IDENTIFICATION AND SELECTION OF MODULES

Conveying Affectiveness in Leading-edge Living Adaptive Systems

CALLAS

Project IST-34800

Deliverable D111 WP1.1

Deliverable
Version 1.0 – 31 October 2007
Document. ref.: callas.D111.ICCS.WP1.1.V1.0

Programme Name:	IST
Project Number:	34800
Project Title:	CALLAS
Partners:	Coordinator: ENG (IT) Contractors: VTT, BBC, Metaware, 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.D111.ICCS.WP1.1.V1.0
Work-Package:	WP1.1
Deliverable Type:	Document
Contractual Date of Delivery:	31.10.2007
Actual Date of Delivery:	31.10.2007
Title of Document:	Identification and selection of modules
Author(s):	FPMS, HMW, ICCS, Paris 8, UOA and VTT

Approval of this report .....

Summary of this report:.....

History:.....

Keyword List: .....

Availability..... This report is: public

Table of Contents

EXECUTIVE SUMMARY 1

1. AUDIO ANALYSIS 2

1.1 EMOTIONAL SPEECH RECOGNITION 2
1.2 SOUND (AUDIO) ANALYSIS 3
1.3 EMOTION RECOGNITION FROM SPEECH 4
1.4 EMOTION RECOGNITION FROM LINGUISTIC FEATURES 5
1.4.1 Statistical Affect Sensing 5
1.4.2 Semantic Affect Sensing 5

2. VISUAL ANALYSIS 6

2.1 VIDEO FEATURES 6
2.2 FACIAL FEATURES DETECTION 6
2.3 GAZE/ POSE ESTIMATION 7
2.4 HAND DETECTION / TRACKING AND GESTURE EXPRESSIVITY FEATURES EXTRACTION 7

3. OTHER SENSORS 9

3.1 GESTURE RECOGNITION 9
3.2 MOTION CAPTURE 9

4. SYNTHESIS/ INTERACTION 14

4.1 EMOTIONAL ATTENTIVE ECA 14
4.2 EMOTIONAL NATURAL LANGUAGE GENERATION 16

5. REFERENCES 17

Executive Summary

WP 1.1 is responsible for establishing a process that will identify and select candidate technologies for inclusion in the CALLAS Shelf. Deliverable 111 describes the identification and selection of modules.

Section 1 tackles the problem of audio analysis. Audio analysis encompasses emotional speech recognition, sound analysis, emotion recognition from speech and emotion recognition from linguistic features.

Correspondingly, section 2 concerns visual analysis, which comprises video features, facial features detection, gaze / pose estimation and hand detection and tracking combined with gesture expressivity features extraction.

Section 3 describes other sensors used for gesture recognition and motion capture, while section 4 tackles the issue of synthesis and interaction and more precisely issues concerning the components for emotional attentive ECA and for emotional natural language generation.

The last section includes the references.

1. Audio Analysis

1.1 Emotional Speech Recognition

For several decades, a lot of efforts were put into the Speech Recognition research. In consequence, there are many available products in this field.

The classical architecture of a full Automatic Speech Recognition (ASR) system was described in D.1.2.1. In short, the initial sound wave is mapped to a series of “acoustic vectors” summarizing the time-frequency information of the signal. This sequence of acoustic vectors is used to “decode” the speech, which is typically performed through the combination of:

- Hidden Markov Models (HMMs), to model the time evolution of speech waves,
- a classifier (most often a Multi-Layer Perceptron or a Gaussian Mixture Model) to estimate the acoustic probabilities of the basic units (typically phonemes),
- and a grammar to constrain the search of possible spoken utterances to contextually realistic ones.

It is important to note that some parts of the ASR (HMMs, classifiers) must be trained in order to efficiently model the properties of the speech. To optimize the device, the training phase must be done with utterances similar to those that will be faced by the recognizer in its utilization phase.

Each block of the architecture is important and was, as itself, the topic of a lot of studies. Among the open-source products, we can mention the following ones:

- HTK, a world-known toolkit for research in ASR developed by the Cambridge University Speech Department. HTK provides very efficient tools for training and using HMMs, which lead this software to be used in fields outside its initial purpose, like research in speech synthesis, character recognition or synthesis of movements.
- Julius, a large-vocabulary continuous recognition system, initially built for Japanese. It was reported to work well with other languages such as English, French or Slovenian and they are currently developing open-source acoustic models for English. This system was developed by the Kawahara Lab Of Kyoto University, the Shikano Lab Of Nara Institute of Science and Technology and the Julius project team of Nagoya Institute of Technology.
- SPHINX, designed at Carnegie Mellon University (CMU), also provides high-level components to perform the full training and decoding, enabling the users to build ASR systems for their own recognition applications. Users can also use default models trained by the CMU group.
- The ISIP Production System, developed by the Institute for Signal and Information Processing (ISIP) of Mississippi State University, which has already been used in many different recognition contexts.

Besides these open-source products, there are also companies selling full speech-to-text software. The most famous are dictation systems like Nuance’s Dragon NaturallySpeaking, IBM’s ViaVoice or Ultimate Interactive Desktops Inc.’s Voice Studio. These dictation devices achieve very good recognition rates after being trained by the user. Acapela Group is another company marketing speech recognition products. Among others, it builds ASR products based on EAR and STRUT, conjointly developed by Multitel and FPMs (see D1.2.1). Apart from large dictation systems, these companies also develop specific recognition devices for

various fields like Health, Law or Education.

ASR systems are typically trained with neutral speech and their performance will thus decrease when speech is influenced by emotions, Lombard effect, etc.

Moreover, we would like to cite some crucial problems encountered in real-life automatic speech recognition:

- speaker adaptation (ideally the system should perform equally with any speaker, but performance is increased when the speaker was involved in training the system)
- vocabulary size (the recognition task is easier when the vocabulary that can be used by the speaker is limited)
- continuous speech processing (with hesitations, coughs, laughers, influence of emotions,… as opposed to isolated word recognition)
- robustness to noise (recognition is harder in a noisy real-life environment than in a quiet lab)

No current ASR system is able to deal correctly with all the adverse conditions that can be encountered in real-life. ASR systems are trained under certain conditions and will perform well if used in the same conditions. Depending on the application, different trade-offs will be chosen: for example, one application might request speaker-independent speech recognition (i.e. the user does not need to train the system) and accept to restrict the vocabulary, while another application, like a dictation system, needs a large vocabulary and therefore typically asks for a training phase by the user to work optimally.

All the above mentioned products are well-known products but they are integrated systems and not components that can be integrated into the CALLAS framework.

Concerning the specific criteria of selection procedure, we can distinguish three types of them:

- Performance of the speech recognition: accuracy of the ASR device, conditions under which the system performs well (see above).
- Computational efficiency of the recognition: how fast is the recognition, what CPU load does it need?
- Format of the recognizer: is the system linked to a specific programming language and/or OS, or is it easy to integrate it in any application?

For the targeted applications, we needed a fast, speaker-independent ASR device, but a small vocabulary was sufficient (keyword-spotting application). When EAR is used with a rather small vocabulary, it fulfils these requirements. Furthermore, EAR is a device we know well and that can be ported to the most common OS and programming languages, which is an asset for future integration into the CALLAS framework and showcases.

For these reasons, we selected EAR as first ASR component. Further precisions about EAR can be found in D1.2.1.

1.2 Sound (Audio) analysis

Audio analysis has traditionally been a research topic on multimedia content domain, where the interest is to automatically annotate audio visual material. In recent years audio analysis has also started to interest context awareness and security researchers.

Typically audio content is classified first into a basic set of main audio categories. Generally these classes contain speech, music, silence and additionally different noise classes or mixed classes, i.e. speech with music. Many of the recent research on audio content analysis and segmentation concentrate on material from news archives, digital libraries and TV

programs/movies [1], [2], [3]. Analysing the amateur created video material with mobile phones has also increasing interest among the researchers, [4], [5]. Such a data has a new challenge since the lack of professional structure and quality of the data.

In field of general audio analysis there is known cap between low level features. Recent EU projects like SIMAC(www.semanticaudio.org) and AUDIOCLAS (http://audioclas.iua.upf.edu/) has put the effort for extracting high-level human readable and understandable metadata.

The audio analysis is also utilised in context aware applications by analysing environmental audio. Korpipää & al. used multiple sensors for context awareness, but with plain audio they reached accuracy 87.6% of correct positive recognition. They used large set of audio features and classified them with HMM's [6]. HMM based environmental audio analysis in [7] reached overall accuracy of 92% for 11 acoustic environments. Also surveillance applications have growing interest in audio analysis for detecting suspicious events as in [8].

Some of the showcases are interested in general audio analysis, particularly on recognising human made sounds like laughter and clapping in different environments. These environments consist home like environment, public space or theatre and they might give their own requirements for the component from the different sound environments. For these reasons the component has to have some flexibility and possible reconfiguration properties behalf of the developer. So, a component developed by a partner of CALLAS consortium, meeting the requirements of integration to both framework and showcases and having the capability of readjustment will be the most efficient selection for CALLAS purposes.

There is no available well known audio analysis software that can be used as component into the CALLAS framework, except VTT's component. Further precisions about it can be found in D1.2.1.

1.3 Emotion recognition from speech

As the recognition of emotions from speech is a relatively new area of research there are not yet any commercial products available. Research so far has mainly been concerned with the offline analysis of speech for emotional clues and has only recently shifted to applications. In the call centre domain, these include the jerk-o-meter, that monitors attention (activity and stress) in a phone conversation, based on speech feature analysis, and gives the user feedback allowing her to change her manners if deemed appropriate [9] and the emotion-aware voice portal currently under development at T-Systems [10]. For the AT&T "How May I Help You?" spoken dialogue system it has been investigated how the user's emotional state affects the accuracy of the system [11].

Recently, methods for the recognition of emotions from speech have also been explored within the context of computer-enhanced learning. For instance, Ai and colleagues [12] consider features extracted from the dialogue between the tutor and the student, such as the prosody of speech, as well as features relating to user and system performance for the emotion recognition process in the ITSpoke tutoring system.

Emotion recognition from speech in cars has so far been investigated e.g. in the FERMUS project, a cooperation with the automobile industry (DaimlerChrysler and BMW) [13], and in the Emotive Driver project [14]. Experiments have, however, been restricted to data collection and evaluation of driving simulators scenarios.

Hegel et al. [15] have enhanced a humanoid robot with the capability to mirror a conversation partner's emotion as conveyed from the voice in the face of the robot, thus creating an empathic robot.

All these systems are unavailable for public use and least of all open source, and thus, except for the humanoid robot project which was built with UOA's emotion recognition component, unsuitable for use within the project. Furthermore, though all have been evaluated in the context of applications, except for [9], [15] it is not possible or remains unclear whether they are suitable for real-time processing.

A first requirement to the emotion recognition from speech component for the showcases is to be able to give feedback to the user in real-time. Furthermore, not only a classical set of emotions should be available as target classes but also other affect related categories or non-Ekmanian emotions, and the recogniser should be adaptable to different scenarios. Therefore, a configurable recognition component is needed that can be trained with audio data similar to the respective application scenario and that allows to use arbitrary classes. Especially the need of adaption to different scenarios renders the UOAs component the only one to be integrated into CALLAS framework. A more detailed description about it can be found in D1.2.1.

1.4 Emotion recognition from linguistic features

To our knowledge there are no commercial products for lexical affect sensing available. However, there is the ConceptNet application from a MIT research project ([16]) that provides an emotional recognition. The system analyzes a text on the base of English sentences from the Open Mind Common Sense corpus (OMCS), and builds a structured semantic net. Calculation of the emotional meaning of a text is done by propagating emotional meaning of analyzed text from the affective seed words. Emotional meaning is a six-element vector containing affective estimates for Ekman emotions *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise* ([17]). Evidently, the approach conceals the following drawback – it calculates a several-component vector that estimates the emotional meaning of a text and not the final scalar estimation. There is no tip on how this meaning can be used for calculating the final estimate and, hence, it is unsuitable for an end-back application.

In the following, two approaches to emotional text analysis are described: the statistical and the semantic. The statistical approach makes a scalar decision on the emotional meaning of the text by using standard data mining techniques as SVM without scrutinizing the semantics of text words. In contrast, the semantic approach analyzes semantically the words of the text and decides about its emotional meaning using a rule-based framework.

1.4.1 Statistical Affect Sensing

To our knowledge there are no commercial products for statistical affect sensing available. However, research has been realised regarding this issue ([18], [19], [20], [21]).

The required component should meet a number of requirements in different categories. Concerning language, it should cover every language, the text could be spontaneous, grammatically incorrect, with repairs, repetitions and inexact words. The text length could vary, e.g. the text could be a single word, while the result would be 9 classes in the movie review scenario (from zero to four stars in half-star step).

The selected component is described in D121.

1.4.2 Semantic Affect Sensing

To our knowledge there are no commercial products for semantic affect sensing available. However, relevant research is described in [22], [23], while [24], [25], [26], [27] and [28] describe the affect dictionaries that can be used in semantic affect sensing - Whissell's Dictionary of Affect Language (DAL), LIWC2001 Dictionary, WordNet-Affect Database and General Inquirer (GI) respectively.

The length of the text to be analyzed can be a single sentence for the sake of simplicity. It should be a grammatically correct text while the result of the component will consist of five classes (high/low arousal, positive/negative valence, neutral).

Since there are no relevant commercial products available, the selected component is developed by a partner of CALLAS, it is easily integrated into framework and it is described in D121.

2. Visual Analysis

2.1 Video Features

The video analysis is vast field and it is utilised in many different purposes computer vision, medical imaging, multimedia content analysis, HCI applications etc. Due to the broad field of applications and many topics that video analysis holds with in the research activities has lead to many different open source developments, like MPT (Machine Perception Toolbox - library that offers among other functionalities face detection for real time video [29]), CamTrack Face Tracker (face tracking system for user interface control and development. This can be utilised to determine face position and orientation in real time using a webcam [30]), Mimas Toolkit (real time C++ machine vision platform that includes variety of algorithm i.e. for object tracking and recognising methods [31]). We should also mention OpenCV, Intel's open source project on computer vision [32]. It is well known and extensively employed in different research and prototyping projects of computer vision applications on a single computer, but, being a completed project, cannot be integrated into CALLAS framework.

For CALLAS purposes, the first requirement for the video features is to perform in real time; the interest lies in getting cues from group behaviour which is obtained through face detection and tracking, as well as object movement and general movement on the camera field. OpenCV open library is targeted for real time computer vision and includes a variety of different algorithms that can be utilised a new way to achieve information from the group behaviour. So, although we cannot use OpenCV as a whole, the selected component is developed by a partner of CALLAS consortium and based on OpenCV open source library which has a rich set of video/image content analysis algorithms. The OpenCV algorithms are combined so that they will provide the information needed for creating component for Showcase purposes. The requirements for live input and easy to use will be met as well, while the responsible partner will keep adapting the current component in order to meet newer requirements. More details regarding the component can be found in D121.

2.2 Facial features detection

There have been published a great deal of techniques that search for facial characteristics at predefined scales and almost known positions of the face. Such techniques are dedicated to specific applications, like drivers attention recognition [39]. Furthermore, a very big amount of methods existing in bibliography make usage of color information for facial characteristics to be found [39], or take advantage of the fact that certain characteristics have certain shape. A typical characteristic of such methods are those that first detect circular characteristics and match them with iris regions [40]. A big disadvantage of such methods is that the placement of the face needs to be quite close to the camera for such shapes to be clear enough.

The facial feature detection component chosen for the CALLAS applications was the one offered by the ICCS. It is based on an idea that was initiated two years ago and has continuously been under research ever since then. The different states of the work through time have been published at international conferences [33], [34], [35], while a latest version is submitted for publication at an international journal. Extensive experimentations have proved that the algorithm outperforms a great amount of the latest algorithms published in the recent past [36], [37], [38].

The chosen method has also been compared with a software developed in ICCS for facial feature detection [41]. The reasons of preference of the chosen technique is that it was proved to process videos at a faster frame rate, due to the former's attribute to make usage of Neural Networks. Also, the work in [41] was written in Matlab and its transfer to C would

necessitate a longer period to have it available for showcase and framework testings.

For the above reasons, the component developed in ICCS has been considered as the most appropriate for the needs of an integrated project, where various showcases might require different environments of testing. That is, the technique provided does not suffer from limitations such as specific skin color requirements, very large resolution cameras or known dimensions of the face. Furthermore, an advantage of the technique chosen is that it is real-time - constituting itself appropriate for live applications – and, of course, easily modified by the developers – since they are members of CALLAS consortium – adapting to different lighting conditions and different showcases or scenarios. For a detailed description of the component, refer to D121.

2.3 Gaze/ pose estimation

In recent bibliography, most gaze detection and pose determination techniques need special hardware setup. Examples of such cases are the work described in [42], where a large resolution image of the iris is necessary and the work in [43], where a specific architecture has to be followed. In other cases, intrusive devices have to be worn by the user [44], making the system less appropriate for wide-range applications. In most systems available in the market, special hardware and equipment is also necessary. Further difficulties that have risen in the choice of specific software was their restrictions regarding operating environments (e.g. availability only for Linux [45], [46]), need for infrared cameras, or availability in non suitable programming languages (e.g. Matlab [46]). The system documented in [47] gives good real-time performance at the expense of using, two cameras. In case one camera is needed, knowledge of its internal parameters is needed or calibration is needed every time it is used with a new camera.

There is a huge variety of products, especially in the field of gaze detection. However, their cost is disproportional to our needs and, in total, they require extra equipment that burdens the application, or could constitute it impractical.

The gaze/pose determination 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 - since they are both developed by the same partner –, thus, making alterations, usage and experimentations more flexible.

In the current work, features are detected and tracked, allowing for a relative freedom of the user, under good lighting conditions. Under these circumstances, the gaze and pose directionality can be approximately determined, which is enough for attention recognition purposes, as well as for general decisions regarding one's gaze. The technique used does not need multi-camera systems or special calibration procedures and the try is towards the aim of developing a system for tracking a user's visual attention without the need of special hardware restrictions. A more detailed description can be found in D121.

2.4 Hand detection / tracking and gesture expressivity features extraction

There are several approaches in the literature concerning hand modelling. On the other hand very few studies have tackled with gesture expressivity analysis, mainly psychological ones.

Hand modelling is roughly divided into two categories vision based and motion capturing techniques. Vision based approaches are very well reviewed on by Wu and Huang on [48]. All these methods attempt to construct hand models' kinematical structure. These models could be cardboard, wireframe or polygon-mesh models. To capture articulate hand motion in full DOF, both global hand motion and local finger motion should be determined from video sequences. It is a challenging problem to analyze and capture hand motion, because the hand is highly articulate. Different methods have been taken to approach this problem. One

possible method is the appearance-based approaches, in which 2-D deformable hand shape templates are used to track a moving hand in 2-D. However, this method is insufficient to recover full articulations, because it is difficult to infer finger joint angles based on appearances only. Another possible way is the 3-D model-based approach, which takes the advantages of *a priori* knowledge built in the 3-D models. This approach aligns a 3-D model to images or even range data by estimating the parameters of the model. In 3-D model-based methods, image features could be looked as the image evidence or image observation of a 3-D model that is projected to the image plane. A 3-D model with different parameters will produce different image evidence. Model-based methods recover the joint angles by minimizing the discrepancy between the image feature observations and projected 3-D model hypotheses which is a challenging optimization problem.

Although, various methods reviewed by Wu [48] seem promising and accurate we have rejected them mainly for three reasons: 1.None could be implemented by a real-time or quasi-real-time robust application for a showcase scenario within the project, 2. The detail level of the set of hand features extracted from the majority of these algorithms is much more farfetched than the requirements of the gesture expressivity feature extraction module, 3. A large number of these methods, especially the ones using motion capture techniques, could be intrusive to the user and thus adding noise to the actual underlying expressivity

Concerning the model of expressivity features we have reviewed two approaches before concluding to the Hartmann et al. model [49].

The first one is the EyesWeb Expressive Gesture Processing Library. Castellano et al. [51] extracted a set of five expressivity features, quite similar to the Hartmann model, namely quantity of motion and contraction index of the body, velocity, acceleration and fluidity of the hand's barycentre, using the EyesWeb Expressive Gesture Processing Library. Although the system works quite well it has a major disadvantage which makes it unusable for the project. The image processing library being used is mostly based on background subtraction to extract the human silhouette, thus, for unknown settings and environments the whole process collapses.

The second one is Wallbott's study [50]. He reported an attempt to demonstrate that body movements and postures to some degree are specific for certain emotions. Although the study itself is interesting it provides no quantitative measurements of gesture expressivity making it impossible to implement as is in an application.

Taking into account all the above-mentioned constraints and the, as easy as possible, integration into CALLAS framework, we chose ICCS component that combines actually both sub modules, namely head and hand detection and tracking and gesture expressivity parameters extraction. In related literature very few works combine similar modules and there aren't any relevant commercial products. A more detailed description of the component can be found in D121.

3. Other sensors

3.1 Gesture recognition

The gesture recognition research has mainly concentrated on systems based on video analysis. Using wireless sensing device with accelerometers for gesture recognition is less researched but still there exists few ready applications based on that, like the wireless game controller Wii Remote, introduced by the game console manufacturer Nintendo [52]. Other commercial product that utilises the motion sensing is Gyration's remote controls for Media Centre R4000 LCD Music Remote and PC with M2000 Travel Air-Mouse [53].

The above mentioned solutions will not offer API for accessing the motion data, and for that the utilising them is not feasible.

There is also SoapBox (Sensing, Operating and Activating Peripheral Box), which is a sensor device developed by VTT, for research activities in ubiquitous computing, context awareness, multi-modal and remote user interfaces, and low power radio protocols [54]. In recent research it has also been utilised in gesture recognition [55].

For wireless gesture recognition there are not many available sensor solutions. Nokia model 5500 has embedded accelerometers with in [56]. They are utilising the movement information for sports applications, but offering API for third parties as well. Other phone manufacturers are embedding accelerometers: NTT DoCoMo FOMA 904i Series [57] utilises them for gaming and Samsung SPH-S4000 [58] health and leisure purposes. The Nokia model 5500 was chosen to be the starting point for following reasons: it is moderate price compared to the soap box and easily available for all partners as well as fully designed product, not a prototype. As familiar equipment, mobile phone, it is also easily approached and accepted by the users.

3.2 Motion capture

Regarding motion and gesture detection, we can list three main categories:

Inside-in technologies, in which both the transducers and eventually the source of the field to be measured lie in the device (sensing suits or exoskeletons with Hall effect based sensors, potentiometers, magnetoresistors, optical fibers or even sensing tissues)

Inside-out technologies, in which the transducers are on board the sensing device but they sense the magnetic or gravitational field of the Earth or a generated external (magnetic) field

Outside-in technologies, in which the sensors are not on the links or on the joints but located in the surrounding environment. In some cases these technologies make use of active or reflective markers.

According to the previous classification, what follows is a short list of some of the Inside-In motion tracking devices available on the market.

Among the inside-in technologies we can list

- the suits Gypsy 5 and Gypsy 6 by Animazoo UK Ltd, exoskeletons using potentiometers and gyroscopes.

The manufacturer's website does not provide accuracy or bandwidth. The price is around 60'000€.

- the glove Cyberglove II by Immersion Corp., in which resistive sensors detect bends.

  Number of sensors: up to 22

  Sensor resolution: 0.5 degrees

  Sensor repeatability: 1 degree

  Sensor linearity: 0.6%

  Sensor data rate: 90 records/sec

  Price: 14'000€

- The glove 5DT Dataglove distributed by VR Logic gmbh. In this glove optical fibres detect bends (only one flexion for each finger).

  Number of sensors: 5 or 15

  Sensor resolution: N.A. (12 bit AD)

  Sensor repeatability: N.A.

  Sensor linearity: N.A.

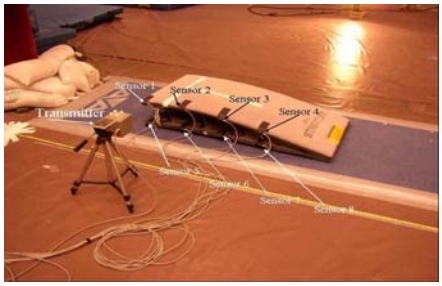
  Sensor data rate: 75Hz

  Price: 400€

Among the inside-out technologies we can list

- the suit Moven by Xsens Technologies B.V., using inertial sensor on the different links the manufacturer's website does not provide accuracy or bandwidth. The price is around 30'000€

- the magnetic system Fastrack by Polhemus, a device that uses a magnetic field produced by a stationery transmitter to determine the real-time position of moving receiver elements. It uses the attenuation of oriented electromagnetic signals to determine the absolute position and orientation of a tracker relative to a source. The accuracy is about 0.15deg and the update rate is 120 Hz. The price is around 22'000€



Among the outside-in technologies we can list

- the MX40 system by Vicon System Motion Ltd, an optical system performing real-time tracking of a subject (wearing reflective markers) from multiple synchronized infrared video cameras. The accuracy is about 0.5mm; the maximum frame rate (at max sensor resolution) is 160 Hz. The price is 200'000€

- the 3d motion analysis system Zebris by Zebris Medical GmbH, using ultrasound emitter and markers. The method for analysing movement is based on the principle of the travel time measurement of ultrasound pulses. The method is based on a kinematical analysis of 3-dimensional position data of markers attached to specific body parts (e.g. wrist, elbow) selected for investigation. The price is around 9'000€.

In the Callas project, motion trackers can be used for puppeteering, for character animation (in videogames, movies, cartoons or ECA) or for emotion extraction from gestures.

The selection criteria can change according to the specific use: a puppeteer can easily accept to wear gloves or a suit to move the virtual marionettes and these interfaces can be more natural than traditional ones (mouse, keyboard, etc.). Motion capture of actors wearing a suit is not new in fantasy or action movies and great results have been achieved. Nevertheless, people will not be keen of wearing a glove to extract emotions from their gestures while watching TV or visiting a museum.

So we can distinguish the case of a performer using a specific interface from the case of people performing daily activities which would need non intrusive devices.

CALLAS requirements involve portability, cost reductions, small set-up time and non structured environments, then the wearable devices using inside-in technologies are almost mandatory. Inside out technologies with external sources not-artificially generated (i.e. inertial sensors) are also suitable.

Unfortunately, the above mentioned systems currently on the market have several drawbacks. The Gypsy suit is definitely bulky and needs too much time to be donned and tuned, and it is rather expensive because there are not many competitors in the market. Similarly the Cyberglove is also expensive because Immersion system has no real

competitors. For example, the 5Dt Dataglove is very poor in performance. The Moven suit suffers from coordinates drift problems and all the other problems related to accelerometers.

Moreover, CALLAS needs components and not completed systems, in order to integrate them into the framework, while the necessary modifications will be easily realised if the developer of the component is a member of CALLAS consortium.

For these reasons we planned to develop our own devices, counting on Humanware's know-how, being already experienced in motion capture using Hall effect based sensors, gloves and suits as soft exoskeletons. Furthermore, usually the suits on the market do not include a multi-dof glove device and the data should be collected from different components and re-formatted. With an ad hoc developed modular device (including the required dof), the data will be collected as requested (in terms of number dof and accuracy) and formatted complying with the standard set by the Callas consortium. Obviously we will pursue the optimum choice taking into account performance and costs.

The first release of the developed components is extensively described in D1.2.1.

4. Synthesis/ Interaction

4.1 Emotional Attentive ECA

Here, we consider some of the technological progress making it possible to build an autonomous embodied conversational agent (ECA). An ECA is a computer-generated animated character that is able to carry out a natural, human-like dialogue with human users. For this purpose, agents use *multi-modal* communication.

Many of agent systems have been developed by research institutions. An architecture proposed by K.R. Thórisson [93] is an example of an agent that offers multimodal bidirectional communication. In the same laboratory, a humanoid called *Rea* was created [67]. In comparison with other communicative skills, the nonverbal behaviour of Rea is somewhat simplified, for example regarding facial expression [67]. Another example of an agent that enables multimodal bidirectional communication is *Max*, which was created at the University of Bielefeld [81]. The emotional states of Max are mapped into the set of Ekman's basic expressions [60], [62]. An agent called *Virtual Human* created in the Geneva MiraLab laboratory is a three-dimensional humanoid full-body character equipped with a personality model. It is able to conduct a conversation and to express emotional states to a human user [72]. Another embodied conversational agent, Baldi, is a three-dimensional character [84], [85]. In this architecture the main emphasis was put on the correct synchronisation of speech and lip movement. The agent called *Artificial Person* (*Finnish Talking Head*) was created by the researchers of Helsinki University of Technology [74] and is a virtual three-dimensional face animated in real-time. The model of facial expressions is based on Ekman's FACS. This embodied agent uses six of Ekman's basic expressions and their blends [79]. Embodied conversational agents have also been created in the Centre for Speech Technology at the Royal Technical University of Stockholm. These agents have been applied in many different applications: The agents like Urban [75] were used in dialogue systems in order to fulfil tasks such as assisting users to find an apartment, or assisting hearing impaired persons.

Commercial embodied conversational agent systems are still rarely offered on the market. One example of a commercially available agent is Alex, offered by Lexicle [82]. Other companies such as Haptek [76] and Cantoche [68], do not offer full agent systems but rather offer tools to create and animate virtual characters to be placed on the web or inside programs as ActiveX controls. These animated characters usually offer limited possibilities to model facial expressions and/or other nonverbal behaviours, and the realism of the characters is limited due to their intended application. More realistic behaviour can be obtained using the tools offered by the Visage SDK [96], since the animated characters created with this technology follow MPEG-4 standard of animation. Finally, it is also possible to build simple animated agents using Verbots software [94].

A number of tools are also available for easing the creation and animation of virtual characters. In particular, and of interest here, a number of face editing and animation tools are available. *Xface* is an open-source project developed at the ITC-irst [59]. First of all, it is a tool for creation of three-dimensional ECAs that are compliant with the MPEG-4 standard. *HeadEd* is an interactive tool being developed in University of Paris 8 [89] for the purpose of animating the head and facial area of 3D agents. The tool is compliant with the MPEG-4 animation standard, supporting a direct import / export capability for the Greta system [88], support for a variety of other auxiliary formats allowing easy export to real-time applications and also embedding its own proprietary format. *AnimEdit*, *FaceEditor* and *EmotionDisc* [86], [92] are a suite of interactive face animation and modelling tools for 2D characters. *Visage|Interactive* [96] is a commercial facial animation toolkit for real-time animation based on a VRML face model. It accompanies the *Visage|SDK*. The *Facial Animation Engine* [73] is commercial software for the creation of face animations driven by MPEG-4 specified

animation parameters. *Digital Puppets* [77] allow for the specification of body configurations and gestures for 3D characters. These are based on a range of possible communicative contexts and the output is composed of animated gesture coordinated with synthetic speech.

All of the presented embodied conversational agents are able to communicate with the user using a variety of verbal and nonverbal methods, and a number of them have the ability to pay attention to their surroundings. While some of them focus on the multi-modality of interaction, others put the emphasis on the realism and believability of agent. For the purpose of CALLAS, we require an agent that is able to generate multi-modal nonverbal behaviour such as facial expressions, gestures, and head movements etc. In order to influence users, we must also be assured that the agent correctly communicates its emotional states. The agent should be able to display various emotional facial expressions that are accompanied by gestures, head movements and other appropriate non-verbal behaviours. Moreover, it should be able to express a rich set of emotional states using both verbal and nonverbal channels - in particular an agent capable of displaying emotional states like *tension* or *relief*. It should also be modular to allow the easily addition of new facial expressions and gestures that can be requested in particular scenarios. Finally, the agent should be able to show interest and attention towards its surroundings, and more specifically, should be seen to be attentive by the user to them and to the environment.

The architecture should offer a detailed animation parameterisation as we intend to use various emotional states that need to be interpreted unambiguously by users. Among other factors, this means that we need a model of the face that is characterised by a high number of parameters, as certain facial expressions can be distinguished by only minute details. The architecture of our agent also needs to be modular; it should be possible to integrate it with other modules and external programs to be integrated into CALLAS framework. It is expected that standard protocols and procedures can be used to streamline communication and integration. Last but not least, we need an agent that uses powerful and standardised languages to describe verbal and nonverbal behaviours.

The embodied conversational agent called *Greta* ([88], [66]) fulfils all the aforementioned conditions and is developed by a CALLAS partner, so it can be adapted to CALLAS requirements. It is therefore our component of choice. Greta is a three-dimensional animated agent that implements MPEG-4 animation standard [87] and is able to communicate with the user by means of both verbal and nonverbal content. This standard furnishes a detailed parameterisation of the face. The MPEG-4 parameterisation allows one to work at a high level when dealing with behaviour generation: one does not need to engage in the low-level graphics generation; at the same time, even small details of any facial expression can be modelled. As a consequence, we expect that various facial expressions generated with Greta can be recognised by users. Greta can talk to the user and display facial expressions, gestures, gazes, and head movements. Greta uses nonverbal cues during dialog, for example, to accentuate the verbal content, to disclose some cognitive processes or to signal internal emotional state [90]. In particular, it has a rich repertoire of facial expressions of emotions and gestures that are defined according to the psychological literature.

Greta uses powerful languages to describe her behaviour. The *Affective Presentation Markup Language* (APML) [71] is a high-level language that can be used to specify communicative functions of the behaviour. The second, *Behavior Markup Language* (BML) [95], is a low-level language that is used to define Greta's behaviour at the signal level. Thus one can choose which concrete verbal or nonverbal behaviour should be displayed by Greta.

In the latest version of Greta, the animation can be alternatively described using *Behavior Markup Language* (BML). It is XML based standardised representation language that can be used for specifying verbal and nonverbal behaviours for embodied virtual agents. The elementary unit of BML is the *signal*: that is any behaviour produced by the agent like a head movement, a facial expression, a gesture, and so on. The main difference between APML and BML is that the latter allows, for each communicative function, to define explicitly its duration.

Greta's architecture is highly modular: modules can be simply exchanged or modified. At the

moment it uses a number of external software systems, for example a speech synthesiser. The new skills of our agent can be defined in separate modules and integrated easily into CALLAS framework. For example, Greta visual attention capabilities will be integrated as a specialised module. Greta is based on Psyclone [91], it allows for easy and standardised communication of different modules.

4.2 Emotional Natural Language Generation

To our knowledge, there are no commercial Emotional Natural Language Generators available. Much research has been conducted in the field of Emotional / Affective Natural Language Generation (see, for example [102] for an overview). However, within this field, little work has been done on emotional variations in the generated utterances. The main contributions in this area include (cf. [98]): Hovy ([100]), where both content selection and realisation are based on various factors, the attitude of the speaker and the desired emotion of the hearer among them, Fleischman and Hovy ([99]), who consider the speaker's attitude and only deal with content realisation, not selection (however, undesired parts can be left out), Walker et al. ([104]), where utterances of an agent are synthesized using acoustic features of the desired emotion. Other important contributions are Loyall and Bates ([101]), in which, based on the speaker's current emotional state, different ways of content realisations are chosen and de Rosis and Grasso ([103]), where certain parts of content selection (some information is intentionally left out) and realisation (aggregation) are based on properties of the hearer, the emotional state being on of them.

Being part of the Callas Shelf, a component for Emotional Natural Language Generation has to meet the following requirements:

- Efficiency: Output of the component has to be generated in real time
- Extensibility: Extending the component, e.g. with new conversational topics, should be achieved easily
- Portability: Porting the component to another domain, language or set of emotions should be achieved easily

Since there are no commercial components available and none of the few implementations of the related work described above fulfil the Shelf requirements, we implemented our own component, EmoNLG. EmoNLG uses a corpus-based approach to generate utterances from semantic representations and as such fulfils the requirements addressed above. Being developed by a CALLAS partner, it is modifiable and adaptable to CALLAS requirements. Emotional variation is achieved by using ideas from ([99]). See Deliverable 1.3.1 for the specification and detailed description of EmoNLG.

5. References

[1] L. Lu, H.-J. Zhang and H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans on Speech and Audio Processing*, vol. 10, no 7, pp 504 – 516, Oct. 2002.

[2] C.-H. Wu, C.-H. Hsieh, "Multiple Change-Point Audio Segmentation and Classification Using an MDL-Based Gaussian Model", *IEEE Transactions on Speech and Audio Processing,* Accepted for future publication, vol PP, is 99, pp.1 – 11, 2005.

[3] C.-H. Lin, S. –H. Chen, T.-K. Truong, Y. Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," *IEEE Transactions on Speech and Audio Processing,* vol. 13, is. 5, Part 1, pp. :644 – 651, Sept. 2005

[4] Mäkelä S.-M., Peltola J., Myllyniemi M.,"Mobile Video Capture Targeted Narrowband Audio Content Classification", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006, May 15-19, 2006, Toulouse, France

[5] Vuorinen O., Peltola J., Mäkelä S.-M."Unsupervised Speaker Change Detection for Mobile Device Recorded Speech", IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, April 15-20, Honolulu, Hawaii, USA

[6] P. Korpipää, M. Koskinen, J. Peltola, S-M Mäkelä and T. Seppänen, "Bayesian Approach to Sensor-Based Context Awareness", Pers Ubiquit Comput, 2003, 7:113 - 124.

[7] Ling Ma, Ben Milner and Dan Smith,"Acoustic Environment Classification", ACM Transactions on Speech and Language Processing, Volume 3 , Issue 2 (July 2006), pp 1-22, 2006

[8] Radhakrishnan, R.; Divakaran, A.; Smaragdis, A.; "Audio analysis for surveillance applications", Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on 16-19 Oct. 2005 Page(s):158 - 161

[9] Madan, A.: Jerk-O-Meter: Speech-Feature Analysis Provides Feedback on Your Phone Interactions. http://www.media.mit.edu/press/jerk-o-meter/ (2005) retrieved: 11.10.2007.

[10] Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R.: An emotion-aware voice portal. In: Electronic Speech Signal Processing Conference, Prague, Czech Republic (2005)

[11] Riccardi, G., Hakkani-Tür, D.: Grounding emotions in human-machine conversational systems. In: Proceedings of Intelligent Technologies for Interactive Entertainment, INTETAIN, Madonna di Campiglio, Italy (2005)

[12] Ai, H., Litman, D.J., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A.: Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proceedings of Interspeech 2006 — ICSLP, Pittsburgh, PA, USA (2006)

[13] Schuller, B., Rigoll, G., Grimm, M., Kroschel, K., Moosmayr, T., Ruske, G.: Effects of in-car noise-conditions on the recognition of emotion within speech. In: Proc. of the DAGA 2007, Stuttgart, Germany (2007)

[14] Jones, C., Jonsson, I.: Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In: Proceedings of the 19th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: citizens online: considerations for today and the future, Canberra, Australia (2005)

[15] Hegel, F., Spexard, T., Vogt, T., Horstmann, G., Wrede, B.: Playing a different imitation

game: Interaction with an empathic android robot. In: Proc. 2006 IEEERAS International Conference on Humanoid Robots (Humanoids06) (2006).

[16] Liu, H., Lieberman, H., Selker, T. 2003. A Model of Textual Affect Sensing Using Real-World Knowledge. Pages 125-132 of: Proceedings of IUI-03, the 8th international conference on intelligent user interfaces. Miami, US: ACM Press.

[17] Ekman, P. 1993. Facial expression of emotion. American Psychologist, 48, 384-392.

[18] Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Pages 417-424 of: Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, US: Association for Computational Linguistics.

[19] Pang, B., Lee, L. 2004. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. Pages 271-278 of: Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics. Barcelona, ES: Association for Computational Linguistics.

[20] Riloff, E, Patwardhan, S., Wiebe, J. 2006. Feature Subsumption for Opinion Analysis. Pages 440-448 of: Proceedings of EMNLP-06, the Conference on Empirical Methods in Natural Language Processing. Sydney, AUS: Association for Computational Linguistics.

[21] Whitelaw, C., Garg, N., Argamon, S. 2005. Using Appraisal Taxonomies for Sentiment Analysis. In: Proceedings of MCLC-05, the 2nd Midwest Computational Linguistic Colloquium.

[22] Wiebe, J., Riloff, E. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. Pages 475-486 of: Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science, vol. 3406. Mexico City, MX: Springer-Verlag.

[23] Yi, J., Nasukawa, T., Bunescu, R:, Niblack, W. 2003. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. Pages 427- 434 of: Proceeding of ICDM-03, the 3ird IEEE International Conference on Data Mining. Melbourne, US: IEEE Computer Society.

[24] Whissell, C.M. 1989. The dictionary of affect in language. In Robert Plutchik and Henry Kellerman (Ed.), *Emotion: Theory, Research, and Experience* (pp. 113–131). New York: Academic Press.

[25] Pennebaker, J.W., Francis, M.E., Booth, R.J. 2001. Linguistic Inquiry and Word Count (LIWC): LIWC2001. Mahwah, NJ: Erlbaum Publishers.

[26] Valitutti, A., Strapparava, C., Stock, O. 2004. Developing Affective Lexical Resources. PsychNology Journal. Volume 2, Number 1, 61-83.

[27] Stone, P. J., Dunphy, D. C., Smith, M.S., Ogilvie, D. M. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press.

[28] The plain list of emotional words in [www.eqi.org].

[29] http://sourceforge.net/projects/mptbox/

[30] http://face-tracking.qarchive.org/

[31] http://www.shu.ac.uk/mmvl/research/mimas/

[32] http://www.intel.com/technology/computing/opencv/

[33] S. Asteriadis, N. Nikolaidis, A. Hajdu, I.Pitas, An Eye Detection Algorithm using Pixel to Edge information, Proceedings of the 2nd IEEE-EURASIP International Symposium on Control, Communications, and Signal Processing (Marrakech, March 2006)

[34] S. Asteriadis, N. Nikolaidis, A. Hajdu, I. Pitas, A novel eye detection algorithm utilizing edge-related geometrical information, EUSIPCO 06 (European Signal Processing

Conference, Florence, Italy, September 06)

[35] S. Asteriadis, N. Nikolaidis, I. Pitas, M. Pardas, Detection of facial characteristics based on edge information, VISAPP 2007 (International Conference on Computer Vision Theory and Applications), Barcelona, Spain, 2007

[36] Z.H. Zhou and X. Geng, Projection functions for eye detection, Pattern Recognition 37, no 5, pp. 1049-1056, 2004

[37] Jesorsky, K. J. Kirchberg, and R. W. Frischholz, Robust face detection using the hausdorff distance, 3rd International Conference on Audio and Video-based Biometric Person Authentication, Halmstad, Sweden, pp. 90-95, 2001

[38] Cristinacce, T. Cootes, and I. Scott, A multi-stage approach to facial feature detection, Proc. Of BMVC, pp. 231-240, London 2004

[39] P. Smith, M Shah, N. da Vitoria Lobo, Determining Driver Visual Attention with One Camera, IEEE Trans. on Intelligent Transportation Systems, Vol 4, No. 4, pp. 205-218, 2003.

[40] T. D"Orazio, M. Leo, G. Cicirelli, and A. Distante, An algorithm for real time eye detection in face images, 17th International Conference on Pattern Recognition(ICPR"04), Vol. 3, pp. 278-281, 2004.

[41] S. Ioannou, G. Caridakis, S. Kollias, K. Karpouzis, Robust Feature Detection for Facial Expression Recognition, EURASIP Journal On Image and Video Processing (to be published), 2007.

[42] J.G. Wang, E. S. and Venkateswarku, R., "Eye gaze estimation from a single image of one eye", 9th IEEE International Conference on Computer Vision, 2003

[43] SensoMotoric Instruments. EyeLink Gaze Tracking. www.smi.de

[44] Beymer, D. and Flickner, M., "Eye gaze tracking using an active stereo head", Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 451–458, 2003

[45] http://www.inference.phy.cam.ac.uk/opengazer/

[46] http://thirtysixthspan.com/openEyes/software.html

[47] http://groups.csail.mit.edu/vision/vip/watson/

[48] Ying Wu; Huang, T.S., "Hand modeling, analysis and recognition," Signal Processing Magazine, IEEE , vol.18, no.3, pp.51-60, May 2001

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

[50] Wallbott, H. G. (1998). Bodily expression of emotion. European Journal of Social Psychology, 28, 879-896

[51] G. Castellano, S. D. Villalba, A. Camurri Recognising Human Emotions from Body Movement and Gesture Dynamics. Proc. Intl. Conf. ACII 2007, Lisbon.

[52] http://wii.nintendo.com/controller.jsp

[53] http://www.gyration.com/

[54] Tuulari E, Ylisaukko-oja A (2002) SoapBox: A Platform for Ubiquitous Computing Research and Applications. First International Conference, Pervasive 2002, pp 26-28, 2002

[55] Kallio S, Kela J, Korpipää P, Mäntyjärvi J,. User independent Gesture interaction for small handheld devices. International Journal of Pattern Recognition and Artificial Intelligence. Vol.20 (2006) No: 4, 505 – 524.

[56] http://www.nokia.co.uk/link?cid=PLAIN_TEXT_52722

[57] http://www.nttdocomo.com/pr/2007/001335.html

[58] http://news.softpedia.com/news/Samsung-Releases-S4000-13495.shtml

[59] Balci, K., Xface: Open Source Toolkit for Creating 3D Faces of an Embodied Conversational Agent, Smart Graphics 2005, pp. 263-266, 2005.

[60] Becker, C., Kopp, S., Wachsmuth, I., Simulating the Emotion Dynamics of a Multimodal Conversational Agent, In:

[61] André, E., Dybkjær, L., Minker, W., Heisterkamp, P., (eds.), Affective Dialogue Systems, Springer Verlag, pp. 154-165, 2004.

[62] Becker, C., Prendinger, H., Ishizuka, M., Wachsmuth, I., Evaluating Affective Feedback of the 3D Agent Max in a Competitive Cards Game, First International Conference on Affective Computing & Intelligent Interaction, Beijing, China, pp. 466-473, 2005.

[63] Beskow, J., Cerrato, L., Granström, B., House, D., Nordenberg, M., Nordstrand, M., Svanfeldt, G., Expressive Animated Agents for Affective Dialogue Systems. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P., (eds.), Affective Dialogue Systems, Springer Verlag, pp. 301-304, 2004.

[64] Beskow, J., Elenius, K., McGlashan, S., The OLGA Project: An Animated Talking Agent in a Dialogue System. Proceedings of Eurospeech - 5th European Conference on Speech Communication and Technology, Rhodes, Greece, pp. 1651-1654, 1997.

[65] Beskow, J., Cerrato, L., Granström, B., House, D., Nordenberg, M., Nordstrand, M., Svanfeldt, G., Expressive Animated Agents for Affective Dialogue Systems. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P., (eds.), Affective Dialogue Systems, Springer Verlag, pp. 301-304, 2004.

[66] Bevacqua, E., Mancini, M., Niewiadomski, R., Pelachaud, C., An Expressive ECA showing complex emotions, in: Proceedings of AISB'07: Artificial and Ambient Intelligence, Newcastle University, Newcastle upon Tyne, UK, April 2nd-4th 2007.

[67] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjlmsson, H., Yan, H., Embodiment in conversational interfaces: Rea., In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 520-527, 1999.

[68] Cantoche, www.cantoche.com

[69] Cassell, J., Embodied Conversation: Integrating Face and Gesture Into Automatic Spoken Dialogue Systems, In: Luperfoy (ed.), Spoken Dialogue Systems, Cambridge, MA: MIT Press, 1999.

[70] Cassell, J., Bickmore, T., Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents, 13 (1-2), pp. 89-132, 2003.

[71] De Carolis, B., Pelachaud, C., Poggi, I., Steedman, M., APML, a Mark-up Language for Believable Behavior Generation, In: Prendinger, H., Ishizuka, M. (eds.), Lifelike Characters. Tools, Affective Functions and Applications, Springer, 2004.

[72] Egges, A, Kshirsagar, S., Magnenat-Thalmann, N., Imparting Individuality to Virtual Humans, First International Workshop on Virtual Reality Rehabilitation (Mental Health, Neurological, Physical, Vocational), Lausanne, Switzerland, pp. 201-108, 2002.

[73] Eptamedia, www.eptamedia.com

[74] Frydrych, M., Kätsyri, J., Dobsík, M., Sams, M., Toolkit for Animation of Finnish Talking Head. ISCA Tutorial and Research Workshop on Audio Visual Speech Processing (AVSP'03), St Jorioz, France, pp. 199-204, 2003.

[75] Granström, B., House, D., Multimodality and speech technology: Verbal and nonverbal communication in talking agents. Proceedings of EuroSpeech 2003, Geneva,

Switzerland, pp. 2901-2904, 2003.

[76] Haptek, www.haptek.com

[77] ISI 2007, E. Shaw, Digital Puppets, www.isi.edu/isd/carte/

[78] B. Fogel, J. C. Bezdek Ed.), Vol. 5200, pp. 64-78, Bellingham, WA:SPIE Press, Aug 2003.

[79] Kätsyri, J., Klucharev, V., Frydrych, M., Sams, M., Identification of synthetic and natural emotional facial expressions, ISCA Tutorial and Research Workshop on Audio Visual Speech Processing (AVSP'03), St. Jorioz, France, pp. 239-244, 2003.

[80] Kim, Y., Hill R.W., and Traum, D.R. A computational model of dynamic perceptual attention for virtual humans. 14th Conference on Behavior Representation in Modeling and Simulation (BRIMS), 2005.

[81] Kopp, S., Jung, B., Leßmann, N., Wachsmuth, I., Max - A Multimodal Assistant in Virtual Reality Construction, KI-Künstliche Intelligenz, 4, pp. 17-23, 2003.

[82] Lexicle, www.lexicle.com

[83] Marsella, S.C., Lewis Johnson, W., LaBore, C., Interactive Pedagogical Drama. Proceedings of the 4th International Conference on Autonomous Agents, pp. 301-308, 2000.

[84] Massaro, D.W., Liu, Y., Chen, T.H., Perfetti, C.A., A Multilingual Embodied Conversational Agent for Tutoring Speech and Language Learning. Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP, September, Pittsburgh, PA, pp. 825-828, 2006.

[85] Massaro, D.W., Cohen, M.M., Beskow, J., Daniel, S., Cole, R.A., Developing and Evaluating Conversational Agents. First Workshop on Embodied Conversational Characters (WECC'98). Lake Tahoe, CA, 1998.

[86] Noot, H. and Ruttkay, Z. Gesture in Style. Gesture Workshop 2003: pp. 324-337, 2003.

[87] Ostermann, J., Face Animation in MPEG-4. In: Pandzic, I.S., Forchheimer, R., (eds.), MPEG-4 Facial Animation - The Standard Implementation and Applications, Wiley, England, pp. 17-55, 2002.

[88] Pelachaud, C., Bilvi, M., Computational Model of Believable Conversational Agents. In: Huget, M.-P. (ed.), Communications in Multiagent Systems, Springer-Verlag, Lecture Notes in Computer Science, vol. 2650, pp. 300-317, 2003.

[89] Peters, C., Presentation: HeadEd Tool, CALLAS Meeting, Paris, September 2007.

[90] Poggi, I., Pelachaud, C., De Rosis, F., Eye communication in a conversational 3D synthetic agent, AI Communications, 13 (3), pp. 169 - 181, 2000.

[91] Psyclone, www.cmlabs.com/psyclone/

[92] Ruttkay, Z., Noot, H. and ten Hagen, P. Emotion Disc and Emotion Squares: Tools to Explore the Facial Expression Space. Computer Graphics Forum 22(1): 49-54, 2003.

[93] Thórisson, K.R., Communicative Humanoids A Computational Model of Psychosocial Dialogue Skills, Ph.D thesis, Thesis, Massachusetts Institute of Technology, 1996.

[94] Verbots, www.verbots.com

[95] Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thorisson, K.R., van Welbergen, H., van der Werf, R., The Behavior Markup Language: Recent Developments and Challenges. In: Proceedings of the 7th International Conference on Intelligent Virtual Agents, Paris, France, 2007.

[96] Visage, www.visagetechnologies.com

[97] Wilhelmsson, K., Talking Heads and Hearing Impaired Persons, www.speech.kth.se/~rolf/gslt_papers/KennethWilhelmsson.pdf

[98] Belz, A.: *And Now with Feeling: Developments in Emotional Language Generation*, Technical Report, University of Brighton, 2003

[99] Fleischman, M. and Hovy, E.: *Towards emotional variation in speech-based natural language generation*, In *Proceedings of the Second International Natural Language Generation Conference (INLG02)*, 2002

[100] Hovy, E. H., *Pragmatics and natural language generation*, 1990

[101] Loyall, A. B. and Bates, J.: *Personality-rich believable agents that use language*, In *Proceedings of the first International Conference on Autonomous Agents*, 1997

[102] Piwek, P.: *An Annotated Bibliography of Affective Natural Language Generation*, Technical Report, University of Brighton, 2003

[103] de Rosis, F. and Grasso, F.: *Affective natural language generation*, In Paiva, A. M.(editor): *Affective Interactions*, Springer, 2000

[104] Walker, M. A., Cahn, J. E., and Whittaker, S. J.: *Improvising linguistic style: Social and affective bases of agent personality*, In *Proceedings of the First International Conference on Autonomous Agents*, 1997