SHELF: SELECTION OF NEW MODULES

Conveying Affectiveness in Leading-edge Living Adaptive Systems

CALLAS

Project IST-34800

Deliverable D112 WP1.1

Deliverable
Version 1.0 – 31 October 2008
Document. ref.: callas.D112.ICCS.WP1.1.V1.0

Programme Name: ..... IST
Project Number: 34800
Project Title:..... CALLAS
Partners:..... Coordinator: ENG (IT)
Contractors:
VTT Electronics, BBC, Studio Azzurro, XIM, Digital Video, Humanware, Nexture, University of Augsburg, ICCS/NTUA, University of Mons, University of Teesside, Helsinki University of Technology, Paris 8, Scuola Normale Superiore di Pisa, University of Reading, Fondazione Teatro Massimo, HITLaboratory New Zealand

Document Number: callas.D112.ICCS.WP1.1.V1.0
Work-Package: WP1.1
Deliverable Type: ..... Document
Contractual Date of Delivery: ..... 31 October 2008
Actual Date of Delivery: ..... 31 October 2008
Title of Document: ..... SHELF: Selection of New Modules
Author(s): .....

Approval of this report .....

Summary of this report: .....

History:.....

Keyword List: .....

Availability ..... This report is: public

Table of Contents

Executive Summary

WP 1.1 is responsible for establishing a process that will identify and select candidate technologies for inclusion in the CALLAS Shelf and for managing this process throughout the project. Deliverable 111 describes the selection of new modules and the improvements and adaptations made for the modules already selected during the first year and described in D111.

D111 defined the characteristics of main components that could be candidates to be included in the Shelf and identified leading edge technologies matching these criteria. New and available tools have already been integrated into showcases and requirements were tackled with shelf components developed specifically for this purpose by CALLAS partners. Apparently, new requirements have been aroused by both existing and new showcases, making the selection of components an ongoing procedure throughout the project. D112 defines the new requirements of the showcases for the already selected components, examines the new versions of these components and selects new components.

Section 1 is devoted to audio analysis. A new component and updated versions of selected ones –accompanied by the new requirements of the showcases- are presented.

Correspondingly, Section 2 is devoted to visual analysis. Combination of already available components, the need of a component for facial expression recognition and modifications of the components included in D111 are described in this section.

Section 3 describes other sensors used for gesture recognition and motion capture, while section 4 tackles the issue of multimodal fusion. Section 5 refers to synthesis and interaction issues and more precisely issues concerning affective music synthesis, laughter synthesis and emotional attentive ECA.

The last section (6) contains the references.

1. Audio Analysis

1.1 New Components

1.1.1 Laughter Recognition

Market Survey

Several CALLAS showcase developers have shown an interest for a device able to detect laughter acoustically. After a market survey, no such components are found to be available.

There are only a few studies concentrated on discriminating laughter from speech, using typical speech features as inputs of popular classifiers like Neural Networks, Gaussian Mixture Models or Support Vector Machines (for more details, see D122). These components are not open-source and were trained for discriminating pure speech from pure laughter, so they cannot be used for CALLAS purposes neither can be included in the framework.

Specific criteria of selection procedure – Shelf requirements

The component desired by CALLAS partners should run continuously in a performance environment and should thus be able to deal with other natural sounds than pure speech or laughter. The design of such a component is complicated, but FPMS, as specialist in audio analysis, will follow a step by step approach to evolve towards this kind of device. (see D122).

Selected component(s)

No software meeting the requirements is available, so FPMS will develop such a component, adapted to CALLAS requirements. A detailed description of Laughter Recognition Component can be found in D122.

1.2 Updated versions of already selected components

1.2.1 Emotional Speech Recognition

After the selection process of the first year, the component “Emotional Speech Recognition” of FPMS was selected.

New products-Market Survey –progress/overview during the last year

No dramatic change has occurred during the last year, since the publication of D111, where an extensive market survey can be found. The commercial dictation software Nuance’s Dragon Naturally Speaking is still leader on the market. A new version was released, claimed to reach up to 99% accuracy and to be faster than the previous one. The figures and demonstration are impressive, but the settings should not be forgotten: *quiet environment*, *high-quality microphone*, *system trained by the user*, *only one voice can be recognized*. Other concurrents are few. Microsoft provides similar software in Vista, but is less effective and popular. Loquendo is also active in the field, while IBM’s Via Voice is not updated anymore.

Regarding open-source products easily integrable in other applications, HTK is still very

popular. Its success also resides in the fact that, besides enabling to perform speech recognition, its powerful tools may also be used in other fields like speech synthesis, movements modelling, etc.

Advantages of the new release of the component in comparison with the previous-selected-one

The new release of the component provides controls over several parameters of the speech recognition like the allowable pause inside an utterance (which affects the delay of recognition when the user stops speaking) or the garbage level (if the system should try to match every speech sound to one of the vocabulary words or if words not closely corresponding to one entry should better be rejected).

The biggest asset is however the easy switch between languages. The previous released already made possible to change the list of recognizable words during the application, but only one language could be used. Now, the spoken language might be changed on the fly, and we can imagine commanding the switch by voice.

Finally, a graphical user interface has been developed, for simplifying tests of the component.

New requirements of Shelf and/or Showcases regarding the selected component

The access to the pause time was asked by Framework developers. Several Showcase developers asked for the potential use of different languages, which led to offering the possibility to switch from a language to another instantaneously.

Updated component description

An updated and detailed description of the component can be found in the deliverable 122.

1.2.2 (Live) Audio analysis

After the selection process of the first year, the component “Live Audio Analysis” of VTT was selected.

New products-Market Survey –progress/overview during the last year

Detailed overview can be found in earlier release, D111. There has not been any notable development in the field in the last year that would have added to the current situation.

Advantages of the new release of the component in comparison with the previous-selected-one, new requirements of Shelf and/or Showcases regarding the selected component

The audio feature extraction classifies the audio stream into 8 different sound classes. The classes of current version are speech, music, silence, constant noise (i.e. car engine noise), variable noise (i.e. restaurant noise), clapping, whistling and applause. The component output is a corresponding audio class for each audio frame. The component will take as input only live audio.

The improvements of audio component include dynamic silence threshold for live situations, optimization of the HMM parameters for each audio class, and detection of 3 new audio classes (applause, clapping and whistling).

It should be noticed that UDP communication was added as there was such a request by a CALLAS partner.

Updated component description

An updated description of the component for live audio analysis component can be found in D122.

1.2.3 Emotion recognition from acoustic features

After the selection process of the first year, the component “Emotion recognition from acoustic features” of UOA was selected.

New products-Market Survey –progress/overview during the last year

Emotion recognition from acoustic features didn’t present much progress in the last year. There was not released any other, commercial or research, open-source product featuring real-time emotion recognition. Available commercial and research products are described in D111. EmoVoice, the component developed by the University of Augsburg and selected for inclusion in the framework during the first year, was further developed and adapted to meet the requirements of the showcases.

Advantages of the new release of the component in comparison with the previous-selected-one - New requirements of Shelf and/or Showcases regarding the selected component

A new requirement by the showcases, esp. the puppet wall and the interactive opera, was to have access not only to the classification result, i.e. emotion classes, but also to the underlying prosodic measures and features. This should allow a direct mapping from speech features to system reaction that is transparent to the user, e.g. reactions to changing loudness of the voice.

For this reason, more output options were integrated into EmoVoice, and a separate interface to track pitch and speech signal energy continuously and in near-real time was released.

Updated component description

For a more detailed and updated description of EmoVoice, see Deliverable 122, section 1.2.

1.2.4 Emotion recognition from linguistic features

After the selection process of the first year, the component “Emotion recognition from linguistic features” of UOA was selected.

New products-Market Survey –progress/overview during the last year

To our knowledge, in the last year there was no commercial or research, open-source product featuring emotion recognition from linguistic features. So, the available commercial and research products are described in D111.

Advantages of the new release of the component in comparison with the previous-selected-one - New requirements of Shelf and/or Showcases regarding the selected component

Our previous release relied on statistical emotion recognition making difficult real-life, online emotion recognition. Moreover, the statistical engine was incomprehensible for a human observer and classified emotions more reliable in long texts (e.g., in product or movie reviews) and not in short texts as natural-language utterances. Such requirements on emotion recognition are, however, indispensable for some showcases, esp. the puppet wall showcase and the interactive opera showcase.

In order to solve these problems, we developed an engine for semantic emotion recognition that analyzes affect using emotion words under strong consideration of English grammar and its peculiarities, e.g., negations.

Updated component description

The demo version of the previous, statistical release and the demo version of the updated, semantic component are placed for testing purposes on http://emotion.informatik.uni-augsburg.de. A detailed description of the component is included in D122.

2. Visual Analysis

2.1 New Components

2.1.1 People's Behaviour Estimation in Multi-user Environments

This new component is a combination of Head pose estimation component by ICCS and Video feature extraction component by VTT, both selected during the first year of the project and described in D121.

Specific criteria of selection procedure – Shelf requirements

The ICCS head pose estimation component has been integrated with the VTT video feature extraction component. This combination is expected to serve the determination of people's existence in multi-persons environments, eliminating the false alarms. On the other hand the integration also helps to make the head pose estimation more efficient by providing it with a small set of target frame areas instead of having to process the entire frame. In addition, it will help to the extraction of people's attention in environments where more than one person exists.

2.1.2 Facial Expression Recognition (output: ekmanian emotions / dimensional models)

Overview-Market Survey

The majority of efforts in affective computing concern automatic analysis of facial displays. Surveys of studies on machine analysis of facial affect can be found at [6], [7], [8], [9]. These surveys indicate that the capabilities of currently existing components for facial expression recognition are rather limited. More specifically, current facial affect analyzers handle only a small set of volitionally displayed prototypic facial expressions of six basic emotions and adopt strong assumptions (i.e., the systems can handle only portraits or nearly-frontal views of faces with no facial hair or glasses, recorded under constant illumination and displaying exaggerated prototypic expressions of emotions).

Automatic detection of the six basic emotions under these assumptions, that is, in posed, controlled displays can be done with reasonably high accuracy. However detecting these facial expressions in the less constrained environments of real applications is a much more challenging problem. There have been just a few such tentative efforts aimed at detection of cognitive and psychological states like interest [10], pain [11] and fatigue [12].

Specific criteria of selection procedure – Shelf requirements

The facial expression recognition component chosen for the CALLAS applications was the one offered by the ICCS. One of the strong reasons for choosing it was that it was directly related to the facial feature detection and tracking component, thus, making alterations, usage and experimentations more flexible.

A first version of facial expressions recognition has been released. It can handle both Ekmanian and dimensional models. The technique is based on the previous components (Facial Feature Detection & Tracking) and employs Fuzzy modeling in order to map FAPs (Facial Animation Parameters), as extracted by the facial feature tracking step, to certain rules that output emotions. The component can function real time, under unconstrained environments, and is user independent. However, at start up, the user needs to face the

camera frontally, before tracking and expression recognition begin. For more robust Expression recognition, the facial features detection and tracking component has been enhanced with handling the nose-tip and two points on each eyebrow. Fraunhofer face detector [5] from Fraunhofer Institute detects the face and the eyes, while it classifies expressions in 4 categories: Happiness, Anger, Sadness, Surprise. The module can be downloaded for test reasons and testing has shown that tracking cannot handle head rotations of more than approximately 45^{o}.

Selected component(s)

A comprehensive description of the ICCS selected component can be found at D122.

2.2 Updated versions of already selected components

After the selection process of the first year, the component "Video features" of VTT was selected.

2.2.1 Video Features

New products-Market Survey –progress/overview during the last year

Detailed overview can be found in earlier release, D111. There has not been any notable development in the field in the last year that would have added to the current situation.

Advantages of the new release of the component in comparison with the previous-selected-one, new requirements of Shelf and/or Showcases regarding the selected component

Within last year the optical flow implementation was included to video feature component. Optical flow is defined as apparent motion of image brightness and can be used to track objects (faces) after they have been found. Performance of object detection and tracking was improved, so more faces can be detected. UDP socket communication was also implemented from request of the partner, to control the shut down and start up of the component in installations using the component.

Integration with the ICCS head pose estimation component has also started (as described in section 2.2.1). This will improve the performance of both components, aiming at functionality in environments with multiple people, reduction of false alarm rate, improved performance in facial feature tracking and extraction of people's attention in environments where more than one person exists.

Updated component description

An updated description of the component for video feature extraction component can be found in D122.

2.2.2 Gaze/pose estimation

After the selection process of the first year, the component “Gaze and Pose estimation” of ICCS was selected.

New products-Market Survey –progress/overview during the last year - New requirements of Shelf and/or Showcases regarding the selected component

As stated by the requirements of certain showcases, the need for a common framework to infer human attention in front of a camera has appeared. Thus, Gaze and Pose modules have been merged and a neurofuzzy system was developed to characterize a person’s engagement towards a system. So far, attention vs distraction is well distinguished, while initial estimates for nervousness and tiredness have been released. Human attention recognition in an HCI environment is an issue that has not been widely studied, even on an academic level. Not a lot of works have appeared based on both head pose and eye gaze [3], especially within a monocular environment. However, there has been quite a lot of relative research in the field of monitoring meeting participants’ attention in certain set-ups [4].

Advantages of the new release of the component in comparison with the previous-selected-one

The improved version of gaze component is much more stable and reliable and does not limit the user in specific positions with regards to the web-cam (eyes bounding boxes) as in previous versions.

Regarding the new version of pose component, the technique is strongly related to the one presented in D111. However, many improvements have been introduced, which are mostly related to the video processing part. More specifically, a series of rules, based on prototypes of facial geometry and natural human motion have been introduced. These rules have enabled our technique to be even more robust in natural and spontaneous environments. During 2008, a series of new methods regarding the problem of head pose estimation have been published. Most methods use monocular systems. However, the techniques used, either require a large amount of training [1], or depend on initial guesses of camera intrinsic parameters [2].

Updated component description

An updated description of the combined component for pose and gaze estimation can be found at D122.

2.2.3 Hand detection / tracking and gesture expressivity features extraction

After the selection process of the first year, the component “Hand detection / tracking and gesture expressivity features extraction” of ICCS was selected.

New products-Market Survey –progress/overview during the last year

Gesture expressivity analysis and synthesis has been increasingly attracting attention from research areas such as affective computing, pattern recognition and psychological and behavioural studies. As a result the corresponding research community has been very active drifting along the way the technology industry in novel either theoretical or application oriented fields.

Mura in [22] explores self expression via wearable technologies and ways to increase the expressive abilities of clothing, with responsive and reconfigurable components that allow a

better representation and expression of personal choices or emotional states. Presenting Smart Shirt, Frison, Buebelle, the Chimerical Garment, and Halo manufactured by Sensatex, Philips, Co he delves more into how clothing becomes a complementary medium of communicating emotions and naturally enhances the expressive abilities of the self by broadening the channels of communication.

Merola in [20] presents a study about athletes' gestures during the telling of their best and worst performances. While manual annotation was adopted significant conclusions were drawn for the correlation of emotion and gestures. On the other hand Pelachaud in [19] focuses on the synthesis side of the problem and how expressivity features, including gesture expressivity, can be included in the animation of an affective embodied conversational agent (ECA). Interestingly the overall scenario is based on the same expressivity features extracted by the Hand detection/tracking and gesture expressivity features extraction component.

Takala et al. In [23] deal with gesture expressivity extraction within the scope of interactive art and virtual environments and present an early showcase, the virtual orchestra, an animated band conducted with motion of a baton.

Advantages of the new release of the component in comparison with the previous-selected-one

The latest version of the component proves to be more robust and accurate since Kalman filtering was employed in the tracking process and enhanced stability. In terms of expressivity the formulas defining were revised so as to include more than just position features. For example fluidity now incorporates movement direction cues and does not depend solely on the norm of the motion vectors. Thus fluidity is better represented since it refers not only to the quantity of motion but additionally to the direction of motion.

Gesture recognition and prediction was also investigated since incorporating methods to enhance the expressivity features extraction component is quite advantageous. In many cases expressivity features depend on the nature of the performed gesture and thus it is quite crucial for the gesture class to be known. Normalisation processes or gesture class profiling can then be applied in order for the expressivity features to be independent from the gesture class. Of course this is not the case for general gestural behaviour with emotional qualities expressed with hand movement and hand movement cannot be classified as a restricted / closed gesture class.

New requirements of Shelf and/or Showcases regarding the selected component

Gesture expressivity definitions were also adapted to fit the Wii accelerometer output in order to be integrated with the Wii remote. Influence of gravity was reduced and single or multiple integrations were employed to derive distances from acceleration values. A description can be found in the section for *WiiGLE (Wii-based Gesture Learning Environment)* in D122.

Updated component description

D122 contains a more detailed description of the new version of the component.

3. Other sensors

3.1 New Components

3.1.1 Wii-based Gesture Learning Environment (WiiGLE)

Overview-Market Survey (available products-open source or not)

The WiiGLE component allows to use Nintendos Wiimote controller with arbitrary gesture sets. To this end, the whole processing pipeline from recording training gestures, selecting feature sets, training classifiers, validating classifiers to online recognition of new instances has been integrated.

There are several open source approaches that aim at the same functionality, but WiiGLE is so far the only component that has the full flexibility to easily define application dependent gestures and incorporate them seamlessly into a given application. There is one commercial product licensed by Nintendo that provides the same functionality for application developers on Nintendo's Wii console, but the price for one license is around 2500 Dollars.

Specific criteria of selection procedure – Shelf requirements

WiiGLE possesses some very useful features. WiiGLE features a modular approach allowing the developer to either use the prepackaged features and classifiers supplied or to implement his own features and classifiers making use of a specifically designed API. Features as well as classifiers can then be loaded dynamically by the component. To allow for rapidly testing and analysing gesture sets for a given application, WiiGLE features the WEKA arff-format opening up the possibility to use this powerful data mining tool to improve gesture usage in an application.

Another important issue is that apart from the CALLAS community, around 30 institutions worldwide are registered users of the WiiGLE component.

Up to date information can be found on the WiiGLE wiki: http://mm-werkstatt.informatik.uni-augsburg.de/documents/WiiGLE/doku.php

Selected component(s)

WiiGLE has been selected as the only component of this functionality which has full flexibility to easily define application dependent gestures and incorporate them seamlessly into a given application. A more detailed description can be found at D122.

3.2 Updated versions of already selected components

3.2.1 Gesture recognition from mobile phones

After the selection process of the first year, the component “Gesture recognition from mobile phones” of VTT was selected.

New products-Market Survey –progress/overview during the last year

Detailed overview can be found in earlier release, D111, no notable changes took place in the past year.

Advantages of the new release of the component in comparison with the previous-selected-one, new requirements of Shelf and/or Showcases regarding the selected component

Implementation of behavioral cues on Symbian was done during last year. An automatic system for recognizing behavioral cues from gestures is used on accelerometer equipped Nokia mobile phones. The component analyses the movement data from the accelerometer and produces a set of features, from which affective data can be extracted. The feature calculation is implemented on Symbian S60 and calculations are performed on the phone. The calculated features are sent to PC via Bluetooth, where incoming messages are received, re-structured and sent ahead via UDP socket by a C++ implementation.

Updated component description

An updated description of the component for gesture recognition from mobile phones can be found in D122.

3.2.2 HumanGlove

After the selection process of the first year, the component “HumanGlove” of Humanware was selected.

New products-Market Survey –progress/overview during the last year

As already mentioned in D111, an alternative for HumanGlove is still Cyberglove, but there is no evidence of improvement in its development.

(http://www.immersion.com/3d/products/cyber_glove.php)

Advantages of the new release of the component in comparison with the previous-selected-one

In the development of the component, we took into account the requirements of the partners and of the reviewers. The partners expressed the need for more input data (in particular the tracking of wrist movements) and more reliable wiring. The reviewer asked for a component more suitable for extracting emotion. Along these lines, we endowed HumanGlove with wrist sensors (ad/abduction and flex/extension), we improved the material and the soldering of the wires and we embedded an inertial platform in the glove (see Sect. 3.2.3)

New requirements of Shelf and/or Showcases regarding the selected component - Updated component description

For a correct usage of the glove, it must be donned properly (see User Manual) in order to have adequate positioning of the sensors with respect to the fingers/wrist joints. Before being able to use the glove, a user must complete a calibration procedure as described in the user manual. A PC with a Bluetooth connection is also needed due to the new wireless requirements.

During the last year many hardware improvements were achieved: wrist sensors have been added to the device: a new cover and a new tissue have been developed in order to facilitate the donning; cable routing has been reassessed and modified, the cable material and the soldering have also been reviewed and optimized.

At the moment, the component is being evaluated by Studio Azzurro [SAZ] who is using it with success for interactive theatre performances. During the first year review we have been requested to address the issue of emotional cues recognition, thus we have started to develop an inertial platform in order to analyze kinematical data of the forearm/arm especially focusing on acceleration, velocity and jerk of movement which are more closely related to emotional state of the wearer than are the finger postures.

3.2.3 Haptic Tracking

After the selection process of the first year, the component “Haptic Tracking” of Humanware was selected.

New products-Market Survey –progress/overview during the last year

In the last decades, thanks to improvements in the inertial technology and in data processing, there are a lot of tracking devices. Zhou et al. already tracked the upper limb [32]. Foxlin at al. used the InertiaCube (http://www.i-glassesstore.com/inertiacube3.html) for inertial tracking [33].
The market leader is probably Xsens Technologies B.V. (http://www.xsens.com/en/home.php) with its MTx 3DoFs orientation tracker, but there is no evidence of improvement in its development during the last year.
None of the previous mentioned researches or products addressed the emotional issue. For more details on this topic refer to D122.

Advantages of the new release of the component in comparison with the previous-selected-one

After the first year project review, the device development has started from scratch. Currently, the inertial platform embedded in HumanGlove consists of a board with 5 sensors which sense a total number of 9 DoFs. The device provides data of 3-axial acceleration, data of 3-axial gyroscopic angular speed, data of 3-axial magnetometer (i.e. 9 new data to be analysed and correlated for puppetry or for emotion deduction). Each independent channel is user selectable, as is the sampling frequency. The reports are in binary format and are transmitted over a wireless Bluetooth link (using a virtual serial COM port).

The purpose of this device is to use kinematical data both for tracking the arm/forearm movement (integrating kinematical data to get the relative position of the hand from the torso) and for analyzing the user's activity (directly from kinematical data or differentiating the acceleration to obtain the motion jerk).The hypothesis is that the inertial platform together with the finger posture tracker (i.e. the improved HumanGlove device) can be used to detect Arousal and Dominance cues by analysis of hand and arm kinematical data. The device should be able to correlate the above mentioned features (such as distance from body, velocity, acceleration and jerk) to emotional states in PAD space. The hypothesis to be proved is that, for gesticulation and spontaneous gestures, distance of hands from the torso and their elevation can be used to evaluate the Dominance, while energy, velocity and jerk can be used to evaluate the Arousal (or Activity in a non-PAD space).

New requirements of Shelf and/or Showcases regarding the selected component - Updated component description

The inertial platform is embedded in the acquisition board of the glove and they both use the same Bluetooth device for communication. Thus the requirements are shared with the HumanGlove component. A further requirement is the need to calibrate the device: for example the hand and the arm must lie in a known position/posture at the beginning of the acquisition.

At month 24 the first prototypes of the inertial platform have been released as beta version and are under testing. Given the above, no data about behavioral cues can be consistently reported. Nevertheless, we already have a hypothesis regarding the emotion elicitation from hands and arm motions.

4. Multimodal / Fusion

4.1 New Components

4.1.1 Fusion Component for multi-modal Recognition of Emotional States and AdHoc Multimodal Semantic Fusion Component

To our knowledge, in the last year there was no commercial or research, open-source product featuring multi-modal emotion recognition. Emotion recognition, however, is indispensable for some showcases, esp. the puppet wall showcase and the interactive opera showcase.

We study fusion component that uses statistical means: features from the statistical engine in section 1.2.4, prosodic features from section 1.2.3, as well as fusion of features from meta-modalities like deictic, stylometric, grammatical. We study also semantic fusion component considering statistical results of following modalities: lexical modality in section 1.2.4, prosodic modality from section 1.2.3, as well as results of meta-modalities like deictic, stylometric, grammatical. In both cases, we explore five modalities and combine their features.

The fusion result confirms slight enhancement of recognition results compared with results without fusion.

5. Synthesis/ Interaction

5.1 New Components

5.1.1 Laughter Synthesis

Overview-Market Survey - Specific criteria of selection procedure – Shelf requirements

Laughter conveys emotional information and is a very communicative signal. In consequence, laughter production is suited to CALLAS applications, aiming at eliciting emotions.

Few attempts have been made on laughter synthesis. Two interesting approaches are described in D133. However, the generated laughter episodes are not perceived as natural and, in consequence, not truly usable in applications.

The Speech Synthesis Loquendo system enables to include some non-verbal emotional signals inside the utterances to increase the naturalness of the voice. Laughter is one of the options. Nevertheless, laughter utterances are not synthesized: the user has to choose between a few pre-recorded utterances. The feature is interesting, but it needs some extension to maintain its perceived naturalness in “long” applications: if the same laughter samples are always played, the users will rapidly perceive this non-natural, repetitive characteristic and get bored.

In CALLAS, we would thus like to extend the aforementioned approaches. First, we need a system to navigate inside a large laughter database, enabling to find episodes corresponding to a desired effect and to bounce from one episode to another, preventing from always playing the same utterances. Then, really synthesizing laughter episodes could be investigated. More details are given in D133.

5.2 Updated versions of already selected components

5.2.1 Emotional Attentive ECA

After the selection process of the first year, the Emotional Attentive ECA developed by Paris VIII was selected.

New products-Market Survey –Progress/overview during the last year

SAIBA [14] is an international research initiative whose main aim is to define a standard framework for the generation of virtual agent behaviour. It defines three levels of abstraction from the computation of the agent's communicative intention, to behavior planning and realization. The **Intent Planning** module decides the agent's current goals, emotional state and beliefs, and encodes them into the Function Markup Language (FML) [17]. To convey the agent's communicative intentions, the **Behavior Planning** module schedules a number of communicative signals (e.g., speech, facial expressions or gestures) which are encoded with the Behavior Markup Language (BML). It specifies the verbal and nonverbal behaviors of ECAs [14]. Finally the task of the third element of the SAIBA framework, **Behavior Realization**, is to realize the behaviors scheduled by the Behavior Planning. It receives input in the BML format and it generates the animation.

There exist several implementations like SmartBody [15] and BMLRealizer [16] that are SAIBA compatible. SmartBody [15] is a modular, distributed open-source framework for

animating ECAs in real time. It is based on the notion of animation controllers. The controllers are organized in a hierarchical structure. In SmartBody two types of controllers are distinguished. Ordinary controllers manage the separate channels e.g. pose or gaze. Then the meta-controllers manipulate the behaviors of subordinate controllers allowing the synchronization of the different modalities to generate consistent output from the BML code. SmartBody corresponds to the Behavior Realization module of the SAIBA architecture. It takes as input BML code (including speech timing data and the world status updates); it composes multiple behaviors and generates character animation synchronized with audio. The verbal content is generated by an external TTS system. BML used within SmartBody is a subset of the standard.

SmartBody can be used with the Nonverbal Behavior Generator [18] that corresponds to the Behavior Planning in the SAIBA framework. It is a rule-based module that generates BML annotations for nonverbal behaviors from the communicative intent and speech text. On the other hand, SmartBody can be used with different characters, skeletons and even different rendering engines. In was used in many applications for example in Virtual Patient [13] to realize a virtual boy with psychological disorders.

BMLRealizer [16] created in the CADIA lab in Reykjavik is another implementation of the Behavior Realization layer of the SAIBA framework. It is an open source animation toolkit for visualizing virtual characters in 3D environment that is partially based on the SmartBody framework. As input it also uses BML; the output is generated with the use of the Panda3D rendering engine.

Advantages of the new release of the component in comparison with the previous-selected-one - New requirements of Shelf and/or Showcases regarding the selected component

As asked by showcase developers, of great interest would be a combination of ICCS components regarding Head Pose and Eye Gaze estimation with Paris8 ECA. To this aim, a real-time user interface for testing shared-attention behaviors with an embodied conversational agent has been developed. In two-party conversations, shared attention and related aspects, such as interest and engagement, are critical factors in gaining feedback from the other party and allowing an awareness of the general state of the interaction. Taking input from a single standard web-camera, the system is capable of processing the users eye and head directions in real-time. We are using this detection capability to inform the interaction behaviors of Greta and enable it to engage in simple shared attention behaviors with the user and objects within the scene in order to study in more depth some critical factors underpinning engagement.

Updated component description

An updated description for the Emotional Attentive ECA Greta can be found at D133.

6. References

[1] M.D. Cordea, E.M. Petriu, D.C. Petriu, "Three-Dimensional Head Tracking and Facial Expression Recovery Using an Anthropometric Muscle-Based Active Appearance Model," IEEE Trans. Instrum. Meas., vol. 57, no. 8, pp. 1578 – 1588, 2008.

[2] L.-P. Morency, J. Whitehill and Javier Movellan, Generalized Adaptive View-based Appearance Model: Integrated Framework for Monocular Head Pose Estimation, 8th International Conference on Automatic Face and Gesture Recognition (FG 2008), September 2008

[3] Matsumoto, Y., Ogasawara, T., Zelinsky, A.: Behavior recognition based on head pose and gaze direction measurement. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Takamatsu, Japan (2000) 2127–2132

[4] Otsuka, K., Takemae, Y., Yamato, J.: A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In: ICMI. (2005) 191–198

[5] Fraunhofer Face Detector: http://www.iis.fraunhofer.de/EN/bf/bv/kognitiv/biom/dd.jsp

[6] M. Pantic, A. Pentland, A. Nijholt, T.S. Huang, "Human Computing and machine understanding of human behaviour: A Survey", *Proc. ACM Int'l Conf.Multimodal Interfaces*, pp. 239-248.

[7] M. Pantic, L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions – The State of the Art", *IEEE Trans. Pattern Analysis and Machine Intelligence*,Vol. 22, No. 12, pp. 1424-1445, 2000.

[8] Y.L. Tian, T. Kanade, J.F. Cohn, "Recognizing action units for facial expression analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 97-115, 2001.

[9] M. Pantic, "Face for Ambient Interface", *Lecture Notes in Artificial Intelligence*, vol. 3864, pp. 35-66, 2006.

[10] P. Ekman, E.L. Rosenberg, (eds.), *What the face reveals: Basic and applied studies of spontaneous expression using the FACS*, Oxford University Press, Oxford, UK, 2005.

[11] M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, "Measuring facial expressions by computer image analysis", *Psychophysiology*, Vol. 36, No. 2, pp. 253-263, 1999.

[12] D. Goleman, *Emotional Intelligence*. Bantam Books, New York, NY, USA, 1995.

[13] P. G. Kenny, T. D. Parsons, J. Gratch, and A. A. Rizzo. Evaluation of Justina: A Virtual Patient with PTSD. In H. Prendinger, J. C. Lester, and M. Ishizuka, editors, Proceedings of 8th International Conference on Intelligent Virtual Agents, volume 5208 of Lecture Notes in Computer Science, pages 394-408, Tokyo, Japan, 2008. Springer.

[14] H. H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thorisson, H. van Welbergen, and R. J. van der Werf. The Behavior Markup Language: Recent developments and challenges. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, Proceedings of 7th International Conference on Intelligent Virtual Agents, volume 4722 of Lecture Notes in Computer Science, pages 99-111, Paris, France, 2007. Springer.

[15] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann. SmartBody: behavior realization for embodied conversational agents. In L. Padgham, D. C. Parkes, J.

Muller, and S. Parsons, editors, Proceedings of Conference on Autonomous Agents and Multi-Agent Systems (AAMAS08), pages 151-158. IFAAMAS, 2008.

[16] B. P. Arnason and A. Porsteinsson. The CADIA BML realizer. http://cadia.ru.is/projects/bmlr/.

[17] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjalmsson. Why conversational agents do what they do? Functional representations for generating conversational agent behavior. the First Functional Markup Language workshop, 2008. The Seventh International Conference on Autonomous Agents and Multiagent Systems Estoril, Portugal.

[18] J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, editors, Proceedings of 6th International Conference on Intelligent Virtual Agents, volume 4133 of Lecture Notes in Computer Science, pages 243-255, Marina Del Rey, CA, USA, 2006. Springer.

[19] Catherine Pelachaud, Studies on gesture expressivity for a virtual agent, Speech CommunicationIn Press, 17 May 2008.

[20] Merola, G. Emotional gestures in sport Language Resources and Evaluation, Springer, 2007, 41, 233-254.

[21] Rehm, M.; Vogt, T.; Wissner, M. & Bee, N. Dancing the night away: controlling a virtual karaoke dancer by multimodal expressive cues Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3, 2008, 1249-1252

[22] Mura, G. Wearable technologies for emotion communication, METU JFA, 2008, 1, 153

[23] Takala, T.; Ekman, I.; Poikola, A. & Mäkäräinen, M. Interactive emotional embodied experience Proceedings of EHTI'08: The First Finnish Symposium on Emotions and Human-Technology Interaction, , 33

[24] SaxEx: a case based reasoning system for generating expressive musical performances, *Josep Lluís Arcos, Ramon López de Mántaras, Xavier Serra,* 1998

[25] NOOS, *IIIA Spain*, (accessed 12/07/2007) http://www.iiia.csic.es/Projects/NOOS.html

[26] Grachten, M. (2001) JIG: Jazz Improvisation Generator. *In Proceedings of the MOSART Workshop on Current Research Directions in Computer Music, 1-6.* Barcelona, Spain: Pompeu Fabra University Publishers

[27] Composing Music with Case Based Reasoning (SICOM) Pereira, F. C. , Grilo, C. , Macedo, L. , Cardoso, A. http://www.cisuc.uc.pt/view_pub.php?id_p=67

[28] Pd – Pure Data. Available on WWW: http://puredata.info/

[29] Max/MSP – Cycling 74. Available on WWW: http://www.cycling74.com/

[30] Jae-woo Chung, G. Scott Vercoe: The affective remixer: personalized music arranging. CHI Extended Abstracts 2006: 393-398

[31] MPEG Group on SMR Symbolic Music Representation. Available on WWW: http://www.interactivemusicnetwork.org/mpeg-ahg/

[32] Zhou H., Hu H and Harris N., “Wearable inertial sensors for arm motion tracking in home-based rehabilitation” in IOS Press, 2005

[33] Foxlin E., Harrington M. and Altshuler Y, “Miniature 6-DOF inertial system for tracking HMDs”, in SPIE vol. 3362, Helmet and Head-Mounted Displays III, AeroSense 98, Orlando, FL, April 13-14, 1998