

# humaine

**D4c**

**Description of potential exemplars:  
Signals and Signs of Emotion**

*Workpackage 4 Deliverable*



**Date: 30<sup>th</sup> November 2004**

<b>IST project contract no.</b>	507422
<b>Project title</b>	<b>HUMAINE</b> <b>Human-Machine Interaction Network on Emotions</b>
<b>Contractual date of delivery</b>	<i>November 30, 2004</i>
<b>Actual date of delivery</b>	<i>November 30, 2004</i>
<b>Deliverable number</b>	D4c
<b>Deliverable title</b>	Description of potential exemplars: Signals and Signs of Emotion
<b>Type</b>	Report
<b>Number of pages</b>	30
<b>WP contributing to the deliverable</b>	WP 4
<b>Task responsible</b>	ICCS-NTUA
<b>Author(s)</b>	Stefanos Kollias, Noam Amir, Jonghwa Kim, Didier Grandjean
<b>EC Project Officer</b>	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

THE PLACE OF THIS REPORT WITHIN HUMAINE .....	5
1 BRIEF OVERVIEW OF THE WORKPACKAGE, THE EXEMPLAR PROPOSAL, AND RELEVANT RESOURCES .....	6
2 RATIONALE FOR THE EXEMPLAR PROPOSAL .....	10
3 TECHNICAL ASPECTS OF THE PROGRAM OF RESEARCH .....	13
4 PROPOSED ALLOCATION OF TASKS.....	24
5 PROVISIONAL TIMELINE .....	28
6 CONCLUSIONS .....	30



## 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 review and assessment documents were delivered in May, and the consultation phase has been ongoing since, using several channels, notably e-mail exchanges moderated by the coordinator; meetings of workpackage leaders by teleconferencing; meetings between workpackage representatives attending the WP4 workshop and the Summer School; and a consultation meeting of WP leaders in Paris on October 29<sup>th</sup> & 30<sup>th</sup>.

This deliverable is one of the group arising from the consultation phase, whose function is defined as to 'describe in some detail a few alternatives that might realistically be chosen as exemplars in each area'. In general, we believe that we have progressed more quickly than we have expected, and that the alternatives described here are close to the ones that should be pursued. What remains to be completed is largely detailed planning. Given the intricacy of the network, that is not a trivial task.

The following persons have contributed to the deliverable:

*Stefanos Kollias, Noam Amir, Jonghwa Kim, Didier Grandjean, Vered Aharonson, Roddy Cowie, Ellen Douglas-Cowie, Anton Batliner, Laurence Devillers, Tanja Banziger, Aihlbe ni Chasaide, Kostas Karpouzis, Amaryllis Raouzaïou, Spyros Ioannou, Maurizio Mancini, Etienne Roesch, Fiorella de Rosis.*

The institutions that have contributed are:

*ICCS-NTUA, TAU, UA, UNIGE, QUB, TCD, LIMSI, FAU-Erlangen, Paris-8.*

# 1. Brief overview of Workpackage 4, the exemplar proposal, and relevant resources

## 1.1 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.

### 1.1.1 Conception of the area before HUMAINE began

Signal analysis and sign extraction have been a major limiting factor in the development of emotion-oriented systems, whether the modality is visual, auditory, somatic, or other. They affect both reception and synthesis. A system that attempts to interact with humans, taking into account their emotional state or attitude, needs to process, analyse and extract all these cues that are provided through speech, facial expressions, hand and body pose, or physiological signs. Conversely, all of the cues can be used to convey emotional messages to a user. The expected focus of the research activity in WP4 is developing the basis for a coherent treatment of these issues. Originally, it was planned to treat synthesis of emotional signals in WP4 as well.

### 1.1.2 The current conception

The synthesis of signals, e.g., speech, visual/avatars, is closely related to the analysis of them, as far as low level information is concerned; cues about synthesis can be developed based on the results of the emotion analysis/recognition tasks. However, at the higher level, synthesis is closely related to man machine interaction and is thus considered in the framework of Humaine WP6.

The signal analysis and emotional sign extraction work is divided in four tasks corresponding to major approaches regarding emotion analysis sub-problems. Each sub-problem is addressed by a working group:

- WG1: Emotional Speech Analysis and Recognition
- WG2: Face& Gesture Analysis for Emotion Recognition & Expressive ECA Synthesis
- WG3: From Physiological Signals to Emotional Signs
- WG4: Multimodal Emotion Recognition.

## 1.2 The exemplar proposal

Following the consultation period, the exemplar proposed for WP4 is *A Technological Pool of Methods for Multimodal Emotion Recognition*. It has 4 main elements, which correspond to the above-mentioned workgroups and to the respective problems regarding multimodal input analysis and recognition.

### 1.2.1 The significance of the title

This title has been chosen to stress the fact that members of the workgroup have acquired large and different degrees of expertise in the various aspects of emotion characterization, analysis and recognition. The aims of the exemplar are to pool the resources of the different members and share their accumulated experience for generating new knowledge and algorithms in the area; the latter will be shared with the rest of HUMAINE WPs for achieving the project goals.

### 1.2.2 The elements of the exemplar

The exemplar proposed by WP4 aims at generating a (technological) pool of methods for analysing various input signals - speech, faces, gestures and physiological signals- in the framework of multimodal emotion recognition. Its outcome will, therefore, include:

- extraction of emotional features and signs from the multimodal inputs,
- recognition of the user's emotional state.

The above can be used both for recognition of user's emotional state, as well as for generating more expressive ECAs that accurately respond to the extracted user's emotional signs. Research will focus on advancing the current state-of-the-art in emotion recognition:

- provided by each modality in separate,
- combining more modalities,
- taking into account knowledge & demands provided by WP3, WP5, WP6, WP7.

In this framework, the exemplar is decomposed in four elements which are listed below.

#### 1.2.2.1 Element 1: Emotional speech analysis

The tasks related to the emotional speech analysis can be summarised as follows:

*a. Basic tasks* (extract F0, intensity, voice quality, word recognition)

- Review existing techniques
- Conduct comparative studies of existing techniques
- Improve on best previously available

*b. Segmentation* (word/phrase/turn boundaries)

- As above, but also develop coherent strategy for using information at multiple levels.

*c. Feature extraction*

- Consolidate techniques from different partners
- Identify relevant types of features not yet modelled from phonetics
- from music (rhythm is conspicuous by its absence)

*d. Classification*

- Compare and select relevant classifiers

*e. Reference data*

- Establish emotion-relevant benchmarking material
- Acquire relevant labelling ('standard' features, perceptually relevant features).

### 1.2.2.2 Element 2: Visual emotional analysis

#### a. Facial feature extraction

- Extract facial features
- Compute MPEG-4 FAPs
- Compute MPEG-4 visemes
- Extract gaze
- Detect patterns of movements
- Define temporal correlations

#### b. Gesture analysis

- Detect arm, wrist
- Detect hand movements
- Extract visual expressivity parameters

#### c. Classification

- Use a-priori knowledge and rules
- Adapt to specific circumstances

#### d. Reference data

- Establish emotion-relevant benchmarking material
- Acquire relevant labelling.

### 1.2.2.3 Element 3: Physiological Signal Emotional Analysis

#### a. Effective data processing

- Extend sensor parameters depending on physiological priority
- Fuse the extended sensors

#### b. Segmentation and Feature extraction

- Select time intervals
- Extract most appropriate features

#### c. Classification

- Statistical techniques
- Probabilistic techniques
- Hybrid techniques

#### d. Reference data

- Establish emotion-relevant benchmarking material
- Acquire relevant labelling.

### 1.2.2.4 Element 4: Multimodal emotion recognition

#### a. Emotion models and Recognition - Application dependency

- Discrete models
- Dimensional models
- Appraisal theory

#### b. Theoretical models of multimodal integration

- Direct integration

- Separated recognition
- Dominant integration
- Motor space integration

*c. Single mode (Speech, Visual, Physiological) Emotion Recognition*

- Advance the state-of-the-art wrt the three single-mode emotion recognition systems

*d. Modality synchronization*

- Use visemes for audiovisual analysis
- Use EMGs and FAPs for visual and physiological analysis
- Use SC/RSP measurements & speech intonation for speech and physiological analysis
- Focus on temporal evolution and modality sequentiality

*e. Multimodal recognition techniques*

- Examine and select classifiers/reasoners suitable for multimodal recognition
- Take into account context and goals of interaction
- Take into account cognitive cues, attention and modality significance in interaction

*f. Generation of the pool of the developed techniques and of the generated knowledge*

- Define the form of presentation of the developed techniques
- Present performance issues
- Publish research papers, an edited book.

### 1.2.3 The ways in which the exemplar depends on other Workpackages

There are a variety of issues which define the dependence of the WP4 exemplar on other workpackages, and particularly WP3, WP5, WP6, WP7. Input From WP3 is. WP5, on the other hand, must supply the databases to which such a system will be applied.

- WP3: Feature-based representation of the appraisal model, as well as other emotion models, based on particular implementations or applications, should be explored. This is necessary for determining emotional labelling/classification systems, which must serve as the basis for training an emotion recognition system.
- WP5: Provision of common multimodal and single mode databases, following the requirement for annotated data.
- WP6: Provision of tools that use low - or intermediate-level features (e.g. hand positions or BAPs for ECAs) as input to test validity of the analyzed data; description of salient events used to attract ECAs' attention, so that these can be extracted as well as taken into account for context analysis in WP4.
- WP7: Provision of higher level cues on cognition and attention that can be used in multimodal emotion recognition.

### 1.2.4 Proposed final output

The final output will be in several forms. One will be algorithms for solving as many of the subtasks as possible, and their implementation in software. Novel contributions will be documented in the scientific literature and probably an edited book. A separate report should document the problems that remain unsolved at the end of HUMAINE, describing the difficulties encountered in their solution.

## 2. Rationale for the exemplar proposal

The exemplar proposal represents a choice to follow a particular line of development rather than others that are possible (or might seem possible to the outsider). The key reasons for making this particular choice are as follows.

### 2.1 Distinctive features of the approach proposed

The proposed exemplar profits from two advantages found within the HUMAINE framework: the pooled experience of the project members, and the close contact with other workpackages.

The participants in WP4 have accumulated leading expertise on almost every different aspect of the problems involved, including speech, visual, physiological signal processing, signal segmentation, emotional feature extraction, single-mode classification, statistical analysis, multimodal synchronisation, integrated recognition models, fuzzy reasoning, context analysis. Pooling this experience within the framework of an integrated approach can certainly create significant theoretical and practical advances and improvements, on the one hand, to the different components, and on the other hand, to the integration of them for multimodal emotion analysis.

In this framework, the proposed approach is to tackle each input modality both in separate, and in common, using a common theoretical emotional psychological background. This is, on the one hand, provided by WP3 outcome's, but on the other hand ensured by the involvement of psychological groups in all related WP4 tasks. This is a major advantage of the proposed approach which clearly distinguishes it from those of other groups that tackle the specific problems from a single point of view.

The study of emotion is in fact a multidisciplinary endeavor, and the result is that small research groups struggling with this problem are generally found to be lacking in expertise in at least some of the aspects involved. The ability to work alongside WG3, which will supply useful labelling schemes, WG5, which will supply test databases labelled according to such schemes, in interaction with WP6 and WP7, which will provide and receive knowledge about the context and dynamics of HCI will aid WG4 in developing the best possible techniques and framework for the problem of emotion recognition.

### 2.2 Rationale for emphasising this approach

The approach adopted here is based on discussion and deliberation amongst research groups who already have great experience in the field, and have encountered various difficulties in attacking emotion recognition on a small scale. When a single research group must cope with defining a labelling scheme, accumulating emotional data, and carrying out automatic classification, limitations stemming from limited time and expertise are inevitably encountered.

The basic distinctive feature of the proposed approach is the close interaction with the rest of the WPs, not only for providing them with WP4 extracted emotional signs, but also for getting from them (both by obtaining the produced results and by involving members of the main research teams leading other WPs in WP4) crucial information about the models used by them for :

- representing emotion,
- generating expressive ECAs,
- attention generation in multimodal recognition.

These increase the work that WP4 members have to develop, since they have to take into account and provide solutions for :

- many emotional models, e.g., discrete, continuous, temporal,
- a variety of expressivity parameters, e.g., related to face, gestures, gaze,
- including higher level mechanisms in the synchronisation and integration of modalities.

Ignoring these factors would deprive the proposed approach from a great part of its novelty, of its wideness and of the expected prospects.

### 2.3 Rationale for subdividing the task

As already mentioned, exemplar implementation is divided in sub-problems corresponding to the major approaches / problems regarding multimodal input analysis and emotion recognition, with each sub-problem being addressed by a working group:

- WG1: Emotional Speech Analysis
- WG2: Facial and Gestural Analysis for Emotion Recognition and Expressive ECA Generation
- WG3: From Physiological Signals to Emotional Signs
- WG4: Multimodal Emotion Recognition.

Integration will take place at three levels: within working groups, across working groups within the workpackage, and (as described above) across workpackages.

It should be mentioned that the above subdivision does not alter the main focus of the exemplar, which is deriving emotional signs based on multimodal signal analysis. In fact, almost all WP4 participants contribute to WG4, which is the multimodal emotion recognition part of the exemplar, playing the role of the ‘umbrella’ which covers the three separate modalities.

It is evident that the three WGs dealing with the three separate modalities, should first work in separate, trying to investigate and solve a variety of issues/problems that appear in the tasks of each modality (such as processing, feature extraction, significance of parameters), and then in common, under the umbrella of WG4, for integration.

The main features to be used for inter- modality analysis have already been identified, e.g., visemes and pause detection for audiovisual analysis, [SC/RSP-measurements with speech intonation](#) for speech and physiological signal fusion, EMG and FAPs for visual and physiological signal fusion.

It should be noted here that of great importance is that the developments in each WG will be made in the same theoretical/psychological emotional framework, provided by WP3 and explored in the framework of WG4.

## 2.4 Measures taken to ensure coherence across subtasks

As already mentioned in the above section, the nature of research within the Humaine network is inherently multimodal. In WP4, it is required to analyse the available data both with respect to the individual modalities and as a combined input. All subtasks can be seen as parts of one large metatask, ensuring that that these subtasks coalesce into one useful system. Furthermore, the fact that several of the databases are multimodal, will enable the different subgroups to apply their different techniques to these databases and create a common framework for comparing and unifying their results.

For example, an audiovisual database will provide different features from the analysis of the linguistic, paralinguistic and visual information. All three modalities can be individually used to provide feedback for the users' emotional state; however, under certain circumstances, a particular modality may provide better information than the others or supply means to complement information in another channel (e.g. viseme information supporting speech analysis). As a result, it is imperative that we come up with tools to investigate meaningful content in single modalities, as well as integrated approaches, where multimodal information can provide more stable results.

## 2.5 Measures taken to ensure coherence with other exemplars

Most participants in HUMAINE are active in more than one WP. There are many interdependencies in the exemplars of different WPs, and in some case the exemplar of one WP is a necessary input for the exemplar of another workgroup. These two factors contribute greatly to a strong coherence across exemplars.

This coherence will be ensured by using common data and representation schemes; these will be produced by the close collaboration of WP4 and WP5 and will be immediately implemented in the WP6 exemplar. WP4 and WP5 will collaborate so as to generate data that take into account signal processing demands, while WP4 and WP6 will collaborate to extract and investigate features that are meaningful or necessary for expressive ECA generation. It has to be noted though that the raw input signals contain much more information than what is meaningful, emotion-wise. Therefore, we will work closely with the theory WP (WP3) to filter out irrelevant features and avoid cluttering the perception processes, and (more importantly) to perform research in the same theoretical (HUMAINE) framework.

## 3. Technical aspects of the program of research

### 3.1 Element 1: Emotional speech analysis

The technical program for emotional speech analysis will be approached on several different levels. One level is the "meta-task", defining the ideal components of an emotion recognition system, what they should be able to do and how they can contribute to such a system. The second level is the breakdown of this meta-task into subtasks, defining which these we intend to implement within the HUMAINE project. In parallel, an expert listening protocol will be defined and carried out on the benchmark data used with the project, in order to provide a further channel for translating between human assessment of emotional content and the relevant acoustic features carrying the emotional information.

#### 3.1.1 Meta task – defining a speech emotion recognition system

The first task can be a "meta-task": a detailed description of an emotional recognition system, under the assumption that we have a magic wand that can solve any technological problem encountered. We can then address these specific problems using two approaches (bottom-up vs. top-down):

1. Bottom-up: Solve them as best as we can, and attempt to build in some robustness to imperfect solutions
2. Top-down: Leave them unsolved, for future generations, and in the meantime replace them manually if possible

Such a system should be able to utilize all the information that a human listener has access to, consciously or not. This is composed of accurate recognition of:

- text
- F0
- intensity
- voice quality
- pauses – duration: pauses, speaking rate
- Background information on the speaker – age, sex, social background, cultural background, personality.
- information on the interaction, such as: social interaction among friends, among strangers, a call center conversation, boss/employee conversation, interview, etc.

Assuming that all the above information can be accurately obtained, the question remains - how do we utilize it? Many problems must be resolved, among them:

- Feature extraction from F0, intensity and voice quality
- Interpretation of textual information on several levels: disfluencies, filled pauses, linguistic analysis
- Synchronization between text and signal (automatic vs manual): finding where are the pauses in order to compute duration measures, syllables lengthening.
- Synchronization between text and prosody: finding where the prosody serves linguistic purposes such as stress, accent, end of phrase, and how these are modified or added to by the presence of emotion.
- Training and classification.

### 3.1.2 Subtask descriptions

The subtasks to be solved are: obtaining the raw information from the speech signal, and then using it correctly to extract emotional states. The following is a breakdown into subtasks, discussing the problems related to each.

#### **Pitch detection:**

Accurate and robust pitch detection is a well-known problem that has been approached in many ways. In the current context it is actually composed of two sub problems: 1) VAD (Voice Activity Detection) – i.e. deciding whether a segment of the signal is voiced or not; 2) Actual pitch detection – i.e. calculating the actual pitch, once a certain segment is determined to have been voiced.

In the context of emotion recognition, the requirements from the PDA are as follows:

1. Robustness to background noise
2. Avoidance of false positives at erroneous values is far more important than false negatives. To a certain extent, false negatives can be fixed through later interpolation
3. In contrast to the analysis of singing performance, for example, absolute accuracy is not extremely important. Avoidance of gross errors such as pitch halving or doubling is much more important.
4. Detection of problematic passages such as hoarseness and creakiness and identifying them as such is more important than trying to find actual pitch values therein, which are often fluctuating and do not reflect the perceived pitch, if any such perception exists.

It would be highly desirable to develop fully automated robust algorithms for the purpose of emotion recognition, perhaps at the expense of lower accuracy and a higher degree of false negatives. It is certainly important to come up with a set of benchmark speech recordings to be used in order to examine different PDA's.

It will be very useful to accumulate benchmarks of typical F0 detection difficulties (*voice peculiarities: hoarse, creaky, pathological voices and specific characteristics: octave jumps...*). It could be very useful for the research community to diagnose F0 algorithm with such benchmarks. Building such benchmarks from different corpora can be examined as well.

#### **Intensity:**

Intensity of the signal itself is misleadingly easy to calculate. The question is to what extent this reflects the intensity of the speech itself. Intervening between the speaker and the recorded signal are the physical distance of the microphone, which can change with head movement, and the gain settings in the various parts of the recording equipment. Some kind of normalization method should be found to extract this information as closely as possible, despite such limitations. In the same way, we can build benchmarks of *different recording quality*.

#### **Voice quality parameterization:**

The analysis of voice quality is complicated by the fact that segmental and intonational features often mask it. In the frequency domain, for example, differences in the spectrum of the glottal pulse are not easily detected when the vocal tract further modifies this pulse. In the time domain, involuntary changes in the pitch contour, such as jitter, are

superimposed on other, voluntary, changes in pitch. The approaches to extracting useful voice quality features, despite these difficulties, are outlined below.

#### Time domain voice quality features

Though changes in pitch are usually ascribed to prosody, certain micro prosodic effects are commonly regarded as voice quality effects. To these one can also add local changes in intensity. Local fluctuations in pitch, over several pitch periods, are termed jitter, whereas local fluctuations in intensity are termed shimmer. The main difficulty in measuring jitter and shimmer is that they can be biased greatly in the presence of voluntary changes of pitch and intensity. In clinical settings, subjects are usually requested to utter a single sustained vowel when these parameters are to be measured. Clearly this is impossible when dealing with emotional speech.

#### Voice quality features based on short term spectra

The short-term spectrum is determined by segmental information to a very large extent, rather than by the glottal pulse. Though one can conceivably compare similar phones, carrying this out automatically is prohibitively complicated. One common approach is simply to average short time spectra over a relatively long segment of speech, several seconds in length at least. This is known as the LTAS - Long Term Average Spectrum. The LTAS is necessarily a gross measure, smearing local changes over a large interval, with no ability to track small short-term changes in spectra. Nevertheless it can prove useful, giving statistically significant correlation with emotion in the same way as pitch mean, taken over similar intervals.

#### Voice quality features based on Parametric models

It seems very desirable to obtain accurate measures for voice quality which can be used over short as well as long term signal segments. Preferably these should be correlated with perceptual judgment of voice quality.

#### **Text extraction:**

This is an example of a task that is relatively easy to perform manually, yet often very difficult to perform automatically. Since it is a large domain of research in itself, it is unlikely that any work on the subject will be carried out within the framework of HUMAINE. It is probably safe to assume that manual transcription will be available for the databases accumulated within the HUMAINE project. The true difficulty is how to enhance the emotion recognition methods, using textual information.

#### **Feature extraction**

Feature extraction from the raw pitch and intensity contours is a major issue. It involves an intermediate step of smoothing and interpolating the raw pitch contour, and then extracting various features. Some points can be raised here:

- The time intervals over which the features are to be calculated should also be considered. Current approaches vary from fixed length intervals, to "tunes" (intervals between significant pauses) to words. This is a question that has been discussed at length and several different approaches will be examined.
- A multitude of features has been experimented with by several groups. Pooling knowledge and experience in features extraction, in conjunction with input from

expert listening assessments, will greatly help in finding the set of features most appropriate to different types of emotions, different recording environments, and different induction schemes.

- Due to the imperfection of pitch detection algorithms, it is important to compare manually corrected F0 features and sub-optimal automatically extracted F0.

### Training and Classification

Classification algorithms are also a very large domain of research. From the literature it seems that researchers in emotion recognition have been using several standard methods, and that the differences between different classifiers are not very large. This will probably not be a major focus in HUMAINE, though a brief comparison of several different methods (NN, HMM, decision trees, SVM...) will be done.

One issue that is definitely important is how training is carried out. This is directly related to the databases and associated emotional labelling being used. This is a thorny issue involving close cooperation with WP3 and WP5.

Another important point is the study of the combination of textual and prosodic models. Work is to be done in this direction.

### Human listening

A model can be used for assessment by expert human listeners. It would be very useful to have expert listeners assess emotional content, and then describe the various cues they heard that pointed to emotional content. This kind of input, though qualitative, has the potential to be very important in pointing the technologists in fruitful directions, instead of grasping in the dark for as many features as they can calculate.

## 3.2 Element 2: Face and Gesture Analysis for Emotion Recognition and Expressive ECA Generation

Many steps have been recently implemented towards facial analysis, or analysis of user's movements and gestures. However, there is the need for major R&D work for extending, successfully implementing and integrating a variety of visual analysis techniques in the framework of WP4 multimodal emotional sign generation framework. These include:

- *Facial feature extraction* (face detection, facial feature extraction, MPEG-4 FAP computation, gaze analysis, detection of patterns of movements and of temporal correlations)
- *Gesture analysis* (detection and computation of arm, wrist, hand movements and of visual expressivity parameters)
- *Classification* (usage of a-priori knowledge and rules, adaptation, fuzzy representations, knowledge technologies)
- *Reference data* (generating databases with facial expressions, gestures and other modalities, appropriate labelling).

The framework of visual analysis can be defined through analysis of two main targets of WP4 and HUMAINE, i.e., contribution to multimodal emotion recognition and to generation of expressive ECAs

### 3.2.1 Targeting Emotion Recognition

A variety of techniques can be used to recognize the underlying emotional states, based on analysis of the FAP features extracted from the user's face. These include neural network classifiers, clustering techniques and neurofuzzy networks.

Of significant interest is usage of unsupervised hierarchical clustering, since this can form a basis for future merging of different emotional representations (i.e. different hierarchical levels), and categorization in either coarser or more detailed classes (half-plane, quadrants, components, discrete emotions).

Usage of such clustering approaches can provide rules, in terms of cluster centers and cluster standard deviations. An can then be to transform these rules to ones expressed in terms of variations of the FAP variables. Neurofuzzy systems can then accept fuzzified FAP predicates at their input and adapt the a-priori knowledge, i.e., the above-mentioned rules, to each specific user's characteristics.

Gestures can be utilized to support the outcome of the facial expression analysis subsystem, since in most cases they are too ambiguous to indicate solely a particular emotion; the latter can, however, be the case in specific contexts. A gesture classification scheme, based on HMMs will be created, as described in D4b, for assisting the facial analysis subsystem for emotion recognition. Consequently, the facial expression analysis subsystem and the affective gesture analysis subsystem will be integrated into a fuzzy system, which provides as result, the possible emotions of the user, each accompanied by a degree of belief.

### 3.2.2 Targeting Expressivity of ECAs

Expressivity of ECAs can be captured with six attributes described below in qualitative terms:

- Overall activation: amount of activity (quantity of movement) across several modalities during a conversational turn (e.g., simultaneous use of facial expression and gesture to visualize communicative acts -- passive/static or animated/engaged).
- Spatial extent: amplitude of movements (e.g., amount of space taken up by body; amplitude of eyebrow raise)
- Temporal: duration of movements (e.g., quick versus sustained actions)
- Fluidity: smoothness and continuity of overall movement (e.g., smooth, graceful versus sudden, jerky)
- Power: dynamic properties of movement (e.g., weak/relaxed versus strong/tense)
- Repetitiveness: tendency to rhythmic repeats of specific movements along specific modalities.

#### a. Facial Expressivity

A facial expression is characterized by its temporal parameters and its shape, that is the quantity of displacement for all the involved FAPs. Knowing the starting time and duration of an expression, the next step is to calculate the course of the expression intensity, i.e., the amplitude of time-varying facial movements that compose it.

Each expression is characterized by four temporal parameters:

- attack: is the time that, starting from the neutral face, the expression takes to reach its maximal intensity

- decay: is the time during which the intensity of the expression lightly decreases, usually to reach a “stable” value, see the next parameter
- sustain: is the time during which the expression is maintained, usually it represents the more visible part of the expression
- release: is the time that, starting from the maximal intensity, the expression takes to return to the neutral expression

Such parameters are different from expression to expression. For example the “sadness” expression is characterized by a long release (the expression takes more time to disappear), while the “surprise” expression has a short attack.

The attributes for facial expressivity take the following form:

- Spatial extent: the quantity of physical displacement of FAPs involved in the expression showing process.
- Temporal : temporal characteristics of the facial expression.
- Fluidity :overall facial muscle contraction; an abrupt movement implies an increase of the muscles' speed of contraction. The reverse happens in smoother movements.
- Power : lip muscle tension (normal, higher, lower), allowing different intensities of muscular strain (e.g., in fear and anger); closely related to visemes.
- Repetitvity : how often a facial expression is repeated (e.g. counting number of head nods, or eyebrow raisings).

## b. Gesture Expressivity

The basic unit of a gesture is the keyframe, inside which there can be one ore more of the following parameters:

- ARM: defines the arm configuration in space, that means the rotation to apply to the shoulder and elbow to allow the arm to be set up in a certain position.
- WRIST: defines the orientation of the palm using two parameters, the direction of the vector orthogonal to the base of the fingers and the vector orthogonal to the palm plane.
- HAND: selects one of static, known from literature, configurations for the hand.

The attributes for gesture expressivity take the following form:

- Spatial extent: Represented by their center coordinates, the location of sectors (wrist positions) are scaled symmetrically or asymmetrically using a simple scaling matrix on homogenous coordinates. Requires the degree-of-freedom groups in the arm - wrist position, palm orientation, finger pose. The elbow swivel also affects expressivity - extended elbows enlarge the body's silhouette.
- Temporal Features : Starting from the synchronicity constraint on the end of the gesture stroke to coincide with the stressed affiliate in speech, preceding and proceeding frame times can be estimated, taking into account how quickly gesture phases are performed.
- Fluidity : Captures continuity between movements, based on continuity of the arms' trajectory paths as well as on acceleration and deceleration of limbs.
- Power/Energy : Depends on the dynamic properties of gestures and on inter-gestural rest phases.
- Repetitvity : Detecting gestures not composed only by an activation, a stroke and a release phase, but that are a sequence of a variable number of strokes, usually very close one to each other.

### 3.2.3 Analysis of Faces

Based on the above-described targets of facial expressivity, the following list of features would be desired to extract:

- FAP general quantity and quality of movement, related to emotional content.
- FAP interaction, i.e., how the activation of one FAP is temporally related to the activation of the others.
- FAP transition, i.e., how FAPs are moved during direct transition between two consecutive facial expressions.

As has been reported in Deliverable D4b, facial analysis includes a number of processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face.

The left, right, top and bottom–most coordinates of the eye and mouth masks, the left right and top coordinates of the eyebrow masks as well as the nose coordinates, are the facial feature points which are then used for defining the FAP values to be used as inputs to the emotion recognition system.

Due to noise, illumination variations and low resolution capturing devices, the detection of facial features, and consequently, facial feature points, can be inaccurate. For this reason, mechanisms are required that can automatically evaluate the quality of each computed mask, assigning a confidence level to it. The emotion recognition system can take advantage of each feature's confidence level when analyzing them. Exploitation of anthropometric knowledge, in the form of a set of criteria, evaluating the relation of the extracted features can form such a mechanism.

The detected facial features are used to extract the Feature Points considered in the definition of the FAPs, to be used as input features to the recognition system. Each FP inherits the confidence level of the feature from which it derives. FAPs are estimated via the comparison of the FPs of the examined frame to the FPs of a frame that is known to be neutral, i.e. a frame which displays no facial deformations. Of particular importance will be viseme analysis through time, since they can be the main feature to synchronize emotional facial and speech analysis for multimodal emotion recognition.

Moreover, evaluation of FAP changes through time and detected expression in a variety of examples can be used to compute the facial expressivity parameters that were described above and form respective rules that can be useful for generation of expressive ECAs in human computer interactions.

### 3.2.4 Analysis of Gestures

WP4 visual analysis can similarly define rules to find the right values for expressivity gesture parameters; these are all related to arm movement. So, starting by tracing the spatial position of wrists through time it should be possible to determine properties like speed, acceleration, direction variation that contribute to define the expressivity parameters.

There are many hand tracking systems that can be used to extract emotion-related features through hand movement, as was described in D4b. The general process involves the creation

of moving skin masks, namely skin color areas which are tracked between subsequent frames. By tracking the centroid of those skin masks estimates of user's movements can be computed.

The utilized features can then feed (as sequences of vectors) appropriate HMM classifiers, taking advantage of HMM capabilities to deal with time sequential data, to provide time scale invariability, while having learning capabilities.

In a given context of interaction, gestures are associated with a particular expression – e.g. *hand clapping* of high frequency expresses *joy*, *satisfaction* - while others can provide indications for the kind of the emotion expressed by the user. As a result, quantitative features derived from hand tracking, like speed and amplitude of motion, indicate an observed emotion; for example, *satisfaction* turns to *joy* or even to *exhilaration*, as the speed and amplitude of clapping increases. The position of the centroids of the head and the hands over time forms the feature vector sequence that feeds an HMM classifier whose outputs corresponds to particular gesture classes.

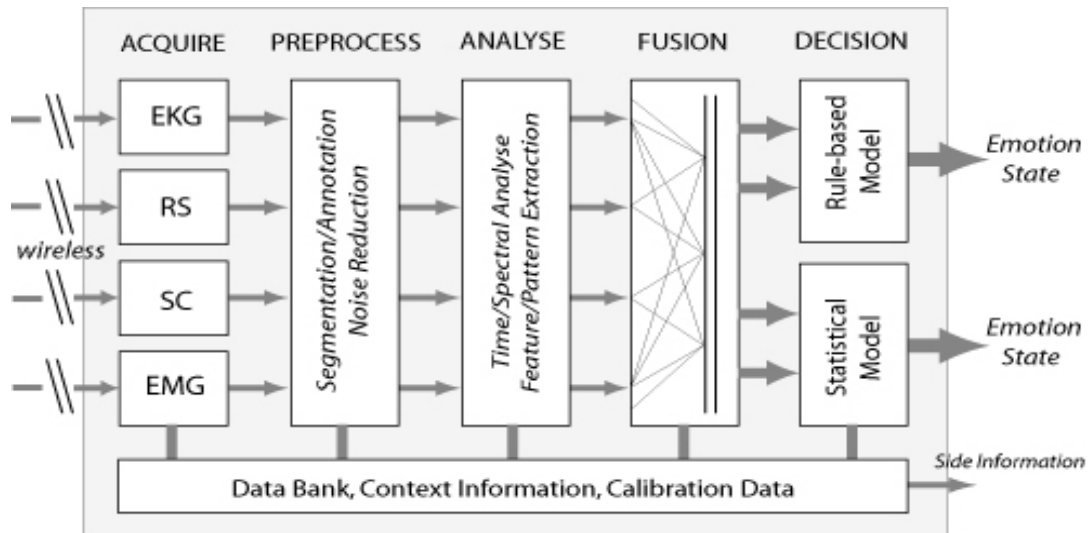
### 3.3 Element 3: Emotional Biosignal Analysis

This part focuses on the analysis of biosignals to extract appropriate features, to recognize respective emotional states and to integrate them into the multimodal recognition system. The issues that are largely resolved and remain to be resolved can be organised around four lines of activity:

#### 3.3.1 Biosignals as single factor for emotion recognition

During the first year of HUMAINE results have been obtained regarding extraction of emotional cues and features from physiological signals (biosignals). Recordings were made using four biosensors, EKG, EMG, SC and RSP, from subjects who were attempting to evoke the strongest possible emotions by imagining highly emotive situations under laboratory conditions. Recognition results with overall accuracy of 88% clearly indicate that in these extreme conditions, emotional signs can be recognized through biosignal analysis, at a level comparable to the results achieved when using extreme facial expressions or speech intonation. However, it remains unknown what level of information can be derived from biosensors when emotion is part of action and interaction, and multiple factors impinge on variables such as heart rate and respiration.

Up to the present, work has focused on finding the emotional cues in biosignals and possible priority of biosensors, by doing empirical and statistical analysis rather than reviewing the cognitive and psycho-physiological issues. Efficient correlation of our experimental experiences with the theoretical psychophysiology is the goal of this part of the exemplar, aiming at improving recognition accuracy and robustness of the recognition framework. In addition to the development of offline training-based classification methods, real-time recognition will be investigated in order to successfully integrate with recognition based on audio/visual cues.



A Systematic Framework for Emotion Recognition using Biosignals

### 3.3.3 Biosignals in multimodal databases

A database (Augsburger Database of Biosignals) has been created and used for recognition of four discrete emotional states, i.e. joy, sadness, anger, and calmness. AuDB has been implemented using musical emotion induction methods to gather elicited emotions from multiple subjects. WP5 aims to provide much more naturalistic data which incorporate recordings of biosignals synchronously with audio/visual recordings in order to generate a multimodal database. Existence of multi-channel recordings including biosignals is a prerequisite for addressing the corresponding WP4 multimodal analysis tasks described in the next section.

### 3.3.2 Biosignals as “baseline-channel” for multimodal emotion recognition

In integrated analysis of audio/visual/physiological channels for multimodal emotion recognition, biosensing has the attractive property that the sensors can provide continuous information, while the performance of the speech analysis, for example, is only triggered when a speech signal is observed from the user. Moreover, speech and visual channels can be intentionally manipulated by human social artifacts, e.g., when people set up a so-called poker face or simply do not speak while being angry. In this case, emotional states of the user take place internally and may not be detected by audio/visual sensing systems. There is a long standing belief that in such cases, biosignal analysis can be significant for multimodal emotion recognition, by providing information when other channels are either inoperative or being deliberately manipulated. Naïve versions of the belief gave rise to the thoroughly discredited technology of lie detection, but the possibility remains that information from biosignals will be more reliable than information from other channels under certain circumstances and given appropriate combination techniques.

An interesting question concerns sensor fusion. Pure statistical models are not able to represent the relationship between different data in a declarative manner. On the other hand, high-level fusion techniques, such as unification grammars, bear the danger that relevant joint information gets lost. To remedy this problem, multi-layer fusion models that are able to integrate multiple modalities at various levels of abstraction seem to provide a promising option.

### 3.3.4 Biosensors as complements to other measurements

A key task is to specify which biosensor type can be deployed to supplement audio/visual measurement with physiological information, e.g. EMG for face recognition and RSP/SC for speech recognition. There are some well known interactions; for example, when user is talking, even if the emotional state is neutral, the value of skin conductivity increases instantly; and similarly, the respiration signal changes into an irregular wave form with lower amplitude. Effects like these prevent any simple interpretation of physiological signals as ‘emotion channels’. However, it is also possible to imagine ways of turning them to advantage. For instance, the respiration wave shows particular abrupt changes corresponding to certain facial muscle-activity. It is also possible that EMG sensors on the jaws or forehead could improve accuracy in computation of FAP-variations in facial emotion analysis. They could also be positioned on the body to measure muscle contraction intensity for gesture recognition.

## 3.4 Element 4: Multimodal Emotion Recognition

Multimodal emotion recognition is a topic that has only partially been tackled up to the present, primarily focusing on audio and visual cues that can be used to provide classification of user’s emotional state to one of the archetypal discrete emotions.

As already described, element 4 is the ‘umbrella’ that covers all specific results provided by the other three ‘separate’ elements of the proposed WP4 exemplar. The basic tasks included in element 4 are the following.

### 3.4.1 Theoretical models of emotion related to different applications

The useful models for recognition of emotion should be identified, investigated and compared. These include:

#### *a. Discrete models for recognition of emotion*

- Description of the different models should be done.
- How much “discrete emotions” should be used.
- Which discrete emotions are more often expressed in specific contexts (like educational context, games context, office context).
- What are the patterns of these discrete emotions on the different channels and how the different channels can we weighted in the recognition of discrete emotion

#### *b. Dimensional models for recognition of emotion*

- An exhaustive description of the different models should be done.
- Which dimensional model would be used.
- How many dimensions would be useful to discriminate as well as possible different emotions.

#### *c. Appraisal models for recognition of emotion*

- Definition of the main criteria of the appraisal models.
- Could the appraisal models be useful to refine the identification of emotion provided with the other models?

- Which “check” could be useful for that?
- Which channels and which pattern could be used to refine the recognition of emotion.
- The theoretical predictions in the literature should be described and some propositions suggested.

### 3.4.2 Theoretical models of multimodal integration and emotion recognition

A description of each model listed below should be done and the main problems and limits of each of them addressed. A proposal for a specific model should be done, probably for different frameworks or context.

- *The Direct Identification model*
  - The input signals are directly transmitted to the bimodal classifier.
- *The Separated Identification Model*
  - The visual, auditory and physiological input are separately identified through two parallel identification processes.
- *The Dominant Modality Recoding Model*
  - A dominant modality drives the perception of other modalities.
  - This modality would be different according to context or the task.
- *The Motor Space Recoding Model*
  - The modalities are projected upon a common motor space where they are integrated before final categorization.
- *Integration of the different Modalities for Emotion Recognition*
  - Integrating audio and/or visual and/or physiological analysis in the recognition process.

### 3.4.3 Synchronization of the different channels

- *Temporality*

The temporal aspects relative to the synchronization of the different channel used in the recognition of emotion will be addressed systematically. The right temporal window to analyze the different channels and the way in which they can be combined during the analysis processes will be defined. The different time course of changes related to an emotional episode for the different channels will be taken into account to extract the synchronized changes.

The integration of signals will also take into account the relative importance of the different channels; the latter will be weighted for detection or refinement of detection of emotions.

➤ *Sequentiality of emotional processes*

The time course of expression of the emotional process is different for each channel and for each emotion; investigating this issue is central for the detection of emotion and will also be addressed. The weighting of the different channels in the process of emotion recognition must change through the time course of emotions according to the detections made in the different channels and their inter-relations.

## 4. Proposed allocation of tasks

There is a common basis in all four elements of the WP4 exemplar described in the previous section, which can be coded in terms of five WP4 tasks, as follows:

### **Task 4.1 Establish reference data**

WP4 will establish reference data capable of acting as a standard within the community as a whole. This data will represent all three modalities considered in WP4, i.e., speech, visual and physiological.

### **Task 4.2 Evaluate existing techniques**

Existing techniques in the fields of speech, face, gesture, and physiological signal emotional analysis, as well as in single-mode and multi-modal recognition, will be applied to standard data, and results will be compared. These techniques will be based on the current experience of WP4 participants and evaluation will be made using the established reference data.

### **Task 4.3 Advance the state of the art**

New techniques will be developed in the domains of speech, face, gesture, physiological signal emotional analysis, and integration of signals from multiple modalities for emotion recognition and generation of signs for expressive ECAs.

### **Task 4.4 Disseminate research results and conclusions**

The results of the integration and joint research activities will be disseminated to the European scientific, technical and user communities through workshop proceedings, reports and other electronic and conventional means.

### **Task 4.5 Provide the rest WPs with emotion analysis results**

New developments and validated techniques will be made available to the rest WPs in HUMAINE, especially to WP3, WP5, WP6, WP7.

Adopting the above task description, each WP4 participant will contribute to every task, since the latter include the research and developments of all four WP4 Working Groups.

In the following, we provide a list of the specific activities that will constitute each WG's contribution to the above WP4 exemplar tasks. Since Tasks 4.4 and 4.5 refer to dissemination and transfer of obtained results to the public and to the rest of HUMAINE WPs, this list mainly refers to Tasks 4.1, 4.2 and 4.3, i.e., to generation of reference material, which will be performed in collaboration with WP5, in the evaluation of existing techniques and the generation of new knowledge in each modality and in modelling the integrated approach, which will be achieved by interacting with WP3 for the theoretical and with WP6 and WP7 for the interaction and attention/cognition frameworks.

#### 4.1 Element 1: Emotional speech analysis

All the groups in WG1 (TAU, QUB, LIMSI, FAU Erlangen, UNIGE, TCD) will be involved in the definition of the speech recognition system. This meta task is composed of the following speech tasks (STs):

- ST1. Pitch detection methods: TAU, QUB
- ST2. Intensity analysis: TAU, QUB
- ST3. Voice Quality: TCD, TAU
- ST4. Text analysis: FAU Erlangen, LIMSI, Bari
- ST5. Prosodic Feature extraction: TAU, QUB, FAU Erlangen
- ST6. Classification: LIMSI, FAU Erlangen, TAU
- ST7. Expert Listening: QUB, TAU, UNIGE
- ST8. Establishing labelling schemes and reference data: all groups.

It can be easily verified that ST8 refers to Task 4.1, while ST1-ST7 involve both the evaluation of existing technologies and tools (Phase A, Task 4.2), and the generation of novel technologies (Phase B, Task 4.3).

#### 4.2 Element 2: Face and Gesture Analysis for Emotion Recognition and Expressive ECA Generation

The main tasks (visual tasks, VT) composing the visual analysis element of the WP4 exemplar are the following:

- VT1. Facial feature extraction: ICCS, Paris8, Newpartner (FER Zagreb)  
This includes face detection, facial feature extraction, MPEG-4 FAP computation, viseme analysis, gaze analysis, detection of patterns of movements and of temporal correlations, FAP interaction and transition.
- VT2. Gesture analysis: ICCS, Paris8  
This includes detection and computation of arm, wrist, hand movements and of visual expressivity parameters.

- VT3. Classification: ICCS, Paris8, Newpartner (FER Zagreb)

This includes usage of a-priori knowledge and rules, adaptation, fuzzy representations.

- VT4. Reference data: ICCS

This includes generation of databases with facial expressions, gestures (and other modalities), with appropriate labelling; this will be done in collaboration with WP5.

It can be easily verified that VT4 refers to Task 4.1, while VT1-VT3 involve both the evaluation of existing technologies and tools (Phase A), corresponding to Task 4.2, and the generation of novel technologies (Phase B), corresponding to Task 4.3.

### 4.3 Element 3: Emotional Biosignal Analysis

The main tasks (biosignal tasks, BT) composing the biosignal analysis element of the WP4 exemplar aim at recognising emotions, such as stress, in realistic situations, such as driving environments. They are the following:

- BT1. Biosignals as single factor for emotion recognition: UA, FAU Erlangen, UNIGE
  - Extending sensor parameters depending on physiological priority
  - Developing new feature types
  - Investigating methods for classifier combination
- BT2. Biosignals as “baseline-channel” for multimodal recognition: UA, TAU, UNIGE
  - Establishing of biosignals as baseline-channel
  - Fusion of features from biosignals and FAPs
  - Feature fusion from biosignals and speech
  - Structural multilayer fusion of audio/visual/physiological channels
- BT3. Biosignals in reference multimodal databases: UA, TAU, FAU Erlangen
  - Establishing annotation and segmentation methods
  - Evaluation of multimodal databases
- BT4. Biosensors as supplementary to external measurements: UA,TAU,FAU Erlangen
  - Sensing of EMG-signals from various facial expressions
  - Correlating SC/RSP-measurements with speech intonation.

It can be seen that BT3 corresponds to Task 4.1, while BT1-BT3 refer to both evaluation of existing technologies and tools (Phase A), corresponding to Task 4.2, and generation of novel technologies (Phase B), corresponding to Task 4.3.

#### **4.4 Element 4: Multimodal Emotion Recognition**

The basic tasks (multimodal recognition tasks, RT) included in element 4 of the exemplar, to which all WP4 partners participate (led by UNIGE) are the following:

##### RT1. Theoretical models of emotion related to different applications

- Investigate discrete models for recognition of emotion
- Examine dimensional models for recognition of emotion
- Investigate appraisal models for recognition of emotion

##### RT2. Theoretical models of multimodal integration for emotion recognition

- The Direct Identification model
- The Separated Identification Model
- The Dominant Modality Recoding Model
- The Motor Space Recoding Model and the Dynamic

##### RT3. Integration of the Different Modalities

- Audiovisual Emotion Recognition
- Multimodal Recognition based on Physiological and External (audio/visual) Signals

##### RT4. Synchronization of the different channels

- Temporality
- Sequentiality of emotional processes

RT1-RT4 refer to both evaluation of existing theory and technologies (Phase A), corresponding to Task 4.2, and generation of novel theoretical and application dependent results (Phase B), corresponding to Task 4.3.

## 5. Provisional timeline

### 5.1 Element 1: Emotional speech analysis

Key actions	Estimated completion	Type of associated deliverable	Estimated date of delivery
Specific exemplar and partners' role description	month 18 <sup>th</sup>	Report	month 18 <sup>th</sup>
Labelling schemes along with associated reference databases	month 48 <sup>th</sup>	Report & Database	a) month 23 <sup>rd</sup> b) month 48 <sup>th</sup>
Comparative study of basic analysis schemes (pitch extraction, pitch smoothing, segmentation), including improved and new methods	month 48 <sup>th</sup>	Report & toolbox	Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Expert assessment of benchmark data, probably only on selected parts	month 30 <sup>th</sup>	Report	month 30 <sup>th</sup>
Feature extraction study	month 48 <sup>th</sup>	Report	a) month 30 <sup>th</sup> b) month 48 <sup>th</sup>
Comparative study of classification methods	month 48 <sup>th</sup>	Report	a) month 30 <sup>th</sup> b) month 48 <sup>th</sup>
Meta task – interim version	month 48 <sup>th</sup>	Report & toolbox	month 48 <sup>th</sup>

### 5.2 Element 2: Face and Gesture Analysis

Key actions	Estimated completion	Type of associated deliverable	Estimated date of delivery
Specific exemplar and partners' role description	month 18 <sup>th</sup>	Report	month 18 <sup>th</sup>
Facial feature extraction	month 48 <sup>th</sup>	Report & toolbox	Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Gesture analysis	month 48 <sup>th</sup>	Report & toolbox	Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>

Classification	month 48th	Report toolbox	&	Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Reference data	month 48th	Report Database	&	a) month 23 <sup>rd</sup> b) month 48 <sup>th</sup>

### 5.3 Element 3: Emotional Biosignal Analysis

Key actions	Estimated completion	Type of associated deliverable	Estimated date of delivery
Specific exemplar and partners' role description	month 18th	Report	month 18 <sup>th</sup>
Biosignals as single factor for emotion recognition	month 48th	Report toolbox	& Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Biosignals as "baseline-channel" for multimodal recognition	month 48th	Report toolbox	& Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Biosensors as supplementary to external measurements	month 48th	Report toolbox	& Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Biosignals in reference multimodal databases	month 48th	Report Database	& a) month 23 <sup>rd</sup> b) month 48 <sup>th</sup>

### 5.4 Element 4: Multimodal Emotion Recognition

Key actions	Estimated completion	Type of associated deliverable	Estimated date of delivery
Specific exemplar and partners' role description	month 18th	Report	month 18 <sup>th</sup>
Theoretical models of emotion related to different applications	month 48th	Report	Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Theoretical models of multimodal integration for emotion recognition	month 48th	Report	Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Integration of the Different Modalities	month 48th	Report toolbox	& Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>
Synchronization of the different channels	month 48th	Report toolbox	& Part A: month 30 <sup>th</sup> Part B: month 48 <sup>th</sup>

## 6. Conclusions

The proposed WP4 exemplar description has been presented in this Deliverable. This exemplar was formulated through interaction and collaboration between the active WP4 partners during the 1<sup>st</sup> year of HUMAINE, as well as through close interaction with the leaders of the rest WPs of HUMAINE, and in particular WP3, WP5, WP6, WP7.

Its form is based on the draft presented in WP4 Deliverable D4b (May 2004), enriched by the discussions and collaboration during the WP4 Workshop held on September 19-21, 2004 in Santorini, Greece, and finalised at the Paris October meeting of HUMAINE WP leaders and representatives.

As described in the text, part of the proposed developments will take place within the time interval of 13-30 months, as foreseen in the new HUMAINE JPA, currently formulated, while the other part will be completed in the end of the four-year R&D developments of HUMAINE.