

# Affective Embodied Conversational Agents for Natural Interaction

Eva Cerezo, Sandra Baldassarri, Isabelle Hupont and Francisco J. Seron  
*Advanced Computer Graphics Group (GIGA)*  
*Computer Science Department, Engineering Research Institute of Aragon(I3A),*  
*University of Zaragoza,*  
*Spain*

## 1. Introduction

Human computer intelligent interaction is an emerging field aimed at providing natural ways for humans to use computers as aids. It is argued that for a computer to be able to interact with humans it needs to have the communication skills of humans. One of these skills is the affective aspect of communication, which is recognized to be a crucial part of human intelligence and has been argued to be more fundamental in human behaviour and success in social life than intellect (Vesterinen, 2001; Pantic, 2005).

Embodied conversational agents, ECAs (Casell et al., 2000), are graphical interfaces capable of using verbal and non-verbal modes of communication to interact with users in computer-based environments. These agents are sometimes just as an animated talking face, may be displaying simple facial expressions and, when using speech synthesis, with some kind of lip synchronization, and sometimes they have sophisticated 3D graphical representation, with complex body movements and facial expressions.

An important strand of emotion-related research in human-computer interaction is the simulation of emotional expressions made by embodied computer agents (Creed & Beale, 2005). The basic requirement for a computer to express emotions is to have channels of communication such as voice, image and an ability to communicate affection over those channels. Therefore, interface designers often emulate multimodal human-human communication by including emotional expressions and statements in their interfaces through the use of textual content, speech (synthetic and recorded) and synthetic facial expressions, making the agents truly "social actors" (Reeves & Nass, 1996). Several studies have illustrated that our ability to recognise the emotional facial expressions of embodied computer agents is very similar to that of identifying human facial expressions (Bartneck, 2001). Related to agent's voice, experiments have demonstrated that subjects can recognize the emotional expressions of an agent (Creed & Beale, 2006) whose voice varies widely in pitch, tempo and loudness and its facial expressions match the emotion it is expressing.

But, what about the impact of these social actors? Recent research focuses on the psychological impact of affective agents endowed with the ability to behave empathically with the user (Brave et al., 2005; Isbister, 2006; Yee et al., 2007; Prendinger & Ishizuka, 2004; Picard, 2003). The findings demonstrate that bringing about empathic agents is important in

human-computer interaction. Moreover, addressing user's emotions in human-computer interaction significantly enhances the believability and lifelikeness of virtual humans (Boukricha et al., 2007). This is why the development of computer-based interfaces capable of understanding the emotional state of the user has been a subject of great interest in recent affective computing researches. Nevertheless, to date, there are no examples of agents that can sense in a completely automatic and natural (both verbal and non-verbal) way human emotion, and respond realistically. In fact, few works related to agent-based human-like affective interaction can be found in literature. In some of them, the user communicates with an affective agent through a dialogue based on multiple choice test, without any kind of non-immersive automated emotional feedback from the user to computer (Prendinger & Ishizuka, 2005; Anolli et al., 2005). In other works, the interaction is enriched by a speech recognition and generation system that allows a minimum instructional conversation with the agent (Elliott et al., 1997) or by an automatic emotion recognizer that transmits the user's emotion to the agent which reacts accordingly (Burleson et al., 2004). In spite of the difficulties and limitations, this type of social interface has been demonstrated to enrich human-computer interaction in a wide variety of applications, including interactive presentations (Seron et al., 2006), tutoring (Elliott et al., 1997), e-learning (Anolli et al., 2005) and health-care (Prendinger & Ishizuka, 2004), and user support in frustrating situations (Prendinger & Ishizuka, 2004; Klein et al., 2002).

Our research focuses on developing interactive virtual agents that support multimodal and emotional interaction. Emotional interaction enables us to establish more effective communication with the user and multimodality broadens the number of potential users by making interaction with disabled users (for example hearing-impaired or paraplegics) and people of different ages and with different levels of education (people with or without a knowledge of computers) possible. The result of our efforts has been Maxine, a powerful engine to manage real-time interaction with virtual characters. The consideration of emotional aspects has been a key factor in the development of our system. Special emphasis has been done in capturing the user's emotion through images and in synthesizing the emotion of the virtual agent through its facial expressions and the modulation of the voice. These two aspects will, therefore constitute the core of the chapter.

The chapter is organized as follows. In Section 2, Maxine, the platform for managing virtual agents is briefly described. Section 3 details the system developed for capturing user's emotion whereas Section 4 is devoted to discuss the natural language communication between the user and the agent. In section 5 evaluations of Maxine agents are commented and, finally, in Section 6, the conclusions are presented and current and future work are outlined.

## **2. A platform for managing affective embodied conversational agents**

### **2.1 Overall description**

Maxine is a script-directed engine for the management and visualization of 3D virtual worlds. In Maxine it is possible to load real-time models, animations, textures and sounds. Even though it is a very generic engine, it has been oriented to the work with virtual characters in 3D scenarios. It has been written in C++ and employs a set of open source libraries.

The modules that conform Maxine are: the Sensory/Perception Modules, that process the inputs of the system, the Generative/Deliberative Modules, in charge of managing the

appropriated reactions according to the inputs, and the Motor Module, that generates and coordinates the final outputs of the system.

While reasoning based on a user's directly input behaviours is important and useful, it is also limited. Therefore, an endeavour is also made to collect the largest possible amount of information on the user by means of body language or facial expression, without requiring him or her to enter data. The ultimate aim is to enhance interaction and establish emotional communication between the user and the virtual character as well as providing the user with different communication modalities. The available inputs are:

- **Voice Interaction.** The user can communicate with the character through voice, formulating an order, a question or any sentence in natural language. One of the requisites of our system was that it should be able to "understand" and speak Spanish. This constraint prevented us from using existing libraries, all of them in English. Details are given in section 4.
- **Image Interaction:** a webcam takes pictures of the user's face. The aim of these pictures is to obtain additional information on the user and, in particular, on his or her emotional state. This kind of input is detailed in section 3.
- **Console (keyboard)/mouse commands:** advanced users can fully control the scene and the agent thanks to the scripting language used, LUA (Lua, 2008). For non-programmer users, it is also possible to associate the execution of a command to the pressing of a certain key or clicking the mouse and, due to the power of the scripting language used, options are very varied.

Regarding the agents' reactions, two kinds of actions are distinguished:

- **Purely reactive:** for example, if the user keys in something, the virtual agent interrupts his/her speech, if a lot of background noise is detected, it requests silence, etc. These reactions are managed in the generative module.
- **Deliberative:** the choice of the reaction of the virtual character calls for more complex analysis. This analysis is done in the deliberative module, which, for example, is in charge of obtaining an answer from the user when interaction is made via voice, as it will be explained later on.

These reactions generally produce outputs, basically facial and body animations and speech with appropriated lip-synchronization.

Detailed description of Maxine is given elsewhere (Baldassarri et al., 2007b).

## 2.2 Maxine virtual agents

The basic aim of the system has been to make it easier to developers the inclusion of these agents in their applications. Therefore, default models, rigged and textured, with basic animations, visemes and expressions are prepared to be loaded in the system. Advanced users' can create their own characters by using commercial software, as appropriated converters have been developed.

In Maxine, the virtual agent is endowed with the following differentiating features:

- it supports interaction with the user through different channels: text, voice (through natural language), peripherals (mouse, keyboard), which makes the use of the generated applications available to a wide range of users, in terms of communication ability, age, etc.
- it gathers additional information on the user and the environment: noise level in the room, image-based estimate of the user's emotional state, etc.
- it has its own emotional state, which may vary depending on the relationship with the user and which modulates the agent's facial expressions, answers and voice.

To study the potential and usefulness of Maxine agents different applications have been developed. In particular:

- Virtual presenters for PowerPoint like presentations: a like-life character presents PowerPoint information on a graphic display. This kind of presenter has demonstrated to be specially useful when the same presentation has to be repeated several times or given in a different language (for example in English by a non-fluent English speaker). The application, MaxinePPT (Seron et al., 2006), is capable of creating and performing a virtual presentation in a 3D virtual scenario enriched with virtual actors and additional information such as videos, images, etc. from a classical PowerPoint file. All the aspects of the virtual presentation are controlled by an XML-type language called PML (Presentation Markup Language). The PML instructions are added to the page notes of the PowerPoint slides in order to determine, for example, the text to be spoken by the virtual presenter. Once the presentation has been created, user intervention is not necessary. Figure 1 shows some screenshots of a virtual presentation.

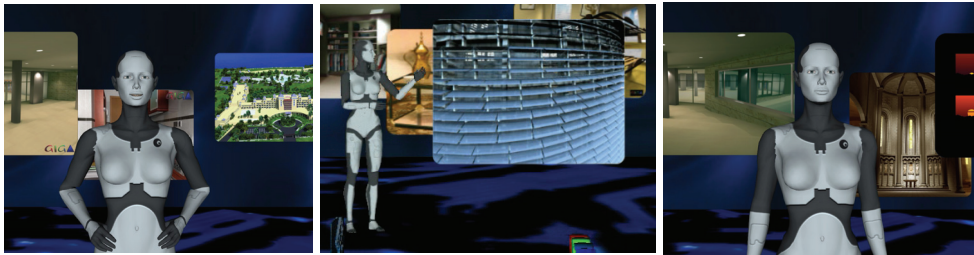


Fig. 1. Some screenshots from the presentation performed by a virtual agent

- Virtual assistants for controlling a domotics environment: a virtual agent, called Max, was created and used as an interactive interface (see Figure 2) for the access and remote control of an intelligent room (Cerezo et al., 2007). The user can ask Max to do different tasks within the domotics environment (to turn on/off the lamps, the tv, the electric kettle, etc.), and, also, may do queries about the different devices of the intelligent room (the state of the door, for example). Presently, adaptation of the agent interface to other devices such as mobile phones and PDAs is being performed.



Fig. 2. User interacting with the domotics environment through Max, the virtual agent

- Virtual Interactive pedagogical agents for teaching Computer Graphics: Maxine has also been used for the development of a learning platform to simplify and improve teaching and practice of Computer Graphics subjects (Seron et al., 2007) in the Computer Science degree. The interactive pedagogical agent helps students in two ways: acting as a virtual teacher to expose some specific topics, and allowing the interaction and handle of a 3D scene to make it easier to understand difficult topics of CG subjects (see Figure 3). Results are promising as it will be discussed in section 5.

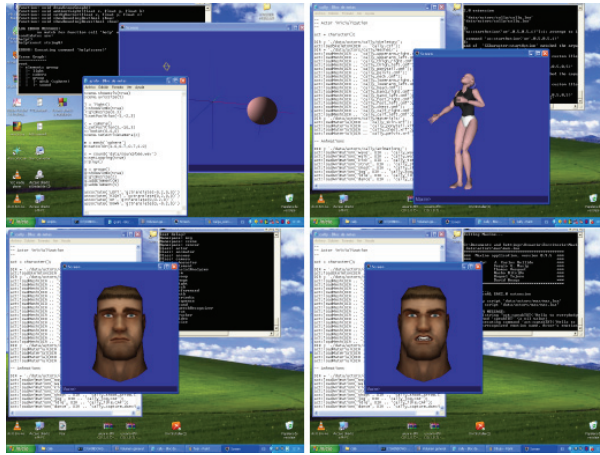


Fig. 3. "Playing" with the Maxine engine: managing a 3D scene (above left), loading and blending animations (above right) and changing characters' expressions (below)

### 3. Capturing user's emotions

As pointed out, the recognition of emotional information is a key step toward giving computers the ability to interact more naturally and intelligently with people. Nevertheless, to develop a system that interprets facial expressions is not easy. Two kinds of problems have to be solved: facial expression feature extraction and facial expression classification. Our work focuses on the second problem: classification. This implies the definition of the set of categories and the implementation of the categorization mechanisms.

Facial expression analyzers make use of three different methods of classification: patterns, neuronal networks or rules. If a pattern-based method is used (Edwards et al., 1998; Hong et al., 1998; Lyons et al., 1999), the face expression found is compared with the patterns defined for each expression category. The best matching decides the classification of the expression. Most of these methods first apply PCA and LDA algorithms to reduce dimensionality. In the systems based on neuronal networks (Zhang et al., 1998; Wallace et al., 2004), the face expression is classified according to a categorization process "learned" by the neuronal network during the training phase. In general, the input to this type of systems is a set of characteristics extracted from the face (points or distances between points). The rule-based methods (Pantic & Rothkrantz, 2000a) classify the face expression into basic categories of emotions, according to a set of face actions previously codified. In (Pantic & Rothkrantz, 2000b) an excellent state-of-the-art on the subject can be found.

In any case, the development of automatic facial classification systems presents several problems. Most of the studies on automated expression analysis perform an emotional classification based on the emotional classification of Ekman (Ekman, 1999). It describes six universal basic emotions: joy, sadness, surprise, fear, disgust and anger. Nevertheless, the use of Ekman's categories for developing automating facial expression emotional classification is difficult. First, his description of the six prototypic facial expressions of emotions is linguistic and, thus, ambiguous. There is no uniquely defined description either in terms of facial actions or in terms of some other universally defined facial codes. Second, classification of facial expressions into multiple emotion categories should be possible (e.g. raised eyebrows and smiling mouth is a blend of surprise and happiness). Another important issue to be considered is individualization. The system should be capable of analyzing any subject, male or female of any age and ethnicity and of any expressivity, which represents a really challenging task.

### 3.1 Image-based interaction process

The stages of the image-interaction process are shown in Figure 4. In the following paragraphs, each of the three stages is explained.

#### Stage 1: Feature extraction

A webcam takes pictures of the user and the tracking of some points enables to extract relevant facial features.

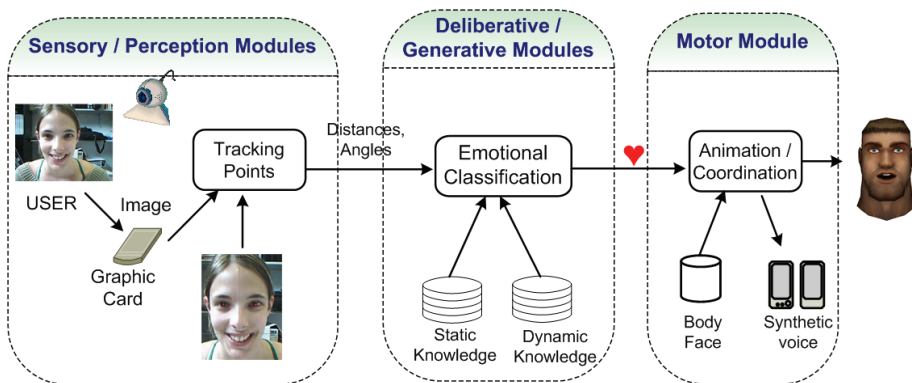


Fig. 4. Stages of the user-avatar image interaction process

#### Feature selection

The first step of the method consists of extracting the 20 feature points of the face that will later allow us to analyze the evolution of the face parameters (distances and angles) that we wish to study. Figure 5 shows the correspondence of these points with the ones defined by the MPEG-4 standard (MPEG-4, 2002). The characteristic points are used to calculate the five distances shown in Figure 6. These five distances can be translated in terms of MPEG-4 standard, putting them in relation to the feature points shown in Figure 5 and with some FAPs defined by the norm. All the distances are normalized with respect to the distance between the eyes (MPEG FAPU "ESo"), which is a distance independent of the expression. This way, the values will be consistent, independently of the scale of the image, the distance to the camera, etc.

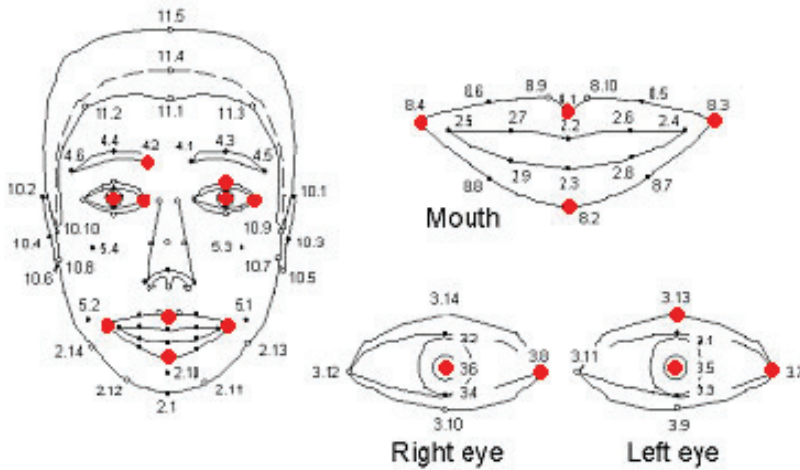
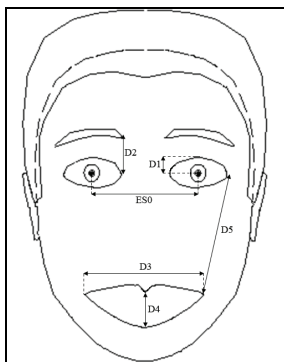


Fig. 5. Facial feature points used for the later definition of the parameters to analyze, according to MPEG-4 standard.



MPEG-4 FAPs NAME	FEATURE POINTS USED FOR DISTANCES
close_upper_l_eyelid close_lower_l_eyelid	$D1=d(3.5, 3.1)$
raise_r_i_eyebrow	$D2=d(4.2, 3.8)$
stretch_l_cornerlip stretch_r_cornerlip	$D3=d(8.4, 8.3)$
open_jaw	$D4=d(8.1, 8.2)$
raise_r_cornerlip	$D5=d(8.3, 3.7)$

Fig. 6. Characteristic distances used in our method (left). On the right, relationship between the five characteristic distances and the MPEG-4 FAPs and feature points.

**Tracking**

The emotional classifier was first developed and tuned based on the tracking of features on static images. Thanks to a collaboration, the Computer Graphics, Vision and Artificial Intelligence Group of the University of the Balearic Islands provided us with a real-time facial tracking module to test our classifier. The features extraction system is non-invasive and is based on the use of a simple low cost webcam (Manresa et al., 2006). The parameters corresponding to the neutral face are obtained calculating the average of the first frames of

the video sequence, in which the user is supposed to be in the neutral state. For the rest of the frames, a classification takes place following the method explained in the next sections.

The automatic features extraction program allows the introduction of dynamic information in the classification system, making it possible the study of the time evolution of the evaluated parameters, and the classification of user's emotions from live video.

Psychological investigations argue that the timing of the facial expressions is a critical factor in the interpretation of expressions. In order to give temporary consistency to the system, a temporary window that contains the emotion detected by the system in each one of the 9 previous frames is created. A variation in the emotional state of the user is detected if in this window the same emotion is repeated at least 6 times and is different from the detected in the last emotional change.

#### Stage 2: Emotional classification

From the extracted facial features, emotional classification is performed in stage 2.

The core of our work has been, in fact, the development of the emotional classifier. It is based on the work of Hammal et al. (Hammal et al., 2005). They have implemented a facial classification method for static images. The originality of their work consists, on the one hand, in the supposition that all the necessary information for the recognition of expressions is contained in the deformation of certain characteristics of the eyes, mouth and eyebrows and, on the other hand, in the use of the Belief Theory to make the classification. Nevertheless, their method has important restrictions. The most important restriction comes from the fact that it is only able to discern 3 of the 6 basic emotions (without including the neutral one). This is basically due to the little information they handle (only 5 distances). It would not be viable, from a probabilistic point of view, to work with many more data, because the explosion of possible combinations would remarkably increase the computational cost of the algorithm.

Our method studies the variation of a certain number of face parameters (distances and angles between some feature points of the face) with respect to the neutral expression. The objective of our method is to assign a score to each emotion, according to the state acquired by each one of the parameters in the image. The emotion (or emotions in case of draw) chosen will be the one that obtains a greater score. For example, let's imagine that we study two face parameters ( $P_1$  and  $P_2$ ) and that each one of them can take three different states ( $C^+$ ,  $C^-$  and  $S$ , following the nomenclature of Hammal). State  $C^+$  means that the value of the parameters has increased with respect to the neutral one; state  $C^-$  that its value has diminished with respect to the neutral one; and the state  $S$  that its value has not varied. First, we build a descriptive table of emotions, according to the state of the parameters, like the one of the Table 1. From this table, a set of logical tables can be built for each parameter (Table 2). That way, two vectors of emotions are defined, according to the state taken by each one of the parameters ( $C^+$ ,  $C^-$  or  $S$ ) in a specific frame. Once the tables are defined, the implementation of the identification algorithm is simple. When a parameter takes a specific state, it is enough to select the vector of emotions (formed by 1's and 0's) corresponding to this state. If we repeat the procedure for each parameter, we will obtain a matrix of as many rows as parameters we study and 7 columns, corresponding to the 7 emotions. The sum of 1's present in each column of the matrix gives the score obtained by each emotion.



	P1	P2
Joy	C-	S/C-
Surprise	C+	C+
Disgust	C-	C-
Anger	C+	C-
Sadness	C-	C+
Fear	S/C+	S/C+
Neutral	S	S

Table 1. Theoretical table of parameters' states for each emotion (example with only two parameters).

Compared to the method of Hammal, ours is computationally simple. The combinatory explosion and the number of calculations to be made have been considerably reduced, allowing us to work with more information (more parameters) of the face and to evaluate the seven universal emotions, and not only four of them, as Hammal does.

		E1 joy	E2 surprise	E3 disgust	E4 anger	E5 sadness	E6 fear	E7 neutral
P1	C+	0	1	0	1	0	1	0
	C-	1	0	1	0	1	0	0
	S	0	0	0	0	0	1	1
		E1 joy	E2 surprise	E3 disgust	E4 anger	E5 sadness	E6 fear	E7 neutral
P2	C+	0	1	0	0	1	1	0
	C-	1	0	1	1	0	0	0
	S	1	0	0	0	0	1	1

Table 2. Logical rules table for each parameter.

### Stage 3: Animating facial expressions

The information about the emotional state of the user can be used to adapt the emotional state of the agent and consequently to modify its facial animations, which are generated and coordinated in this stage.

The technique used for facial animation is the skeletal one and the nomenclature followed is that of the VHML standard (VHML, 2001). Each agent has got the corresponding animation of the 6 Ekman emotions (see Figure 7).



Fig. 7. Neutral face plus Ekman facial emotions: happiness, sadness, anger, fear, surprise and disgust.

### 3.2 Emotional classification: describing emotions

#### Databases

In order to define the emotions in terms of the parameters states, as well as to find the thresholds that determine if parameter is in a state or another (as explained in last section), it is necessary to work with a wide database. In this work we have used two diferent facial emotions databases: the FG-NET database (FG-NET, 2006) that provides video sequences of 19 different caucasian people; and the MMI Facial Expression Database (Pantic et al., 2005) that holds 1280 videos of 43 different subjects from different races (caucasian, asian and arabic). Both databases show the 7 universal emotions of Ekman (Figure 8).

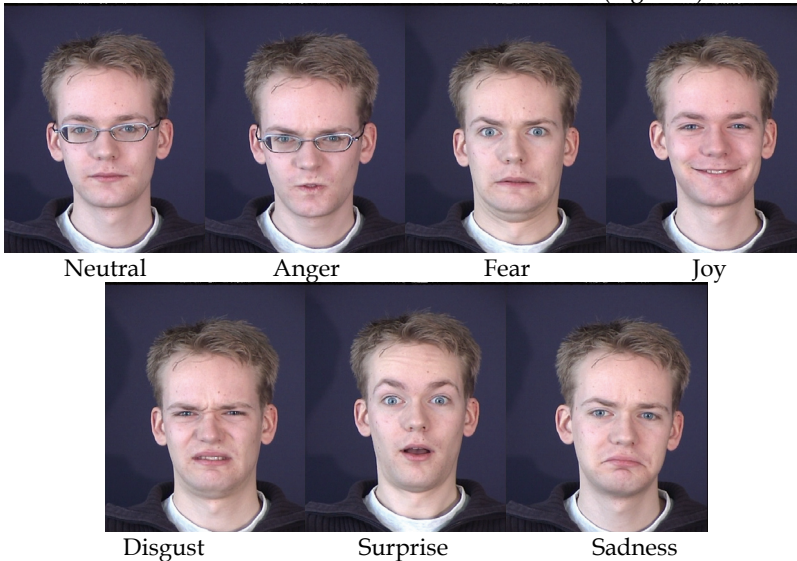


Fig. 8. Example of selected frames of the MMI Facial Expression

#### Emotions' definitions and thresholds

In order to build a descriptive table of each emotion in terms of states of distances, we must determine for each distance the value of the states that define each emotion (C+, C- or S), as well as evaluate the thresholds that separate a state from another. To do this, we studied the

variation of each distance with respect to the neutral one, for each person of the database and for each emotion. An example of the results obtained for distance D4 is shown in Figure 9. From these data, we can make a descriptive table of the emotions according to the value of the states (Table 3).

	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>	Wrinkles	Ang 1	Ang 2	W/H
Joy	C-	S/C-	C+	C+	C-	No	C+	S/C+/C-	S/C-
Surprise	S/C+	S/C+	S/C-	C+	S/C+	No	C-	C+	C-
Disgust	C-	C-	S/C+/C-	S/C+	S/C-	Yes	S/C+/C-	S/C+	S/C-
Anger	C-	C-	S/C-	S/C-	S/C+/C-	Yes	C+	C-	C+
Sadness	C-	S	S/C-	S	S/C+	No	S/C+/C-	S/C-	S/C+
Fear	S/C+	S/C+/C-	C-	C+	S/C+	No	C-	C+	C-
Neutral	S	S	S	S	S	No	S	S	S

Table 3. Theoretical table of the states taken by the different studied characteristics for each emotion. The distances (D1,..D5) are those shown in Figure 6. Some features do not provide any information of interest for certain emotions (squares in gray) and in these cases they are not considered. Note also that the distances D1, D2 and D5 have a symmetric facial distance (one in each eye). Facial symmetry has been assumed after having calculated the high correlation between each distance and its symmetric.

One step necessary for our method to work is to define the values of the thresholds that separate a state of another one, for each studied distance. Two types of thresholds exist: the upper threshold (marks the limit between neutral state S and state C+) and the lower threshold (the one that marks the limit between neutral state S and state C-). The thresholds' values are determined by means of several tests and statistics on all the subjects and all the expressions of the databases. Figure 9 shows an example of thresholds estimation for the distance D4.

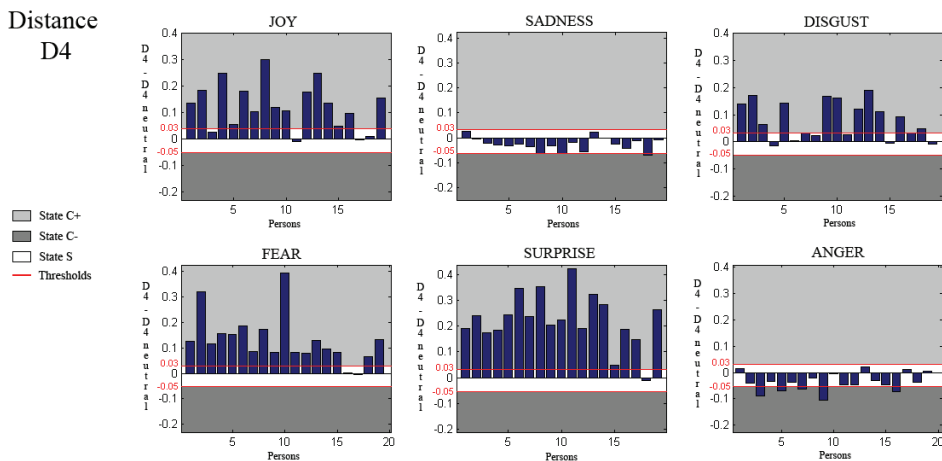


Fig. 9. Statistics results obtained for distance D4. Thresholds estimations are also shown.

### 3.3 Results

#### Classification rates

The algorithm has been proved on the images of the databases. In the evaluation of results, the recognition is marked as "good" if the decision is coherent with the one taken by a human being. To do this, we have made surveys to 30 different people to classify the expressions shown in the most ambiguous images. For example, in the image shown in Figure 10, the surveyed people recognized it as much "disgust" as "anger", although the FG-NET database classifies it like "disgust" exclusively. Our method obtains a draw.

First we considered to work with the same parameters as the Hammal method, ie, with the 5 characteristic distances shown in Figure 6. The obtained results are shown in the third column Table 4. As it can be observed, the percentage of success obtained for the emotions "disgust", "anger", "sadness", "fear" and "neutral" are acceptable and similar to the obtained by Hammal (second column). Nevertheless, for "joy" and "surprise" the results are not very favorable. In fact, the algorithm tends to confuse "joy" with "disgust" and "surprise" with "fear", which comes justified looking at Table 3, where it can be seen that a same combination of states of distances can be given for the mentioned pairs of emotions.



Fig. 10. Frame classified like "disgust" by the FG-NET database (FG-NET, 2001).

EMOTION	% SUCCESS HAMMAL METHOD	% SUCCESS OUR METHOD	% SUCCES WRINKLES	% SUCCESS MOUTH SHAPE
Joy	87,26	36,84	100	100
Surprise	84,44	57,89	63,16	63,16
Disgust	51,20	84,21	94,74	100
Anger	not recognized	73,68	94,74	89,47
Sadness	not recognized	68,42	57,89	94,74
Fear	not recognized	78,95	84,21	89,47
Neutral	88	100	100	100

Table 4. Classification rates of Hammal (Hammal et al., 2005) (second column), of our method with the 5 distances (third column), plus wrinkles in the nasal root (fourth column) plus mouth shape information (fifth column).

In order to improve the results obtained in “joy”, we have introduced a new face parameter: the presence or absence of wrinkles in the nasal root, typical of the emotions “disgust” and “anger”. This way, we are able to mark a difference between “joy” and “disgust”. The obtained success rates are shown in the forth column in Table 4. We observe, as it was expected, a considerable increase in the rate of successes, especially for “joy” and “disgust”. However, the rates still continue to be low for “sadness” and “surprise”, which indicates about the necessity to add more characteristics to the method. A key factor to analyze in the recognition of emotions is the mouth shape. For each one of the 7 basic emotions, its contour changes in many different ways. In our method, we have added the extra information about the mouth behaviour that is shown in Figure 11. Results are shown in the fifth column in Table 4. As it can be seen, the new information has introduced a great improvement in our results. The importance of the mouth shape in the expression of emotions is thus confirmed.

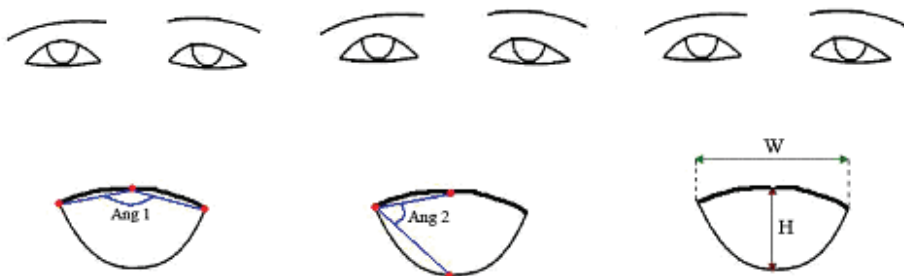


Fig. 11. Extra information added about the mouth shape

The method has also been tested with other databases different from the ones used for the threshold establishment, in order to confirm the good performance of the system. Related to classification success, it is interesting to realize that human mechanisms for face detection are very robust, but this is not the case of those for face expressions interpretation. According to Bassili (Bassili, 1997), a trained observer can correctly classify faces showing emotions with an average of 87%.

Once satisfactory classification rates were achieved, the system has been used to analyze the influence of gender and race in the studied face characteristics. Details of the results can be found in (Hupont & Cerezo, 2006).

#### Analysing video sequences: real-time interaction

The features extraction program captures each facial frame and extracts the feature points which are sent to the emotion classifier. When an emotional change is detected, the output of the 7-emotion classifier constitutes an emotion code which is sent to Maxine’s character. For the moment, the virtual character’s face just mimics the emotional state of the user (Fig. 12), accommodating his/her facial animation and speech. More sophisticated behaviour may be implemented.

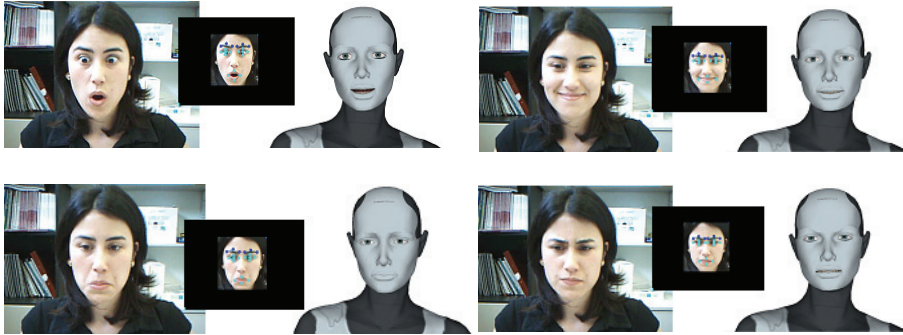


Fig. 12. Examples of the integrated real-time application: detection of surprise, joy, sadness, anger. For each example, images caught by the webcam, small images showing automatic features' tracking and synthesized facial expressions are shown. The animated character mimics the facial expression of the user.

#### 4. Voice-based interaction

Natural language is one of the most used communication methods, and although it has been extensively studied, relevant aspects still remain opened. As stated before, in order to obtain a more natural and trustworthy interaction, HCI systems must be capable of responding appropriately to the users with affective feedback. Within verbal communication it implies the addition of variability in the answers and the synthesis of emotions in speech (Cowie et al., 2000).

The incorporation of emotion in voice is carried out by changes in the melodic and rhythmic structures (Bolinger, 1989). In the last years, several works focus on the synthesis of voice considering the components that produce emotional speech. However, most of the studies in this area refer to the English language (Murray & Arnott, 1993, Shroder, 2001). Related to the Spanish language, the work of Montero et al (Montero et al., 1999) focus on the prosody analysis and modelling of a Spanish emotional Speech Database with four emotions. They make an interesting experiment about the relevance of voice quality in emotional state recognition scores. Iriondo et al. (Iriondo et al., 2000) present a set of rules that describes the behaviour of the most significant speech parameters related with the expression of emotions and validate the model using speech synthesis techniques. They simulate the 7 basic emotions. A similar study was made in (Boula et al., 2002), but getting the expressions of emotions of videos performed by professional actors in English, French and Spanish.

Our system performs emotional voice synthesis in Spanish, but unlike the previous works, it allows interaction, supporting real time communication with the user in natural language. Moreover, in this conversational interface the emotional state (that may vary depending on the relationship with the user along the conversation) is considered and expressed by the modulation of the voice and by selecting the right answer. For this purpose, the system keeps information about the "history" of the conversation.

#### 4.1 User-avatar communication process

The overall process of communication between user and avatar through voice is shown in Figure 13. In the following paragraphs, each of the three stages is explained.

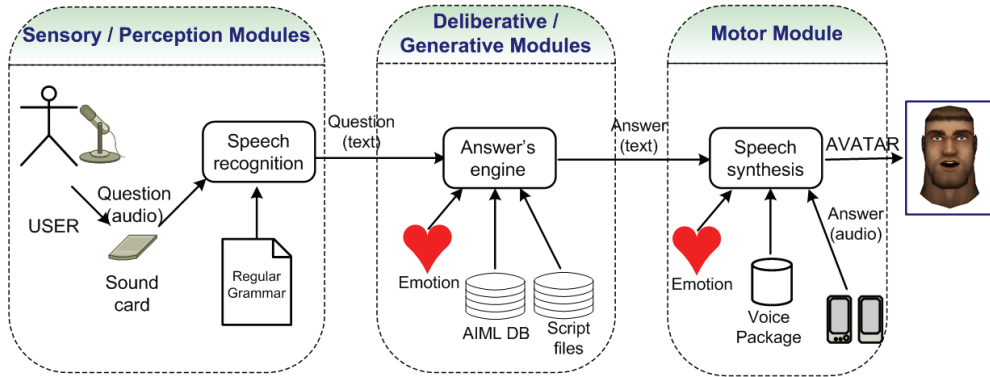


Fig. 13. Stages of the user-avatar voice communication process

##### Stage 1: Audio Speech Recognition (ASR)

The aim of this first stage is to obtain a text chain from the words said by the user in Spanish. To do this, a voice recognition engine has been constructed on the basis of the commercial Loquendo ASR (Audio Speech Recognition) software. The ASR is based on a dynamic library that enables a recognition device to be created and integrated ad hoc within a certain system; however, the disadvantage is that it has to be developed largely from scratch; in particular, it is necessary to pick up the audio and prepare it for processing by the recogniser and to develop a grammar with the words that are going to be recognised. Loquendo ASR only enables three possible context-free grammars, ABNF (Augmented BNF), XMLF (XML Form) and JSGF (Java Speech Grammar Format). We have chosen JSGF syntax as it avoids complex labelling of the XML and is used by a broader community than ABNF syntax.

One of the requisites of our system is that it must be able to “understand” and speak Spanish. This constraint prevented us from using existing open-source libraries, all of them in English. Moreover, during the development of the recogniser, some problems that are specific to Spanish had to be solved: specifically, Loquendo ASR is not capable of distinguishing between words with or without ‘h’ (this letter is not pronounced in Spanish), with ‘b’ or ‘v’, or with ‘y’ or ‘ll’ (these letter pairs correspond to single phonemes).

##### Stage 2: Getting the right answers

This stage is basically in charge of generating the answer to the user’s questions in text mode and it is based on the recognition of patterns, to which fixed answers are associated (static knowledge). These answers, however, vary depending on the virtual character’s emotional state (explained later on), or may undergo random variations so that the user does not get the impression of repetition if the conversation goes on for a long time (dynamic knowledge). In our case, this type of answer’s system has proved to be sufficient because, for the moment, we use it restricted to specific topics (education domains or specific orders for managing domotics systems).

The system we have developed is based on chatbot technology under GNU GPL licences: ALICE (ALICE 2007) and CyN (CyN 2004). However, CyN is only designed to hold conversations in English, so we had to modify the code to support dialogues in Spanish. The main differences lie in the work with accents, dieresis and the “ñ” character, and in the use of opening interrogation and exclamation marks.

The knowledge of the virtual character is specified in AIML -Artificial Intelligence Markup Language- (AIML, 2001). AIML is an XML derivative, and its power lies in three basic aspects:

- AIML syntax enables the semantic content of a question to be easily extracted so that the appropriate answer can be quickly given.
- The use of labels to combine answers lends greater variety to the answers and increases the number of questions to which an answer can be given.
- The use of recursivity enables answers to be provided to inputs for which, in theory, there is no direct answer.

The AIML interpreter has been modified to include commands or calls to script files within the AIML category. These commands are executed when the category in which they are declared is activated, and their result is returned as part of the answer to the user. This makes it possible, for example, to consult the system time, log on to a website to see what the weather is like, etc.

#### Stage 3: Text to Speech Conversion (TTS) and Lip-sync

The synthesis of the voice is made using Windows SAPI5, but the function uses packages of Spanish voice offered by Loquendo. SAPI gives information about the visemes (visual phonemes) that take place pronouncing the phrase wanted to be synthesized, what allows to solve the problem of the labial synchronization: a lip-sync module specially developed for Spanish language has been implemented. In order to avoid the voice sounding artificial, it has been equipped with an emotional component, as it will be described in next section.

## **4.2 Emotional voice generation**

The voice generated by text-voice converters usually sounds artificial, which is one of the reasons why avatars tend to be rejected by the public. To succeed in making the synthesiser appear “alive”, it is essential to generate voice “with emotion”. In our system we work with the six universal emotion categories of Ekman. SAPI5 enables tone, frequency scale, volume and speed to be modified, which is why we have used it as a basis. To represent each emotion, fixed values are assigned to the parameters that enable the relevant emotion to be evoked. The configuration of these emotional parameters is based on several studies (Boula et al., 2002; Francisco et al., 2005; Iriondo et al., 2000). The process carried out to find the values at which these parameters must be fixed for each emotion was voice assessment by users. The three assessment paradigms used were: Forced Choice, providing the subjects with a finite set of possible answers that take in all emotions that have been modelled, Free Choice, where the answer is not restricted to a closed set of emotions and Modified Free Choice in which neutral texts were used together with emotion texts. The values of the emotional parameters validated by the tests are shown in Table 5. Details of the evaluation process are given in (Baldassarri et al., 2007a).



Emotion	Volume (0-100)	Speed (-10 -10)	Pitch (-10 -10)
Joy	80	3	4
Disgust	50	3	-6
Anger	70	3	0
Fear	56	1	2
Neutral	50	0	0
Surprise	56	0	3
Sadness	44	-2	2

Table 5. Setting volume, speed and tone parameters for emotional voice generation.

### 4.3 Emotional management

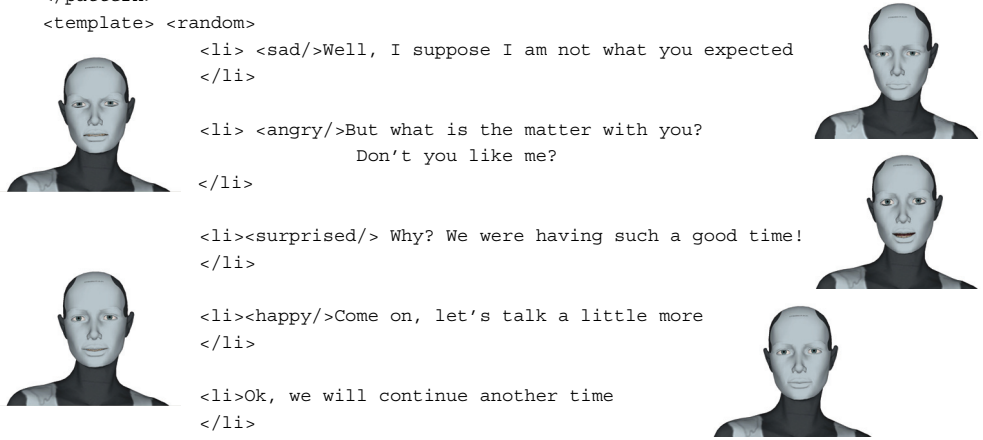
Emotion is taken into account not only in the voice synthesis (as it was previously explained) but also in the generation of the answers at two levels:

- The answer depends on the avatar's emotional state. For this reason, the AIML `<random>` command has been redesigned to add this feature, as it can be seen in the following example. There may be more than one answer with the same label, in this case, one of these answers would be given at random. There must always be an answer (applied to neutral emotional state) that does not have an associated label.

```

<category>
  <pattern> I BELIEVE THAT WE WOULD HAVE TO LEAVE THIS CONVERSATION
  </pattern>
  <template> <random>
    <li> <sad/>Well, I suppose I am not what you expected
    </li>
    <li> <angry/>But what is the matter with you?
      Don't you like me?
    </li>
    <li><surprised/> Why? We were having such a good time!
    </li>
    <li><happy/>Come on, let's talk a little more
    </li>
    <li>Ok, we will continue another time
    </li>
  </random> </template>

```



- Besides, the emotional state of the virtual character may change during a conversation, depending on how the conversation develops. That is, if the conversation is on a topic that pleases the character, it gets happy; if it is given information it was not aware of, it is surprised; if it is insulted, it gets angry; if it is threatened, it gets frightened, etc.

#### 4.4 Results

The system developed makes it possible for the user to maintain a conversation in Spanish with the virtual character. For example, in the virtual presenters application (see Section 2) the user can ask questions about the presentation and get answers from the character. As far as the voice interface is concerned, we have endeavoured to reduce to a minimum the time that elapses between the point at which the user finishes speaking and the point at which the answer begins. Excessive lead time decreases the sensation of interactivity and would not be readily accepted by the user. The duration of each stage in the communication process has been measured with sentences that are liable to be used during a conversation, in which the longest sentence is no longer than 20 words. The voice recognition and synthesis tests were all carried out in Spanish. The time measurements in the search for results were carried out with Alice's brain, which has some 47,000 categories. Table 6 shows a time study carried out through several conversations. In the table, both for synthesis and voice recognition, the maximum time applies to the recognition or synthesis of the longest sentences.

Stages of a conversation	Min Time	Max Time	Average
Speech recognition	1.6s	2.01s	1.78s
Text to Speech	0.18s	0.2s	0.3s
Search of Answers	0.1s	0.17s	0.2s

Table 6. Time measurements of the different stages of a conversation (in seconds).

#### 5. Evaluating Maxine agents

As we mentioned in section 2, Maxine system has been used to develop a learning platform to simplify and improve teaching and practice of Computer Graphics subjects. One of the ways the interactive pedagogical agent helps students is by exposing some specific topics, acting as a virtual presenter. Last term students have been asked to evaluate Maxine agents in these virtual presentations and their usefulness. Two questionnaires, one before and one after, have been done; their objectives are:

- To evaluate the previous knowledge of the subject that will be presented
- To assess the effectiveness of the "information provision" aspect of the message, ie, the Maxine's effect on the presentation subjects' comprehension
- To measure the perceptions about the Maxine agent.

Specifically, an introductory presentation about CG and its applications has been evaluated. The students are asked to evaluate their knowledge about different topics of CG before and after the presentation, evaluating it from 1 (very low) to 10 (very deep). In Figure 14 mean values are shown (classified by gender). It is interesting to realize that female students systematically rate their knowledge lower than male students; explanation of this behaviour cannot be based on objective facts, as they all are in the same university level, having coursed almost the same subjects and having female students usually higher marks.

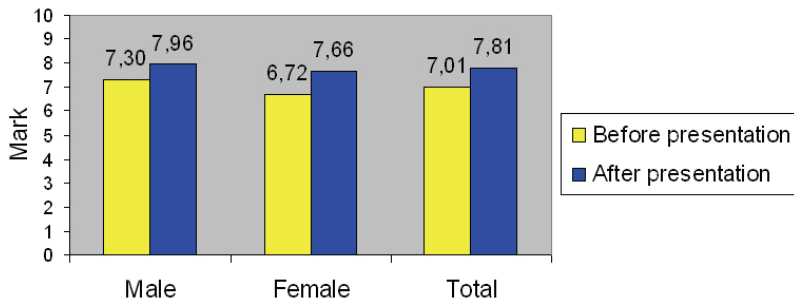


Fig. 14. Effectiveness on the subjects' comprehension after one of Maxine's presentations

Students are also asked about the aspects of the virtual agent that have attracted their attention (results in Figure 15) and which attributes would they use to describe the presenter (see Figure 16). Most of them think that this kind of "virtual teachers" only can be used as a tool and can not replace tutors (75%), but could be a good option for distance training (92%).

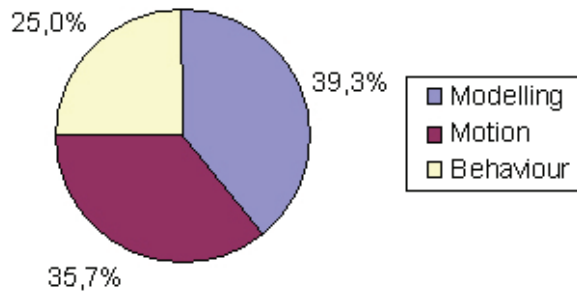


Fig. 15. Remarkable aspects of the virtual presenter

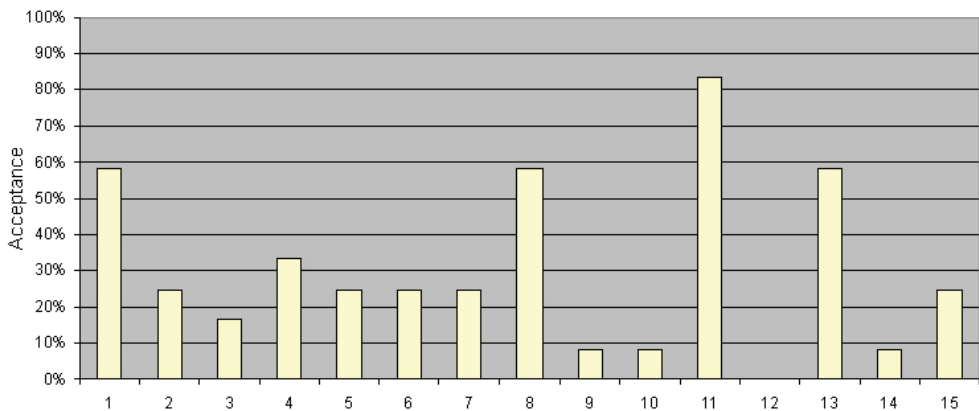


Fig. 16: Acceptance of the following virtual presenter's attributes (1-Helpful, 2-Intelligent, 3-Likable, 4-Reliable, 5-Believable, 6-Competent, 7-Friendly, 8-Clear, 9-Natural, 10- Not very convincing, 11-Stiff, 12-Happy, 13-Neutral, 14-Sad, 15-Coherence expression-message)

The students are also asked to describe the agent (see Table 7) and, by an open question, to compare it with other virtual characters they know (from videogames, programs,...). Their answers are all positive, considering it good, simple but effective. Students are also asked to point out which aspects contribute most to the realism and to the lack of realism of the virtual agent. The answers have been divided into two groups: those corresponding to students being used to videogames and those that are not. It is especially interesting the different consideration about the aspects contributing most to the lack of realism (see Figure 17).

Statement	Rank (1-10)
The virtual actor looks real	7.7
The movements of his/her head look natural	7.7
His/her gaze looks natural	7.0
His/her facial expressions look natural	8.0
Good lip-synchronization is achieved	7.5
Voice modulation is always coherent	6.6
Coherence between facial expression and voice	6.9

Table 7. Description of the virtual actor

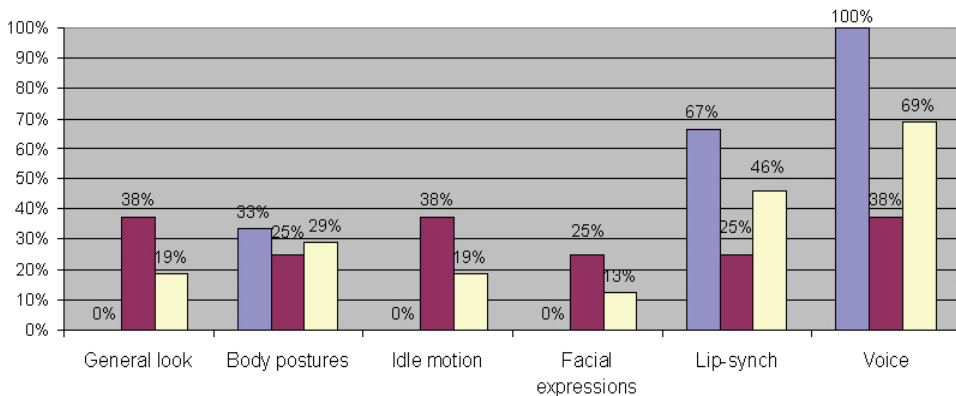


Fig. 17. Aspects contributing to the lack of realism. Opinion of students that: usually don't play videogames (blue), usually play videogames (red), total (yellow)

## 6. Conclusions and future work

This chapter presents a completely automated real-time character-based interface, where a scriptable affective humanoid 3D agent interacts with the user. Special care has been taken in making it possible multimodal natural user-agent interaction: communication is accomplished via text, image and voice (natural language). Our embodied agents are equipped with an emotional state which can be modified throughout the conversation with the user, and depends on the emotional state detected from the user's facial expressions. In fact, this nonverbal affective information is interpreted by the agent, which responds in an empathetic way by adjusting its voice intonation, facial expression and answers. These agents have been used as virtual presenters, domestic assistants and pedagogical agents in different applications and results are promising.

The chapter has focused on two main aspects: the capture of the user emotional state from web cam images and the development of a dialog system in natural language (Spanish) that takes also emotional aspects into account.

The facial expression recognizer is based on facial features' tracking and on an effective emotional classification method based on the theory of evidence and Ekman's emotional classification. From a set of distances and angles extracted from the user images and from a set of thresholds defined from the analysis of a sufficiently broad image database, the classification results are acceptable, and recent developments has enabled us to improve success rates. The utility of this kind of information is clear: the general vision in that is a user's emotion could be recognized by a computer, human computer-interaction would become more natural, enjoyable and productive.

The dialog system has been developed so that the user can ask questions, give commands or ask for help to the agent. It is based on the recognition of patterns, to which fixed answers are associated. These answers, however, vary depending on the virtual character's emotional state, or may undergo random variations so that the user does not get the impression of repetition if the conversation goes on for a long time. Special attention has also been paid in adding an emotional component to the synthesized voice in order to reduce its artificial nature. Voice emotions also follow Ekman's ones and are modeled by means of modifying volume, speed and pitch.

Several research lines remain open:

- Regarding Maxine, next steps are:
  - to allow not only facial expressions but body postures to be affected by the emotional state of the agent,
  - to use the user emotional information in a more sophisticated way: the computer could offer help and assistance to a confused user or try to cheer up a frustrated user and, hence, react in ways more appropriated than simply ignoring the user affective states, as is the case in most current interfaces,
  - to consider not only emotion but personality models for the virtual agents,
  - to give the system learning mechanisms, so that it can modify its display rules based on what appears to be working for a particular user, and improve its responses while interacting with that user, and
  - to carry out a proper validation of Maxine system and characters.

- In the case of the facial emotional classifier the next step is to introduce fuzzy models and fuzzy rules to the classification algorithm in order to obtain wider information from the emotional state of the user. The objective is to obtain the membership percentage of the displayed user emotion to each one of the 7 basic emotions (for example 70% happiness, 20% surprise, 10% neutral, 0% others).
- Regarding the interaction via voice, it will be important to improve the dynamic knowledge of the system because, till now, only the "history" of the conversation is stored. In this way, the system should be able to learn and should possess a certain capacity of reasoning or deduction to manage basic rules of knowledge. An other interesting field is the extraction of emotional information from the user's voice. As we are not specialist in the subject contacts have been established with a voice-specialised group.

## 7. Acknowledgements

The authors wish to thank the Computer Graphics, Vision and Artificial Intelligence Group of the University of the Balearic Islands for providing us the real-time facial tracking module to test our classifier and to David Anaya for his work in the natural language dialog system. This work has been partly financed by the Spanish "Dirección General de Investigación", contract number N° TIN2007-63025 and by the Aragon Government through the WALQA agreement (ref. 2004/04/86) and the CTPP02/2006 project.

## 8. References

- AIML (2001). Artificial Intelligence Markup Language (AIML) Version 1.0.1, <http://www.alicebot.org/TR/2001/WD-aiml/>
- ALICE (2007). Artificial Intelligence Foundation, <http://www.alicebot.org/>
- Anolli, L.; Mantovani, F.; Balestra, M.; Agliati, A.; Realdon, O.; Zurloni, V.; Mortillaro, M.; Vescovo, A. & Confalonieri, L. (2005). The Potential of Affective Computing in E-Learning: MYSELF project experience. *International Conference on Human-Computer Interaction (Interact 2005), Workshop on eLearning and Human-Computer Interaction: Exploring Design Synergies for more Effective Learning Experiences*, Rome, Italy September 2005
- Baldassarri, S.; Cerezo, E. & Anaya, D. (2007a). Interacción emocional con actores virtuales a través de lenguaje natural (in Spanish). *Proceedings VIII Congreso Internacional de Interacción Persona-Ordenador*, ISBN 978-84-9732-596-7, Zaragoza, Spain, September 2007, Thomson
- Baldassarri, S.; Cerezo, E. & Seron, F. (2007b). An open source engine for embodied animated agents. *Proceedings CEIG'07: Congreso Español de Informática Gráfica*, pp. 35-42, ISBN 978-84-9732-595-0, Zaragoza, September 2007, Thomson
- Bartneck, C. (2001). How convincing is Mr. data's smile: Affective expressions of machines. *User Modeling and User-Adapted Interaction*, Vol. 11, No.4, pp. 279-295, ISSN 0924-1868

- Bassili, J.N. (1997). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, N° 37, pp. 2049-2059, ISSN: 0022-3514
- Bolinger, D. (1989) *Intonation and its uses: melody and grammar in discourse*, Stanford University Press, ISBN-13: 978-08-0471-535-5
- Boukricha, H.; Becker, C. & Wachsmuth, I. (2007). Simulating Empathy for the Virtual Human Max, *Proceedings 2nd International Workshop on Emotion and Computing in conj. with the German Conference on Artificial Intelligence (KI2007)*, pp. 22-27, ISSN 1865-6374, Osnabrück, Germany.
- Boula de Mareuil, P. ; Celerier, P. & Toen J. (2002). Generation of Emotions by a Morphing Technique in English, French and Spanish. *Proceedings of Speech Prosody 2002*, pp. 187-190, ISBN 9782951823303, Aix-en-Provence, France, April 2002
- Brave, S.; Nass, C. & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent, *International journal of human-computer studies*, Vol. 62, Issue 2, pp. 161-178, ISSN 1071-5819
- Burleson, W.; Picard, R.W.; Perlin, K. & Lippincott, J. (2004). A platform for affective agent research. *Workshop on Empathetic Agents, Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*, ISBN 1-58113-864-4, New York, United States, August 2004
- Casell, J.; Sullivan, J.; Prevost, S. & Churchill, E. -eds.- (2000). *Embodied Conversational Agents*, MIT Press, ISBN 0-262-03278-3, Cambridge Massachusetts
- Cerezo, E.; Baldassarri, S.; Cuartero, E.; Seron, F.J., Montoro, G.; Haya, P.A. & Alamán X. (2007). Agentes virtuales 3D para el control de entornos inteligentes domóticos (in Spanish). *Proceedings VIII Congreso Internacional de Interacción Persona-Ordenador*, ISBN 978-84-9732-596-7, Zaragoza, España, Septiembre 2007
- Cowie, R.; Douglas-Cowie, E. & Shroder, M. -eds- (2000). *Proceedings ICSA Workshop on Speech and Emotion: a Conceptual Framework for Research*, Belfast
- Creed, C. & Beale, R. (2005). Using Emotion Simulation to Influence User Attitudes and Behavior, *Proceedings of the 2005 Workshop on the role of emotion in HCI*, September 2005, Edinburgh, UK
- Creed, C. & Beale, R. (2006). Multiple and Extended Interactions with Affective Embodied Agents, *Proceedings of the 2006 Workshop on the role of emotion in HCI*, September 2006, London, UK
- CyN (2004). Project CyN, <http://www.daxtron.com/cyn.htm>
- Edwards, G.J.; Cootes, T.F. & Taylor, C.J. (1998). Face Recognition Using Active Appearance Models, *Proceedings of the European Conf. Computer Vision*, Vol. 2, pp. 581-695, ISBN 3540646132, Freiburg, Germany, June 1998, Springer
- Ekman, P. (1999). Facial Expression. In: *The Handbook of Cognition and Emotion*. John Wiley et Sons, pp. 45-60, ISBN: 0471978361, Sussex, UK
- Elliott, C.; Rickel, J. & Lester, J.C. (1997). Integrating affective computing into animated tutoring agents. *Proceedings of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent*, pp. 113-121, Nagoya, Japan, August 1997

- FG-NET (2006). <http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html>
- Francisco, V.; Gervás, P. & Hervás, R. (2005). Expression of emotions in the synthesis of voice in narrative contexts (in Spanish). Proceeding of the Symposium on Ubiquitous Computing & Ambient Intelligence (UCAmI'05), pp.353-360, Granada, Spain, September 2005
- Hammal, Z.; Couvreur, L.; Caplier, A. & Rombaut, M. (2005). Facial Expressions Recognition Based on the Belief Theory: Comparison with Different Classifiers, *Proceedings of the 13th International Conference on Image Analysis and Processing*, ISBN: 3-540-28869-4, Cagliari, Italy, September 2005, Lecture Notes in Computer Science, Vol. 3617, Springer Verlag
- Hong, H.; Neven, H. & von der Malsburg, C. (1998). Online Facial Expression Recognition Based on Personalized Galleries, *Proceedings of the. Int'l Conf. Automatic Face and Gesture Recognition*, pp. 354-359, ISBN: 0818683465, Nara Japan, April 1998, IEEE
- Hupont, I. & Cerezo, E. (2006). Individualizing the new interfaces: extraction of user's emotions from facial data. *Proceedings of SIACG'06 (Iberoamerican Symposium on Computer Graphics)*, pp. 179-185, ISBN: 3-905673-60-6, Santiago de Compostela, July 2006, Spain
- Iriondo, I.; Gaus, R.; Rodriguez, A.; Lázaro, P., Montoya, N., Blanco, J. M.; Bernadas, D.; Oliver, J.M.; Tena, D. & Longth, L. (2000). Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. *Proceedings of ISCA Workshop on Speech & Emotion*, pp.161-166, Belfast, Northern Ireland, 2000
- Isbister, K. (2006). *Better game characters by design: A psychological approach*, Elsevier/Morgan Kaufmann, ISBN-13: 978-1-55860-921-1, Boston
- Klein, J.; Moon, Y. & Picard, R. (2002). This computer responds to user frustration: theory, design and results, *Interacting with Computers*, Vol. 14, pp. 119-140, ISSN 0953-5438, Elsevier
- Lyons, M.J.; Budynek, J. Akamatsu, S. (1999). Automatic Classification of Single Facial Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, n°12, pp. 1357-1362, ISSN: 0162-8828, IEEE
- Lua (2008) <http://www.lua.org/>
- Manresa-Yee, C.; Varona J. & Perales, F.J. (2006). Towards hands-free interfaces based on real-time robust facial gesture recognition. In: *Lecture Notes in Computer Science*, N°4069, Perales, F.J. & Fisher B. (Eds.), pp. 504-513, ISBN 103-540-36031-X, Springer
- Montero, J.M.; Gutierrez-Arriola, J.; Colas, J.; Enriquez, E. & Pardo, J.M. Analysis and modelling of emotional speech in Spanish. *Proceedings of the 14th International Conference on Phonetic*, pp. 957-960, San Francisco, United States
- MPEG-4 (2002). MPEG-4 Overview - (V.21), ISO/IEC JTC1/SC29/WG11 N4668, March 2002 <http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm>
- Murray, I. & Arnott, J. (1993). Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of the Acoustical Society of America*, Vol. 93, N°2, pp. 1097-1108, ISSN: 0001-4966



- Pantic, M. (2005). Affective Computing, *Encyclopedia of Multimedia Technology and Networking*, M. Pagani (ed.), Vol. 1, pp. 8-14, Idea Group Reference, USA.
- Pantic, M. & Rothkrantz, L.J.M. (2000a). Expert System for Automatic Analysis of Facial Expression. *Image and Vision Computing J.*, Vol. 18, N<sup>o</sup>. 11, pp. 881-905, ISSN: 0262-8856
- Pantic, M.; Rothkrantz, L.J.M. (2000b). Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, Vol. 22, Issue 12, pp. 1424-1445, ISSN 0018-9340
- Pantic, M.; Valstar, M.F.; Rademaker, R. & Maat, L. (2005). Web-based Database for Facial Expression Analysis. *Proceedings of the. IEEE Int'l Conf. Multimedia and Expo (ICME'05)*, July 2005, IEEE
- Picard, R.W. (2003). Affective Computing: Challenges, *International Journal of Human-Computer Studies*, Vol. 59, No. 1, pp. 55-64, ISSN 1071-5819
- Prendinger, H. & Ishizuka, M. (2004). What Affective Computing and Life-like Character Technology Can Do for Tele-Home Health Care, *Workshop on HCI and Homecare: Connecting Families and Clinicians (Online Proceedings) in conj. with CHI-04*, Vienna, Austria, April 2004
- Prendinger H. & Ishizuka, M. (2005). The Empathic Companion: A Character-Based Interface That Addresses Users' Affective States. *Applied Artificial Intelligence*, Vol. 19, N<sup>o</sup>. 3-4, pp. 267-285, Taylor & Francis, ISSN 0883-9514
- Reeves, B. & Nass, C. (1996). *The media equation: How people treat computers, televisions and new media like real people and places*, CLSI Publications, ISBN-10 1-57586-053-8, New York.
- Seron, F.J.; Baldassarri, S. & Cerezo, E. (2006). MaxinePPT: Using 3D Virtual Characters for Natural Interaction. *Proceedings WUCAml'06: 2nd International Workshop on Ubiquitous Computing and Ambient Intelligence*, pp. 241-250, ISBN 84-6901744-6, Puertollano, Spain, November 2006
- Seron, F.J.; Baldassarri, S. & Cerezo, E. (2007). Computer Graphics: Problem-based Learning and Interactive Embodied Pedagogical Agents. *Proceedings Eurographics 2008, Education papers*, Crete, Greece, April 2007 (to appear)
- Shroder, M. (2001). Emotional Speech Synthesis: A review. *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, Vol. 1, pp.561-564, Aalborg, Denmark, September 2001
- Vesterinen E. (2001). Affective computing. *Tik-111.590 Digital media research seminar*, Finland.
- VHML, (2001). Virtual Human Markup Language, VHML Working Draft v0.3, <http://www.vhml.org>
- Wallace, M.; Raouzaoui, A.; Tsapatsoulis, N. & Kollias, S. (2004). Facial Expression Classification Based on MPEG-4 FAPs: The Use of Evidence and Prior Knowledge for Uncertainty Removal, *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Vol. 1, pp. 51-54, Budapest, Hungary, July 2004
- Yee, N.; Bailenson, J.N.; Urbanek, M.; Chang, F. & Merget, D. (2007). The Unbearable Likeness of Being Digital: The Persistence of Nonverbal Social Norms in Online

Virtual Environments. *The Journal of CyberPsychology and Behavior*, Vol. 10, No. 1, pp. 115-121, Mary Ann Liebert Inc Publ, ISSN 1094-9313.

Zhang, Z.; Lyons, M.; Schuster, M. & Akamatsu, S. (1998). Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron, *Proceedings. Int'l Conf. Automatic Face and Gesture Recognition*, pp. 454-459, Seoul, Korea, May 2004