humaine

D5e

Proposal for exemplar and work towards it:

Data and Databases

Workpackage 5 Deliverable

Date: 30th November 2005

IST project contract no.	507422
Project title	**HUMAINE** **Human-Machine Interaction Network on Emotions**
Contractual date of delivery	*November 30, 2005*
Actual date of delivery	*November 30, 2005*
Deliverable number	D5e
Deliverable title	Proposal for exemplars and work towards it: Data and Databases
Type	Report
Number of pages	67
WP contributing to the deliverable	WP 5
Task leader	QUB
Author(s)	Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Ian Sneddon, Margaret McCrory, Edelle McMahon, Laurence Devillers, Jean-Claude Martin, Noam Amir, Katri Oinonen, Dennis Hofs, Anton Batliner, Marc Schroeder, Kostas Karpouzis, Catherine Pelachaud, Elisabeth Andre, , Ailbhe Ni Chasaide, Jina Lee
EC Project Officer	Philippe Gelin

Address of lead author: Ellen Douglas-Cowie
School of English
Queen's University Belfast
Belfast
United Kingdom

Table of Contents

1 THE PLACE OF THIS REPORT WITHIN HUMAINE……3

2 BRIEF OVERVIEW OF WORKPACKAGE 5 AND THE EXEMPLAR PROPOSAL 8

2.1 The field covered by Workpackage 5 8

2.2 The research objectives 8

2.2.1 Main elements of the exemplar 11

2.2.2 How the subtasks link to each other 14

2.2.3 How the subtasks link to other aspects of HUMAINE 14

3 THE PLANNED PROGRAM OF RESEARCH 15

3.1 Element 1: development and selection of data collection techniques 15

3.1.1 Leader: QUB 15

3.1.2 Main participants : QUB, TAU, CNRS-LIMSI, UT 15

3.1.3 Main steps planned towards producing element 1 15

3.2 Element 2: Collecting Records 16

3.2.1 Leader: QUB 16

3.2.2 Main participants : QUB, TAU, CNRS-LIMSI 16

3.2.3 Main steps planned towards producing element 2 16

3.3 Element 3: Developing labelling schemes for – (a) emotion content, (b) signs (c) context 17

3.3.1 Leader: CNRS-LIMSI 17

3.3.2 Main Participants : CNRS-LIMSI, QUB, FAU-Erlangen, UNIGE, ICCS 17

3.3.3 Main steps planned towards producing element 3 17

3.4 Element 4: Delivering data labelled for (a) emotion content, (b) signs, (c) context 17

3.4.1 Leader: UT 17

3.4.2 Main Participants: UT, QUB, CNRS-LIMSI 17

3.4.3 Main steps planned towards producing element 4 17

3.5 Delivering documentation for the techniques and tools so that they can be used throughout the research community 18

3.5.1 Leader: QUB 18

3.5.2 Main Participants: QUB, CNRS-LIMSI 18

3.5.3 Main steps planned towards producing element 5 18

3.6 Making recordings and tools available for use in the medium term 19
3.6.1 Leader: QUB 19
3.6.2 Main participants : QUB, CNRS-LIMSI, FAU-Erlangen 19
3.6.3 Main steps planned 19

3.7 Steps to ensure co-ordination 19

3.8 Steps to ensure dissemination 20

4 RESEARCH ACHIEVEMENTS TO DATE 22

4.1 Achievement 1: development and selection of data collection techniques 22
4.1.1 Development of the SAL technique 22
4.1.2 Development of mood induction techniques (for use in Driving Simulator experiments) 24
4.1.3 Development of Emotion in Action/Reality Television-type elicitation & recording techniques 25
4.1.4 Participants 27
4.1.5 Output (demonstrations, resources, etc) 27
4.1.6 Follow-up in progress 27

4.2 Achievement 3: Development of labelling schemes 27
4.2.1 Databases Labelled 27
4.2.2 Development of emotion labelling tools 30
4.2.3 Development of labelling of Signs 32
4.2.4 Development of Physiological labels 34
4.2.5 Development of Context labels 36
4.2.6 Participants 36
4.2.7 Publications 36
4.2.8 Other output (demonstrations, resources, etc) 38

4.3 Providing resources for immediate use across the network 38
4.3.1 Recordings provided 38
4.3.2 Tools provided 39
4.3.3 Participants 40
4.3.4 Publications 40
4.3.5 Other output (demonstrations, resources, etc) 40
4.3.6 Follow-up in progress 40

5 CONCLUSION 41

5.1 Obstacles encountered or foreseen 41

5.2 Relation to the state of the art 42

5.3 Evidence of esteem 42

6 REFERENCES 43

7 APPENDIX: THE REMOTE SAL SYSTEM 45

REQUIREMENTS FOR CLIENT APPLICATION INTERFACE FOR OPERATOR 47

REQUIREMENTS FOR CLIENT APPLICATION INTERFACE FOR RESPONDENT 48

COMMUNICATION PROTOCOLS 56

JAVA APPLICATIONS 59

SalServerApp 59

SalOperatorApp 59

SalRespondentApp 60

1 The place of this report within HUMAINE

The HUMAINE Technical Annex identifies a common pattern that is followed by most of the project's workpackages:

> The measure of success will be the ability to generate a piece of work in each of the areas which exemplifies how a key problem in the area can be solved in a principled way; and which also demonstrates how work focused on that area can integrate with work focused on the other areas. We call these pieces of work *exemplars*. The exact form of an exemplar is not prespecified: it may be a working system, but it might also be a well-developed design, or a representational system, or a method for user-centred design. (p 4)
>
> To that end, each thematic group will work out a proposal for common action, embodied in one or more exemplars to be built during the second half of the funding period (p.16)
>
> The process will begin with production by each thematic group of a review of key concepts achievements and problems in its thematic area; and drawn from the review, an assessment of the key development goals in the area. This review and assessment will be circulated to the whole network for discussion and comment, aimed both at building understanding of basic issues across areas, and at identifying the choices of goal that would be most likely let the different groups achieve complementary developments. That consultation phase will provide the basis for deliverables in month 11, which describe in some detail a few alternatives that might realistically be chosen as exemplars in each area, and their linkages to issues in other thematic areas. A decision and planning period will follow, involving consultation within and between thematic areas, leading to presentations at the second plenary conference, which will describe a single exemplar that has been chosen for development in each area, and the way work on the exemplar will be divided across institutions. The remainder of the project will be absorbed in developing the chosen exemplar. (p. 21)

The consultation phase has now ended. Near-final plans were presented to the whole network at the Plenary in May 2005, and adjustments have been made accordingly.

This deliverable reports the plans that have now been set out for the remaining 25 months of the project. They are necessarily provisional, because they will be subject to two reviews (in 2006 and 2007) before they are completed.

Work has begun on several aspects of the planned programme. It is also reported.

Ethical issues affect the whole of HUMAINE, but rather than repeating essentially similar points in multiple deliverables, they will be handled coherently in a single document, D0o (Science and Society).

The following persons have contributed to the work reported in the deliverable:

Sarkis Abrilian, Vered Ahronsson, Noam Amir, Elisabeth Andre, Tanja Baenziger, Anton Batliner, Roddy Cowie, Cate Cox, Laurence Devillers, Ellen Douglas-Cowie, Dennis Hofs, Spiros Ioannou, Kostas Karpouzis, Loic Kessous, Jina Lee, Jean-Claude Martin, Margaret McCrory, Edelle McMahon, Ailbhe Ni Chasaide, Katri Oinonen, Catherine Pelachaud, Christian Peters, Mannes Poel, Amaryllis Raouzaiou, Marc Schroeder, Ian Sneddon, Stefan Steidl

The institutions that have contributed are:

QUB, DFKI, ICCS, UNIGE, Paris 8, UA, FAU Erlangen, CNRS-LIMSI , TAU, TCD, UT, USC

2 Brief overview of Workpackage 5 and the exemplar proposal

2.1 The field covered by Workpackage 5

Workpackage 5 is concerned with data and databases in the HUMAINE context. Two fundamental considerations impact on the way HUMAINE approaches the topic of data and databases. The first is the general domain of HUMAINE which is emotion-sensitive machine interfaces. The second is HUMAINE's commitment to data that represents the type and range of emotional colouring that characterise naturally occurring emotional states in a multimodal context. These two considerations do not always sit easily together. It is an ambitious principle to focus on relatively naturalistic data (breaking with a long tradition of using acted data), and it brings with it practical problems for people working on interfaces. Naturally occurring data is noisy, and it is much easier to develop machine interfaces on the basis of simplified but clean data. But it is necessary to go in this direction if emotional interfaces are to develop in appropriate ways, i.e. in ways which transfer from the laboratory to real applications. Hence the core issue and challenge for WP5 is to develop techniques relevant to eliciting and labelling everyday emotional data so that the data is sufficiently tractable to be used in the development of emotion-sensitive machine interfaces. As a way of addressing this fundamental issue, WP5 takes as a principle the need to consider and provide a range of data - from relatively manageable data (supportive data) to data chosen to highlight the challenges that the field faces (provocative data).

2.2 The research objectives

There are six key research objectives:

- *To discover and describe what there is in terms of existing relevant emotion databases, to disseminate this information, and to establish a small core of databases for shared use across the network pending the development of new HUMAINE data*

The development of databases in line with the requirements of HUMAINE will span the lifetime of the project, and so it makes sense as an interim measure to make existing databases more widely available. In particular it makes sense for researchers in other workpackages in HUMAINE to be working on the same core sets of data. Ideally these core sets should involve data that is as naturalistic as possible in line with the core thrust of HUMAINE.

It might seem like a simple research objective to describe and make available the existing databases but in fact it is quite problematic. There are three types of problem:

(i) The descriptors across databases are not consistent, and in many cases description is sketchy and incomplete (e.g. size of data, how emotion labels were assigned).

(ii) The more naturalistic datasets of the type that are of special interest to HUMAINE are in a small minority, and are usually not freely available because of copyright or ethical reasons.

(iii) The field is expanding quite quickly and so the exercise cannot be a static one-off exercise. There needs to be room for growth.

WP5 has made a first attempt to address this objective in a special deliverable (D5b) and continues to address it as part of the exemplar reported in this deliverable.

- *To develop a set of principles and standards that should underpin new databases of emotion*

It is rare to find principles and standards stated as part of existing databases of emotion. Databases often seem to have been put together on the basis of what is readily available (e.g. call-centre data). There is not much consideration of why particular emotions have been chosen or explicit discussion or explanation of what types of labels might be relevant to particular types of data. Ethical considerations or issues of availability to the wider research community are rarely discussed explicitly. There need to be principles to address these issues and standards governing the material collected (primary records) - what these records should include (e.g. what modalities, what assessments of the person's feelings), what techniques for establishing primary records should be used (e.g. ethological considerations, induction methods), what structure primary records should have (e.g. sample size, avoidance of confounding samples, factorial design, repeated measures), what format records should be in (e.g. concerning compression, unit size and name, platform, language conventions).

This objective is addressed in detail in D10c (Interim report on prospects for establishing standards). The principles and standards set out there underpin many aspects of the exemplar described in this deliverable and are referred to when relevant.

- *To develop methods for eliciting or acquiring data that represents the type and range of emotional colouring that characterise naturally occurring emotional states in a multimodal context*

Research related to this objective forms a major part of the exemplar (see below). Before HUMAINE there was very little concentration on naturally occurring emotional states in a multimodal context, and rather little focused work on developing appropriate methods for eliciting or acquiring this type of data. HUMAINE moves forward on this objective by adopting a principle of gradation, i.e. developing a graded variety of methods for elicitation/acquisition ranging from those that produce relatively controlled data in laboratory conditions (more tractable) to those that produce totally naturalistic complex data (noisy and difficult from a computing perspective). This allows us to explore the possibilities/limitations of laboratory induced data beside truly naturalistic data, and allows those developing algorithms to make progress and consolidate at each stage in a set of progressively more stretching targets.

- *To develop methods for labelling this data, both in terms of emotional content and signs of emotion*

This objective has two main parts: (i) labels for emotion, (ii) labels for some signs of emotion. Research related to this objective forms a major part of the exemplar.

In terms of labels for the emotional content of data, the aim is to develop labels appropriate to naturally occurring emotional states. Much prior work has focused on labels for acted or posed data and has used labelling schemes based on traditional divisions of emotion into supposedly 'primary' or 'basic' categories. It is evident that that type of labelling does not transfer. Labelling will be developed in line with the psychological literature on emotion (but

will not be afraid to challenge psychologists' assumptions if the data demands it). It will also be developed on a range of scales, from coarse to fine resolution, applying the same principle of gradation as in the development of methods for eliciting data. This allows engineers to use the data in varying ways, for varying purposes to achieve varying levels of accuracy.

In terms of labels for the signs of emotion, hand labelling on a large scale is not an option. There are two broad levels of sign - low order, automatically extracted features and high order, perceptually determined signs. The aim is to co-operate with WP4 on the lower order signs providing hand labelled data to validate automatic feature extraction developed in WP4.

WP5 will also address some higher order signs of emotion, developing some limited labelling of complex properties.

- *To explore how methods for elicitation and labelling can be applied cross-culturally*

There is very little work on cross cultural expression of emotion. The objective is to explore in a limited way how elicitation methods work across cultures and languages and how labels developed apply across cultural groups. This is taken up in the exemplar.

- *To demonstrate the application of elicitation and labelling techniques to a subset of data*

The objective is not to produce large amounts of data, rather to demonstrate the outcomes of the elicitation methods and to show how labels have been applied for good practice. This demonstration forms part of the exemplar.

2.2.1 Main elements of the exemplar

As stated in D5d, the exemplar will consist of a library of recordings that can underpin meaningful research into emotion as it appears in action and interaction in everyday life. The recordings will cover signs in visual, audio, and physiological modalities, and provide data across a range of cultures. The exemplar will incorporate descriptive schemes for labelling both the emotion expressed and the signs of emotion: these schemes will be applied to a subset of the data to show good practice.

There are a few refinements to the exemplar since D5d, and these are referred to as relevant below.

There are five core elements:

2.2.1.1 Development and selection of data collection techniques to be used in building the library

Two types of data will be used to build the library. The first is induced data. The second is wholly naturalistic data.

Appropriate induction techniques will be developed for emotion elicitation in controlled contexts. Three main techniques are proposed. They cover emotion in monologue, emotion in dialogue and emotion in action. The three induction techniques (described in full in D5d) are the Recall Technique piloted by the Tel Aviv team (emotion in monologue), the SAL (Sensitive Artificial Listener) Technique piloted by the QUB team (emotion in dialogue) and the Driving Simulator Technique piloted by QUB (emotion in action). All address a range of requirements for the type of data required by HUMAINE: they are based on observation of naturalistic data; they are multimodal (though they currently only involve two modalities each); they have undergone reasonable testing; they produce data that is reasonably tractable compared to wholly naturalistic data; and the data illustrates different levels of tractability (from emotion data produced in monologic situations to data in action). However all require substantial development.

Since D5d was delivered, it has become apparent that a new database of induced persuasive communication is needed specifically to address the needs of WP8. Exploratory work is under way to develop appropriate induction methods and collect a small dataset.

With regard to naturalistic data, several bodies of data are owned by members of the consortium. Work will focus on three main bodies: The Belfast Naturalistic Database, the EMoTV database and the Reality TV/Emotion in Action collection of recordings from the QUB team.

2.2.1.2 Collecting records

This element involves making the recordings that will eventually be annotated. Given the focus on developing appropriate induction techniques, resources will be focused on development and not on the collection of large amounts of data. Representative sets of data will be collected in a limited number of laboratories to cover a range of languages and cultures. Since D5d it has been decided that the SAL induction technique will be used for

cross cultural/language exploration and will be developed for Hebrew and records will be collected in Israel.

2.2.1.3 Developing labelling schemes for – (a) emotion content, (b) signs (c) context

Emotion content: Three main labelling schemes will be developed in line with the psychological literature on descriptors of emotion: verbal categorical labelling, broad dimensional labelling, appraisal related labelling. In addition (and since D5d) there will be labelling schemes to address presence/absence of emotion on a dimensional scale, whether the data appears acted or realistic and the degree of masking of emotion (also on dimensional scales).

Signs: The second part is developing schemes to label the signs of emotion. There are two broad levels of sign - low order, which it is realistic to envisage being extracted automatically; and high order, which will need to be perceptually determined for the foreseeable future. The aim is to co-operate with WP4 on the lower order signs, by providing hand labelled data that can be used to validate automatic extraction techniques developed in WP4. WP5 will also address some higher order signs of emotion, developing some limited labelling of properties that cannot be labelled automatically, particularly in areas that are of interest to ECA designers (particularly with respect to gestures and body movement) or to speech engineers, perhaps particularly in the area of synthesis (for example with respect to voice quality, intonation patterns, rhythmic patterns). A coding system for the auditory signs of emotion is being piloted at Belfast (Douglas-Cowie et al, 2003b) and will be developed. This categorises perceived auditory signs of emotion in naturally occurring data (e.g. paralinguistic signs, voice qualities) as perceived by trained phoneticians. The multimodal coding scheme which is being piloted at CNRS-LIMSI will be developed (see http://emotion-research.net/ws/summerschool1). This deals with multimodal signs of emotion (particularly gesture, body movement, torso position etc) in naturally occurring data.

Context: The third part is developing schemes to label the context in which the emotional data occurs. Pilot work on the naturalistic data collected by QUB (Belfast Naturalistic Database) and CNRS-LIMSI (EmoTV1) suggests that the expression of emotion is dependent on context at multiple levels and that appropriate labels need to be developed.

2.2.1.4 Delivering data labelled for – (a) emotion content, (b) signs (c) context

Given the focus on developing appropriate labelling schemes, the intention is to focus resources on development and not on labelling large amounts of data. A subset of data will be labelled to show good practice. Labelling will entail the mounting of the data on a platform, relevant programming to implement the labelling schemes and design of the whole to take into account sensible interrogation of the database by the users.

2.2.1.5 Delivering documentation for techniques and tools

Much of the effort in the database exemplar is focused on the development of appropriate induction techniques and on labelling methods. These will be described in the academic literature and at conferences as appropriate, but practical guidelines and information on how to use them will also be released so that they can be widely used throughout the research community.

2.2.1.6 Making recordings and tools available for use in the medium term

This part of the WP's function is not strictly part of the exemplar, but it carries comparable weight. Reviewers rightly raised the need for researchers across HUMAINE to be sharing the same data as soon as possible. A number of 'pilot' datasets have been distributed across the network or shared across specific groups for specific purposes under certain agreements. These focus on a range of relatively naturalistic material and are:

SAL pilot dataset: This dataset has been released to all partners under an agreement governing the use of the data. It uses the SAL induction method (see below) to elicit a range of emotional behaviour and consists of a pilot set of recordings used to test the concept. There are 4 speakers, with around 20 minutes of speech each and a range of non acted emotional behaviour. The data is audiovisual. It is labelled for emotion using FEELTACE which gives labels on two dimensions related to emotion (activation and evaluation).

AIBO dataset: This dataset has been made available under the CEICES agreement to a number of partners. It consists of unimodal (speech only) data and consists of 51 ten to twelve year old children communicating with Sony's Aibo pet robot. Partners are committed to using the database in a pilot experiment to compare recognition performances of different systems on the data. Under the agreement they have to exchange and share results. The HUMAINE teams using it are involved in WP4, specifically in emotion recognition from speech features (QUB, UNIGE, ITC, CNRS-LIMSI, UA, TAU).

CNRS-LIMSI EmoTV data subset / Belfast Naturalistic Database subset: These subsets are taken from two naturalistic audiovisual databases collected by partners. They are subject to copyright restrictions, but subsets are able to be released under certain terms of agreement. Two such sets are shared by CNRS-LIMSI and QUB for use in the development of labelling schemes and in cross cultural comparisons. The CNRS-LIMSI set is also being shared with Paris 8 as the basis for joint work on an ECA.

Reality TV/Emotion in Action datasets: There are two datasets in this category. Both are concerned with emotion in action. The first is from television Reality TV footage - of people taking part in testing outdoor activities - and is available under certain conditions to the QUB team and CNRS-LIMSI team for the development of labelling. The second dataset is being collected at QUB. It consists of pilot experiments aimed to elicit emotion in action in naturalistic contexts (e.g. in outdoor physical challenges such as mountain bike riding).It is audiovisual. Some initial data has already been shared and new data will be shared freely across the network. It is a source that is valuable for comparison with induced data in laboratory conditions, and is a rich source for ECAs (in terms of gesture and body movement).

UNIGE acted subset: This is a subset of a new audiovisual acted database created by UNIGE under another project. It has been shared with several partners from HUMAINE as an example of well acted data, and it has proved a useful comparator for data recorded from more naturalistic sources.

Complementing those activities, the exemplar will include a web page that provides links to other selected databases, assembled either within and outside HUMAINE – see forward to 4.3.

The FeelTrace labelling tool has also been distributed along with training material and instructions to all interested partners. This will be made more widely available via the Toolbox on the portal.

2.2.2 How the subtasks link to each other

The relation between the core elements of the exemplar are self explanatory: they move from development of methods and techniques to application. Within each element there is also a structure governed by a principle of gradation - moving from tractable to complex data and from coarse to fine resolution in terms of labelling. As explained above, this allows people to use the data for varying purposes and aiming for varying degrees of accuracy.

2.2.3 How the subtasks link to other aspects of HUMAINE

Databases are fundamental to HUMAINE as a whole, and are particularly central to WP4, WP6 and WP8. Different types of data and different levels are needed for these workpackages, and the exemplar is designed with this variety in mind.

WP4 (From Signals to Signs) essentially needs data that can be used to develop algorithms for automatic analysis of signals and test emotion recognition systems. The data that will be provided through the exemplar contains a sufficient range of data (from tractable to complex) to facilitate progress in the development of algorithms and a sufficient range of emotion labels (from coarse labels for emotion to fine) to test the capabilities of emotion recognition systems.

WP6 with its focus on ECAs has a particular interest in gestures. The existing naturalistic databases (Belfast Naturalistic Database and especially EmoTV1 which is partially labelled for high order gesture signs) can be used to inform work on gesture, but they are much too noisy for technical use. The induction techniques do not lend themselves easily to the capture of gesture – the scenarios involve sedentary interaction which does not give gesture movements of the type wanted. .

WP8 needs a rather specialised database of persuasive communication, and plans to develop that have been indicated above.

3 The planned program of research

3.1 Element 1: development and selection of data collection techniques

3.1.1 Leader: QUB

3.1.2 Main participants : QUB, TAU, CNRS-LIMSI, UT

3.1.3 Main steps planned towards producing element 1

Subtask	Carried out by	Start / end dates
Further refinement of Recall Technique for use in audiovisual and cross cultural settings	TAU	M25-30
Development of the SAL technique to move from prototype to semi automated Wizard of Oz presentation	QUB, UT	M20-24
Translation of SAL into Hebrew, testing & amendment	TAU	M18-24
Refinement and further development of SAL	TAU, QUB	M25-30
Developing mood induction techniques suitable for use in Driving Simulator experiments (driving in range of moods through range of tasks)	QUB	M18-27
Design of physiological measures to be used in Driving Simulator experiments	QUB, TAU, UA	M18-30
Development of techniques to induce persuasive communication for WP8 needs	QUB & ITC	M25-36
Development of techniques to collect Reality/Emotion in Action data	QUB	M18-30

3.2 Element 2: Collecting Records

3.2.1 Leader: QUB

3.2.2 Main participants : QUB, TAU, CNRS-LIMSI

3.2.3 Main steps planned towards producing element 2

Subtask	Carried out by	Start / end dates
Collection of audiovisual data using Recall technique	TAU	M30-36
Collection of audio data using Recall technique (in Irish)	TCD	M30-36
Collection of data using Wizard of Oz semi automated SAL	QUB, UT	M25-26
Collection of SAL Hebrew data using Wizard of Oz semi automated SAL	TAU	M30-36
Collection of more naturalistic data to extend the EmoTV1 database	CNRS-LIMSI	M25-36
Collection of QUB Driving Simulator data	QUB	M30-36
Collection of QUB Reality/Emotion in Action data	QUB	M30-36
Collection of Persuasive communication dataset	QUB/ITC	M36-42

3.3 Element 3: Developing labelling schemes for – (a) emotion content, (b) signs (c) context

3.3.1 Leader: CNRS-LIMSI

3.3.2 Main Participants : CNRS-LIMSI, QUB, FAU-Erlangen, UNIGE, ICCS

3.3.3 Main steps planned towards producing element 3

Subtask	Carried out by	Start / end dates
Development of emotion labelling (FeelTrace, ETrace, MaskTrace; appraisal rating; agreed verbal categories)	QUB, CNRS-LIMSI, UNIGE	M18-36
Development of context labelling	CNRS-LIMSI, QUB	M18-36
Development of high order speech labels	QUB, TCD	M18-36
Development of high order gesture/body movement labels	CNRS-LIMSI, Paris 8	M18-36
Hand labelling to allow validation of automatic extraction techniques	QUB, ICCS	M18-36
Development of physiological data analysis and labeling of physiological data	QUB, UA, FAU-Erlangen	M30-36

3.4 Element 4: Delivering data labelled for (a) emotion content, (b) signs, (c) context

3.4.1 Leader: UT

3.4.2 Main Participants: UT, QUB, CNRS-LIMSI

3.4.3 Main steps planned towards producing element 4

Subtask	Carried out by	Start / end dates
Labelling of subsets of data	QUB,CNRS-LIMSI	M37-42

Development of technical platform for delivery	UT, DI-BARI	M30-42
The data will be made available via the portal.	DFKI	M42-48

3.5 Delivering documentation for the techniques and tools so that they can be used throughout the research community

3.5.1 Leader: QUB

3.5.2 Main Participants: QUB, CNRS-LIMSI

3.5.3 Main steps planned towards producing element 5

Subtask	Carried out by	Start / end dates
Preparation and delivery of instructions for Recall Induction technique	TAU	M18-42
Preparation and delivery of instructions for SAL Induction technique	QUB, UT	M18-42
Preparation and delivery of instructions for Mood Induction Techniques (as used in Driving Simulator)	QUB	M27-42
Labelling techniques: instructions for users	QUB, CNRS-LIMSI	M18-42

3.6 Making recordings and tools available for use in the medium term

3.6.1 Leader: QUB

3.6.2 Main participants : QUB, CNRS-LIMSI, FAU-Erlangen

3.6.3 Main steps planned

Subtask	Carried out by	Start / end dates
Preparation of pilot SAL data for distribution (labelling, user agreement, transfer to cd) Delivery of pilot SAL data to all partners	QUB	M13-16 M17
Release of AIBO dataset under CEICES agreement to QUB, UNIGE, ITC, CNRS-LIMSI, UA, TAU To other partners as requested	FAU-Erlangen	M18-M48
Release of EmoTV& Belfast Naturalistic Database subsets between CNRS-LIMSI and QUB (for labelling development) and between CNRS-LIMSI & Paris8 (for ECA development)	QUB, CNRS-LIMSI	M18
Release of subset from TV Reality dataset (to CNRS-LIMSI for labelling development) Release of QUB pilot Reality/Emotion in Action recordings	QUB	M24 M18-48
Release of UNIGE acted database samples	UNIGE	By agreement
Release of FeelTrace labelling tool to partners on request	QUB	M1 onwards

3.7 Steps to ensure co-ordination

Method Used	Partners/Place	Time	Function
Interchanges under HUMAINE	QUB to CNRS-LIMSI TAU to ICCS	December 2005 2006	Detailed work on emotion labelling Collaboration on multimodal analysis (video and sound) and recognition of emotion

			using the SAL database
Workshops at conferences	Interspeech 2005 LREC 2006	September 2005 May 06	HUMAINE organised workshop on Emotion and Speech in a Multimodal Context with focus on databases– integrated HUMAINE participants and wider research community HUMAINE organised workshop on Emotional Databases– aim to integrate HUMAINE participants and wider research community
Other meetings at major conferences	e.g. Interspeech 06, 07, ACII 07, LREC06, 07		To meet up and share work achieved
Short visits/more regular meetings between geographically related partners	Between QUB & TCD Between CNRS-LIMSI & Paris8	06 06	Work on higher order speech signs & Irish data Work on ECA driven by EMoTV
Phone conferences	Key WP5 partners	Every 6 weeks throughout project	To ensure progress on targets
Specified HUMAINE meetings	Meeting with other WP key researchers Plenary	06 07	To ensure common goals of HUMAINE

3.8 Steps to ensure dissemination

Dissemination across the network and beyond the network to the wider research community will take place in a number of ways:

- Special sessions & workshops led by HUMAINE members at major international conferences
- Publications in conference proceedings and journals
- Via the portal – especially through the Toolbox (where labelling tools and induction techniques will be made available), and through the interactive database (as part of WP1) summarising available data.

4 Research achievements to date

4.1 Achievement 1: development and selection of data collection techniques

4.1.1 Development of the SAL technique

4.1.1.1 Prototype remote version of the SAL system.

Twente are developing a version of the SAL system which advances it in two main ways. First, it allows the voices of the 'sympathetic listeners' to be generated from recordings rather than by an operator reading expressively. Second, it allows the human operator to be removed to a remote location, creating a true 'Wizard of Oz' system. The key advantages gained by these developments are that the vocal skills or the operator are no longer crucial, and that it is no longer necessary for the operator to be present in the same room as the user (the presence of the 'experimenter' acting as the SAL persona tended to inhibit the emotional freedom of the user).

Twente have produced a specification for a system with three components: a central SAL server, a client at the respondent side and a client at the operator side. The system will have an interface for the respondent and the interface for the operator. An operator can log on to the central server any time to show that s/he is available for interaction with a respondent. A respondent can log on to the central server any time s/he wishes to interact with the system. The server keeps track of available operators and arranges connections between operators and respondents. Session data will be logged at the respondent side to avoid large amounts of recorded audio and video being transmitted over the network.

Video recordings are saved in maximum uncompressed quality to a tape in the digital video camera and they are captured to smaller size compressed mpeg videos on the respondent machine. Audio recordings from the respondent microphone are captured to high quality wav files on the respondent machine and added to a lower quality audio track in the mpeg file. The lower quality video with audio track is transmitted over the network to the operator machine.

The operator selects wav files and may speak into the microphone. The audio (from the wav file or microphone) is transmitted over the network to the respondent machine where it is played back and added to an audio track in the mpeg file. Thus the mpeg file contains all recorded video and audio from the entire session. In addition the operator sends messages to the respondent to indicate which wav files were selected at what time and when the respondent state or the speaker was changed. This data is logged into a script that contains all selected SAL utterances, speakers and respondent states with times.

During the SAL experiment as much data as possible will be collected for further automated or manual analysis, for example facial emotional expressions and speech. SAL data will for the time being be saved in a database of audio tracks. SAL utilities (replies and time chosen by operator, emotion state of respondent) are synchronized.

Further technical specifications are outlined in 'The Remote Sal System' documentation provided by Twente (Appendix 1).

4.1.1.2 Translation of SAL for use in different cultural/linguistic settings, and plans for refinement of SAL

TAU has translated SAL into Hebrew, explored the implications of translation into another language and for a different cultural group and carried out some pilot testing.

The English language version of SAL includes numerous expressions that are specific to Anglo/American culture and of course to the English language itself. Many of these cannot be translated word by word. In order to come up with a Hebrew language version that would be appropriate for young to middle aged Israelis, SAL was initially translated separately by three students (undergraduates in the Department of Communication Disorders, Tel Aviv University) in their twenties. It was interesting to observe that several difficulties cropped up at this phase, since some of the expressions were totally foreign to them. After some consultation with the local HUMAINE coordinator (Dr. Noam Amir) and the developers of SAL on such issues, these three students came up with three complete translations. They then compared and discussed their different versions, and submitted a preliminary version to Dr. Amir. After going over the translation and contributing a series of comments, a final version was hammered out.

The same three students then started to conduct some pilot experiments. As experimenters, they found SAL somewhat awkward in getting used to. In contrast to the QUB experimenter (Cate Cox), they are not researchers, with far less of a background in psychology and emotion research methods, and therefore required some time in getting into the appropriate frame of mind, without having too much of a feeling of being in an artificially contrived situation.

After several pilot sessions, the experimenters still reported some difficulties. The first was in finding an appropriate response rapidly enough in order to keep the conversation going, without too many awkward pauses. The second difficulty was that they found themselves severely constrained by the list of responses in the SAL protocol. Often they found themselves searching for an appropriate response to a specific issue, but couldn't in fact find one. Interestingly, when viewing the taped sessions, they do not appear to be failures. Sometimes the experimenter's responses provoked unexpected laughter from the subjects, where they were somewhat incongruous, but there were definitely segments in which true "emotional coloring" could be observed.

This is where things stand at the present. They are currently awaiting a digital video camera that should arrive in the coming weeks, to tape several more sessions at higher quality audio and video. These will comprise an initial SAL Hebrew database.

In parallel, they are discussing means in which to improve the experiment. Though SAL is loosely modelled after the ELIZA program for conducting typed conversations with a computer, it lacks several degrees of freedom that are present in ELIZA. Since ELIZA is text based and run by a computer, it can respond very rapidly and insert parts of the subject's dialog into its responses. They are currently examining methods of doing this while still keeping the protocol as simple as possible. When they come up with a preliminary version we will discuss it with other groups using SAL (QUB and NTUA), and decide whether to go along directly to trying it out. Currently they expect to accumulate useful extracts of SAL of 1 to 2 hours in length, and have as much of them FeelTraced as possible. These will serve as raw material to be analyzed for speech and facial expressions by WP4.

4.1.2 Development of mood induction techniques (for use in Driving Simulator experiments)

Driving Simulator experiments are being developed to look at the influence of emotion on a range of driving tasks. Physiological measurements will be recorded for the subjects taking part. In order to set up the experiments, subjects need to be induced into a range of moods prior to the particular driving tasks they are asked to perform. The development of appropriate mood induction techniques for this purpose is taking place at QUB.

The development of techniques for the induction of emotion and emotion-related states has received a good deal attention in the psychological literature over the last three decades. A wide variety of techniques have been developed, including standard classical mood induction procedures (MIPs) such as the Velten method (Velten, 1968), techniques employing stories or films (Gross & Levenson, 1995), music (Clark et al., 2001; Kenealy, 1997), or pictures (Öhman et al., 2001; Ito et al., 1998; Lang et al, 1993) and, more recently, techniques employing new technology such as computer games (van Reekum, et al., 2004).

Reviews and meta-analyses (Westermann, et al., 1996; Gerrards-Hesse, et al., 1994; Martin, 1990; Clark, 1983) have generally reported that Film and Music MIPs are the most effective, especially when participants are instructed to try to experience the emotional state being induced. Choosing suitable MIPs presented the QUB team with two challenges. The first was to investigate which technique or combination of techniques would induce states that were robust. The team felt that in order to induce emotional states of sufficient intensity and duration in the laboratory, it would be necessary to find ways to reinforce the induced emotional state. They tested different combinations of mood induction procedures and developed some ideas of their own, including an interview technique intended to facilitate participants moving into a particular emotional state. The second challenge was to consider how they would adapt induction procedures for use in the driving simulator. For example Film, although widely regarded as one of the most effective MIPs, would not necessarily be of use for practical reasons.

Focus has been narrowed to three emotional states: happiness, sadness, and anger. The first pilot study (N=4) developed the interview/discussion technique which aimed to induce anger. Two experimenters were used. Experimenter 1 greeted the participant and fully briefed him/her with regard to the aims of the experiment, ethics, his/her rights, informed consent and data protection, etc., following a carefully prepared script. Experimenter 2 was then introduced and began the first stage of the actual induction, which was to compile a list (with the help of the participant) of situations, people, or things that made the participant angry. Using this information, Experimenter2 then proceeded to lead a provocative discussion on these topics. After 30 minutes Experimenter 2 left the room and Experimenter 1 returned. The participant rated his/her mood using the Self Assessment Manikin (SAM) (Lang, 1980) and was then fully debriefed. The entire procedure was captured on digital video tape, although the semantic content of the discussion was masked using a filter. While the video data clearly showed that the participants did become angry and various stages of the discussion, the SAM results and verbal feedback indicated that the participants finished up in a positive mood. This would indicate that they experienced a sort of cathartic process whereby they ‘got it off their chest’ and felt better as a result.

The team then looked at ways of combining existing MIPs to induce emotional states. They carried out a small study (N=5) using the Velten technique (Velten, 1968) along with music. Gerrards-Hesse, et al., (1994) provides a list of classical pieces which have been found to be effective in evoking emotion. From this list they chose Delibes’ *Coppelia*, used in 4 studies to evoke elation. Participants first did some relaxation exercises in order to achieve a state of

calm. They then underwent the Velten procedure to induce elation and immediately afterwards listened to *Coppelia*. Participants filled in a SAM both before and after the induction. In addition to the SAM, participants were allowed to choose up to 3 words from a list of categorical labels (Douglas-Cowie et al., 2005; Cowie et al., 2001). The SAMs revealed a definite shift to a positive mood state and participants reported that the Velten predisposed them to be affected by the music.

The QUB team are currently working on the design of an experiment to compare a combination of several MIPs (Velten, music, and discussion of emotive topics) with film (which has been found to be particularly effective – see Westermann, et al., 1996; Gerrards-Hesse, et al., 1994; Martin, 1990; Clark, 1983). In a preliminary survey, they asked people to list films, pieces of music and events in the news/recent history that made them feel either happy, sad, or angry. The news items section has been particularly informative and the resulting list has been used to generate emotive topics for discussion in the a full induction experiment. Full results should be available by the end of the year, but preliminary indications are encouraging.

4.1.3 Development of Emotion in Action/Reality Television-type elicitation & recording techniques

A group at QUB have developed techniques to elicit emotion in a natural situation and have carried out a pilot study. They have been specifically concerned with how context influences expressed emotion. Understanding this is arguably essential for understanding and recreating emotional expression. In order to gain a long term understanding of the processes involved, they attempted to identify the most important variables which would help generalize from one situation to another.

Their interest lies not in the classic emotional states (e.g. anger, sadness, etc.), but in the much more subtle emotional colour prevalent in almost all behaviour. The subtlety and ambiguity of such behaviour requires sophisticated techniques to sample and measure fluctuating emotional temperament.

Currently, there are technical (and ethical) issues with existing footage of naturalistic behaviour. Very few databases of non-sedentary naturalistic behaviour focus on an individual for long enough to allow us to measure the frequency of different behaviours. The major motivating factor in the QUB group gathering their own data was to provide a statistically valid means of sampling behaviour.

A group of 4 female friends participated in a series of outdoor physical challenges. Behaviour was recorded using 2 hand held cameras, and also a helmet-mounted camera aimed at participants' faces. In a pre-organized series of 10-minute sessions, each camera focused on either the face or body of the participants. The resulting footage comprised samples of behaviour from the group, both when participating in the tasks and observing the challenges.

Preliminary observations suggest that individuals observing real events may behave quite differently to those watching a movie (a common paradigm for eliciting emotion in the past). Scherer & Ceschi (1997) have argued for the need to change the span of attention when analyzing emotion. This is an important observation in that we found the speed of expressed emotional transitions to be extremely fast. Subsequent issues to be addressed are:

- Why do we express so many fleeting emotional states?

- Are we capable of detecting all the subtle nuances in such rapid sequence?
- Are some individuals better at detecting the subtleties than others?

As the brain processes environmental stimulation it tracks changes rapidly, e.g. surprise may give way to the more negative fear/horror, which in turn may be replaced by relief and humour. Although it is not yet certain that ECA's will need to be capable of producing these rapid transitions and blends in order to be convincing, it is certain that machine systems will need to be capable of interpreting such subtleties in order to make sense of human emotion.

Study of the context in which emotional expression occurs can seem a daunting task. The infinity of possible contexts begs the question "Where do we start?" There are serious limitations on what purely descriptive analyses can provide. We can analyze frequency and nature of transitions from one emotion to another; also analyze relative frequency of emotionally expressive elements such as smiles, eye movements, body gestures, etc. However the pattern of data is only relevant for a particular group in a specific context. If any one of a huge number of variables (e.g. mood, gender, task, etc.) is changed, then the pattern will change too - and we are back to our infinity of contexts. One way forward is to begin to construct a matrix of contextual variables which would facilitate comparison of the relative frequency, form and duration of behaviours across different situations.

Short sequences of natural behaviour from the present study were transformed into a series of still photos (80ms and 240ms frame interval) and compared with similar photo sets from video examples of acted behaviour (TV and Geneva database). The graph below shows the results of a 'FeelTrace' type analysis of the photos. Figures refer to the total level of change measured between adjacent frames. It is clear that the Natural photos show a greater degree of change between photos, indicating more rapid transitions in natural than in acted behaviour.

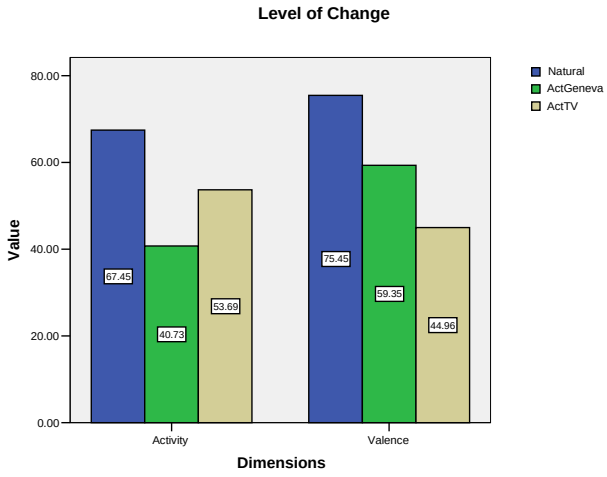


Figure 1. Comparison of acted and natural behaviour : level of change between adjacent frames

The participants in this study reported being so absorbed in the tasks that they forgot about the cameras. There were, however, differences in behaviour depending on whether or not they were interacting with each other. The filmed data highlights this important distinction between observed and unobserved behaviour. In one task, it was clear from the helmet camera that the participant was quickly switching from 'putting on a brave face' to 'fear and anxiety' as other participants turned toward or away from her. A future requirement is to ensure there is a continuous record of both face of the individual and what he/she is looking at.

The social mix in any situation – whether interaction is between friends or strangers - is also important. This may seem obvious, but the detail of how such differences translate into the subtleties of emotional expression is by no means obvious.

Future directions are to identify tasks/situations which will provide a fairly standardised set of experiences. Variables to be manipulated include observation conditions and social mix; appraisal type variables such as novelty and goal conduciveness; also formality of situation and level of competitiveness.

When focusing on the emotional colouring which pervades behaviour we seem to be dealing with a more ambiguous landscape than that of clear cut emotional episodes. Emotional peaks of anger, etc. are clear to see; however in the lower reaches of the landscape the task of disentangling background emotional colour from mood, stress, etc. is more difficult. In addressing these issues, the task of emotional analysis will need to interface with the various literatures of Sociolinguistics, Social Psychology, Anthropology, etc.

4.1.4 Participants

QUB, TAU, UT

4.1.5 Output (demonstrations, resources, etc)

- SAL prototype + description (see Appendix 1)
- Mood induction package
- Posters and demonstrations, see http://emotion-research.net/ws/wp5

4.1.6 Follow-up in progress

- SAL – implementation program by UT
- SAL – further experimentation by TAU
- Mood induction – further development
- Reality data – further development

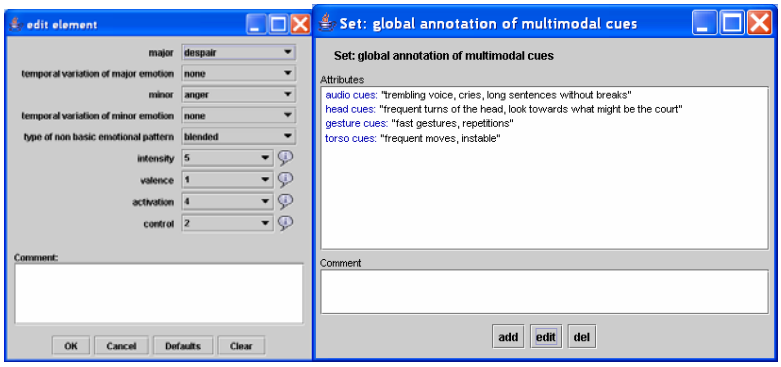
4.2 Achievement 3: Development of labelling schemes

4.2.1 Databases Labelled

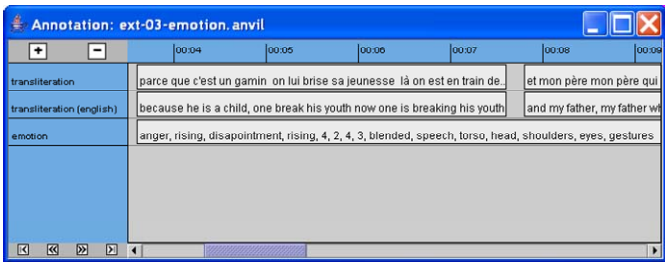
Work on labelling has used four main collections, the CNRS-LIMSI EmoTV material; the reality TV database at QUB; the Belfast Naturalistic database (also QUB); and the AIBO database (at Erlangen). These are summarised below. Several other significant databases are held by HUMAINE partners, and it is a medium term goal to integrate lessons from all of them in a cohesive way.

4.2.1.1 EmoTV

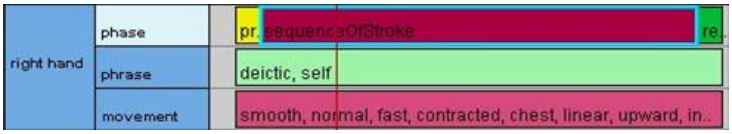
The most systematic development has taken place at CNRS-LIMSI using the EmoTV database. This is a naturalistic database of emotional TV audiovisual interviews, which covers a wide range of positive and negative emotions and of emotional intensities. It consists of 48 videotaped people in 51 videos. Coding schemes have been defined for annotating the observed emotional behaviours at multiple levels of abstraction and temporality via a discrete approach based on the definition of emotional segments and the ANVIL tool (Kipp 2004). The coding schemes combine verbal labels, abstract dimensions (intensity, activation, self-control and valence), appraisal dimensions and multimodal behaviors (Figure 2).



(a) Annotation at the global level of a whole video: emotion labels and dimensions (*left*), free text high-level multimodal cues (*right*)



(b) Annotation at the local level of a non-basic emotion segment by one of the coders with a combination of 2 categorical labels (anger and disappointment), classical dimensions (intensity, valence, activation, control) and emotionally relevant modalities



(c) Lowest level of annotation: time-based annotation of behaviors in several modalities including gesture expressivity (smooth, fast, …)

Figure 2. Multilevel annotation of emotional behavior in the EmoTV corpus

Several coding schemes have been defined and applied to subsets of the EmoTV corpus in order to explore different directions with respect to the annotation of real-life emotional behaviors (Table 1).

These various coding schemes and annotations have been used for computing representations of real life multimodal emotions at several levels of details such as soft vectors representing the contribution of annotated dimensions (verbal label, modality activation, expressivity profile) to the perception of the complex emotional behaviors.

Table 1. Coding schemes applied to subsets of EmoTV corpus

Number of EmoTV clips	Number of coders	Coding scheme	Reference
51	2	Coding Scheme Emotion and Context (V1): One verbal label per segment (14 verbal labels obtained from 176 free labels), comparison between audio only, video only and audio-video annotation, context annotated.	S. Abrilian, L. Devillers, J-C. Martin (a) (HCII 05)
36	2	Coding Scheme Multimodal (V1) Multimodal behaviors (speech, head moves, facial expressions, gestures, torso movements)	S. Abrilian, L. Devillers, J-C. Martin (b) (HCII 05)
11	3	Coding Scheme Emotion and Context (V2): Two verbal labels per segment (18 verbal labels, 3 macro classes), type of complex emotion (blend, masked, sequential), high level multimodal cues, temporal description of intensity variation per segment, context annotation refined.	L. Devillers, S. Abrilian, J-C. Martin (ACII 05)
3	1	Coding Scheme Multimodal (V2) Refinement of the previous multimodal coding scheme	J-C. Martin, S. Abrilian, L. Devillers (ACII 05)
1	40	Coding Scheme Emotion and Context (V2): Perceptive test, 2 labels per segment, type of complex emotion	LREC 06 submitted

4.2.1.2 Reality TV dataset

Reality TV footage held by QUB has been used to explore continuous labelling. The footage is of people taking part in outdoor activities in exacting conditions. It has been labelled in a pilot experiment using a first version of EmoTrace, a program which allows a user to record in a continuous fashion the degree of emotional content perceived in long stretches of data.

4.2.1.3 Belfast Naturalistic database

The Belfast Naturalistic Database had been labelled prior to HUMAINE using the 2 dimensional continuous labelling tool FeelTrace which allows listeners to trace the perceived emotion of an episode in continuous fashion on two dimensions (activation and evaluation). It has been used to look at compatibility with EmoTV.

4.2.1.4 AIBO

The AIBO database (FAU Erlangen) has been made available through the CEICES initiative and is central to other projects (such as the work on entropy mentioned in the following section).

4.2.2 Development of emotion labelling tools

Work on emotional labelling has been developing on a number of levels. In line with the approach outlined above, we have been working with naturalistic data (on samples from the EmoTV database and the Belfast Naturalistic database) and with a graded approach to labelling (from coarse labelling to fine resolution).

A suite of programs has been developed to deal with a labelling approach that is broad brush. These are part of the FeelTrace family which was instigated prior to HUMAINE, but which forms the basis for the new suite of programmes.

FeelTrace itself is a computer programme which allows listeners to trace the perceived emotional content of an episode in real time on two broad dimensions – evaluation (from negative to positive) and activation (from active to passive). The user moves a mouse round a circle which is composed of two dimensions and the co-ordinates of the mouse are recorded at regular intervals. Work has been done in tidying up FeelTrace and it will be released via the portal (Toolbox).

The other programs which have been developed in this vein are: EmoTrace, MaskTrace, ActTrace. Each consists of one dimension, but the same method applies as in FeelTrace – users trace the perceived presence of emotion (strong, weak, emotionally coloured) in EmoTrace, the perception of whether the data is acted (in ActTrace) and the perception that the speaker is masking emotion in MaskTrace. These three types of emotional label were considered necessary when looking at naturalistic data. Much naturalistic data is not at emotional extremes, indeed much of it does not seem to be particularly emotional at all, and so it is sensible to explore this unknown territory of presence of emotionality in naturalistic data and degree of emotionality. Some naturalistic data from TV shows (as in the EmoTV database and the Belfast Naturalistic one) may also have elements of acting in it (perhaps particularly in chat show material) and ActTrace is a tool that allows exploration of listeners' perception of the emotion in such data. Much of the data is the Reality TV database to which QUB has access appears to be masked – listeners feel that the speaker is emotional but is trying to mask that fact. MaskTrace is an attempt to address that issue. These programs are

currently being refined and validated. Appropriate scales are being developed on the dimensional representations.

Side by side with this broad brush labelling, there has been development of finer resolution labelling in terms of everyday verbal labels and a set of appraisal descriptors. This work has been mainly carried out on the EMoTV database. Batliner has also developed a verbal labelling scheme to apply to the AIBO robot data. Two key issues are being addressed (i) what labels are appropriate to naturalistic data and (ii) how can we reduce them to a reasonable number.

The work on EmoTV suggests that the nature and interaction of labels appropriate to naturalistic data is quite complex and quite different from some of the traditional work on labelling emotion in acted or more artificial data.

Two raters used the Anvil tool (Kipp 2004) to annotate all the videos of the EmoTV database for perceived emotions in three conditions: 1) Audio only, 2) Video only and 3) Audio-visual. As a first step towards finding an appropriate set of emotional labels, the two annotators labelled the emotion they perceived in each emotional segment by selecting one label of their choice (free choice). This resulted in 176 fine-grain labels. These were classified into 14 broader categories. But even after the reduction to 14 classes, the inter-coder agreements on emotion labels were low. The kappa statistics were: 0.37 for audio and video (281 segments), 0.43 for video only (295 segments), and 0.54 for audio only (181 segments).Contrary to expectation, agreement was lowest in the multimodal case. That is a first indication that the relationship between modalities is not straightforward and that the nature of naturalistic data is quite complex. Close analysis showed that the low kappa measures do not appear to be due to bad labelling but are rather a reflection of the nature of real-life data. A large part of the data consists of blended emotions and contradictory multimodal cues, by example cry to bring relief, “looking contented” despite a deception. Different modalities are often – but not always –associated with different aspects of a complex state (for instance tears with signs of relief, a positive face with deeply negative words).

The prevalence of these complex states is critical both for approaches to labelling and for understanding the contributions of different channels, and so the issue was followed up in a second study on only audio-visual condition. Three labellers used a more sophisticated scheme, one of whose elements was describing complex global pattern of emotion in terms of five categories – Single label, blended, masked, sequential (very rapid transition between overt patterns of expression, which suggests that both underlying states are in some sense present) and cause-effect (one kind of emotional event, eg crying, leads to another, e.g relief). On that criterion, the proportion of segments with no agreed label is 21%. The key point is that segments rated as showing a specific non-basic pattern (33%) were nearly as common as segments showing a pure emotion (46%). In that situation, asking raters to assign a single label is not only unlikely to yield agreement: it misrepresents the situation, and pre-empts key questions about the roles of different modalities.

This work is being followed up in an interchange between QUB and CNRS-LIMSI in order to develop an appropriate set of labels and to reduce the labels to a meaningful number that capture the essence of emotion in naturalistic data. Scherer’s set of appraisals is also being applied to the data to see how it fits with naturalistic data.

The QUB and CNRS-LIMSI teams will discuss the issues that arise in the co-ordinated labelling exercise, identifying the successes and difficulties that arise from it, and allowing

agreement on a homogeneous framework for describing both databases.. Key subtasks are 1) to assess the scale of cross cultural issues in dealing with naturalistic emotion , 2) to achieve agreement on a set of everyday emotion term that are suitable for use as a default set in labelling naturalistic data 3) to compare and eventually combine other elements of the labelling schemes, including discrete/continuous coding schemes for valence/activity, and appraisal dimensions.

On a different level, FAU Erlangen has developed an entropy – based tool for comparing ratings of databases, and more specifically for comparing human and automatic labellings.
The traditional method of evaluating an automatic labelling system is to use (classwise averaged) recognition rates. That approach is subject to several problems.

- a (hard) reference needed for it to be applicable;
- scores are dependent on the number and the similarity of the classes
- confusions of similar emotions are as wrong as confusions of totally different emotions
- scores are extremely sensitive to the ease or difficulty the of the reference material

The FAU Erlangen tool makes it easy to adopt a more satisfying approach, which is to compare the entropy of a machine human with the entropy of a comparable human labelling.

4.2.3 Development of labelling of Signs

4.2.3.1 High level speech signs

At QUB Steffi Stronge is developing a set of higher order prosodic descriptors of emotional speech as her PhD thesis. She has applied these to two subsets of data: a naturalistic dataset from the Belfast Naturalistic Database and an acted dataset based on the naturalistic one. She is working with descriptors under 6 main headings – paralanguage, pitch, voice quality, timing, articulation and intensity.

4.2.3.2 Gestures and application to ECA

CNRS-LIMSI have been developing their gesture/body movement labelling scheme with application to the EmoTV database. This scheme has been used in work in progress by CNRS-LIMSI and Paris8 on an affective Embodied Conversational Agent as part of WP6.

CNRS-LIMSI has grounded the coding scheme on requirements collected from both the parameters known in the literature as perceptually relevant for the study of emotional behavior, and the features of the emotionally rich TV interviews. They defined the coding scheme at an abstract level and then implemented it as a XML file for use with the Anvil tool. Each track is annotated one after the other while playing the audiovisual clip (e.g. the annotator starts by annotating the 1st track for the whole video and then proceeds to the next track). In the videos only the upper body of people is visible. Torso, head and gestures tracks contain both a description of pose and movement. Pose and movement annotations thus alternate. The direction of movement, its type (e.g. twist vs. bend) and the angles can be computed from the annotations in the pose track. Movement quality is annotated for torso, head, shoulders, and hand gestures. The attributes of movement quality that are listed in the studies reported in the previous section and that we selected as relevant for our corpus are: the number of repetitions, the fluidity (smooth, normal, jerky), the power (soft, normal, hard), the speed (slow, normal, fast), and the spatial extent (contracted, normal, expanded). A first annotation phase revealed that such sets of three possible values for each expressive parameter were more appropriate than a larger set of possible values. The head pose track contains pose attributes adapted from the FACS coding scheme: front, turned left / right, tilt

left / right, upward / downward, forward / backward. Head primary movement observed between the start and the end pose is annotated with the same set of values as the pose attribute. A secondary movement enables the combination of several simultaneous head movements which are observed in EmoTV (e.g. head nod while turning the head). Facial expressions are coded using combinations of Action Units. As for gesture annotation, CNRS-LIMSI kept the classical attributes but focused on repetitive and manipulator gestures which occur frequently in EmoTV. The coding scheme enables the annotation of the structural description ("phases") of gestures as their temporal patterns might be related to emotion: preparation (bringing arm and hand into stroke position), stroke (the most energetic part of the gesture), sequence of strokes (a number of successive strokes), hold (a phase of stillness just before or just after the stroke), and retract (movement back to rest position). The following set of gesture functions ("phrases") were selected as they revealed to be observed in the EmoTV corpus: manipulator (contact with body or object), beat (synchronized with the emphasis of the speech), deictic (arm or hand is used to point at an existing or imaginary object), illustrator (represents attributes, actions, relationships about objects and characters), emblem (movement with a precise, culturally defined meaning). Currently, the hand shape is not annotated since it is not considered as a main feature of emotional behavior in our survey of experimental studies nor in our videos. Direction of movement for shoulders is also annotated as some of them are observed. Whereas the annotations of emotions have been done by three coders and lead to computation of agreement, the current protocol used for the validation of the annotations of multimodal behaviors is to have a second coder check the annotations done by a first coder followed by brainstorming discussions. CNRS-LIMSI has started the validation of the multimodal annotations by the automatic computation of inter-coder agreements from the annotations by several coders.

The data extracted from the manual annotations are used to compute a model of multimodal emotional expressive behavior along three dimensions: emotion, activation of head/torso/hand, and gesture expressivity.

4.2.3.3 Automatic labelling for WP4

In the work on labelling signs that has been described so far, the focus has been on assigning labels by hand. One of the key functions of that work is to provide ways of validating automatic annotations (see deliverable WP4, LREC 2006 submitted). Work has also been done on the development of automatic annotation methods. That task overlaps with WP4. Techniques for recovering facial features and gestures were developed in related projects. The main developments in the context of HUMAINE have dealt with the movements of other body parts, particularly hands.

Image processing of TV videos : challenges

The task of locating body regions in image sequences is based on detecting continuous areas of skin color. According to the representation of color distribution in certain color spaces, current techniques of skin detection can be classified into two general approaches: nonparametric and parametric. For the given application, a very coarse and simple model is sufficient, since there is no need for detailed interpretation or recognition of precise gestures. An important issue is that the videos contained in the specific corpus are of relatively low resolution and, therefore, the skin regions are relatively small and possess very low detail; in addition to this, color resolution and fidelity can suffer from analog to digital conversion (e.g. in the case of digitizing a VHS tape). As a result, skin detection must be performed after a user-assisted initialization step, where the system suggests possible skin regions to be approved by the annotator; after that, since lighting and color conditions do not usually change within the clip, detection and tracking are performed automatically.

Another usual impediment to this process is the fact that camera movement can be uncontrolled and may result in skin regions moving abruptly within a clip without the subject showing the relevant activity. In our approach, this can be tackled by taking into account the change of the relevant positions of the skin regions, since they will not change in the event of sudden camera movement.

Algorithm for processing TV videos

Processing algorithms and modeling of hand and head activity depend primarily on the intended application and the requirements of the specific material. Human hand motion is highly articulate, since hands consist of many connected parts that lead to complex kinematics. At the same time, hand motion is also highly constrained, which makes it difficult to model. In our approach, a skin color probability matrix is computed for each given frame by calculating the joint probability of the Cr/Cb image values and thresholded to provide the skin color mask. Possible moving areas are found by thresholding the difference pixels between the current frame and the next, resulting to the possible-motion mask. This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color and motion masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. All described morphological operations are carried out with a disk structuring element with a radius of 1% of the frame width. The distance transform of the color mask is first calculated and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation.

In the following, the moving skin mask is created by fusing the processed skin and motion masks, through the morphological reconstruction of the color mask using the motion mask as marker. The moving skin mask consists of many large connected areas. For the next frame a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In the case of hand object merging and splitting, e.g. in the case of clapping, a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand is established.

The final processing step is to calculate a measure of motion or activity in the video. In the proposed approach, since we are not interested in gesture recognition, there is no need to employ detailed recognition procedures and use specific area and position information. As a result, the measure of activity in subsequent frames is calculated as the sum of the moving pixels in the moving skin masks, normalized over the area of the skin regions. Normalization is performed in order to discard the camera zoom factor, which may make moving skin regions appear larger without actually showing more vivid activity.

4.2.4 Development of Physiological labels

To combine efforts for processing of physiological user states, Augsburg University organized a workshop at Augsburg University from Oct. 10-11 2005. Humaine members from FAU, ICCS-NTUA, QUB and FAU participated in the workshop where the following activities were conducted:

Hands-on-Training

During the workshop, participants conducted an experiment with a car simulator provided by FAU where one subject was connected to multi-channel biosensors to record electromyogram, electrocardiogram, skin conductivity, respiration change and skin temperature. The subject had to conduct a driving task under two different conditions: In condition (1), the subject essentially just had to change lanes as indicated by road signs. In condition (2), the subject had to solve arithmetic calculations in addition to the driving task in order to induce a higher amount of stress in him. The recorded data were jointly analyzed by the workshop participants using the Augsburg biosignal toolbox AuBT leading to a recognition rate of 100% for a simple two-class problem (stress vs. no-stress) Given the available amount of time, it was not possible to collect a sufficient amount of data. The primary objective of the exercise was, however, to demonstrate and discuss all steps from data acquisition, feature calculation, feature extraction and classification to an interdisciplinary audience.

Discussion of Experiment for Physiological Data Analysis and Labeling of Physiological Data

The workshop participants discussed a position paper by Roddy Cowie describing lessons learnt at QUB over about five years of work in the area of physiological data analysis. Due to the long-term experience by QUB in the design of emotion-inducing experiments, the workshop participants decided that the Humaine partners specialized in the area of physiological data analysis (UA and FAU) should base their studies on data to be collected by QUB in car simulator experiments in the next few months. The results of the data analysis provided by UA and FAU will then be interpreted by QUB in order to study the effect of induced emotion on physiological indicators. We hope that a closer collaboration will enhance our knowledge on the relationship between the dependencies of stress and emotions and bridge the gap between controlled conventionally designed experiments and more realistic settings. Anton Batliner from FAU presented first ideas on the labeling of physiological data. It was agreed that for the time being the annotation of the corpus to be collected by QUB should be done automatically taking advantage of a structured experimental design.

Discussion of Fusion Problems

As suggested by ICCS-NTUA, participants also discussed problems arising during the fusion of multimodal signals.

Concrete Steps for Future Collaboration

At the end of the workshop, concrete steps for future collaboration were discussed

QUB and FAU	*Experimental design:* setting up a driving simulator scenario eliciting stress and emotional states, at least one or two pilot experiments (data to be distributed among the partners), to start with, we aim at "automatic labelling" via "structured design" and/or sequencing
FAU and UA	*Analysis:* Computation of features and first classification results
QUB	*Interpretation:* Interpretation of classification results and of impact of features
ALL	*Dissemination:* Joint paper at a relevant conference in 2006

4.2.5 Development of Context labels

The main work on context labelling has focused on developing a working set of context descriptors and on pilot work on the EmoTV database.

A list of context descriptors has been compiled as follows:

Agent characteristics (age, gender, race)
Recording-context (recording style, acoustic quality, video quality)
Intended audience (kin, colleagues, public)
Overall communicative-goal (to claim, to sway, to share a feeling, etc).
Social setting (none, passive other, interactant, group)
Spatial focus (physical focus, imagined focus, none)
Physical constraint (unrestricted, posture constrained, hands constrained)
Social constraint (pressure to expressiveness, neutral, pressure to formality)

The EmoTV is rich in different topics (politic, law, sport, etc). In order to study the meaningfulness of the context, five sets of attributes represent the context in a labelling scheme for EmoTV. The first set is the "emotional context" and includes some appraisal dimensions describing the reasons that cause the emotions: degree-of-implication (low, normal, high), cause-event (free-text), person-event relation (society subject, true story by himself or by kin) and time of event (near, recent, present, future). The other sets are the "interview context": theme, place, the video-taped person: age, gender, race, the overall communicative goal of the video-taped person, which combines consequence-event and communicative function: what-for (to claim, to share a feeling, etc.), to-whom (kin, colleagues, public, etc.); and the "recording context": camera (static, dynamic), character (static, dynamic), acoustic quality, video quality. Integrating the other appraisal dimensions of Scherer's model is currently being studied.

4.2.6 Participants

QUB, CNRS-LIMSI , Paris8, FAU Erlangen, UNIGE, UA, ICCS

4.2.7 Publications

S. Abrilian, L. Devillers, and J.-C. Martin. EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. In HCI International, Las Vegas, July 2005.

S. Abrilian, J.-C. Martin, and L. Devillers. A Corpus-Based Approach for the Modeling of Multimodal Emotional Behaviors for the Specification of Embodied Agents. In HCI International, Las Vegas, July 2005.

A. Batliner, S. Steidl, N. Amir, R. Cowie, L. Devillers, L. Kessous, K. Laskowski, B. Schuller, D. Seppi, L.Vidrascu (submitted to LREC 2006) CEICES: Combining Efforts for Improving automatic Classification of Emotional user States - a "forced co-operation" initiative.

L. Devillers, S. Abrilian and JC. Martin, Representing Real-life Emotions in Audiovisual Data with Non Basic Emotional Patterns and Context Features, in ACII, Bejing, October 2005

J-C. Martin, S. Abrilian and L. Devillers, Annotating Multimodal Behaviors Occuring during Non Basic Emotions, in ACII, Bejing, October 2005.

E. Douglas-Cowie, L. Devillers, J-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal Databases of Everyday Emotion: Facing up to Complexity. In InterSpeech, Lisbon, September 2005.

4.2.8 Other output (demonstrations, resources, etc)

The CNRS-LIMSI coding scheme and its annotation guides have been uploaded on the Humaine Portal (CNRS-LIMSI)

FeelTrace has been made available to partners. It will be part of the ToolBox.

4.3 Providing resources for immediate use across the network

4.3.1 Recordings provided

- Pilot SAL dataset (labelled for emotion using FeelTrace) released to all partners
- AIBO dataset released under CEICES agreement
- TAU Hebrew Recall dataset released to QUB
- EmoTV & Belfast Naturalistic Database subsets released between CNRS-LIMSI and QUB for labelling
- Subset from TV Reality dataset released by QUB to CNRS-LIMSI for labelling development
- UNIGE acted database samples to limited number of partners
- Two video clips from the EmoTV database (one video featuring a superposition of anger and despair, one video featuring the masking of a felt disappointment by a fake serenity) have been used for the collaboration with WP6 (University Paris 8) on replaying annotated multimodal behaviour by an expressive ECA (LIMSI-CNRS: Martin et al. 2005 IVA). The frames shown below illustrate how the original (left hand frame) is reproduced by and ECA (right hand frame) on the basis of the LIMSI-CNRS annotation.


- Four video clips from the EmoTV database have been used for a collaboration with WP4 (ICCS) aiming at assessing the potential of automatic image processing techniques for TV resolution video (e.g. global estimation of movements for skin areas).

4.3.2 Tools provided

- FAU Erlangen entropy-based evaluation tool (for assessing annotations)

http://emotion-research.net/restricted/wp5/slides/EBE.pdf/view?searchterm=Evaluation%20of%20Decoders

http://www5.informatik.uni-erlangen.de/Forschung/Projekte/HUMAINE/software.xml?language=en

- The CNRS-LIMSI coding scheme and annotation guides on Humaine Portal (CNRS-LIMSI)

- FeelTrace made available

Interactive database on Portal which provides details on available resources, allows people to add as datasets developed. In cooperation with WP1, work is underway to make information gathered in WP5 more accessible. In an earlier deliverable (D5b), we collected a list of databases of emotions, along with details about their availability, the modalities covered, scientific papers describing them, etc. This information had been represented at the time as an Excel spreadsheet.

As the first piece in a more general "HUMAINE Tool- and Resource-Box", WP1 has made this data available in a more accessible form, as a searchable, structured and extendable collection of database descriptors (see screenshot in Figure 3).

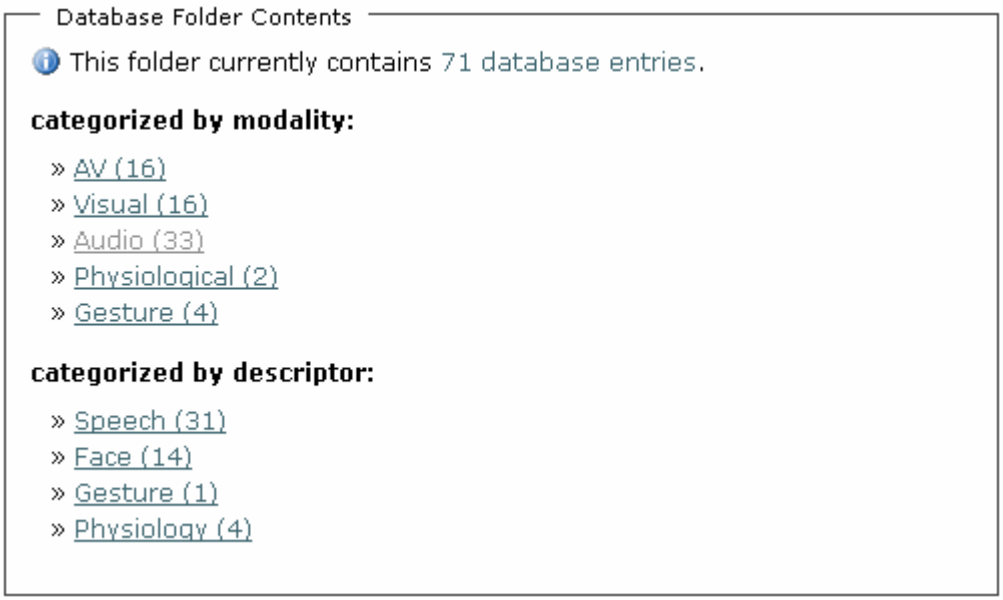


Figure 3. Screenshot of portal-based collection of emotional databases

This newly created software provides a structured search mechanism, detailed listing of individual database descriptions. All portal users can add new entries, to allow for community participation; however, as a means of quality control, all new entries have to pass a review step
before becoming publicly visible. We are currently testing and evaluating the software, prior to its public launch.

4.3.3 Participants

DFKI, QUB, CNRS-LIMSI, TAU, FAU Erlangen, UNIGE, ICCS

4.3.4 Publications

Cowie, R., E. Douglas-Cowie, C. Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks* 18, 371-388.

Douglas-Cowie, E., Devillers, L., Martin, J-C., Cowie, R., Savvidou, S., Abrilian, S. and C. Cox. Multimodal Databases of Everyday Emotion: Facing up to Complexity. In InterSpeech, Lisbon, September 2005.

Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M. and Pelachaud, C. (2005). Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. 5th International Working Conference On Intelligent Virtual Agents (IVA'2005), Kos, Greece, September 12-14 Springer. 405-417 http://iva05.unipi.gr/

Batliner, A., Hacker, C., Steidl,S., Nöth, E., D'Arcy, S., Russell, M. and Wong, M. "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus, in *Proc. LREC 2004*, Lisbon, 2004.

Batliner, A., S. Steidl, C. Hacker, and E. Nöth, "Private Emotions vs. Social Interaction - towards New Dimensions in Research on Emotion," in *Proc. Workshop on Adapting the Interaction Style to Affective Factors, User Modelling 2005*, Edinburgh, 2005.

Batliner, A., S. Steidl, C. Hacker, E. Nöth, and H. Niemann. Tales of Tuning - Prototyping for Automatic Classification of Emotional User States, in *Proc. Interspeech 2005*, Lisbon 2005.

Devillers, L., L. Vidrascu, and L. Lamel, Challenges in real-life emotion annotation and machine learning based detection, in *Journal of Neural Networks, 18/4*, Special Issue "Emotion and Brain", July 2005.

Steidl, S., M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency, in *Proc. ICASSP 2005, Philadelphia, U. S. A.*, 2005.

Vidrascu, L. and L. Devillers, Real-life Emotions Representation and Detection in Call Centers, in *Proc. ACII*, Bejing, October 2005

Vidrascu, L. and L. Devillers. Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center. In *Proc. ICME*, Amsterdam, June 2005.

Vidrascu, L. and L. Devillers. Detection of Real-Life Emotions in Call Centers. In *Proc. Interspeech 2005*, Lisbon 2005.

4.3.5 Other output (demonstrations, resources, etc)

FeelTrace training pack; SAL protoptype

4.3.6 Follow-up in progress

- Interchange QUB/CNRS-LIMSI on shared naturalistic data to pursue labelling

- Interchange TAU & ICCS working on SAL pilot data- collaboration on multimodal analysis (video and sound) and recognition of emotion using the SAL database. One of the major current challenges is the synchronization and segmentation of data from audio recordings and videos. This is problematic when they are recorded on different media, i.e., for example on DV video and on DAT for audio. To be able to do it automatically, or to provide a version of the existing database suitable for audio-video synchronous analysis are part of the goals of this exchange.

- CEICES group working on AIBO material

- USC using SAL pilot data to code facial and body movement

5 Conclusion

5.1 Obstacles encountered or foreseen

There have been a number of difficulties which were foreseen to some extent but which were more problematic than might have been expected.

The first has to do with the labour intensive nature of labelling. Even piloting labelling schemes takes a lot of time and co-location of people working on the same data to compare and discuss methods there and then. The interchange between QUB and CNRS-LIMSI (December 2005) arises from the need to spend concentrated time on the task, and should help address the problem. Linked to the issue of labelling is the problem that labelling schemes which seem rational on paper or in principle may not turn out to be reliable in practice. For example, we have experimented with using the Scherer appraisal scheme for labelling, but pilot work suggests that labellers find this to difficult to apply, leading to unreliability as a practical labelling tool.

The second difficulty is overcoming differences of approach, expectation and understanding across partners with respect to databases. Some people want large amounts of data, others need small amounts of closely labelled data; some want word by word labelling, others are satisfied with much broader labelling; some need very constrained data (heads relatively static etc), others want people with gestures and movement visible. It has been a long process developing understandings of what partners might need, why initial expectations might not be realistic, and in general to come to agreements on the range of data and labelling that needs to be made available.

The third difficulty is one which we are only beginning to address: that is the technological issues surrounding data format and the mounting and distribution of data. The assignment of the new partner UT to the workpackage provides some of the technical support needed, but it is recognised that even by the end of HUMAINE, there will still be many longer term technical issues to address in emotional multimodal databases. That is why an application has been made involving partners from within HUMAINE and beyond to the EU for funding for a STREP called EmoBase - Interdisciplinary Integrating Framework and Database for Emotion and Behaviour Analysis.

5.2 Relation to the state of the art

Work in WP5 is making a leading contribution to the state of the art on emotional databases, especially in terms of naturalistic emotion – its induction and its labelling. The markers of esteem listed below provide evidence that the researchers are considered to be leaders in the field.

5.3 Evidence of esteem

- Workshops leading the area at major conferences – key examples being the Interspeech 2005 workshop on Emotional Speech which was organised by members from QUB and CNRS-LIMSI and had a strong emphasis on databases, and the forthcoming workshop specifically on Corpora for Research on Emotion and Affect in LREC 2006, and organised by members from QUB, CNRS-LIMSI and FAU-Erlangen.

- Other international workshops (organised by HUMAINE partners and contributed to by HUMAINE) with a focus on data - Tutorial and Research Workshop Affective Dialogue Systems 2004.

- Special Issues of journals relating to emotion and featuring database issues – key example is Neural Networks 18 (4), Special Issue on Emotion and Brain.

- Members of WP5 HUMAINE partner institutions on Scientific Committees of major conferences, e.g. ACII 2005, LREC, Interspeech

- Members reviewing for key journals in the field

- Many requests from across the world for emotional databases and for labelling tools

6 References

Abrilian, S., L. Devillers, and J.-C. Martin (a).. EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. In HCI International, Las Vegas, July 2005

Abrilian, S., J.-C. Martin, and L. Devillers. (b). A Corpus-Based Approach for the Modeling of Multimodal Emotional Behaviors for the Specification of Embodied Agents. In HCI International, Las Vegas, July 2005

Boone, R. T. and Cunningham, J. G. (1998). "Children's decoding of emotion in expressive body movement: The development of cue attunement." Developmental Psychology 34(5): 1007-1016.

DeMeijer, M. (1989). "The contribution of general features of body movement to the attribution of emotions." Journal of Nonverbal Behavior(13): 247 – 26

Devillers, L., S. Abrilian and JC. Martin, Representing Real-life Emotions in Audiovisual Data with Non Basic Emotional Patterns and Context Features, in ACII, Bejing, October 2005

Ekman, P. (1999). Basic emotions. Handbook of Cognition & Emotion. T. Dalgleish and M. J. Power. New York, John Wiley: 301–320.

Ekman, P. and Friesen, W. V. (1975). Unmasking the face. A guide to recognizing emotions from facial clues., Prentice-Hall Inc., Englewood Cliffs, N.J.

Ekman, P. (2003). The Face Revealed. London, Weidenfeld & Nicolson.

Clark, L., Iversen, S.D., & Goodwin, G.M. (2001). The influence of positive and negative mood states on risk taking, verbal fluency, and salivary cortisol. *Journal of Affective Disorders, 63*, 179-187.

Clark, D.M. (1983). On the induction of depressed mood in the laboratory: Evaluation and comparison of the Velten and musical procedures. *Advanced Behavior Research and Therapy, 5*, 27-49.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J.G., (2001) Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processing Magazine*, (January 2001).

Douglas-Cowie, E., Devillers, L., Martin, JC., Cowie, R., Savvidou, S., Abrilian, S. & Cox, C. (2005) Multimodal Databases of Everyday Emotion: Facing up to Complexity. *Interspeech, Lisbon, Sept 2005*.

Gerrards-Hesse, A., Spies, K., & Hesse, F.W. (1994). Experimental inductions of emotional states and their effectiveness: A review. *British Journal of Psychology, 85*, 55-78.

Gross, J.J. & Levenson, R.W. (1995). Eliciting emotion using films. *Cognitions and Emotion, 9 (1)*, 87-108.

Ito, T.A., Cacioppo, J.T., & Lang, P.J. (1998). Eliciting affect using the International Affective Picture System: Trajectories through evaluative space. *Personality and Social Psychology Bulletin, 24 (8),* 855-879.

Kenealy, P. (1997). Mood-state-Dependent Retrieval: The Effects of Induced Mood on Memory Reconsidered. *The Quarterly Journal of Experimental Psychology, 50A (2),* 290-317.

Kipp, M. (2004). Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. Florida, Boca Raton, Dissertation.com. 1581122551. http://www.dfki.de/~kipp/dissertation.html

Lang, P.J., Greenwald, M.K., Bradley, M.M., & Hamm, A.O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology, 30,* 261-273.

Lang, P.J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications. In J.B. Sidowski, J.H, Johnson, & T.A. Williams (Eds.), *Technology in medical health care delivery,* Norwood NJ: Ablex. (pp. 119-137)

Martin, M. (1990). On the induction of mood. *Clinical Psychology Review, 10,* 669-697.

Martin, J-C., S. Abrilian and L. Devillers, Annotating Multimodal Behaviors Occuring during Non Basic Emotions, in ACII, Bejing, October 2005.

Newlove, J. (1993). Laban for actors and dancers. New York, Routledge. 1 85459 160 6.

Öhman, A., Flykt, A., & Esteves, F., (2001). Emotion Drives Attention: Detecting the Snake in the Grass. *Journal of Experimental Psychology; 130*, (3), 466-478.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*. *80* (3), 381 -396.

Scherer, K. R. Analyzing Emotion Blends. 10th Conference of the International Society for Research on Emotions, Wrzburg, Germany, Fischer, A.142-148

van Reekum, C.M., Johnstone, T., Banse, R., Etter, A., Wehfle, T., & Scherer, K.S. (2004). Psychophysiological responses to appraisal dimensions in a computer game. *Cognition and Emotion, 18 (5),* 663-688.

Velten, E. (1968). A laboratory task for inductions of mood states. *Behaviour Therapy and Research, 6,* 473-482.

Wallbott, H. G. and Scherer, K. R. (1986). "Cues and Channels in Emotion Recognition." Journal of Personality and Social Psychology 51(4): 690-699.

Westermann, R., Spies, K., Stahl, G., & Hesse, F.W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology, 26,* 557-580

7 Appendix: The Remote SAL System

This document describes a possible configuration of the remote SAL system. Remote SAL builds up the current SAL system by removing the human operator, who chooses the replies of the 4 characters to a remote location. Since the respondent and the operator will be in different rooms, the most flexible solution will be the use of a network connection between two computers, one in each room. The connection should be fast enough to process real-time video and audio. A central server at another location allows for a system with different operators. As there is no human operator in the same room with the respondent, to whom the respondent can talk to, the whole set-up of the experiment becomes quite artificial. The respondents are expected to have emotional experiences similar to emotions arousing in real-life conversations, while talking to an imaginary character, the SAL speaker.

We will describe the latter option of a system with three components: a central SAL server, a client at the respondent side and a client at the operator side. We distinguish two interfaces for the remote SAL system, the interface for the respondent and the interface for the operator. An operator can log on to the central server any time to show that she is available for interaction with a respondent. A respondent can log on to the central server any time he wishes to interact with the system. The server keeps track of available operators and arranges connections between operators and respondents. Session data will be logged at the respondent side to avoid that large amounts of recorded audio and video need to be transmitted over the network.

At the respondent side, there should be a camera and a headset or microphone with speakers connected to a computer. The computer will run a client application that can communicate with the central server and with an operator. The central server will be used to find an available operator and set up a connection. During the session, recorded video and audio will be sent to the operator continuously. At the same time, the respondent will receive data from the operator consisting of audio, data for synchronised animation of the avatar and possible changes in the virtual environment. This will require a fast network connection between the respondent and the operator.

The operator will have a headphone or speakers and a microphone connected to a computer. During the interaction the operator will mainly select pre-recorded wav files, but the welcome text and instructions for the system need to be spoken into the microphone. The computer will run a client application that at first will be similar to the existing SAL interface, but will be configurable and extendible. The application will continuously render the video and audio recordings received from the respondent. Any relevant interaction data that needs to be logged is sent to the respondent.

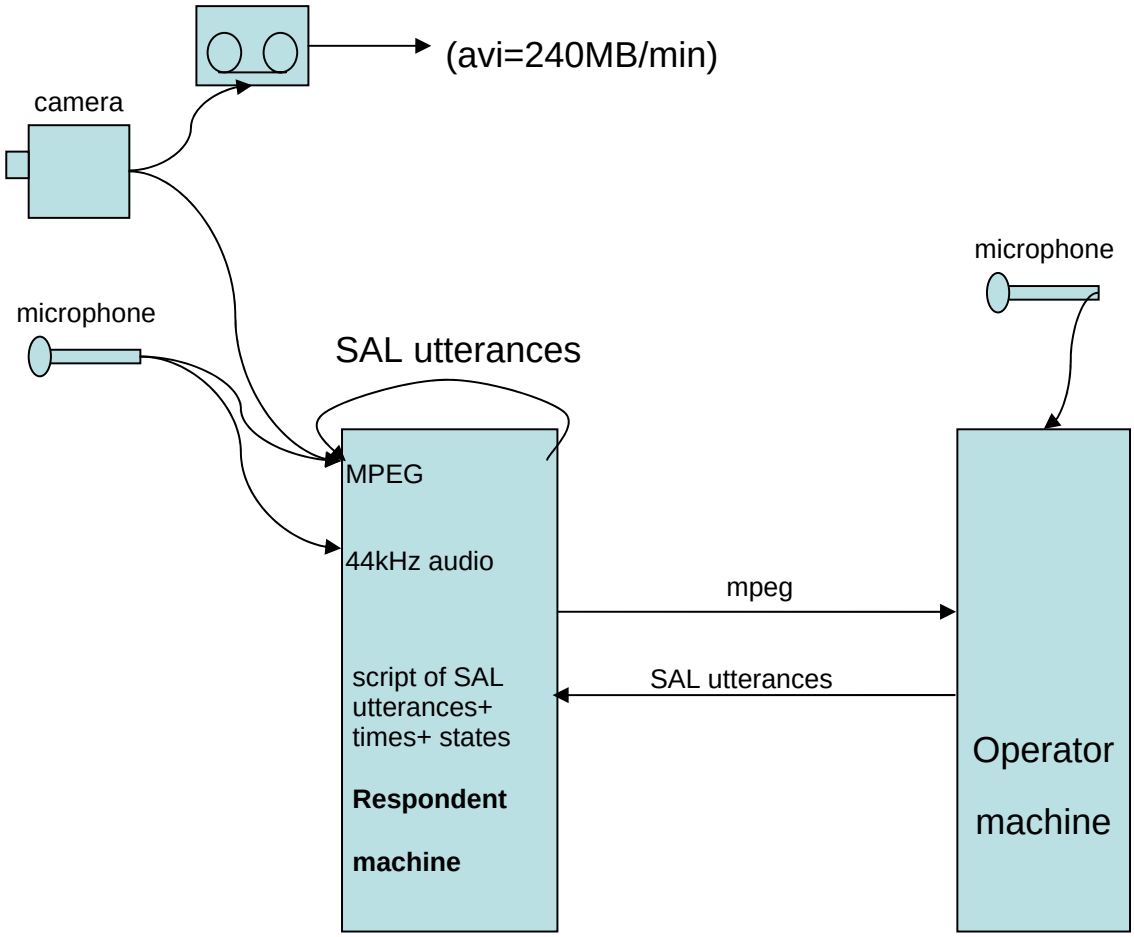


Figure 1: Data recording

Figure 1 shows what data is recorded and exchanged between the operator and respondent and what data is saved to log files at the respondent machine. A camera and microphone are connected to the respondent machine, whereas only a microphone is connected to the operator machine. Video recordings are saved in maximum uncompressed quality to a tape in the digital video camera and they are captured to smaller size compressed mpeg videos on the respondent machine. Audio recordings from the respondent microphone are captured to high quality wav files on the respondent machine and added to a lower quality audio track in the mpeg file. The lower quality video with audio track is transmitted over the network to the operator machine.

The operator selects wav files and may speak into the microphone. The audio (from the wav file or microphone) is transmitted over the network to the respondent machine where it is played back and added to an audio track in the mpeg file. Thus the mpeg file contains all recorded video and audio from the entire session. In addition the operator sends messages to the respondent to indicate which wav files were selected at what time and when the respondent state or the speaker was

changed. This data is logged into a script that contains all selected SAL utterances, speakers and respondent states with times.

During the SAL experiment we collect as much data as possible for further automated or manual analysis of for example facial emotional expressions and speech. Saving of the data on the SAL database, like for example the integrated video and audio track in high quality mpeg, the separate audio track captured direct to disc, and SAL utilities (replies and time chosen by operator, emotion state of respondent) is synchronized.

Requirements for client application interface for operator

The main usability requirements of the interface of the remote SAL system are the simplicity and flexibility. In the interface only the functionalities are shown, that are used for running the SAL experiment. The different functionalities are separated on different adjustable windows or frames. We think, that opening only the windows with functionalities, that the operator knows he needs, and letting the operator to self organize items on the desktop, is an intuitive way to learn to understand and to use the interface.

Below is a list of the requirements for the operator interface:

1. Real-time rendering of the integrated video and audio stream recorded on the respondent site on the desktop of operator. To speed up the rendering of the video from the respondent site through the network, we use low quality mpeg format.
2. The window frame rendering the video from the respondent site should be part of the interaction frame, so that it does not disappear behind the interaction window each time the operator click on that.
3. Classification of emotional state of the respondent by a human operator into one of the four categories of emotional states (-ve act, +ve act, -ve pass, prag), by pressing buttons on the button frame ‘Change State to’. The button of the active state is replaced with a green label.
4. In the remote SAL system, the replies are selected by human operator by clicking on the text samples, which launches waw-files of recorded human speech.
5. The four SAL speakers own set of replies, which we name speaker scripts. Speaker scripts of each speaker are divided into four ranges of replies that correspond to the four possible emotional states that the respondent can be classified into, which we name state scripts. The state scripts have some or all of the following four response categories of replies: 1. statement, 2. question/request, 3. change speaker dialogue, and 4. repair. One more script category 'start up' is needed to show only during the introduction of the SAL character. The selection of scripts by chosen speaker and emotion state of respondent we name speaker state scripts.
6. The state scripts are automatically laid out in one or two columns so that the amount of white space is limited and all scripts fit in one window.

7. Operator should have at least 19 inch monitor for a comfortable arrangement of the columns which scripts and the size of the text of speaker state scripts on the desktop or interaction window.
8. Menu option to change the size of the text on the interface.
9. Operator has button set for changing the speaker.
10. Operator has an option to let the respondent to change the speaker with button set to change the speaker on the respondent site.
11. Button set for still to definable script sets as laugh and 'aha' replies in the interaction window frame.
12. After one reply is used for one respondent, it should not be removed from the list shown to the operator, but marked. Otherwise the set of possible replies would become too little. The best would be to have a frequency count visualized about how many times a reply is spoken to one respondent.
13. Operator needs a microphone for reading introduction texts for which there are no wav files.

Requirements for client application interface for respondent

1. Button set on the desktop interface of the respondent for changing the SAL speaker by the respondent. The button set can be put on or off by the operator. The button of the active speaker is replaced with a green label. The dynamics of the whole experiment would change, if the SAL speaker was changed by the operator instead of the respondent. When respondents change the speaker, the emotional state of the respondent is not directed or controlled with any other way than the choice of the operator for a reply script from the chosen SAL speakers script.
2. For the remote SAL system the respondent interface should have abstract visuals that catch the attention and the focus of the respondent and supports the mood and personality of the SAL speakers. The visual image on the screen in the respondent interface changes as the SAL speakers change, and it should display the emotion in the voice of the SAL speaker on the simplest possible way. We can use or implement one or more basic algorithms for visuals that move and change colour according some parameters that can be extracted by speech recognition algorithms from the wav format of speaker scripts. Implementation of the visual display of the emotion in the SAL speakers voice will take place after the implementation of more crucial parts of the remote SAL system.
3. The camera should be placed above the computer screen, but not too high, so that the whole face of the respondent is in the picture.

Interface design

Start SAL system

The key component of the system is the SAL server, which ideally should always be running at a fixed computer. By pressing ‘Start server’ in the **SAL-server** application, the server is opened to receive connections from operators and respondents, which

can be located anywhere. The server has a graphical user interface that keeps track of available operators, waiting respondents and ongoing sessions.

The server holds a repository of SAL projects (possibly only one). When a respondent connects, he receives a list of the projects and needs to select one. As soon as a respondent has selected a project, he is available for a new session. When there is an available operator, the server arranges a new session between the respondent and operator.

The **SAL-operator** application allows an operator to connect to the SAL server. A respondent runs the **SAL-respondent** application.

Connect to the server from the operator site

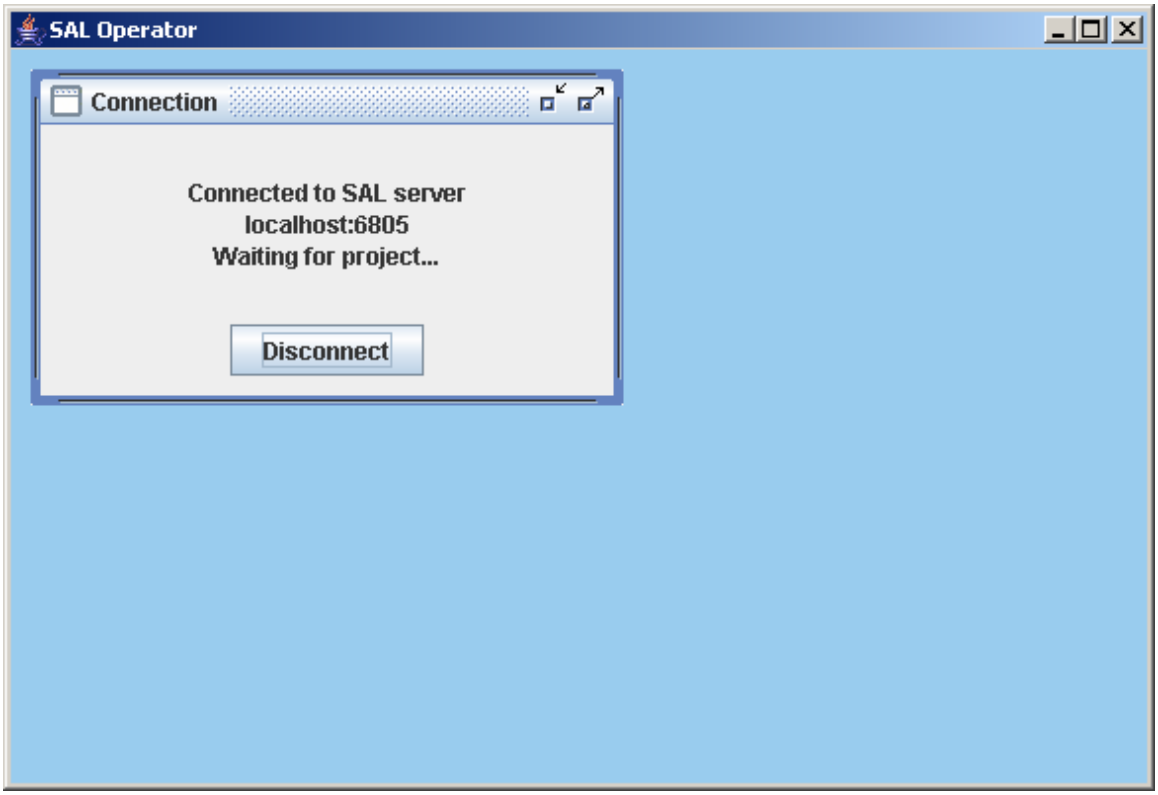


Figure 2: SAL operator available for a new session

When the **SAL-operator** application is started, the interface will only show a connection window. The operator can click 'Connect' to connect to the SAL server. The window will show that the operator is connected to the SAL server and if no respondent is waiting for a session, the window will show 'Waiting for project' (see **Figure 2**). The connect button changes into 'Disconnect'. If something goes wrong with the establishment of the connection, an error message appears. In this case contact the system administrator to solve the connection problem.

If a respondent was waiting for a session or as soon as a respondent connects to the SAL server and selects a project, thus becoming available for a session, the SAL server may arrange a new session between the operator and the respondent.

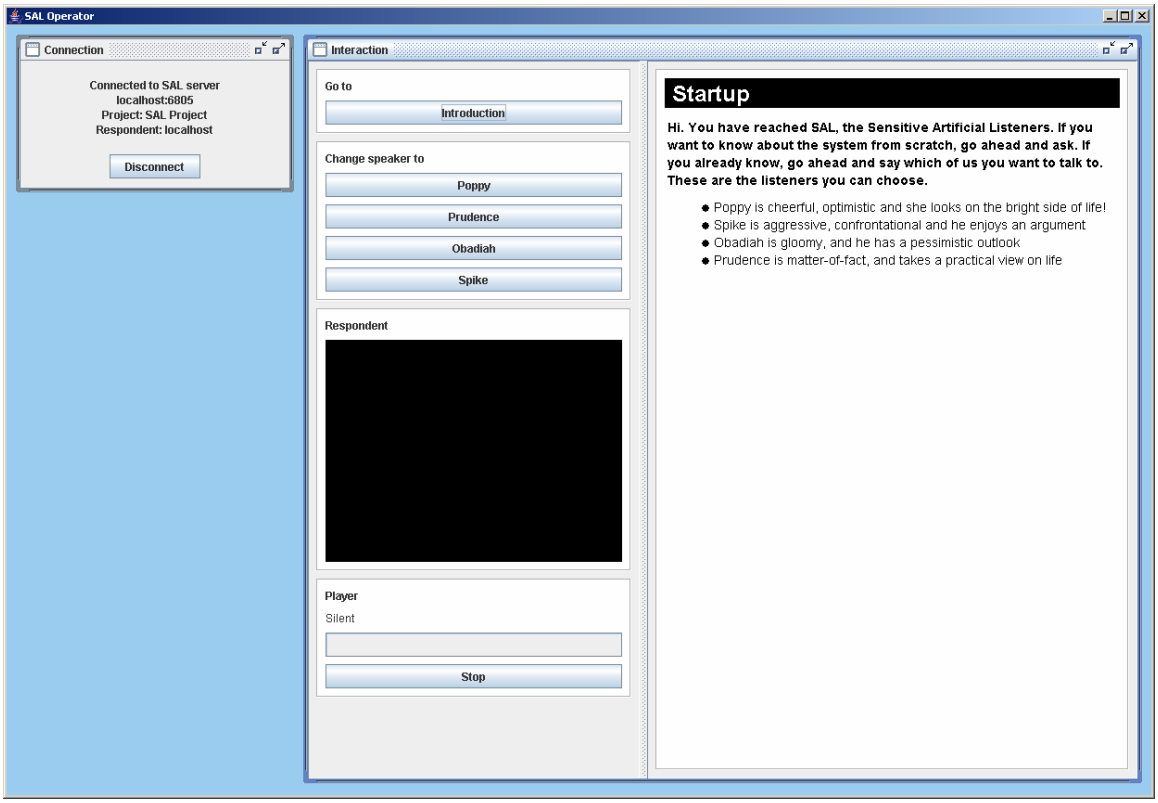


Figure 3: SAL operator at start of new session

When the operator is set up for a new session, the connection frame will show the name of the project and the location of the respondent. An interaction window appears showing the start up content for the SAL project. During the session the operator application will continuously play back the video and audio received from the respondent and it will send selected phrases, recorded audio and interaction data (selected speakers and states) to the respondent in response to the operator actions in the interaction window. The session ends when the operator clicks 'Disconnect' or closes the application, when the respondent closes the session or when the SAL server is closed. All session data is saved at the respondent side.

At the start of a new SAL experiment session with a respondent, the interaction window looks as shown in **Figure 3**. It has the following panels at the left.

Go to: For viewing information about the SAL experiment (start-up and introduction), to be read into the microphone for the respondent.

Change Speaker to: For choosing the SAL speaker.

Respondent: Shows the video of the respondent.

Recorder: (to be implemented, replacing the player panel in this screen): Allows the operator to open/close the microphone to read the introduction texts for the respondent.

After the operator has read the introduction of the SAL experiment for the respondent, she can choose the SAL speaker to start the session with.

Connect to the server from the respondent site

The **SAL-respondent** application shows a similar connection window at start-up as the operator application. When the respondent clicks 'Connect', it connects to the SAL server. The SAL server will return a list of available projects. If there is only one project, the respondent application automatically selects that project. Otherwise the list of projects will be presented to the user to select one. After project selection, if no operator is available yet, the connection window will show 'Waiting for operator'. As in the operator application, the connect button changes into 'Disconnect', and an error message appears if something goes wrong while establishing the connection.

If an operator was available, or as soon as an operator becomes available for a new session, the SAL server may arrange a new session between the operator and the respondent. When the session has been set up, the connection window will show the location of the operator.

As SAL does not yet have virtual embodied conversational characters, the respondent site has no further presentation or interaction content. At the moment, that we have an abstract visualization for the emotion in the voice of the SAL speakers, a window frame can be opened to present the visualization. We are at the moment designing and implementing an algorithm, which visualizes the emotion in the voice by colour, form, movement, and tempo of the movement.

The interface will include a video window, which shows the video of the respondent. This allows the respondent to see what is actually recorded and sent to the operator and to position himself properly in front of the camera.

During the session the camera and microphone recordings will be continuously sent to the operator. The respondent will occasionally receive audio from the operator, which will be played back instantly. All recorded and received media data as well as received interaction data, is saved to files. All media data will be written to one mpeg file. The microphone recordings will additionally be written to a high quality wav file. The interaction data will be written to XML files. The exact format is as yet unknown.

The session is closed when the respondent clicks 'Disconnect' or closes the application, when the operator closes the session, or when the server is closed.

Interaction window frame of the operator

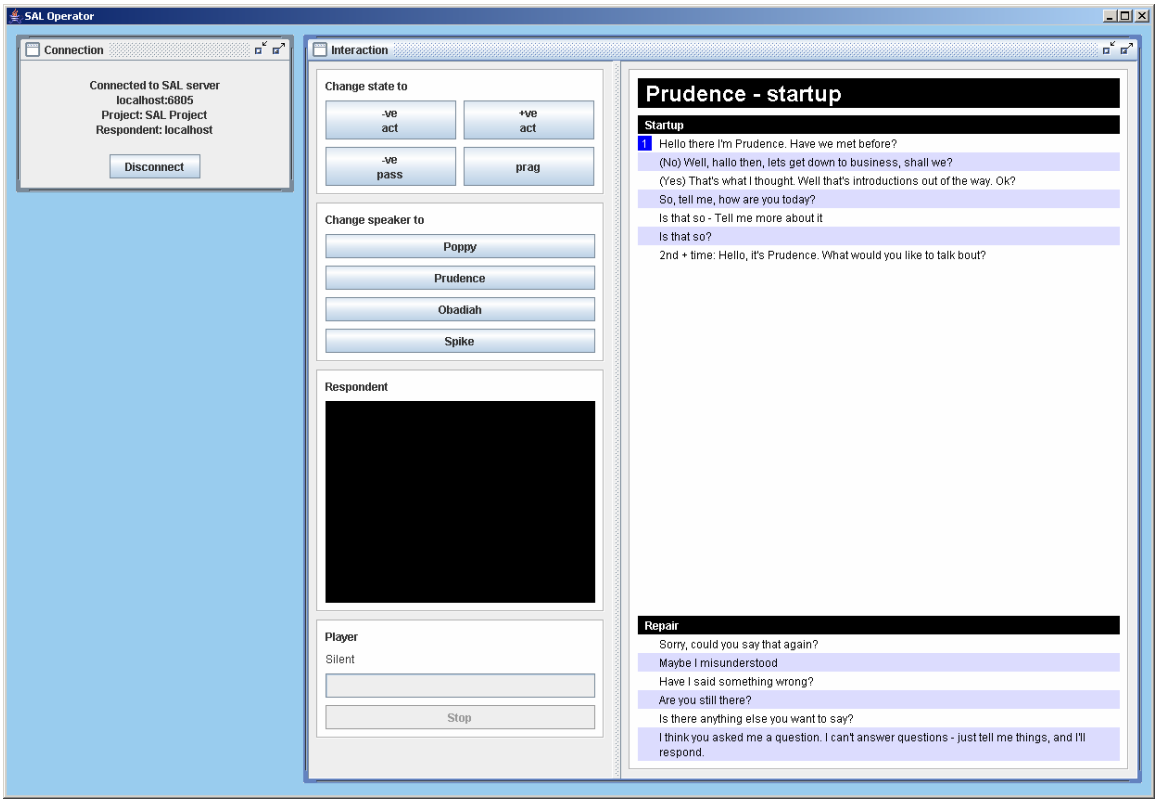


Figure 4: SAL operator at the start of interaction with Prudence

Speaker - start-up

Each time the SAL speaker is switched, the Start-up content screen, see Figure 4, is displayed for the operator. The start-up content screen has **two response categories**: **Startup** and **Repair.** The operator clicks on the responses to launch the wav-files with the corresponding speaker recordings. The following panels are shown at the left.

Change state to: For classifying the current emotional state of the respondent into the four categories. Pressing 'Change state to' to classify the respondent emotion state, launch a new content 'range of replies', corresponding to the respondent emotion state to be displayed on the operator interaction window.

Change Speaker to: For choosing the SAL speaker.

Respondent: Shows the video of the respondent.

Player: Shows the status of a wav file being played back and sent to the respondent. It allows the operator to stop playback and transmission.

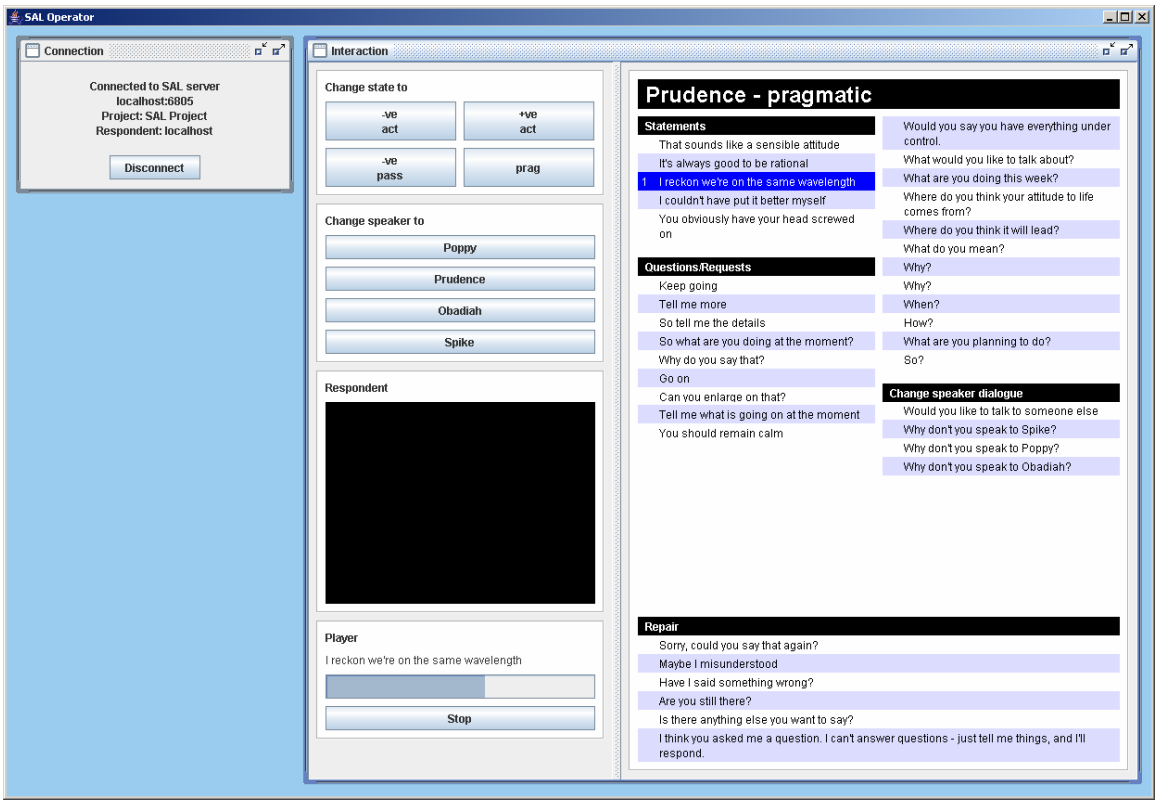


Figure 5: SAL operator controls Prudence while respondent is pragmatic

Speaker – respondent state

For each speaker, all four categories of respondent states have their own range of replies divided into three categories: **Statements**, **Questions/Requests**, **Change speaker dialogue**. The set of responses in **Repair** does not change when respondent emotion state is switched. An example is shown in **Figure 5**.

Underlying configurations of the SAL system

In SAL-system each network connection is configured to have only 2 sites, one for the respondent and one for the operator application and interface. To run other experiments with the software developed for the SAL system, it should be possible to configure connections between any amount of sites with own set of interface functionalities and interaction possibilities. This enables for example recording of a virtual meeting. Ready to use functionalities can be implemented for the future 'authoring tool' for SAL like experiments, to make it for the researchers easy to configure any amount of sites. Later, functionalities as automated analysis of facial expression with some algorithm and visualization of the results real-time on the

operator window frame can be configured next to the items we implement now for the remote SAL system.

Functionalities for the respondent site in the remote SAL system:

Enable video stream

Enable video recording

Enable audio stream

Enable audio recording

Enable rendering audio stream from operator site

Functionalities for the operator site in the remote SAL system:

Enable audio stream

Enable audio recording

Enable rendering video stream from respondent site

Enable rendering audio stream from respondent site

Enable interactive frame (Interactive frame is for button sets for selections, multiple choices, editable text- boxes, and display of content material)

- Enable recording of interaction in interactive frame

SAL project Implementation

SAL projects consist of the following elements:

- ✓ ID and name of one or more speakers.
- ✓ ID and name of one or more emotional states.
- ✓ A welcome text.
- ✓ An instruction text.
- ✓ Start-up phrases for each speaker.
- ✓ Repair phrases for each speaker.
- ✓ Sets of phrases for each combination of speaker and emotional state, ordered by categories such as statements or questions.

A project can be specified in an XML file like the following:

```
<sal_project>
   (<speaker id="..." name="..."/>)+
   (<state id="..." name="..."/>)+
   <phrases>
      <welcome>...</welcome>
      <instructions>...</instructions>
      (<speaker id="...">
         <startup>
            (<phrase wavfile="...">...</phrase>)+
         </startup>
         <repair>
            (<phrase wavfile="...">...</>...</phrase>)+
         </repair>
         (<state id="...">
            (<category name="...">
               (<phrase wavfile="...">...</>...</phrase>)+
            </category>)+
         </state>)+
      </speaker>)+
   </phrases>
</sal_project>
```

The project specification could be extended with questionnaires or information about the hardware configuration.

Communication protocols

The remote SAL system consists of three components: a central SAL server, a client at the operator side and a client at the respondent side. The communication protocol is illustrated in the following figures.

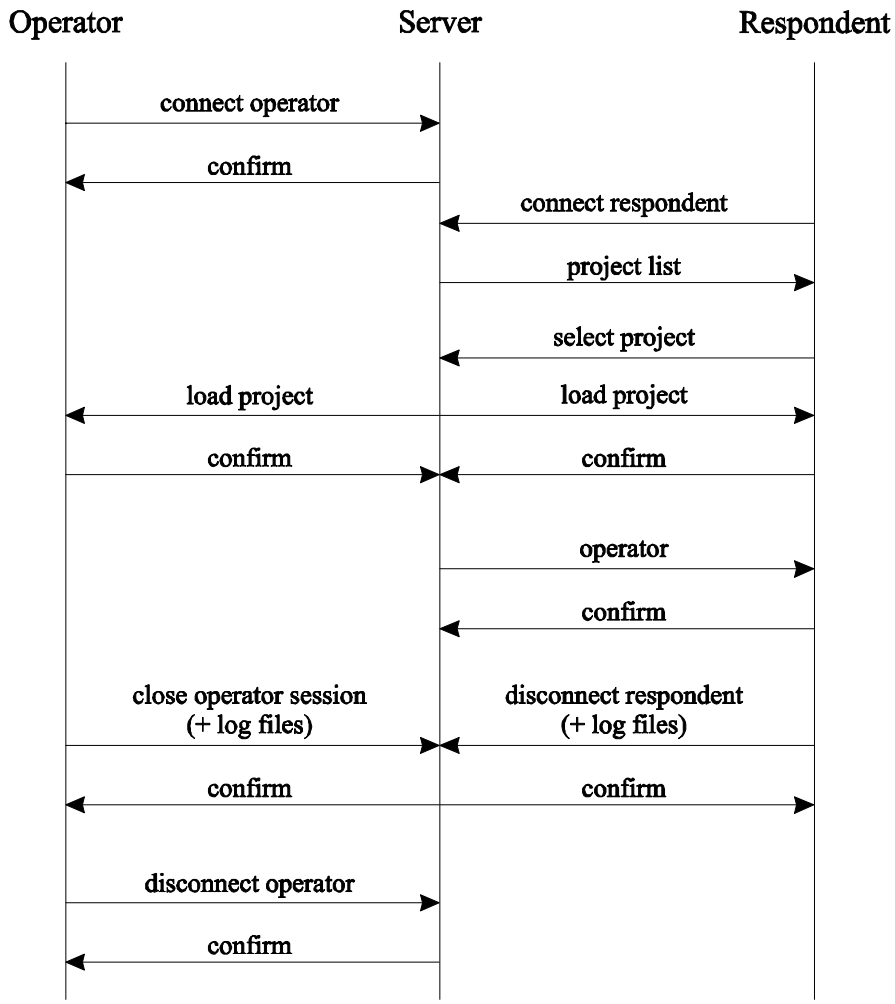


Figure 6 SAL Server

The SAL server should always be running. It waits for connecting operators and respondents. If there is no free operator, a connecting respondent needs to wait. As soon as an operator is free, it is bound to the connecting respondent for a new session. The SAL server has a list of projects. The list with project names is sent to the respondent and the respondent should select one of the projects. Then the server sends the project (in XML code) to the operator and the respondent. After successful project loading, the server sends the host name and port number of the operator to the respondent, so they can make a direct connection. At the end of the session, both the operator and the respondent should send a close/disconnect message with the

log files. An operator can disconnect to show that he is not available anymore for new sessions.

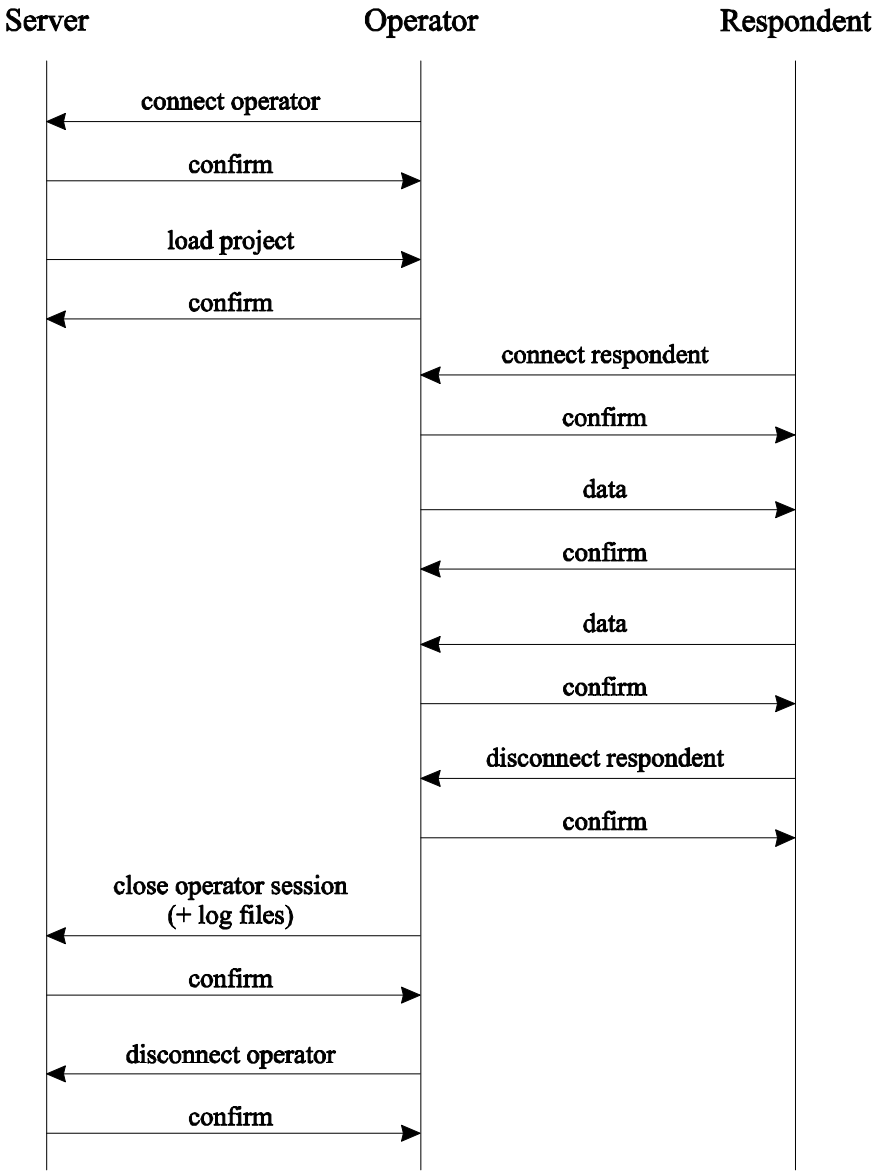


Figure 7 Operator protocol

The operator should first connect to the server to show that he is available to accept respondent connections. When the server binds a respondent to the operator and the respondent has selected a project, the operator will receive the selected project from the server. Then it can get a connection from the respondent. During a session, data (audio, video) can be exchanged between the operator and the respondent. All data recorded at the operator side is logged. At the end of the session, the operator sends a close message with the log files to the server. The operator can disconnect to show that he is not available anymore for new sessions.

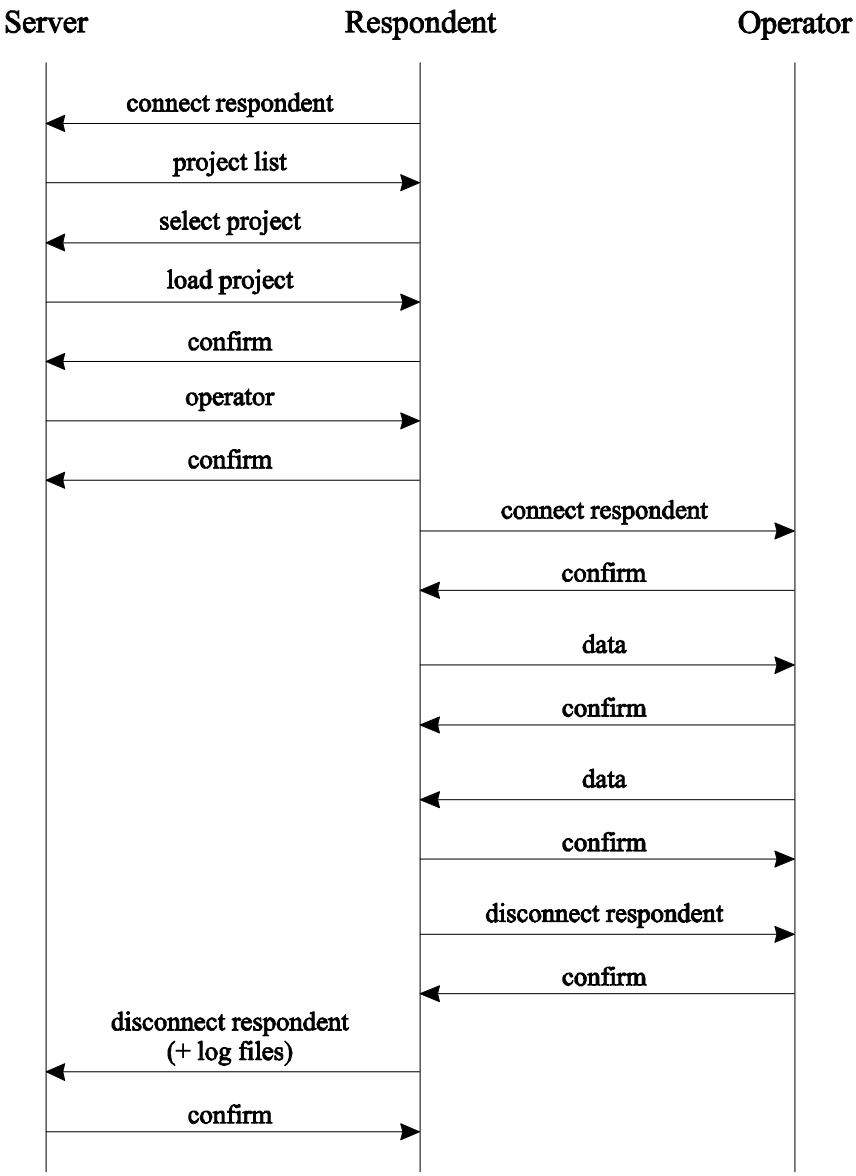


Figure 8 Respondent protocol

A respondent should first connect to the SAL server. The server will try to bind the respondent to a free operator, but if no free operator is available, the respondent needs to wait. The server will send a list with project names and the respondent needs to select one. Then it receives the project data and the host name and port number of a free operator. The respondent needs to connect to the operator directly. Then they can exchange data. All data recorded at the respondent side is logged. At the end of the session, the respondent needs to disconnect from the respondent and from the server. It will also send the log files to the server.

Java applications

This section lists the main Java classes for each of the three applications in the remote SAL system.

SalServerApp

SalServerApp

The main application implementing a GUI that shows the free operators, waiting respondents and ongoing sessions.

SalServer

Handles connections from operators and respondents and exchanges messages with them.

LogWriter

Writes all received log files to disk and orders them in folders for each session.

SalOperatorApp

SalOperatorApp

The main application implementing the GUI with the menu bar and pop-up windows.

RecorderFrame

A pop-up window that allows the operator to start or stop the recording of audio.

MediaRecorder

Records audio, controlled by a RecorderFrame.

PlayerFrame

A pop-up window that can render received audio and video.

MediaPlayer

Processes received audio and video so it can be rendered in a PlayerFrame.

SalClient

Handles the connection to the SAL server and exchanges messages with it.

OperatorServer

Handles the connection from a respondent and exchanges messages with it.

LogWriter

Writes all recordings to log files so they can be sent at the end of a session.

SalRespondentApp

SalRespondentApp

The main application implementing the GUI with the menu bar and pop-up windows.

RecorderFrame

A pop-up window that allows the respondent to start or stop the recording of audio and video, and that can render recorded video.

MediaRecorder

Records audio and video, controlled by a RecorderFrame.

PlayerFrame

A pop-up window that can render received audio.

MediaPlayer

Processes received audio so it can be rendered in a PlayerFrame.

SalClient

Handles the connection to the SAL server and exchanges messages with it.

OperatorClient

Handles the connection to an operator and exchanges messages with it.

LogWriter

Writes all recordings to log files so they can be sent at the end of a session.

Ideas and Future Work

For authoring interface for developing new research experiments

During the design of the interface for running remote SAL experiments, we came up with the idea to design and implement an authoring tool for new research experiments with conversational agents and through a remote network. The authoring tool makes it as well possible to design different running interfaces for naïve and sophisticated respondents and operators. For example an experiment with

a conversation between respondents and a storytelling wizard, who leads the conversation by telling a story, or an experiment for analysing conversation in virtual meeting of several respondents by using avatars through a network connection could be implemented with the same tool as the remote SAL experiment. Remote SAL like experiments creates many new ideas for experiments or changes in the experiments, which could be rapidly implemented or changed with an authoring and implementation tool for research experiments through network connection. Still, the development of such an interface should not cost too much time, as it is not a project task for SAL or HUMAINE at the moment. Anyway, thinking about such an authoring interface will not stand on the way of the implementation of the current remote SAL system.

Design of the authoring interface

With authoring interface all the configuration for remote network connection, hardware setup configuration, functions for the operator and respondent site interface, filling in the experiment content as text and organizing these into interactive multi-user presentations, pictures, movies, real-time rendering of recordings on the other site, recording on the database, electronic questionnaires embedded in the rendered research project, and maybe later real-time speech, facial expression, and hart beat measurement analysis (and visualization of results by graphs in pop-up windows) and rendering of virtual characters with facial and speech expressions etc... Can be made for a research project.

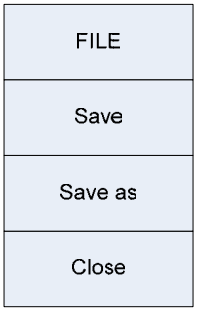


Figure 9: Frame 2

Main Menu:

- File with sub menu

[The name of the project saved or opened is visualized on the Main Menu bar]

1. Save
2. Save as

[Open pop-up window to give name, browse, open folder, save, and close]

3. Close

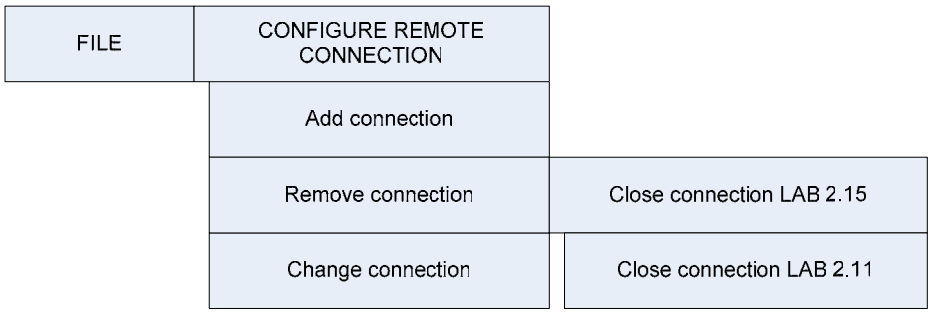


Figure 10: Frame 2 open window for configuration of remote connection

- Configuration Network Connection with sub menu

[Configuration of network connection and hardware setup define what functions are possible to configure for the respondent and the operator site. For example without camera and recordings of the face of the respondent, no real-time analysis to classify emotional state according facial expressions can be made]

[In SAL only connections with two sites, respondent and operator are defined. In other projects the tool should enable to add any amount of sites into one connection and to define for each site own interface functions.]

- add

  [Click opens pop-up window for configuring the connection]
- remove opens list of existing connections

  [Click on connection deletes configuration]
- change opens list of existing connections

  [Click on connection opens configuration in pop-up window for configuring the connection]

[What all information can be filled in through the configuration interface, we still need to think about]

[There should also be a test option for the configured connection in the configuration window]

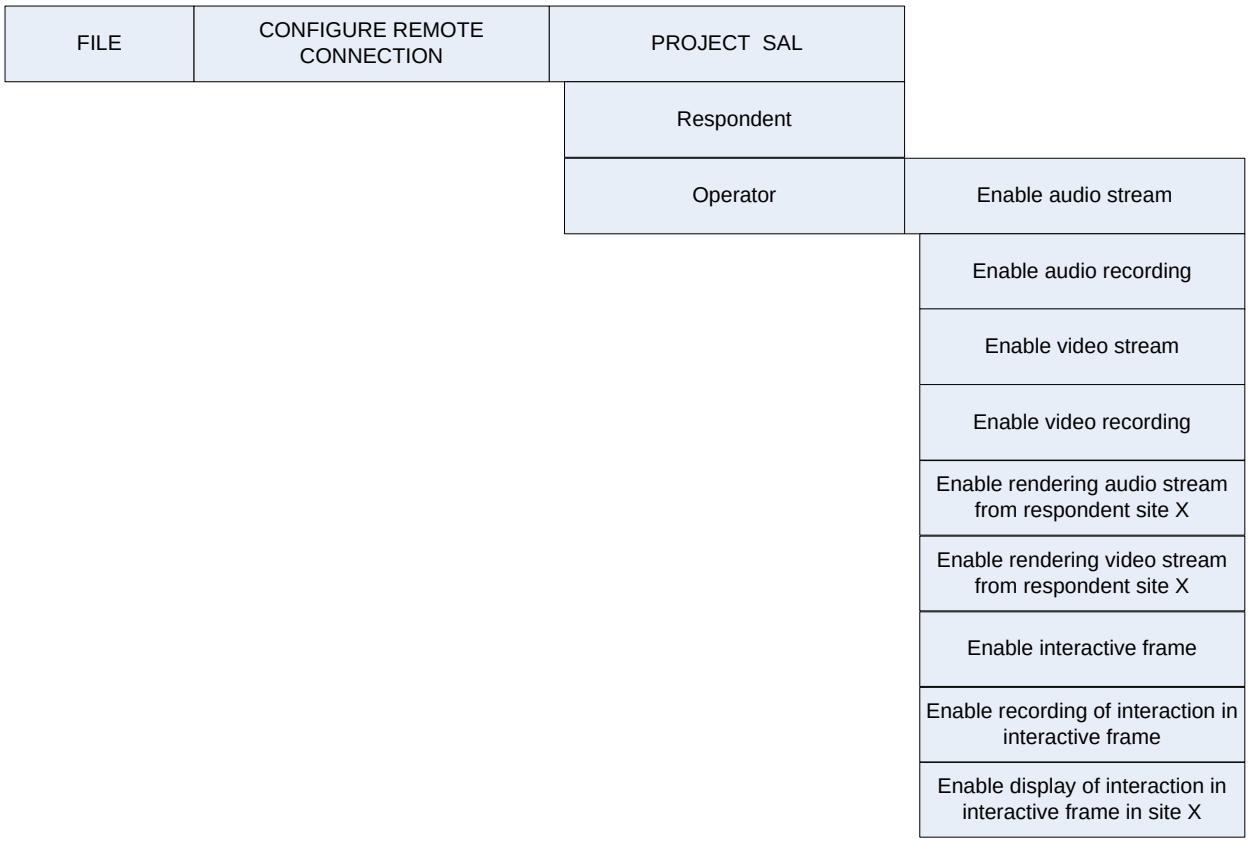


Figure 11: Frame 2 for configuring functionalities on the interface and underlying software for the sites connected

- Configuration functions with sub menu

[Configuration of functions for interfaces and underlying programs for the sites configured for a connection that is open with the hardware setup available.]

1. Respondent site with sub menu of the (interface) functionalities for the respondent site

[Click on functions on sub menu open pop-up window to configure the function]

1. Video stream
2. Video recording
3. Audio stream
4. Audio recording
5. Rendering audio stream from site X
6. Rendering video stream from site X

7. Interactive frame

[To define interactive frame with button sets (for interactive content selection during experiment), multiple choices, editable text- boxes, and display of content material based on the interaction]

8. Recording of interaction in interactive frame

2. Operator site with sub menu of the (interface) functionalities for the operator site

[Click on functions on sub menu open pop-up window to configure the function]

1. Video stream

2. Video recording

3. Audio stream

4. Audio recording

5. Rendering audio stream from site X

6. Rendering video stream from site X

7. Interactive frame

8. Recording of interaction in interactive frame

- Edit research experiment content (as pop-up questionnaires)

[With edit option the content of the experiment as text that belongs to a content category defined by functions as in SAL a reply set]

[For SAL the reply sets (as statements of poppy for positive active respondent state) can be organized hierarchically (first categorize by SAL speaker, then by respondent state) or as a function of two variables (SAL speaker, then by respondent state) chosen by operator]

General ideas for developing remote SAL system

1. At the current and the remote SAL system, the operator classifies the emotional state of the respondent. The operator might use for example the content of text spoken by respondent, voice, facial expressions, and body movement for predicting the emotional state of the respondent. Visualization of the history of the emotion processes experienced by the respondent during the conversation with the SAL speakers would be a nice assistance for the operator to better classify the emotional state of the respondent.

2. In the future the SAL system can be extended by automated detection, analysis and classification of the speech, content of text spoken by the respondent, facial expressions, and body movement. The detection, analysis, and visualization could be done on real-time or off-line by analysing the database. Real-time analyzed data about the emotional state of the respondent could be used for assisting the operator to choose better replies, or for automated selection of the replies. Real-time or off-line analysis of emotional state of the respondents can be visualized on the separate windows.

3. Appraisal is strongly related to arousal of the following vague mental and physiological phenomena; emotion, feeling, mood, and personality. Most of the emotions of the respondents arise from their social and work related problems (might be different in Africa). According the classification of the respondents' content of speech into the coping strategies, better selection of the possible replies can be provided for the operator. We are interested to study how the every-day conversations, as in SAL experiment, can be divided into common occurring social and emotional situation scenarios as for example argument with a boss, loss of dear person, getting divorce etc... There is little literature and publications about which emotional process patterns can be detected in everyday conversations. If we can detect enough situation categories where about the daily conversations are held, the replies of the SAL characters can be thought and organized on a way, that the SAL character intensifies the emotions that are predicted to rise from the current conversation. This will be a step towards more automated SAL speakers. The basic models for emotion dynamics should be taken into account while designing SAL experiment. In the paper of Gratch (A domain-in depended Framework for Modelling Emotion) is shown how range of human emotions and their dynamics can be modeled. We are interested on the models of emotion dynamics. Appraisal of emotion seems to follow some patterns. For example rejection of love from the object of the love results into depressive disappointed state of mind in the subject of love emotion, which again can turn into anger with higher probability in human who's personality does not seem to let them easily to cope with defeat or rejection, or by personalities that tend to be emotionally dependent on other humans. An other example of emotion dynamics given by Gratch is how fear aroused by aggressive suppressor tends to turn into anger towards the suppressor. Instead of letting respondent to choose self the characters they like to interact with, we think it is better to analyze their current mental emotional state, and use characters with replies that with highest certainty transform the respondent into the following emotion according their personality and models for emotion dynamic. For example, after suppressing respondent with depressive and negative comments by Obadiah, respondent can easier turned angry by Spike than directly happy by Poppy. After being angry and relieved by that, it might be easier to make respondent laugh again and try to be happy (like, a little bit irony and a joke to lighten the mind up after loosing nerves seems to be a pattern of natural mood management as well). Further, in the whole HUMAINE network is interest for the recognition of more situational context in order to understand emotions more as processes and patterns corresponding to situational semantics, instead of seeing them as discrete states.

4. The SAL system is supposed to recognize in the future automatically the conversation situation from the natural language of the speech of the respondent, and to use the more 'suitable' replies from the best matching category for the current conversation.

5. SAL character can try to get more information about persons and personalities involved in the situation the respondent is talking about by asking questions as; Is your boss friendly or a terrible person ? Is your daughter easy to handle or a problematic personality? It might not be important, if the actual personality of the persons involved in respondents emotions are correctly classified, if the subjective interpretation of the respondent about those persons is well understood by the system.

6. As human with different personality react or response with different strategies, we think it is interesting to analyze and classify the respondent with help of a short questionnaire in the beginning of the experiment. How human react emotionally during conversation is largely determined by their personality, that influence how the input about the external world situation is mapped into emotional and cognitive state.

7. In future work, we would like to have a better system for the classification of the emotional state of the respondent, which is at the moment many times a guess of the operator. Sometimes and operator just choose a new state by random, so that something changes in the experiment, when the conversation is drying out. It would be handy to give confidence level (uncertainty degree) for the classification of the emotional state. The possibility to choose multiple options to describe an emotion blending is useful as well.

8. For directing or controlling the emotional process of the respondent into some wanted state (like getting the respondent into angry emotion state after being very happy), a new button set and organization of content have to be created. We can design such a directing or controlling of the emotions of the respondent, after analysis of the results of the experiments with the remote SAL system.

9. We want the SAL speakers to be automated conversational and (realistic) animated avatars.

10. There is still not much knowledge about how to stimulate the mood of the human with different types of ambiances, which correspond to the personalities of the 4 SAL characters, created by surround sound system, flickering lights, and visual images. But it is very clear that with sound (high noise, low bass), visual (scary and cute images), light (blue flickering, yellow peacefully), and touch (soft surface, cold surface) the primitive emotion, feeling, and mood processes of the human can be stimulated. For example, with irritating noise, the mental state of a happy positive person can be disturbed to make him nervous or angry. By designing transition from comfortable ambiance to irritating ambiance well, the respondent can, next to other factors, become extra angry through disappointments after the positive experience was disturbed.

11. It might be, that black decoration in a laboratory intensifies emotion and imagination more than for example blue decoration. It is not relevant for SAL experiment to create an artificial experiment conditions, but it is interesting for other research on human emotions.