end="1.27425s">

        <param name="jitter" value="0.00493"/>
        <param name="shimmer" value="0.03646"/>
        <param name="voice_breaks" value="0"/>
        <param name="hnr" value="14.408"/>
        <param name="prominence" value="3"/>
        <param name="syllable_text" value=""/>
        <param name="vowel" value="Undefined"/>

        <lg:intensity clip-peak="1.21551s"
begin="60.4151" peak="61.5663"
end="60.335" mean="61.264"/>
    </seq>
  </par>
</body>
```

```

        <lg:pitch clip-peak="1.25236s"
              begin="95.1949"                peak="95.8278"
end="95.0216" mean="95.1483"/>
    </audio>
    <audio id="vowel_3" src="061107_Gaming00269.wav"
          clip-begin="1.38639s" clip-end="1.42715s">

        <param name="jitter" value="0.02238"/>
        <param name="shimmer" value="0.05894"/>
        <param name="voice_breaks" value="0"/>
        <param name="hnr" value="6.417"/>
        <param name="prominence" value="1"/>
        <param name="syllable_text" value=""/>
        <param name="vowel" value="Undefined"/>

        <lg:intensity clip-peak="1.4011s"
              begin="57.3812"                peak="57.9387"
end="55.3379" mean="57.3454"/>

        <lg:pitch clip-peak="1.38639s"
              begin="94.8373"                peak="94.8373"
end="87.7477" mean="91.2107"/>
    </audio>
    <audio id="vowel_4" src="061107_Gaming00269.wav"
          clip-begin="1.53167s" clip-end="1.62716s">

        <param name="jitter" value="0.01388"/>
        <param name="shimmer" value="0.09641"/>
        <param name="voice_breaks" value="0"/>
        <param name="hnr" value="8.79"/>
        <param name="prominence" value="2"/>
        <param name="syllable_text" value=""/>
        <param name="vowel" value="Undefined"/>

        <lg:intensity clip-peak="1.57863s"
              begin="61.9276"                peak="62.8451"
end="54.8375" mean="62.1097"/>

        <lg:pitch clip-peak="1.53167s"
              begin="89.0087"                peak="89.0087" end="999"
mean="87.3693"/>
    </audio>
    <audio id="vowel_5" src="061107_Gaming00269.wav"
          clip-begin="1.7514s" clip-end="1.82762s">

        <param name="jitter" value="0.05302"/>
        <param name="shimmer" value="0.06945"/>
        <param name="voice_breaks" value="0"/>
        <param name="hnr" value="4.815"/>
        <param name="prominence" value="2"/>
        <param name="syllable_text" value=""/>
        <param name="vowel" value="er"/>

        <lg:intensity clip-peak="1.80589s"
              begin="51.8781"                peak="59.8835"
end="59.0946" mean="58.8144"/>

        <lg:pitch clip-peak="1.77252s"
              begin="999"                    peak="90.6853" end="75.3577"
mean="83.4883"/>
    </audio>
    <audio id="vowel_6" src="061107_Gaming00269.wav">
```

```
clip-begin="1.90815s" clip-end="1.93579s">
  <param name="jitter" value="0.01043"/>
  <param name="shimmer" value="999"/>
  <param name="voice_breaks" value="0"/>
  <param name="hnr" value="9.143"/>
  <param name="prominence" value="0"/>
  <param name="syllable_text" value=""/>
  <param name="vowel" value="uw"/>
  <lg:intensity clip-peak="1.91419s"
    begin="59.1181" peak="59.1573"
end="58.7332" mean="59.0484"/>
  <lg:pitch clip-peak="1.93579s"
    begin="86.791" peak="88.2936"
end="88.2936" mean="87.4613"/>
</audio>
<audio id="vowel_7" src="061107_Gaming00269.wav"
clip-begin="2.01216s" clip-end="2.09962s">
  <param name="jitter" value="0.14888"/>
  <param name="shimmer" value="0.20225"/>
  <param name="voice_breaks" value="0"/>
  <param name="hnr" value="3.794"/>
  <param name="prominence" value="1"/>
  <param name="syllable_text" value=""/>
  <param name="vowel" value="er"/>
  <lg:intensity clip-peak="2.03689s"
    begin="58.1845" peak="58.4559"
end="55.0496" mean="57.3811"/>
  <lg:pitch clip-peak="2.05612s"
    begin="999" peak="91.9353" end="88.0385"
mean="84.0047"/>
</audio>
<audio id="vowel_8" src="061107_Gaming00269.wav"
clip-begin="2.27405s" clip-end="2.36577s">
  <param name="jitter" value="0.04825"/>
  <param name="shimmer" value="0.12062"/>
  <param name="voice_breaks" value="0"/>
  <param name="hnr" value="9.82"/>
  <param name="prominence" value="1"/>
  <param name="syllable_text" value=""/>
  <param name="vowel" value="uh"/>
  <lg:intensity clip-peak="2.32452s"
    begin="54.8816" peak="55.7971"
end="52.5751" mean="54.9399"/>
  <lg:pitch clip-peak="2.31501s"
    begin="86.3812" peak="89.2378"
end="88.7219" mean="85.3493"/>
</audio>
<audio id="vowel_9" src="061107_Gaming00269.wav"
clip-begin="3.24792s" clip-end="3.30127s">
  <param name="jitter" value="0.10946"/>
  <param name="shimmer" value="0.05978"/>
  <param name="voice_breaks" value="0"/>
  <param name="hnr" value="6.703"/>
  <param name="prominence" value="1"/>
```

```
<param name="syllable_text" value=""/>
<param name="vowel" value="er"/>

<lg:intensity clip-peak="3.28028s"
begin="44.9706" peak="57.672"
end="55.2175" mean="55.8207"/>

<lg:pitch clip-peak="3.26788s"
begin="999" peak="100.316" end="81.1273"
mean="91.9717"/>
</audio>
<audio id="vowel_10" src="061107_Gaming00269.wav"
clip-begin="3.39297s" clip-end="3.47368s">

<param name="jitter" value="0.06359"/>
<param name="shimmer" value="0.0393"/>
<param name="voice_breaks" value="0"/>
<param name="hnr" value="5.594"/>
<param name="prominence" value="2"/>
<param name="syllable_text" value=""/>
<param name="vowel" value="Undefined"/>

<lg:intensity clip-peak="3.41772s"
begin="57.6965" peak="59.8192"
end="56.1029" mean="59.0589"/>

<lg:pitch clip-peak="3.45146s"
begin="79.4335" peak="90.882"
end="82.2306" mean="84.328"/>
</audio>
<audio id="vowel_11" src="061107_Gaming00269.wav"
clip-begin="3.59798s" clip-end="3.70349s">

<param name="jitter" value="0.01183"/>
<param name="shimmer" value="0.10007"/>
<param name="voice_breaks" value="0"/>
<param name="hnr" value="13.678"/>
<param name="prominence" value="2"/>
<param name="syllable_text" value=""/>
<param name="vowel" value="uh"/>

<lg:intensity clip-peak="3.61576s"
begin="58.7007" peak="60.5001"
end="55.7876" mean="58.9905"/>

<lg:pitch clip-peak="3.59798s"
begin="92.1476" peak="92.1476"
end="86.9788" mean="88.3738"/>
</audio>
</seq>
<seq id="intensity_boundaries">
<lg:contour_shape>Unrated</lg:contour_shape>

<audio id="intensity_initial"
src="061107_Gaming00269.wav" clip-begin="1.03165s">
<param name="peak" value="64.2944"/>
</audio>
<audio id="intensity_highest"
src="061107_Gaming00269.wav" clip-begin="1.03165s">
<param name="peak" value="64.2944"/>
</audio>
<audio id="intensity_lowest"
src="061107_Gaming00269.wav" clip-begin="2.32452s">
<param name="peak" value="55.7971"/>
```

```

        </audio>
        <audio id="intensity_final"
src="061107_Gaming00269.wav" clip-begin="3.61576s">
            <param name="peak" value="60.5001"/>
        </audio>
    </seq>
    <seq id="pitch_boundaries">
        <lg:contour_shape>Unrated</lg:contour_shape>

        <audio id="pitch_initial"
src="061107_Gaming00269.wav" clip-begin="1.01284s">
            <param name="peak" value="112.862"/>
        </audio>
        <audio id="pitch_highest"
src="061107_Gaming00269.wav" clip-begin="1.01284s">
            <param name="peak" value="112.862"/>
        </audio>
        <audio id="pitch_lowest" src="061107_Gaming00269.wav"
clip-begin="1.93579s">
            <param name="peak" value="88.2936"/>
        </audio>
        <audio id="pitch_final" src="061107_Gaming00269.wav"
clip-begin="3.59798s">
            <param name="peak" value="92.1476"/>
        </audio>
    </seq>
    <seq id="linguistic_analysis">
    </seq>
    <seq id="emotional_dimensions">
    </seq>
</par>
</body>
</smil>
```

10References

- [1] S. Chung, "Vocal expression and perception of emotion in Korean," in *14th International Conference of Phonetic Sciences*, San Fransisco, USA, 1999, pp. 969–972.
- [2] Scherer and Ceschi, "Geneva Airport Lost Luggage Study " *INCOMPLETE REF*, 2000.
- [3] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," *Speech Communication Special Issue Speech and Emotion*, vol. 40, pp. 33–60, 2003.
- [4] B. Katz, *Mastering Audio: The Art and the Science*. Burlington, MA: Focal Press, 2002.
- [5] I. S. Enberg, Hansen, A.V., Anderson, O., Dalsgaard, P., "Design, recording and verification of a Danish Emotional Speech Database," in *Eurospeech '97*, Rhodes, Greece, 1997.
- [6] M. Kienast, Sendlmeier, W.F., "Acoustical analysis of spectral and temporal changes in emotional speech," in *ISCA ITRW on Speech and Emotion*, Newcastle, Belfast, 2000, pp. 92-97.
- [7] C. Pereira, "Dimensions of emotional meaning in speech," in *ISCA ITRW on Speech and Emotion*, Newcastle, Belfast, 2000, pp. 25-28.
- [8] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, pp. 614-636, 1996.
- [9] N. Amir, Ron, S., Laor, N., "Analysis of an emotional speech corpus in Hebrew based on objective criteria," in *ISCA ITRW on Speech and Emotion*, Belfast, 2000.
- [10] C. Johns-Lewis, "Prosodic differentiation of discourse modes," in *Intonation in Discourse*, C. Johns-Lewis, Ed. San-Diego: College Hill Press, 1986, pp. 199-220.
- [11] T. Johnstone, "Emotional Speech Elicited using computer games," in *Spoken Language, ICSLP 96. Proceedings., Fourth International Conference on*, Philadelphia, PA, USA, 1996.
- [12] D. Pugmire, "Real Emotion," *Philosophy and Phenomenological research*, vol. 54, pp. 105-122, 1994.
- [13] G. Stemmler, "The vagueness of specificity: Models of peripheral physiological emotion specificity in emotion theories and their experimental discriminability.," *Journal of Psychophysiology* vol. 6, pp. 17-28, 1992.
- [14] G. H. Stemmler, M., Pauls, C.A., Scherer, T., "Constraints for emotion specificity in fear and anger: The context counts," *Psychophysiology*, pp. 275-291, 2001.
- [15] P. Greasley, Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., Horton, D., "Representation of prosodic and emotional features in a spoken language database," in *XIIIth ICPhS*, Stockholm, 1995, pp. 242-245.
- [16] P. Roach, R. Stibbard, J. Osborne, S. Arnfield, and J. Setter, "Transcription of Prosodic and Paralinguistic Features of Emotional Speech," *Journal of the International Phonetic Association*, pp. 83-94, 1998.
- [17] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in *ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000, pp. 39-44.
- [18] M. Gottdiener, "Field Research and Video Tape," *Sociological Inquiry*, vol. 4, pp. 59-66, 1979.
- [19] B. a. H. S. B. Geer, "Participant Observation and Interviewing: A Comparison," *Human Organization*, vol. 16, pp. 28-32, 1957.
- [20] F. Erickson, and Schultz, J., "The Counsellor as Gatekeeper: Social Interaction in Interviews," in *Language, Thought and Culture: Advances in the Study of Cognition*, E. Hammel, Ed. New York: Academic Press, 1982.

- [21] B. L. Bellman, & Bennetta Jules-Rosette., *A Paradigm for looking*. Norwood: Ablex Publishing, 1977.
- [22] A. Gerrards-Hesse, K. Spies, and F. W. Hesse, "Experimental inductions of emotional states and their effectiveness: A review," *British Journal of Psychology*, vol. 85, pp. 55-78, 1994.
- [23] F. Weiss, Blum, G.S., Gleberman, L., "Anatomically based measurement of facial expressions in simulated versus hypnotically induced effect.," *Motivation and Emotion*, vol. 11, 1987.
- [24] F. Strack, Schwarz, N., Gschneidinger, E., "Happiness and reminiscing: The role of time perspective, effect, and mode of thinking," *Journal of Personality and Social Psychology*, vol. 49, pp. 1460-1469, 1985.
- [25] R. M. B. Banos, C. ; Liaño. V. ; Rey, B. ; Guerrero, B., Alcaniz, M., "Virtual Reality as Mood Induction Procedure," in *PRESENCE 2003, 6th Annual International Workshop on Presence*, 2003, pp. 1-4.
- [26] A. H. Baumgardiner, Arkin, R.M., "Affective state mediates causal attributions for success and failure," *Motivation and Emotion*, vol. 12, pp. 99-111, 1988.
- [27] M. R. Cunningham, Schaffer, D. R., Barbee, A.P., Wolff, P.L., Kelley, D.J, "Separate processes in the relation of elation and depression to helping : Social versus personal concerns " *Journal of Experimental Social Psychology*, vol. 26, pp. 13-33, 1990.
- [28] L. T. Worth, Mackie, D. M., "Cognitive mediation of positive affect in persuasion.," *Social Cognition*, vol. 5, pp. 76-94, 1987.
- [29] J. P. Forgas, "Affective influences on individual and group judgements ." *European Journal of Social Psychology* vol. 20, pp. 441-453, 1990.
- [30] G. K. B. Manucia, D. J. & Cialdini, R. B., "Mood influences on helping: Direct effects or side effects?," *Journal of Personality and Social Psychology*, vol. 46, pp. 357-364, 1984.
- [31] J. J. W. James D. Laird., Mark Halal., and Martha Szegda, "Remembering What You Feel: Effects of Emotion on Memory," *Journal of Personality and Social Psychology*, vol. 45, pp. 646-657, 1982.
- [32] R. W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Trans Pattern Analysis & Machine Intelligence*, pp. 1175-1191, 2001.
- [33] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition and Emotion*, pp. 87-108, 1995.
- [34] A. Iida, N. Campbell, and M. Yasumura, "Design and Evaluation of Synthesised Speech with Emotion," *Journal of Information Processing Society of Japan*, vol. 40, pp. 479-486, 1998.
- [35] R. Kehrein, "The prosody of authentic emotions," in *Speech Prosody*, Aix-en-Provence, France, 2002.
- [36] R. Fernandez and R. Picard, "Modelling drivers' speech under stress," in *ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000, pp. 219-224.
- [37] T. Johnstone, C. M. v. Reekum, K. Hird, K. Kirsner, and K. R. Scherer, "Affective speech elicited with a computer game," *Emotion*, pp. 513-518, 2005.
- [38] F. J. a. S. Tolkmitt, K., "Effect of Experimentally Induced Stress on Vocal Parameters.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 12, pp. 302-313, 1986.
- [39] R. Jain, "Quality of Experience," in *IEEE Multimedia*, 2004.
- [40] R. Westermann, Spies, K., Stahl, G., & Hesse, F. W., "Relative effectiveness and validity of mood induction procedures: a meta analysis," *European Journal of Social Psychology*, vol. 26, pp. 557-580, 1996.
- [41] H. Kooijman. V, P ,Cutler. A, "Electrophysiological evidence for prelinguistic infants' word recognition in continuous speech," *Cognitive Brain Research* vol. 24 pp. 109-116, 2005.

- [42] F. Ramus, "Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues.," in *Annual Review of Language Acquisition* 2 vol. 2: John Benjamins Publishing Company, 2002.
- [43] C. Cullen, Vaughan, B. ,Kousidis, S., Wang, Yi ., McDonnell, C. and Campbell, D. , "Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction " in *International Conference on Multidisciplinary Information Sciences and Technologies* Extremadura, Merida, 2006.
- [44] Digidesign, "Digidesign Mbox 2 Pro Site." vol. 2007, <http://www.digidesign.com/index.cfm?langid=51&navid=100&itemid=4956>, Ed., 2007.
- [45] RealVNC, "RealVNC software home page," in *Real VNC*. vol. 2006, R. VNC, Ed., 2006, p. The home page of RealVNC.
- [46] ISLE, "IMDI (ISLE Metadata Initiative), Metadata Elements for Session Descriptions," Draft Proposal Version 3.0.3 ed, 2003.
- [47] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proceedings of Speech Prosody*, Aix-en-Provence, France, 2002.
- [48] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, vol. 1, pp. 87-90, 2001.

11 Glossary

Partner Acronyms

AM	Activa Multimedia, ES
BLITZ	Blitz Games, UK
DIT	Dublin Institute of Technology, IE
DTS	Digital Theatre Systems, UK
FBM-UPF	Fundació Universitat Pompeu Fabra, ES
GVG	Grass Valley Germany, DE
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT
LFUI	Leopold-Franzenzs Universtät Innsbruck, AT
UPF	Fundació Universitat Pompeu Fabra, Music Technology Group, ES
PGP	Pepper's Ghost Productions Ltd., UK
TAIK	Taideteollinen Korkeakoulu, FI
UG	University of Glasgow, UK
URL	Universitat Ramon Llull, ES