humaine

D4d: Proposal for exemplars and work towards them: Signs and Signals

Workpackage 4 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	D4d
Deliverable title	Proposal for exemplars and work towards them: Signs and Signals
Type	Report
Number of pages	31
WP contributing to the deliverable	WP 4
Task leader	ICCS-NTUA
Author(s)	S. Abrilian, N. Amir, E. Andre, A. Batliner, R. Cowie, L. Devillers, D. Grandjean, S. Ioannou, K. Karpouzis, J. Kim, S. Kollias, M. Mancini, E. McMahon, J.C. Martin, I. Pandzic, C. Pelachaud.
EC Project Officer	Philippe Gelin

Address of lead author:

Stefanos Kollias
Computer Science Department
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou 15773, Athens, Greece

Table of Contents

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).

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

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

S. Abrilian, N. Amir, E. Andre, A. Batliner, R. Cowie, L. Devillers, D. Grandjean, S. Ioannou, K. Karpouzis, J. Kim, S. Kollias, M. Mancini, E. McMahon, J.C. Martin, I. Pandzic, C. Pelachaud.

The corresponding institutions that have contributed are:

ICCS, TAU, FAU Erlangen, LIMSI-CNRS, UA, UNIGE, FER, Paris 8, QUB.

2 Brief overview of Workpackage 4 and the exemplar proposal

2.1 The field covered by Workpackage 4

The area covered by this workpackage is described in the Technical Annex, particularly in Section 6.2, and in more depth in the review and assessment document for the workpackage. We summarise the area here partly so that the deliverable can be read as a stand-alone document, and partly to draw attention to changes of emphasis that have taken place during the first period of HUMAINE. In general, it became apparent that the fields originally covered by WP4, WP5, WP6 and to some extend WP7 overlap in many aspects and cannot be clearly separated. In this framework, several facets of the research outlined in the deliverables and conducted from the different teams are attributed to more than one WP4 and indeed have to rely on participants of different workpackages to function.

It has to be mentioned that the foreseen activities and tasks have taken into account all comments made by the reviewers in the first annual review of HUMAINE. Particular aspects considered in this framework include emotional models and multimodal emotion analysis, gestural visual analysis, context and emotion analysis in affective interactions. Moreover, the presented exemplar proposals have been created based on specific activities, selected (or extended) among the ones presented in D4C, by the WP4 participants.

2.2 The research objectives

The following sections describe the outline of the work of each subgroup, the results so far and the expected outcomes of joint work, as well as provisions to other workpackages.

2.2.1 Main elements of the exemplar

2.2.1.1 Emotional Speech Signal Analysis

FAU Erlangen initiated CEICES (***C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates - a "forced co-operation" initiative*). Partners so far are from within HUMAINE TAU, QUB, LIMSI-CNRS, ITC-irst, and UA, and from outside HUMAINE, interACT, University of Karlsruhe, and TUM (Technical University of Munich). The first paragraph in the "Agreement of Use" of this initiative reads like this: "The classification performance of emotional user states found in realistic, spontaneous speech is not very high, compared to the performance reported for acted speech in the literature. This might be partly due to the difficulty of providing reliable annotations, partly due to suboptimal feature vectors used for classification, and partly simply to the difficulty of the task. CEICES aims at improving this state of affairs by combining the competence found at different sites that deal with this topic within a "forced co-operation" initiative under the guidance of HUMAINE."

The database used is a German corpus with recordings of 51 ten to twelve year old children communicating with Sony's Aibo pet robot; conceptualization, design and recordings have been done at the University of Erlangen-Nuremberg; recordings were made at two different schools in Erlangen. The AIBO (controlled by a Wizard over WLAN) behaved often disobediently by following its own script irrespective of the child's commands. By that, spontaneous reactions of different children to the same sequence of Aibo's actions could be obtained. For the word-based annotation of *emotion-related user states*, we employed five labellers. This annotation was fully data driven and not based on some emotion theory or model; actually, only two ‘traditional’ emotional labels (*joyful* and *angry*) were used and had a rather low frequency. Moreover, in our UM05 paper, we could show that the traditional emotional dimension *arousal* does not adequately represent the data; instead, besides *valence*, a new dimension which we call *interaction* was found, by applying NonMetrical multiDimensional Scaling.

CEICES is thus in some ways rather orthogonal to two points that are considered to be pivotal in HUMAINE’s activities: multimodality and emotion models: The database is unimodal, and the labels to be used so far are categorical, not dimensional. Results achieved so far at the "originator site", i.e., University of Erlangen-Nuremberg, are documented in papers published in the proceedings of LREC 04, ICASSP 05, UM 05, and Interspeech 05. However, it is well known that competence at one single site concentrates on specific approaches, and can get fossilized w.r.t. extraction and computation of features etc. - we can draw a comparison with pure-bred vs. half-breed dogs: the latter tend to be more alert and intelligent than the former ones. In the same vein, while pursuing this initiative, we do not want to stick to only one approach, or to a competition between different approaches and/or sites: creativity is not necessarily facilitated by strict competition but by co-operation, cf. T. Amabile's fifth myth "Competition beats collaboration". CEICES addresses strict co-operation and benchmarking which is for the moment, in our opinion, only possible if we confine ourselves to unimodality. To our knowledge, a comparable initiative does not exist.

The approach to be followed within CEICES looks like this: the originator site provides speech files, phonetic lexicon, manually corrected word segmentation, emotional labels, definition of train and test samples, etc. Moreover, a balanced sub-sample has been defined which contains roughly the same number of four different cover labels (*Angry, Motherese, Emphatic, Neutral)*; to start with, experiments will be done only for this well-defined sub-sample *AMEN* which contains 6070 words or 3996 turns respectively with manually corrected word segmentation. Annotation has been word-based, thus we aim at two different classification tasks: word-based classification, and turn-based classification; for the latter, the word-based labels will be converted into turn-based ones. Two-fold cross-classification to ensure strict speaker-independence will be applied: the first school class as training set and the second school class as test set, and vice versa. All partners commit themselves to share with all the other partners extracted feature values together with the necessary information (which feature models which acoustic or linguistic phenomenon, format of feature values, classifier used, etc.). The files containing the feature values will be exchanged via up- and download facilities at the HUMAINE server. Thus each site can assess the features provided by all other sites, together with their own features, aiming at a repertoire of optimal features. We want to look not only at acoustic but at linguistic features as well.

Points that we want to address after these two basic benchmark experiments are:

- Taking into account more context, e.g., acoustic tri- or five-gram context information for word-based classification
- Automatically computed word-segmentation vs. manual word segmentation
- Benchmarking different classifiers with other things being equal
- Hard vs. soft (dimensional) labelling

- Comparison of classification systems
- Assessment of manual vs. automatically extracted pitch values

From this co-operation, we certainly expect improved recognition rates. Note, however, that the answers to such everlasting questions as addressed above need not be given in terms of higher performance: it is an educated guess that, for instance, manual segmentation yields more reliable and by that, better results – but we simply do not know yet whether and to what extent this will turn out to be a fact.

At the moment, the different sites have started to adapt feature extraction to the CEICES speech data and to compute acoustic (for instance, new spectral) features; for instance, as most sites have only dealt with adult speech, thresholds appropriate to children's voices have to be found.

2.2.1.2 Emotional Visual Signal Analysis

Manual Annotation and Image Processing of Multimodal Emotional Behaviours in TV Interviews

There has been a lot of psychological researches on emotion and nonverbal communication of facial expressions of emotions (Ekman 1999).

Static postures were recorded by De Silva et al. (De Silva et al. 2005) using a motion capture system during acted emotions (two nuances for each of four basic emotions ; e.g. upset and angry as nuances of anger). In Gunes et al. (Gunes and Piccardi 2005) the video processing of facial expression and upper body gestures are fused in order to recognize 6 acted emotional behaviours (anxiety, anger, disgust, fear, happiness, uncertainty). A vision-based system that infers acted mental states (agreeing, concentrating, disagreeing, interested, thinking, and unsure) from head movement and facial expressions is described in el Kaliouby et al. (el Kaliouby and Robinson 2005). Choi et al. (Choi and Kim 2005) describe how video processing of both facial expressions and gaze are mapped onto combinations of seven emotions (neutral, surprise, fear, sadness, anger, disgust, happiness).

Thus, real-life multimodal corpora are indeed very few despite the general agreement that it is necessary to collect audio-visual databases that highlight naturalistic expressions of emotions as suggested by Douglas-Cowie et al. (Douglas-Cowie et al. 2003).

Regarding hand gestures and body movements during acted data, it had been observed that the perception of emotions was related to the expressivity of movement (DeMeijer 1989; Newlove 1993; Boone and Cunningham 1998; Wallbott 1998). It led Pelachaud et al. (Hartmann et al. 2005) to propose six parameters (overall activation, spatial extent, temporal extent, fluidity, power, and repetition) for implementing gestures expressivity in the Greta expressive agent. But setting adequately the values and the temporal variation of these expressive parameters to achieve realistic behaviours requires some experimental data.

Building a multimodal corpus of real-life emotions is challenging since it involves subjective perception and requires time-consuming annotations of emotion at several levels. Manual annotation might benefit from image processing via the objective and automatic detection of emotionally relevant segments, the validation of the subjective annotation of expressivity parameters such as speed or spatial extension, and possibly the reduction of annotation time for example by providing segments of movement to the manual annotation process.

Yet, most of the work in image processing of emotional behaviour has been done on high quality videos recorded in laboratory situations where emotions might be less spontaneous than during more spontaneous TV interviews. But the performances of automatic processing techniques for processing emotional behaviour in TV quality videos remains to be explored.

Furthermore manual annotation and image processing provide information at quite different levels of abstraction and their integration is not straightforward.

In this framework, LIMSI-CNRS are leading the creation of a multimodal corpus of real-life emotions, based on non staged TV interviews (Devillers et al. 2005; Abrilian et al. 2005a). The EmoTV corpus features 50 videos samples of emotional TV interviews. The images are 720*576 and the image rate is 25 images/sec.

The goal of this corpus is to provide knowledge on the coordination between modalities during non-acted emotionally rich behaviours. The video is annotated at several temporal levels (whole video, segments of the video, behaviours observed at specific moments) and at several levels of abstraction (multimodality, emotion, context). The global behaviour observed during the whole video is annotated with contextual information, emotions and multimodal cues. The segments are annotated with emotion labels, the modalities perceived as relevant with regards to emotion, and also include an annotation of the temporal variation of the intensity of the emotion during the segment.

The definition of these coding schemes were grounded on requirements collected from both the parameters described in the literature as perceptually relevant for the study of emotional behaviour, and the features of the TV interviews found in the corpus.

A more detailed description of multimodal behaviours in each segment includes tracks for each visible modality: torso, head, shoulders, arms, facial expressions, gestures and global body (Martin et al. 2005; Abrilian et al. 2005b). These tracks contain a description of the pose, and of the movement. Pose and movement annotations thus alternate. The attributes of movement quality for these modalities that we considered as relevant for the corpus are: the number of repetitions, the fluidity, the strength, the speed, the spatial expansion. Regarding the overall activation, there is an annotation of multimodal cues at the level of the whole video.

Image processing might provide automatic estimations of head and hand movements. The task of head and hand localization in image sequences is based on detecting continuous areas of skin colour. For the given application, a very coarse model is sufficient, since there is no need for recognition of hand shape. Furthermore, the videos considered in the EmoTV corpus are of low resolution and, therefore, the skin regions are small and possess very low detail; in addition to this, colour resolution and fidelity can suffer from analogue to digital conversion. 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 colour conditions do not usually change within the clip, detection and tracking are performed automatically. Another usual impediment to image processing of TV videos 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.

The measure of movement 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. 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 further tracking only in moving image areas. Both colour and motion masks contain a large number of small objects due to the presence of noise and objects with colour similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. In the following, the moving skin mask is created by fusing the processed skin and motion masks, through the morphological reconstruction of the colour mask using the motion mask as marker.

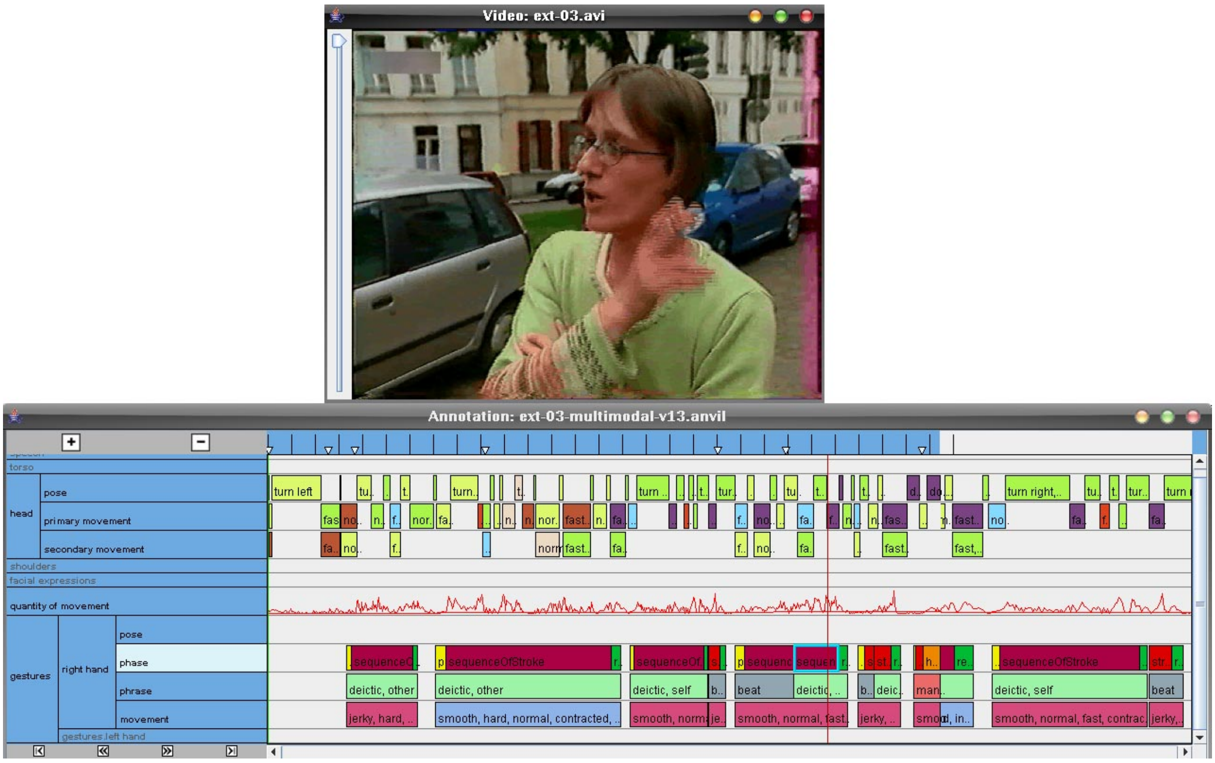

Figure 1: Integration of manual annotation (coloured blocks) for head movements (up) and gesture (low) and results of image processing for the estimation of movement quantity (continuous line in the middle).

We illustrate our approach on the combination of image processing and manual annotation on video #3 of the EmoTV corpus (duration 29 seconds, frame rate 25 fps, 722 frames). The image processing module provides information related to emotional behaviour at several levels. At the level of the whole video, an estimation of the overall activation is computed. For the video #3, this overall activation normalised by the number of frames is 1340. It was compared with the results obtained with two laboratory recorded videos with different behaviours but similar viewpoint. The overall activation for a video with fewer movements (showing a single gesture) was smaller (44). For a video with more activation (showing several repetitive gestures), this value was higher (2167). Preliminary tests enabled to investigate the thresholding of this image processing results to produce an estimation of segments of movements. Different values of estimated movement quantity for hand movement detection and the sampling rate of these thresholded values were observed to leads to different values of agreement between the manual annotations and the image processing. The next steps will be to proceed to a thorough evaluation of the correlations between image processed estimations of movement quantity and : the different phases of gestures expected to be enriched in movement (preparation, retraction, repetitive strokes) as well as the annotation of not only hand gestures but also movements of torso and head.

Facial expressions and the component process model

The component process model approach aims to decode the underlying mechanism of emotion. In this perspective, emotion is considered as a theoretical construct that consists of five components corresponding to five distinctive functions; cognitive component for evaluation of objects and events, the peripheral efferent component for system regulation, the motivational component for preparation and direction of action, the motor expression component for communication of reaction and behavioural intention, and the subjective feeling component for monitoring of internal sate and organism-environment interaction. As these components are part of the psychobiological endowment of higher organisms, one might ask how emotional states are to be distinguished from non-emotional states in the flow of experience of an organism. In the framework of the component process model, emotion is defined as an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism. Concretely, the term "emotion" is reserved for those periods of time during which several organic subsystems are coupled or synchronized to produce an adaptive reaction to an event that is considered as central to the individual's well-being. The major features of this definition are discussed in greater detail below.

According to cognitive theories of emotion, emotions are closely related to the situation that is being experienced (or, indeed, imagined) by the agent. Specifically, emotions are connected to mental representations that emphasize key elements of a situation and identify them as being either positive or negative. These representations have generally been called appraisals. An appraisal can be thought of as a model which is selective and valenced – i.e., highlights key elements of a situation and their values for good or ill. Appraisals are not necessary conscious, thus the evaluation processes can occur also by an unconscious way as demonstrated by a important corpus of study in cognitive neuroscience, with different methods as subliminal presentations of stimuli or by clinical neuropsychology.

Scherer has developed an appraisal model of emotion in which emotions are conceptualized as the outcome of a fixed sequence of checks. According to Scherer’s view, emotion serves an important function as “…an evolved phylogenetically continuous mechanism that allows increasingly flexible adaptation to environmental contingencies by decoupling stimulus and response and thus creating a latency time for response optimization”.

The appraisal is the sequence of Stimulus Evaluation checks (SECs), which represent the smallest set of criteria necessary to account for the differentiation of main groups of emotional states. These checks are not necessarily binary and are subjective (i.e. they depend on both the appraising individual’s perception of and inference about the specific characteristics of the event.

The individual SECs can be grouped together in terms of what are called Appraisal Objectives, of which there are four: 1) Relevance Detection: comprising Novelty Check, Intrinsic Pleasantness Check, and Goal Relevance Check; 2) Implication Assessment: comprising Causal Attribution Check, Discrepancy from Expectation Check, Goal/Need Conduciveness Check, and Urgency Check; 3) Coping Potential Determination: comprising Control Check, Power Check, and Adjustment Check (can the event be controlled, if so by how much power do I have to exert control, and if not can I adjust?); 4) Normative Significance Evaluation: comprising Internal Standards Check, and External Standards

Check. A major assumption of Scherer's SEC Theory is that the sequence of the checks and of the groups is fixed. However, this does not rule out parallel processing as, in theory, all of the SECs are processed simultaneously (see Sander, Grandjean, & Scherer for a discussion about the sequential view).

Representations of emotional states using this model of emotion are explained in terms of cognitive appraisals of the antecedent situation, and these appraisals account for the differentiated nature of emotional responses, individual and temporal differences in emotional responses, and for the range of situations that evoke the same response. Appraisals also make appropriate emotional responses likely, and conflict between automatic, unconscious appraisals and more consciously deliberated ones may explain some of the more irrational aspects of emotions. To the cognitive neuroscience perspective, Sander, Grafman, and Zalla (2003) recently suggested, in a review paper, that amygdala could subserve a mechanism of relevance detection, a central concept of appraisal theory. Thus, progressively, different scientific fields take into account the appraisal view of emotion to better understand the emotional processes into their entire complexity. To modelling the intra and inter individual differences (including the personality view) into the genesis of emotional processes the appraisal theory is probably the best way to give an opportunity of a realistic implementation, even this will not be easy (Sander, Grandjean, & Scherer, 2005).

2.2.1.3 Emotional Physiological Signal Analysis

Augsburg University designed and implemented a toolbox for physiological data analysis called AuBT encompassing all steps from feature calculation, feature extraction and classification. The toolbox includes several feature selection/reduction methods, such as analysis of variance (ANOVA), sequential forward selection (SFS), sequential backward selection (SBS), principal component analysis (PCA) and Fisher projection as well as three well-known classifiers, linear discriminant function (LDF), k-nearest neighbour (kNN) and multilayer perceptron (MLP).

Among other things, the toolbox has been used to analyse a data set we recorded at Augsburg University based on the musical induction method. Recognition rates of about 80% were achieved for all three classifiers. By applying appropriate feature reduction, results could be improved up to 92%. Our recognition rates are comparable to those achieved in the literature. However, in contrast to earlier work, we rely on an automatically selected feature set. To compare the results, a second data set was used which had been recorded at MIT Media Lab. It turned out that for both data sets it was easier to distinguish emotions along the arousal axes than the valence axes. Statistical methods were used to find out which features are significant for a specific emotion for each data set. Although differences in the physiological response of the subjects were noticed we also found similarities, e.g. a higher breathing rate for emotions with a negative valence.

A major objective is the improvement of the user interface of AuBT to make it available to the Humaine members in 2006. Among other things, AuBT will be used to analyze the data acquired for a car driving simulator experiment to be conducted by QUB. Based on the test results by the Humaine partners, an extended version of AuBT including further classification as well as visualization methods will be developed. Furthermore, links will be established to the FP6 IP project Sensation which aims at the provision of unobtrusive nano-sensor technology for analysis of the human physiological state. A first meeting is planned for January 2006 to discuss concrete ideas for the exchange of technology.

Work at QUB started with initial laboratory experiments that induced emotion-related states, which affect performance on a range of tasks (boredom/ drowsiness, excitement/ frustration/

irritation, calm vs. pressurised states, states affected by mental loads), and examined their effects on physiological indicators (HR, GSR, RSP, SKT), and in some cases (boredom & excitement), their vocal correlates.

To illustrate, in one study calm states were induced by playing tranquil music (by Chopin) for about a minute. Pressurised states were induced by presenting a difficult discrimination task, with loud threatening noises (‘phaser fire’) that continued until the task was completed correctly. Each time a subject responded with an incorrect solution to the discrimination task, they were required to count aloud to ten before responding again. Subjects experienced sixteen cycles of calming music followed by pressurized states. During one of the tranquil periods, they were subjected to a sudden, loud, crack close behind them.

The graph below shows results for a representative subject. The vertical lines (coloured pink) mark experimental events. The vertical lines above the horizontal axis are in groups of three. The first (of height one) marks the time of a warning tone which followed the end of a Chopin excerpt. The second (of height two) marks the time when the screen containing the experimental task appeared. The third (of height three) marks the time when the screen was replaced by a mask of random dots and the phaser fire noise began. The last vertical line in each group, which extends below the horizontal axis, shows when a correct response was made. Its length shows the number ot attempts that needed to achieve a correct answer.

The physiological responses are clearly associated with these events. The yellow line shows skin conductance. It rises relatively quickly after the warning tone. The dark blue line shows heart rate (in beats/sec). It rises slightly more slowly after the warning, and tends to show a peak for each response (right or wrong). The light blue line shows respiration. It dips below the axis, showing expiration, during the test period. Because the response is oral, it does require chest movements, but not of that magnitude. The skin temperature data are not shown because they were not systematically related to emotion.

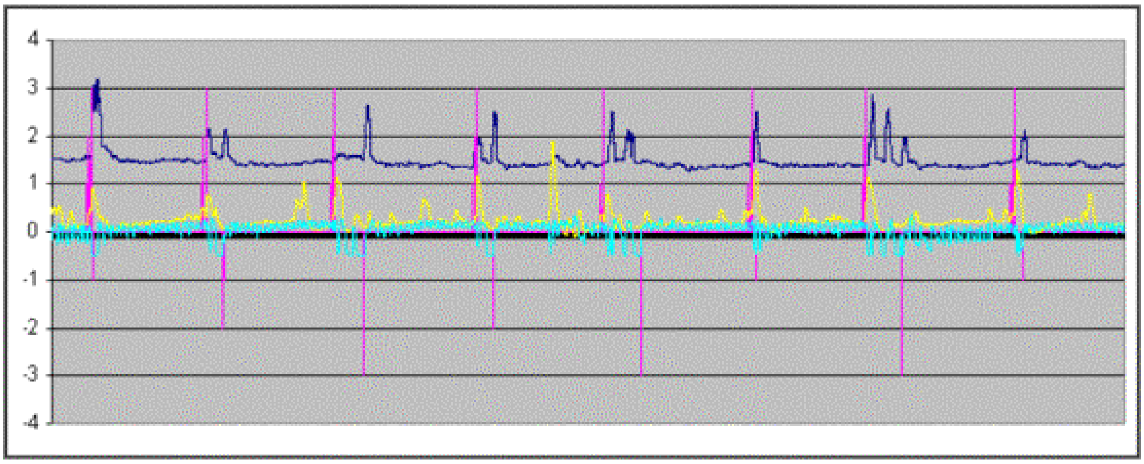

Figure 2: ‘Chopin’ experiment results

In the driving simulator scenario, we adapted the laboratory techniques to induce stress-related states in a setting that is more closely related to real life situations, i.e. a driving simulator. We have developed simulations which present conditions designed to induce these states, either by the nature of the driving or by imposed loads. We considered stress rather than emotion per se on the assumption that the issues would be simpler. Data recorded from each experiment was of three types:

- Psychophysiological: Raw measures were ECG, GSR, chest circumference, skin temperature, sampled at 200Hz.

- Driving performance: Raw measures were: lane position; velocity & acceleration, lateral & longitudinal; use of brakes, wheel and throttle; inherently dangerous actions (driving off-road, crashes)

- Audio/visual: Video cameras were placed to record drivers' facial expressions and body movements (expt 1) and head mounted microphones were used for audio recordings of drivers' speech (expts 2-4).

Concluding experiments at the QUB site:

- We have developed the STISIM Driving Simulator as a platform for running psychological experiments. The platform lets us develop bridges between fully realistic settings and conventional, designed experiments, and helps us to see how complex the situations are

- We have considered physiological data which could reasonably be expected to indicate stress-related states. The relationships were less stable and more complex than one would have hoped.

- We have explored performance measures, in the form of 'exceedances' (going beyond a limit that would be considered safe). These do show dependences on stressors, but they are (quite reasonably) linked to situation in quite complex ways.

2.2.1.4 Multimodal Emotion Recognition

Quite significant research results have been obtained using individual modalities for emotion recognition. In the case of vision, most approaches work on the facial area, while hand and body gestures have only recently been taken into account, and again in controlled environment; this may well be attributed to that fact that gesture analysis requires algorithms with high computational complexity and, therefore, can hardly be performed in real-time and to the occurrence of multiple occlusions in the palm/finger area. Regarding aural information, prosody is always a reliable source of emotion-related characteristics, while linguistic information is usually subject to culture- and personality-related interpretation and requires syntactic and context analysis. However, achieving reliable assessment typically requires the concurrent use of multiple modalities (i.e. speech, facial expression, gesture, and gaze) occurring together.

Relatively few papers have focused on implementing emotion recognition systems using multimodal information, usually either audiovisual (e.g. Potamianos et al) or face and gestures (Balomenos et al). Fusion of the results of each modality can either occur at feature level or, later, at the decision level after separate identification of emotion from each modality.

Feature-level fusion is performed by merging extracted features from each modality into one cumulative structure and feeding them to a single classifier, generally based on multiple HMMs or neural networks. In this framework, correlation between modalities can be taken into account automatically via learning. In general, feature fusion is more appropriate for closely coupled and synchronized modalities, such as speech and lip movements, but tends not to generalize very well if modalities differ substantially in the temporal characteristics of their features, as is the case with speech and facial expression input. Due to the high dimensionality of input features, large amounts of data must be collected system and labelled for training purposes

Decision (semantic) level fusion caters for integrating asynchronous but temporally correlated modalities. Here, each modality is first classified independently and the final classification is based on the fusion of the outputs from the different modalities. Designing optimal strategies for decision level fusion is still an open research issue. Various approaches have been proposed, e.g. sum rule, product rule, using weights, max/min/median rule, majority vote etc. As a general rule, semantic fusion builds on individual recognizers, followed by an integration process; individual recognizers can be trained using unimodal data, which are easier to collect. However, this approach still fails to model the interplay between the different modalities, a fact which one can take exploit to fortify the results obtained from an individual modality (e.g. between visemes and phonemes) or resolve uncertainty in cases where one or more modalities are not dependable (e.g. speech analysis in the presence of noise can be assisted by visually extracting visemes and mapping them to possible phonemes).

Fusion of physiological signals with audiovisual channels

Augsburg University conducted a Wizard-of-Oz experiment in order to acquire a corpus of spontaneous vocal and physiological data that reveal information on the user's emotional state. In the work three different fusion methods are evaluated and compared with the unimodal recognition approaches. In feature-level fusion, the features of both modalities were simply merged and provided as input to a single classifier. Thereby, we also attempted to extract the most significant features from the fused features by sequential forward selection (SFS) to compare the results. In decision-level fusion, the outputs of two unimodal classifiers for speech and biosignals were integrated according to a given set of criteria. We employed posterior probability criteria as well as majority voting to the decision process, according to the recognition rates of each emotion from unimodal classifiers. Finally, we employed a new hybrid scheme of the two fusion methods in which the output of feature-level fusion is also fed as an auxiliary input to the decision-level fusion stage.

The best results were obtained by feature-level fusion in combination with feature selection. In this case, not only user-dependent, but also user-independent emotion classification could be improved compared to the unimodal methods. We did not achieve the same high gains that were achieved for audio-visual data. The results seem to indicate that speech and physiological data contain less complementary information. In fact, in a natural setting like ours, we cannot exclude that the subjects are inconsistent in their emotional expression. Inconsistencies are less likely to occur in scenarios where actors are asked to deliberately express emotions via speech and mimics. This adds the importance of fusion algorithms adapting to given signal types to be combined. First results of this work have been presented at Interspeech 2005.

2.2.2 How the subtasks link to each other

Regarding work conducted within WP4 itself, the first three objectives coincide with the thematic areas of the first subgroups dealing with the processing of audio, visual and physiological information, respectively. Here, the main task is to improve the robustness and effectiveness of unimodal processing in each case, in order to come up with useful feature-extraction and recognition schemes; on top of that, each subgroup must also work towards the goal of the fourth subgroup as well, since its theme, multimodal emotion recognition, is one of the features essential but still lacking in current recognition infrastructures. To this goal, some of the existing feature-extraction and representation concepts need to be redesigned, to cater for the cooperation of the different modalities in a single, multimodal recognizer. Besides this, exploring the multimodal case is expected to provide notions of fall-back or reinforcement solutions, when more than one modalities are present, but not all of them can provide useful information; an example of this would be an audiovisual speech-to-text system, where visual lip reading can prove useful in noisy situations, where the audio channel may become incapacitated.

2.2.3 How the subtasks link to other aspects of HUMAINE

Workpackage 4 can be thought of as supplying the ‘ground truth’ to other WPs that produce synthetic ECAs or utilize high-level knowledge, since it processes real, naturalistic data from everyday human-machine or human-human intercourse to provide measurable definitions of events and concepts related to emotion and affect. In this framework, one would think of WP4 as building on the results of WP5, that provides the actual data and annotation schemes, and on emotion representation models defined within WP3. The first outcome of this work is measurable feature definitions, utilized by WP6 to render expressivity on ECAs; besides this unimodal and multimodal recognition infrastructures can be utilized by WPs working on higher-level concepts, such as WP7, to supply the much-wanted bridge between high-level concepts such as task-oriented behaviour, with features automatically detected from aural, visual or other elementary steams. Section 7 in deliverable D4b provides an extensive list of specific interactions between WP4 and other WPs.

It has to be noted that for some purposes of rendering expressivity, for example to design and recreate expressive gestures on an ECA, specific values and temporal evolution of features from real, naturalistic data are not indispensable; indeed, high-end animation productions utilize motion capture techniques or talented animation designers for this job, without resorting to automated or even semi-supervised signal processing algorithms or high-level emotion representation or filtering concepts. Although this approach is still feasible within Humaine in the sense of extending Wizard-of-Oz environments with ECAs acting on predefined animations and synthetic speech, the vision of truly interactive, emotion-aware systems need to be built on actual feature-based definitions of affect signs, so that these are recognized on the input side and recreated truthfully by ECAs on the output side. Joint research across WP4/WP5/WP6 enables to explore required levels of representation of the emotional multimodal behaviour and the required levels of fidelity in the replay via perceptual studies.

3 The planned program of research

3.1 Element 1: Emotional Speech Signal Analysis

3.1.1 Leader

Noam Amir, Tel Aviv University

3.1.2 Main participants

FAU Erlangen, LIMSI-CNRS, Queen’s University of Belfast, University Augsburg, ITC-irst

3.1.3 Main steps planned towards producing element 1

Subtask	Carried out by	Start / end dates
1.1 CEICES first benchmark experiment: Word-based classification	FAU, TAU, QUB, LIMSI-CNRS, ITC-irst, UA	June 2005 – June 2006
1.2 CEICES second benchmark experiment: Turn-based classification	FAU, TAU, QUB, LIMSI-CNRS, ITC-irst, UA	June 2005 – June 2006
1.3 CEICES further actions	FAU, TAU, QUB, LIMSI-CNRS, ITC-irst, UA	March 2006 - December 2006

Task description:

Task 1.1 CEICES first benchmark experiment: Word-based classification

The experiment will be based on generated speech files, phonetic lexicon, manually corrected word segmentation, emotional labels, definition of train and test samples. A balanced sub-sample has been defined which contains roughly the same number of four different cover labels (*Angry*, *Motherese*, *Emphatic*, *Neutral)*; in particular it contains 6070 words or 3996 turns respectively with manually corrected word segmentation.

Task 1.2 CEICES second benchmark experiment: Turn-based classification

The second experiment runs in parallel with the first. However, in this case, the word-based labels will be converted into turn-based ones, so that data are appropriately annotated. The results will be compared with those obtained in the first experiment.

Task 1.3 CEICES further actions

The actions to be taken following completion of the two basic benchmark experiments will include first investigation of context, e.g., acoustic tri- or five-gram context information for word-based classification, while automatically computed word-segmentation vs. manual word segmentation. Various classifiers will be compared, with benchmarking experiments. Hard vs. soft (dimensional) labelling will also be investigated. An assessment of manual vs. automatically extracted pitch values will be carried out as well.

3.2 Element 2: Emotional Visual Signal Analysis

3.2.1 Leader

Stefanos Kollias, ICCS

3.2.2 Main participants

Paris8, LIMSI-CNRS, GERG

3.2.3 Main steps planned towards producing element 2

Subtask	Carried out by	Start / end dates
2.1 Manual Annotation and Image Processing of Multimodal Emotional Behaviours in TV Interviews	LIMSI-CNRS, Paris8, ICCS	June 2005 – December 2006
2.2 Facial expressions and the component process model	GERG, Paris8, ICCS	June 2005 – December 2006
2.3 Expressivity in gestures	Paris8, ICCS, LIMSI-CNRS	June 2005 – December 2006

Task description:

Task 2.1 Manual Annotation and Image Processing of Multimodal Emotional Behaviours in TV Interviews

Annotation of EmoTV material is a large project within Humaine, combining research groups from WP3, WP4 and WP6 to provide labelled material for recognition and synthesis purposes. In this framework, automated or user-assisted techniques can assist in labelling large video databases, by providing measurable definitions of facial and hand expressivity; these definitions can be used, for example, for motion cloning in ECAs or to provide higher-level emotion representations for the relevant WPs.

Task 2.2 Facial expressions and the component process model

The aim of this task is to test the theory of appraisal checks, using facial expression definitions from real data, as well as investigate several concepts of this theory which are not clear, by producing synthetic videos for subjective evaluation. These concepts include the actual method of superimposing the different appraisal processes (purely sequential, additive, linear or other combination), temporal and duration aspects, the means of producing the final facial expression, etc. A graphics animation environment is being developed within the Greta system. It will allow the manual specification of signals (facial expression, gesture, etc) along time. The animation will be computed from the set of signals that could happen sequentially, overlapping each other or in parallel. Its aim is to provide a visualization tool for facial expressions arising from appraisal processes.

Task 2.3 Expressivity in gestures

In this task, participating research teams process videos of expressive gestures to provide actual measurements of the expressive parameters used in the Greta ECA. Several videos of acted gestures have been recorded, segmented and processed with the algorithms described in Task 2.1 to provide values for the six expressive parameters proposed by Paris8. In addition to this, Hidden Markov Models (HMMs) have been implemented as a gesture recognition environment for a relatively small gesture vocabulary (seven everyday gestures). This recognition environment will be extended to cope with expressivity and include more gestures, described both in low-level, as well as by higher-level representations, derived and analyzed from the EmoTV videos.

3.3 Element 3: Emotional Physiological Signal Analysis

3.3.1 Leader

Jonghwa Kim, University of Augsburg (UA)

3.3.2 Main participants

FAU, QUB, ICCS-NTUA

3.3.3 Main steps planned towards producing element 3

Subtask	Carried out by	Start / end dates
3.1 Extending sensor parameters depending on physiological priority	UA, QUB, FAU	June 2005 – June 2006
3.2 Fusion of extended biosensor types	UA, ICCS	December 2005 – December 2006
3.3 Adding correlation-based features	UA, ICCS, QUB	December

		2005 – December 2006
3.4 Refining classifier combination	UA, FAU, ICCS	March 2006 – December 2006
3.5 Online recognition engine	All	March 2006 – December 2006

Task description:

At the end of the Augsburg workshop presented in the following section, 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

Further issues of common interest are:

- AuBT: evaluation at the three sites (UA, QUB, FAU)
- feature evaluation and interpretation in general
- defining analysis windows
- concepts for fusion with speech (and facial expressions)
- defining standards for such multi-modal databases

Task 3.1 Extending sensor parameters depending on physiological priority

In addition to the sensors we used so far we will concentrate on new biosensor types, such Temp, BVP, Respiration Noise, etc. This task includes the calibration of the sensors and the estimation of the sensor priority based on empirical pre-knowledge and psycho physiology in order to establish a robust sensor configuration.

Task 3.2 Fusion of extended biosensor types

To cope with incomplete and potentially inconsistent information potentially arising from the extended biosensor set (Task 3.1), this task covers the development of a probabilistic framework which fuses heterogeneous sensorial data at various levels of abstraction. In particular, we will investigate methods for signal-, feature- and symbol-level fusion. This task comprises the design and implementation of a modular, flexible and scalable architecture for multi-layer biosensor fusion which allows for the easy integration of an open set of sensors.

Task 3.3 Adding correlation-based features

Efficient feature selection should make a multimodal emotion recognition system more reliable. In addition to features, such as derivative, difference, and mean values of raw signals which were used in most works so far, we will extract a cross-correlation feature set between the various biosignal types.

Task 3.4 Refining classifier combination

We will use a hybrid approach to combine various statistical and probabilistic pattern recognition methods for improved classification and make use of a consensus theory to achieve a consistent decision.

Task 3.5 Online recognition engine

Using online-training methods, we will develop an online recognition engine.

3.4 Element 4: Multimodal Emotion Recognition

3.4.1 Leader

Didier Grandjean, GERG

3.4.2 Main participants

GERG, ICCS-NTUA, UA, QUB, TAU, FER Zagreb, Paris-8

3.4.3 Main steps planned towards producing element 4

Subtask	Carried out by	Start / end dates
4.1 Joint audiovisual processing	GERG, ICCS, TAU, FER, QUB, UA	June 2005 – December 2006
4.2 Fusion of physiological signals with audiovisual channels	All	December 2005 – December 2006

Task description

The adoption and usage of different emotional models in emotion recognition will be investigated within the three separate WP4 subgroups. In parallel, there will be two main subtasks for achieving multimodal emotion recognition, which are listed below.

Task 4.1.1 Synchronizing different window lengths and units in audio, visual and physiological channels

Experiments will be conducted with different window sizes in order to shed light on the question of how they influence the recognition rates. Based on the results we will investigate how to combine results when different and overlapping window lengths have been used that are adapted to the specifics of the single input channels.

Task 4.1.2 Audiovisual speech recognition

A serious obstacle of speech recognition and transcription systems is their dependence on good speech quality and the absence of noise. While this requirement can be easily met in human-computer interaction setups, e.g. when users sit in front of their computer, it can rarely be enforced in uncontrolled, everyday environments. In this framework, research teams will work on the ‘bartender problem’, e.g. how to use visual information in the form of visemes to resolve speech when audio information is unclear. ICCS, TAU and FER will initially work on specialized databases, recorded for this reason and then move on to naturalistic data produced from WP5.

Task 4.1.3 Evaluation of integrated feature set

This task covers activities to evaluate the efficiency of new speech and facial feature sets supplemented by biosensor data and to arrange and select the most relevant features using the classification error-criterion.

Task 4.2.1 Feature fusion of biosignal with audiovisual signals

An empirical approach to fusing features from bio-speech-visual channels will be applied using the classification error-criterion to find the most relevant fused feature set for emotion classification. Moreover, we will need a model of fusion that is able to explicitly represent and update the belief of multiple sensory observations and efficiently combine potentially incomplete data gained from different channels.

Task 4.2.2 Possibility of biosignal as baseline-channel

We will consider the possibility to give priority to bio signals over audio/visual channels since they allow us to continuously gather affective information from users. This information can be used to estimate the “baseline” for analyzing signals from audio channels, similar to the variation of the emotional intensity under a certain “mood”.

Task 4.2.3 Correlating SC/RSP-measurements with speech intonation

We will investigate how skin conductivity and breathing change depending on speech intonation.

Task 4.2.4 Fusion of features from biosignal and FAP

Fusion of features from biosignal and FAP will be performed to closely correlate the physiological/visual modalities for improved recognition accuracy.

3.5 Steps to ensure co-ordination

The basic means of collaboration will be phone conferences, visits between labs, some of which taking advantage of the interchange funding provided by HUMAINE, and meetings. Such a meeting has already taken place within the speech subgroup, within the physiological subgroup, while an interchange visit is planned in the audio-visual subgroup. In addition to this, two meetings were held to start the cooperation between ICCS and LIMSI-CNRS (during HCII & AVI conferences). A joint paper has been submitted to LREC'2006 where the next meeting will be held in May 2006.

3.6 Steps to ensure dissemination

Report on Workshop "Hands-On Training on Physiological Data Processing" in Augsburg

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, we 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. We 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, we also discussed problems arising during the fusion of multimodal signals including:

- *Window Size:* Different modalities require different window sizes to improve classification: emotions show-up at different speeds in different modalities. For example heart rate changes slowly while a speech feature can be detected in a sub-second duration. How to combine results from different modalities when different and overlapping window lengths have been used?
- *Facial Analysis:* What about slow-evolving emotions and fast-evolving expressions? Should we use different-sized overlapping windows?
- *Fusion Level:* When combining results from different modalities, should the combination be at the feature level, decision level, a combination, or something else?
- *Conflicting Results:* What should we do in case of contradicting results from different modalities? How should we deal with such a case when selecting the training set for a multimodal classifier?

Report on Special session organized in ICME-2005 by Humaine WP4

A special session was organised in the IEEE International Conference on Multimedia and Expo (ICME-2005) in Amsterdam in June 2005; many HUMAINE partners participated in this and in the Conference, including ICCS-NTUA, LIMSI/CNRS, University of Augsburg, FER Zagreb, Paris8, QUB, and KCL presenting their own and collaborative work in the processing, analysis of input signals for emotion recognition and affective interaction.

The special session which was one of the four special sessions accepted in the Conference, also included a 40 min speech from the organisers, S. Kollias and K. Karpouzis, of ICCS-NTUA, presenting the main issues in multimodal emotion recognition, as derived from the first year work of HUMAINE WP4. The framework of this presentation was the investigation of the best possible techniques for multimodal emotion recognition and expressivity analysis in human computer interaction, based on a common psychological background. The session mainly dealt with audio and visual emotion analysis, with physiological signal analysis serving as supplementary to these modalities. Specific topics that were examined included extraction of emotional features and signs from each modality in separate, integration of the outputs of single-mode emotion analysis systems and recognition of the user's emotional state, taking into account emotion models and existing knowledge or demands from both the analysis and synthesis perspective. Various labelling schemes, supplies of accordingly labelled test databases, as well as synthesis of expressive avatars and affective interactions, were issues brought up and examined in the proposed framework.

All presentations have been included in the Proceedings of the ICME Conference and constitute point of reference for international related research. In addition to this, a dedicated web page with copies of the presentations and references to related papers from the conference has been setup up at http://www.image.ece.ntua.gr/icme2005

Special sessions related to WP4

A special session related to WP4 is planned for June 2006 during the 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI 2006 – http://www.icsd.aegean.gr). This session is organized by Dr. K. Karpouzis, Humaine WP4 ICCS postdoc, Dr. D. Tzovaras, researcher at the ITI/CERTH Institute and member of the 'Similar' FP6 Network of Excellence and Dr. J. Soldatos, researcher at the Athens Institute of Technology and member of the CHIL (Computers in the Human Interaction Loop) FP6 IP. Besides serving as a valuable discussion forum between projects from the 'multimodal interaction' action line (papers from the AMI – Augmented Multiparty Interaction IP will also be presented), this session will ensure broader dissemination of Humaine since its proceedings will be published by Springer in the main IFIP series, as well as in a special issue of one of its journals.

In addition to this, a special session devoted to multimodal emotion recognition will be featured in ICANN 2006, the international conference on Artificial Neural Networks (http://www.icann2006.org), organized in September 2006 in Athens by ICCS and chaired by Prof. Stefanos Kollias. ICANN is a premium conference on machine learning and semantics, research fields which are central in WP 4. It will be attempted to invite and extend discussions with researchers in HUMAINE and in other related projects in the field.

Regarding unimodal work, a special session on 'Prosody and Affective Computing' is planned for the Speech Prosody International Conference and is co-organized by Noam Amir.

4 Research achievements to date

The research achievements follow the developments produced by each one of the four subgroups working in the framework of WP4. The produced achievements have been reported in the two former deliverables D4b and D4c. In summary these refer to the following:

4.1 Achievement 1: Emotional Speech Analysis and Recognition

4.1.1 Participants

FAU Erlangen, LIMSI-CNRS, Queen’s University of Belfast, University Augsburg, ITC-irst

4.1.2 Publications

These include nos. [1] - [11] in the references section of this report.

4.1.3 Other output (demonstrations, resources, etc)

New developments are related to the formation of the CEICES group, which is leaded by members of HUMAINE and placed the focus of work on comparing, investigating and extending the state-of-the-art in work on emotional speech recognition, as described in section 3.1.3.

4.1.4 Follow-up in progress

The main focus is on performing the theoretical and experimental comparative work based on a common set of emotionally driven speech data.

4.2 Achievement 2: Visual Analysis for Expression Recognition and Affective Interaction

4.2.1 Participants

ICCS-NTUA, LIMSI-CNRS, GERG, Paris-8

4.2.2 Publications

These include nos. [12] - [20] in the references section of this report.

4.2.3 Other output (demonstrations, resources, etc)

Other output (demonstrations, resources, etc) A major event, in the form of a special session devoted to emotion recognition, including visual analysis, with many papers and partners involved was organised at ICME 2005 in Netherlands, July 2005. Similar special sessions are foreseen for 2006 as reported in section 3.2.3.

4.2.4 Follow-up in progress

The main focus is on analysing temporal facial and body gestures, for providing cues for emotion recognition and for expressive avatar synthesis.

4.3 Achievement 3: Emotional Physiological Analysis and Recognition

4.3.1 Participants

UA, FAU, ICCS-NTUA, QUB

4.3.2 Publications

These include nos. [21] and [22] in the references section of this report.

4.3.3 Other output (demonstrations, resources, etc)

The workshop organised in Augsburg has formed the basis for starting a collaborative work in physiological emotion analysis, as reported in section 3.3.3.

4.3.4 Follow-up in progress

The focus of work is twofold. On the one hand to analyse physiological signal processing for emotion recognition purposes, and on the other hand, to combine these results with the other two modalities, i.e., speech and vision.

4.4 Achievement 4: Multimodal Emotion Recognition

4.4.1 Participants

GERG, ICCS-NTUA, UA, QUB, TAU, FER Zagreb, Paris-8

4.4.2 Publications

These include nos. [23] - [32] in the references section of this report.

4.4.3 Other output (demonstrations, resources, etc)

The above reported special sessions and workshops have as main component the (single or multi - modal) emotion recognition task, as reported in 3.4.3)

4.4.4 Follow-up in progress

The main aspect of the foreseen work is on investigation of the psychological models as well as of context in emotion analysis of input signals, on synchronisation and multimodal emotional analysis of these. In the latter framework, a combined activity initiated by TAU and ICCS refers to multimodal (audiovisual) emotional analysis based on a common dataset, annotated by QUB (SAL data)

5 Conclusions

5.1 Obstacles encountered or foreseen

Preliminary work towards the WP4 exemplars has already identified a number of interesting issues that need to be investigated. An important remark has to do with fusing features from different modalities that operate on different time scales. For example, visual processing is usually performed at frame level, while speech and physiological features are calculated for time intervals (windows). In this case, the usual approach is to use a characteristic frame for each window to pick the visual features from or some method of averaging visual features over the same period; in either case, valuable information can be lost.

Another important issue, falling in the gray area between WP4 and WP5, has to do with the production of multimodal data containing physiological information, which is relatively hard to evoke and may also pose interesting ethical questions. Similarly, naturalistic videos containing gestures are quite scarce and may be also subject to culture-related interpretations or clichés. In this framework, one needs to look into the relation between acquiring reliable data, which would cater for robust feature extraction and, hence, recognition, and the degree of usability and intuitiveness of the capturing procedure. For example, facial expression recognition works best when users look steadily and directly into the camera, while speech analysis needs controlled, noise-free audio capture from microphones located directly in front of the user; however, both these requirements seriously hamper the induction of and expression of emotions from the user.

5.2 Relation to the state of the art and evidence of esteem

Work within WP4 is largely defining the state of the art in the area of signal analysis and processing for affective applications, especially if one takes into account the connections and cooperation of WP4 participants with research groups from other ‘multimodal interfaces’ projects. Invited speeches and special sessions at leading international conferences, such as ICME, Web3D, ICANN, AIAI, HCII, LREC and Eurospeech, as well as seminal papers in related journals (Neural Networks, Applied Signal Processing, Speech Communication, etc.) prove that work done within WP4 is particularly valuable for researchers in the field; this fact is also fortified by the numerous citations of this ongoing work.

Especially when it comes to multimodal recognition, research groups from Humaine are among the first to try related approaches systematically and, more importantly, in naturalistic data, free from imposing specific requirements to the people whose expressivity is recorded. In all but a few existing emotion recognition approaches, data capture is controlled with respect to expressivity, and thus, escapes the affective computing scenario that Humaine works upon. Within WPs 4 and 5, special precautions are taken to ensure that both data and feature extraction and fusion processes operate in naturalistic settings, while maintaining high recognition rates.

6 References

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

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

[3] Batliner, 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.

[4] E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach. (2003). Emotional speech; Towards a new generation of databases. Speech Communication (40).

[5] V. Radman, K. Smid, I. Pandzic, Automatic Content Production for an Autonomous Speaker Agent, AISB2005 Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction, University of Hertfordshire, College Lane Campus, Hatfield, England, 2005.

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

[7] L. Vidrascu 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.

[8] T. Vogt and E. André, Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In IEEE International Conference on Multimedia & Expo (ICME 2005), 2005.

[9] G. Zoric, I. Pandzic, A Real-time Lip Sync System Using a Genetic Algorithm for Automatic Neural Network Configuration, in Proceedings of the IEEE International Conference on Multimedia & Expo ICME, Amsterdam, The Netherlands, 2005.

[10] G. Zoric, I. Pandzic, Automatic Lip Sync and its Use in the New Multimedia Services for Mobile Devices, in Proceedings of the 8th International Conference on Telecommunications, ConTEL, Zagreb, Croatia, 2005.

[11] G. Zoric, K. Smid, I. Pandzic, Automated Gesturing for Virtual Characters: Speech-driven and Text-driven Approaches, ISPA 2005 4th International Symposium on Image and Signal Processing and Analysis, Zagreb, Croatia, 2005.

[12] S. Abrilian, L. Devillers, S. Buisine, J.-C. Martin (2005a). EmoTV: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. 11th International Conference on Human-Computer Interaction (HCII'2005), Las Vegas, Nevada, USA, 22 - 27 July.

[13] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, S. Kollias, Emotion Analysis in Man-Machine Interaction Systems, in Samy Bengio, Hervé Bourlard

(Eds.), Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, Vol. 3361, pp. 318 - 328, Springer-Verlag.

[14] R. Cowie, E. Douglas-Cowie, J. Taylor, S. Ioannou, M. Wallace, S. Kollias An Intelligent System for Facial Emotion Recognition, Proceedings of ICME 2005, Amsterdam, The Netherlands, July 6-8, 2005.

[15] B. Hartmann, M. Mancini, C. Pelachaud, C. (2005). Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. Gesture Workshop, Vannes, May 2005.

[16] M. Mancini, B. Hartmann, C. Pelachaud, A. Raouzaiou, K. Karpouzis, Expressive Avatars in MPEG-4, IEEE International Conference on Multimedia & Expo (ICME) 2005, Amsterdam, The Netherlands, July 6-8, 2005.

[17] A. Raouzaiou, E. Spyrou, K. Karpouzis and S. Kollias, Emotion Synthesis: an Intermediate Expressions' Generator System in the MPEG-4 Framework, International Workshop VLBV05, 15-16 September 2005, Sardinia, Italy.

[18] M. Wallace, S. Ioannou, A. Raouzaiou, K. Karpouzis, S. Kollias, Dealing with Feature Uncertainty in Facial Expression Recognition Using Possibilistic Fuzzy Rule Evaluation, International Journal of Intelligent Systems Technologies and Applications, accepted for publication.

[19] G. Zoric, K. Smid, I. Pandzic, Automatic Facial Gesturing for Conversational Agents and Avatars, in Proceedings of the 2005 International Conference on Active Media Technology AMT 2005, Takamatsu, Japan, 2005.

[20] G. Zoric, K. Smid, I. Pandzic, Automated Gesturing for Embodied Agents, JSAI 2005 Workshop on Conversational Informatics, in conjunction with the 19th Annual Conference of The Japanese Society for Artificial Intelligence JSAI 2005, Kitakyushu city, Japan, 2005.

[21] J. Kim, E. Andre, M. Rehm, T. Vogt and J. Wagner, Integrating Information from Speech and Physiological Signals to Achieve Emotional Sensitivity. In Proc. of the 9th European Conference on Speech Communication and Technology, 2005.

[22] J. Wagner, J. Kim and Elisabeth André, From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In IEEE International Conference on Multimedia & Expo (ICME 2005), 2005.

[23] S. Abrilian, J.-C. Martin, and L. Devillers, (2005b). A Corpus-Based Approach for the Modeling of Multimodal Emotional Behaviours for the Specification of Embodied Agents. 11th International Conference on Human-Computer Interaction (HCII'2005), Las Vegas, Nevada, USA, 22 - 27 July.

[24] A. Batliner, 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.

[25] L. Devillers, S. Abrilian, and J.-C. Martin (2005). Representing real life emotions in audiovisual data with non basic emotional patterns and context features. First International Conference on Affective Computing & Intelligent Interaction (ACII'2005), Beijing, China, October 22-24 http://www.affectivecomputing.org/2005

[26] L. Devillers, 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.

[27] S. Ioannou, M. Wallace, K. Karpouzis and S. Kollias, A Robust Scheme for Facial Analysis and Expression Recognition, Measuring Behaviour 2005, 5th International Conference on Methods and Techniques in Behavioural Research, Wagenigen, The Netherlands, 30 August - 2 September 2005.

[28] S. Ioannou, M. Wallace, K. Karpouzis, A. Raouzaiou and S. Kollias, Combination of Multiple Extraction Algorithms in the Detection of Facial Features, Proceedings of the IEEE International Conference on Image Processing (ICIP), Genova, Italy, September 2005.

[29] S. Ioannou, M. Wallace, K. Karpouzis, A. Raouzaiou and S. Kollias, Confidence-Based Fusion of Multiple Feature Cues for Facial Expression Recognition FUZZ-IEEE 2005, May 22-25, Reno, Nevada, USA.

[30] J.-C. Martin, S. Abrilian, and L. D Devillers. (2005). Annotating Multimodal Behaviours Occurring during Non Basic Emotions. 1st International Conference on Affective Computing & Intelligent Interaction (ACII'2005), Beijing, China, October 22-24 http://www.affectivecomputing.org/2005

[31] J.-C. Martin, S. Abrilian, L. Devillers, M. Lamolle, M. Mancini, M. and C. Pelachaud, (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 http://iva05.unipi.gr/

[32] S. Steidl, 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.