EVALUATION OF ECA GESTURE STRATEGIES FOR ROBUST HUMAN-COMPUTER INTERACTION

BEATRIZ LÓPEZ MENCÍA, ÁLVARO HERNÁNDEZ TRAPOTE, DAVID DÍAZ PARDO DE VERA, LUIS HERNÁNDEZ GÓMEZ

Signal, Systems and Radio communications Department, Universidad Politécnica de Madrid (UPM)
Madrid, 28040, Spain

MARÍA DEL CARMEN RODRÍGUEZ GANCEDO, JOSÉ RELAÑO GIL

Telefónica I+D, Spain

Embodied Conversational Agents (ECAs) offer us the possibility to design pleasant and efficient human-machine interaction. In this paper we present an evaluation scheme to compare dialogue-based speaker authentication and information retrieval systems with and without ECAs on the interface. We used gestures and other visual cues to improve fluency and robustness of interaction with these systems. Our test results suggest that when an ECA is present users perceive fewer system errors, their frustration levels are lower, turn-changing goes more smoothly, the interaction experience is more enjoyable, and system capabilities are generally perceived more positively than when no ECA is present. However, the ECA seems to intensify the users' privacy concerns.

Keywords: ECA evaluation; gestures, human-machine interaction, voice recognition, spoken language dialogue systems.

1. Introduction

Human-Machine Interaction (HMI) theorists have long envisioned artificial systems with which human beings could interact, and even relate to, as they do with each other. In short, the world of HMI has been increasingly playing with the idea of *natural* interaction Accordingly, discussions have developed concerning design goals and the related design principles. For instance, what makes any particular sort of interaction natural, and how can we measure its degree of naturalness? And more fundamentally, as not only designers but also philosophers have asked, is it appropriate (i.e., is it an appropriate design goal), or can it be beneficial, for people to in~~terac~~t with an artificial system like they do with other people? How close does this come to personifying a thing, what are the desirable and undesirable, conceptual and practical consequences of it, and does it even make sense to do so?[a]

Designers are probably wise to point out that the answers to most of the above questions depend on what we want to achieve with each particular system, what we want to use it for. In computer system designers' words: it all depends on the application.

Today we are witnessing a proliferation of studies that explore a variety of approaches to human-like HMI, dealing with different aspects of human interaction, most notably spoken dialogue (with its four main elements: speech recognition, speech generation, dialogue design and the pseudo-cognitive processes of "understanding" the messages received and generating meaningful communication goals and their associated messages), body movement and facial gestures (exploiting the visual channel), manipulation of physical objects (touch or haptic interaction), and physical presence (e.g., human-like robots).

In this paper we concern ourselves with spoken dialogue and visual communication. Specifically, we present experimental results of a study we are currently undertaking –in the context of COMPANIONS, a European Union project – to identify effects of incorporating an animated agent onto a spoken language dialogue system (SLDS). Such dialoguing animated agents are commonly referred to in the literature as embodied conversational agents (ECAs) [8]).

The main aspect of the visual communication offered by the human-like figure is primarily in the form of gestures designed as visual cues that, we hope, convey supra-linguistic information –that is, information that accompanies the main content or meaning of the dialogue and which frames it and allows it to flow along the lines constantly negotiated between the interlocutors, and which also carries other levels of meaning not contained in the words themselves: information regarding expectations, mental processes and emotions, for example

Indeed, an animated character, if well designed, may help the user understand the conversation with the system better and make it seem more "natural." But beyond that, interactional information including, for instance, visual cues for turn management, clues as to how well the system is understanding what the user is saying, and even emotional strategies to keep the user relaxed and in a good mood even in the face of errors –this being not only desirable in itself, but also crucial to prevent making error situations even worse – could potentially be very helpful in improving the flow of the dialogue.

A major problem with SLDSs is robustness. Speech recognition difficulties and errors are hard to recover from, and error recovery strategies can cause confusion among users as the dialogue takes unexpected twists. It is interesting to study whether ECAs can help by providing visual cues about what's going on with the system's oral comprehension, for instance by displaying meta-cognitive gestures to show that the system couldn't quite catch something the user said, or apologizing for having previously misunderstood the user (thus implicitly showing the user that he or she has been listened to and understood, and the information corrected). We will consider gestural strategies in more detail later, in conjunction with the dialogue strategies we have designed.

The rise of ECA systems builds on recent advances in disciplines such as those that deal with intelligent agents, multimodal interfaces, natural language, computer graphics and vision processing

There are many areas in which we are still very much in the dark. One such area is the role may ECAs play in improving dialogue robustness

Proponents of mounting ECAs on interfaces claim that interacting with them is natural, intuitive and can simultaneously convey different layers of nuanced information, apart from adding a social dimension to the interaction. Detractors, on the other hand, point out that no interaction benefits have been proved, beyond adding mere aesthetic or entertainment value. Furthermore, ECAs can be misleading, can make the user have false expectations, they can be confusing, distracting, and even increase anxiety and reduce the users' sense of control, which is the opposite of the effect they should have

One problem evinced by the debate is that the claimed benefits of ECAs are still far from being proven in realistic scenarios. We have aimed to put together a 'reasonably realistic' scenario to study how (and indeed if) ECAs can improve HMI parameters and user satisfaction. Research in these areas is still in early days, however, and the dimensions involved are still only vaguely defined.

Knowing what we mean by quality, what factors affect it and how we measure it is, of course, essential to evaluating HMI. We have begun to define (as yet tentatively) an 'evaluation frame' to guide us in categorizing quality parameters that may affect user acceptance. We hope this frame will allow us to propose and test models relating a variety of elements, with a view to measuring how they ultimately affect user acceptance.

One aspect of the quality evaluation frame has to do with the interface, which is related to the interaction metaphor that the interface is designed to bring to life. In our case we will be concerned with natural dialogue with or without an ECA, a metaphor of human dialoguing agent standing in front of the user. Another frame level has to do with what the system is essentially for, i.e., what task is it designed to perform, which determines the goal of the interaction.

Our test system integrates two tasks: The first task is *biometric access* to the system: the users are asked to enroll with a speaker recognition system, and then must verify their identity in order to move on to the second task. The dialogue is very much predetermined, and follows a request-answer scheme. We are interested in the effect an ECA might have on the performance and user acceptance of biometric systems. It is generally considered that there is a trade-off between security and usability in these systems and we wish to see whether ECAs may allow simultaneous improvement of both. The second task is *information retrieval*: it is a task designed solely to motivate a more flexible dialogue that may go through the main stages identified in the literature for automatic dialogue generation Our goal is to see what effects an ECA might have on the flow of the dialogue, especially in situations (e.g., error recovery) where robustness has often been found wanting in dialogue systems.

The information retrieval task we have designed in order to have a realistic experimental scenario is a videotelephony service where users call 'home' using mobile phones (simulated on a computer screen) to check the state of various home appliances.

The secure access through voice authentication technology, which is real (although, as we shall see, the outcome is pre-programmed), adds to the service metaphor, which is simulated (there is no actual home with devices that are controlled by the test users). The remote information retrieval service[b] that the metaphor recreates, is certainly interesting in its own right. Today new videotelephony applications are being developed for mobile terminals with spoken dialog to access a variety of information services, voicemail or videomail. Incorporating ECAs onto this new visual channel affords challenges of its own. For instance, screen space is more limited, so what ECA size, appearance and gestures are best and whether it is appropriate to have an ECA on screen in the first place are all relevant questions for research.

Finally, a few words about our evaluation scheme are in order. Our approach is to combine performance and interaction data with users' responses to questionnaires. These have been inspired by the ITU P.851 recommendation on questionnaire design for the evaluation of spoken dialogue systems for general telephone services, to which we have added dimensions (in the form of sets of questions) as we have seen appropriate to evaluate user perceptions related to the ECA and the secure access.

The rest of the paper is organized as follows: Section 2 explains the dialogue strategies we have designed to increase robustness, and the ECA behavior we associated with them. Section 3 sketches a user-centered acceptability assessment frame we are currently developing that has guided us when preparing our user questionnaires and the responses obtained. In Section 4 we describe the structure of the empirical test. Section 5 presents the main results of the experiment. Finally, conclusions form Section 6.

2. Gestures for robust ECA interfaces

As we mentioned in the introduction, ECAs offer the possibility to combine several communication modes such as speech and gestures, making it possible, in theory, to create interfaces with which human-machine interaction is more natural and comfortable. Unfortunately, we are still a long way from understanding how best to incorporate nonverbal communication, through ECAs, to improve human-machine dialogue. This notwithstanding, ECAs are already being employed to improve interaction

Among the more common critical dialogue situations for which it is worth examining the positive effects an ECA could have are the following:

- Efficient turn management: the body language and expressiveness of agents are important not only to reinforce the spoken message, but also to regulate the flow of the dialogue [17]. In particular, turn taking can be made to work more smoothly with facial feedback provided by avatars
- Improving error recovery: recognition-error recovery processes usually lead to some degree of user frustration Once an error occurs it is common to enter an error spiral, because, as the user becomes increasingly frustrated, her frustration leads to more recognition errors, making the situation worse . ECAs may help to limit such feelings of frustration, thus making error recovery more effective

- Correct understanding of the state of the dialogue: one of the most common problems in dialogue systems is that the user does not know whether or not the process is working normally [30]. This sometimes leads the dialogue to error states that could be avoided. The expressive capacity of ECAs could be used to convey to the user what state the system takes the dialogue to be in at any particular time.

A variety of studies have been carried out on behavioral strategies for embodied conversational agents They deal with behavior in hypothetical situations and in terms of the informational goals of each particular interaction (be it human-human or human-machine). We direct our attention to the overall dialogue systems dynamics, focusing specifically on typical robustness problems and on how to improve dialogue flow. We draw from existing research undertaken to try to understand the effects different gestures displayed by ECAs have on people, and we apply this knowledge to a real dialogue system, including voice authentication related dialogs.

We have implemented a dialogue strategy to deal with various critical dialogue stages, react to different recognition confidence levels and manage error situations. Associated with the dialogue strategy is an ECA gesture scheme, with a set of gestures corresponding to each dialogue stage. The gesture repertoire of our ECA is partially based on relevant gestures to which we have added a few suggestions of our own.

In defining the ECAs behavior we sought to exploit the following supra-linguistic resources: conversational skills (such as beat gestures to emphasize information, nodding and "*don't understand*" gestures), shifts in camera shots and lighting intensity (in order to create "proxemic" effects that might be meaningful for the user), and the portrayal of an empathic attitude (smiling or showing an expression of apology) to try to keep user frustration low when interaction problems occur.

Table 1 shows each dialogue stage, what prompts it, and the associated ECA behavior, both for the main dialogue and the peculiarities of the guided dialogue in the secure access task. The dialogue-gesture scheme is also summarized in Figure 1.

Initiation. The inclusion of an ECA at this stage "humanises" the system This is a problem, first because once a user has such high expectations the system can only end up disappointing her, and secondly because the user will tend to use more natural (and thus complex) communication which the system is unable to handle. The interaction experience will, thus, probably end up being frustrating.

Another concern is that contact with a dialoguing animated character may have the effect of reducing the user's level of attention to the actual information that is being given ([38], [39]), especially in the case of new users (as our test users are). Thus, the goal at initiation is to present a human-like interface that is upon first contact less striking and less distracting, and one that clearly "lays down rules" of the interaction and makes sure that the user keeps it framed within the capability of the system.

In order to try to foster a sense of ease in the user and help her focus we have designed a welcome gesture for our ECA based on the recommendations in Kendon (see Table 1).

Termination. It is confusing if a dialogue concludes without the user being aware of it. It is important to end with a clear farewell message. We have complemented this with typical farewell gestures in human-human interaction .

Table 1. Gesture repertoire for the main dialogue stages

MAIN DIALOGUE		
Dialogue stage	**Description (when it occurs)**	**ECA behavior (movements, gestures and other cues)**
Initiation	At the beginning of the dialogue.	Look straight at the camera, smile, wave hand. Zoom in for task explanation. Zoom out, lights dim.
Turn management	*Take Turn:* when the system starts to speak.	Look straight at the camera, raise hand into gesture space. Camera zooms in. Light gets brighter.
	Give Turn: when the system prepares to listen to the user.	Look straight at the camera, raise eyebrows. Camera zooms out. Lights dim.
Wait	When a timeout occurs.	Slight leaning back, one arm crossed and the other touching the cheek. Shift of body weight.
Help	When the system gives some explanation to the user.	Beat gesture with the hands. Change of posture.
Confirmation (low confidence)	When the system cannot understand something the user has said.	Slight leaning of the head to one side, stop smiling, mildly squint.
Confirmation (high confidence)	The system has recognized the user utterance with a high level of certainty.	Nod, smile, open eyes wide (wider than for neutral expression).
Acknowledgement of misunderstanding	After user informs the system that it has misunderstood what he or she has said.	Sequence: a) apologize; b) request repetition or rephrase from the user. Apology: rotate head slightly rightward and downward, raise inner eyebrow, “eyebrow of sadness” (to show remorse). Request: open eyes, smile gently (to show interest).
Error recovery with correction	When the user has corrected a recognition error and the system confirms the correction.	Lean towards the camera, perform beat gesture with hands.
Termination	The task has finished.	Look ahead, nod, smile, wave hand
ENROLLMENT AND VERIFICATION DIALOGUE (secure access task)		
Verification error	When the user’s identity hasn’t been positively verified (a false rejection has occurred).	Smile and (verbally) express remorse for not having been able to verify the user’s identity.
Wrong sequence of numbers recognized	The system “believes” to have “understood” a sequence of numbers uttered by the user, but it is not the one requested.	Same behavioral sequence as for *Acknowledgement of misunderstanding*.
Marking the tempo	Visual cue indicating the tempo with which the sequence of numbers (which the user is asked to repeat) is given.	Beat gesture with one hand for each number of the sequence.

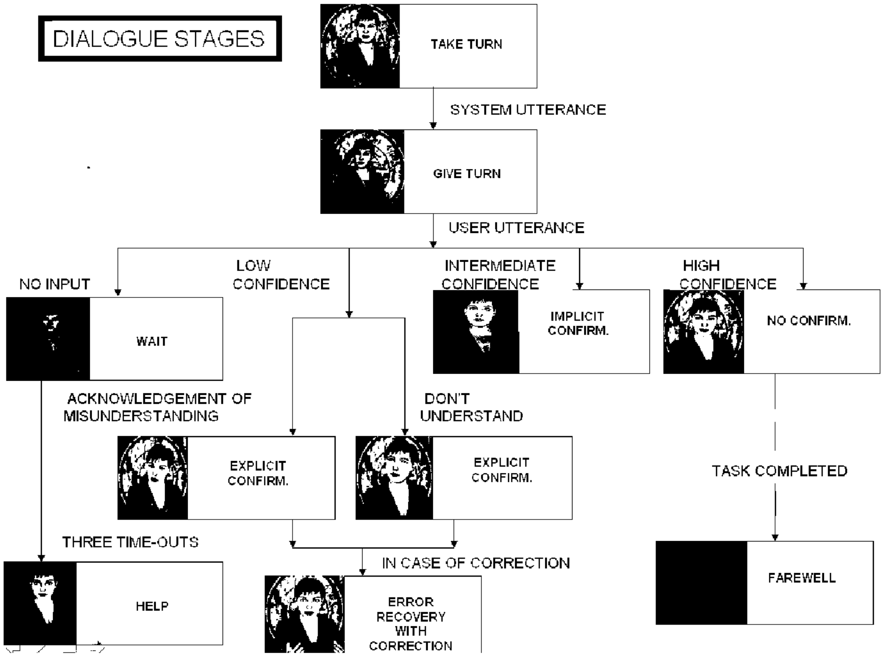

Fig. 1. Main interaction paths. (Each box represents a dialogue stage described in Section 2 –not all are included here–, and features a picture of the ECA performing the characteristic gesture of the stage. Arrows represent stage transitions, and their labels describe what prompts them.)

Turn management. Turn management involves two basic actions: taking turn and giving turn. Dialogue fluency improves and fewer errors occur if alternate system and user turns flow in orderly succession with the user knowing when it is her turn to speak.[c]

Our ECA strategy is as follows: When it's the ECA's turn the camera zooms-in slightly and the light becomes brighter. While the ECA approaches it raises a hand into the gesture space to "announce" that it is going to speak (see Figure 2). When it's the user's turn the camera zooms out, lights dim and the ECA raises its eyebrows to invite the user to speak. Hopefully the user will learn to associate different gestures, camera shots and levels of light intensity with each of the turns.

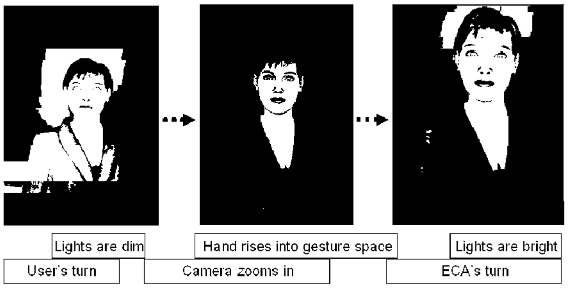

Fig. 2. Visual sequence of turn transition from user to ECA.

Confirmation. Once the user's utterance has been recognised, information confirmation strategies are commonly used in dialogue systems. Different strategies are followed depending on the level of confidence in the correctness of the speech recognition unit's interpretation of the user's utterance Our dialogue scheme and the associated gestural strategies are as follows:

- *High confidence in recognition:* The dialogue continues without confirmation request. The ECA nods her head , smiles and opens her eyes wide to show the user that everything is going well and the system understands her.
- *Intermediate confidence:* The result is regarded as uncertain and the system tries implicit confirmation (by including the uncertain piece of information in a question about something else). This allows the user to correct the system if an error did occur, and to feel everything is going well if what the system understood was correct. No specific ECA gesture was designed for this case. The idea is to keep the user speaking normally and without hyperarticulating (which would make recognition more difficult
- *Low confidence:* The dialogue becomes more guided with the system asking the user to repeat or rephrase. The ECA leans her head slightly to one side, stops smiling and mildly squints (a "*What was that you said?*" gesture; see Figure 3).

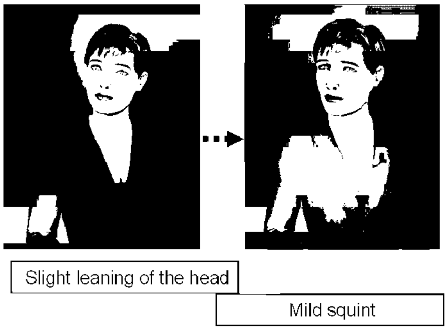

Fig. 3. ECA gesture sequence expressing low confidence in having correctly interpreted the user's utterance.

Acknowledgement of misunderstanding. A particularly delicate situation arises when the system misunderstands the user. If the user tries to correct the system or point out that it has misunderstood, the system will hopefully realise what has happened. It then tries to keep the user in a positive attitude and avoid her distrust while seeking to obtain the correct information. The dialogue scheme to pursue this consists in an apology followed by a kind request for a repetition or rephrase. The ECA gestures accordingly (Table 1), stressing the system's "interest" in getting it right to further motivate the user and preserve her trust.

Error recovery with correction. If the user points out that recognition errors have taken place and gives the correct information at the same time, the ECA repeats the corrected information emphasizing it by leaning towards the camera and marking the relevant words with beat gestures of the hands (up-down movements of the forearms, palms facing each other, fingers extended).

Help. A help message is given either when the user requests it or when the system has failed to hear the user say anything for longer than a reasonable waiting period. The ECA emphasizes the more important information in the help message with beat gestures performed with the hands. The idea is to see whether this captures the interest of the user, makes her more confident and the experience more pleasant, or if, on the contrary, it is distracting and makes help delivery less effective.

Wait. As we mentioned before, it sometimes happens that the user doesn't realize it is her turn to speak. To help the user realize the system is waiting for her to say something the ECA performs a waiting gesture: leaning back slightly with her arms crossed and shifting the body weight from one leg to another.

Verification errors. When the system is unable to verify the identity of the user –a typical problem with voice authentication (called false rejection)–, she may become nervous and, as a consequence, more prone to failure in the next verification attempt (because then her voice is strained and acquires a different quality than that the system knows from the training stage (enrollment). To partly avoid this problem our ECA doesn't tell the user that the system couldn't recognize her. Instead, the ECA kindly asks the user for another voice sample, making it seem simply that another sample is necessary as a normal part of the process. By hiding the fact that a verification error has occurred we hope to keep the user in a calm mood. The corresponding gestural strategy for the ECA is simply to remain smiling while requesting another voice sample.

Wrong sequence of numbers recognized. In order to prevent fraudulent access (e.g., using voice recordings), our system requests a different random sequence of numbers in every verification attempt. If the sequence the system believes the user has uttered is different from the one requested, the user is rejected. This can cause rejections of genuine users (i.e., false rejections) and increase frustration levels in them. In this situation our ECA empathizes with the user with a gesture showing remorse for not having been able to identify her ('taking the blame' for misunderstanding), followed by one expressing interest in order to keep the user confident for the next verification attempt.

Presenting the sequence of numbers. A common situation in speaker verification dialogue is that during training (enrollment) users repeat the sequence of numbers slowly, but once they acquire familiarity with the system they tend to repeat the requested sequence of numbers at a significantly higher pace. This can be a source of errors because verification algorithms perform better when a similar tempo of speech is followed in the training phase and in the verification attempts. The idea is, then, to implement an ECA strategy to try to get users to follow the same constant tempo when repeating the requested number sequence in both enrolment and verification, but without telling them, so that the system doesn't seem overly cumbersome to use. For this purpose our ECA marks the tempo with one beat of the hand for each number of the sequence

3. Some notes on quality evaluation

As was mentioned in the introduction, in order to evaluate a system one obviously needs to have an idea of what is what one actually wants to evaluate. But this isn't a simple task when one touches a concept as slippery as quality. In this section we present the basic outer-structure of a user acceptance-oriented HMI quality evaluation frame that we are developing in an attempt to crystallize (as far a possible) a notion of quality centered on system acceptability (from the user's point of view, insofar as we are able to capture it).[d] We have turned to the literature for inspiration for our user-centered quality model frame (especially the work of Angela Sasse on user acceptance of biometric technology and on Möller et al.'s taxonomy of quality factors for dialogue systems). Our conceptual frame, which we hope will provide us with a workable structure on which to propose and test models associating a variety of quality-related parameters, has guided us when preparing the questionnaire items and analyzing the users' responses to them. (We have also relied on the ITU P.851 recommendation).

Figure 4 shows the basic elements of our HMI evaluation frame. User acceptance, we posit, is influenced by three major classes of factors:

Usefulness (as perceived by the user): This class involves all aspects relating to how well the user believes the system is suited to the pursuit of the goals she would expect or want to achieve by using it. To evaluate a dialogue system a relevant question would be, for instance, how well users believe the system understands them. And for a voice authentication system, how well users believe the system can recognize them.

Likeability: This class includes all factors that have to do with the experience of using the system. For instance, usability-related factors such as pleasantness, dialogue clarity, and ease of use, as well as emotions and other sensations.

Rejection factors: This class is qualitatively different from the other two. While in the latter the user's response may have a positive or a negative valence, rejection factors can only be negative. We believe that when rejection elements are present they may affect user acceptance in a different way to how negative values on likeability factors such as ease of use do. For this reason we choose to study them separately. We have focused on certain aspects of privacy and security that are important in secure access systems in that they may cause rejection in users. Namely, fear that unauthorized people may manage to access the system, fear that the biometric data may be misused, feeling observed and concerns about impersonation.

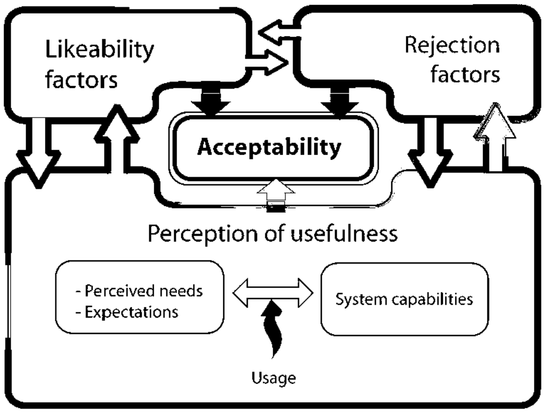

Fig. 4. User acceptance-oriented HMI quality evaluation frame.

Factors in each of these classes may affect factors in the other classes, as is illustrated by the arrows in the figure. The figure also shows an example of intra-class relationship: One aspect of user perception of usefulness has to do with the needs the user perceives she has and her expectations regarding the extent to which the system will fulfill them. Both of these factors are qualified by the system's capabilities as seen by the user, and this relationship of user perceptions itself develops in time with actual experience in use.

Now, user acceptance can be thought of as being composed of various levels in which user perception is exerted (whether the user is aware of distinguishing each level or not), each of which bearing a similar factor-class structure (sharing some of the elements contained in the classes). There is a global system-assessment level, a task-related, or goal, level (what the system or it's various elements are "supposed" to do), and finally there is an interaction-through-interface level, dealing with how things are presented to the user and how the user is expected to handle the system in order to achieve the desired goals. Figure 5 illustrates the layered scheme. In our case the task level is composed of two sub-levels: biometric access and remote domotic control (the dummy task we use as an excuse to test our dialogue system). Our interface level corresponds either to the VOICE-only interaction metaphor or to the ECA metaphor. As Figure 5 shows, it is reasonable to believe that there may be quality aspects (like "Aspect X" in the figure) that have their counterpart on every perception level, all influencing each other and affecting the overall counterpart. There may also be aspects influencing others on different levels that belong to a different class (such as task-related likeability "Aspect Y", in the figure, which affects usefulness "Aspect X" at the overall system evaluation level).

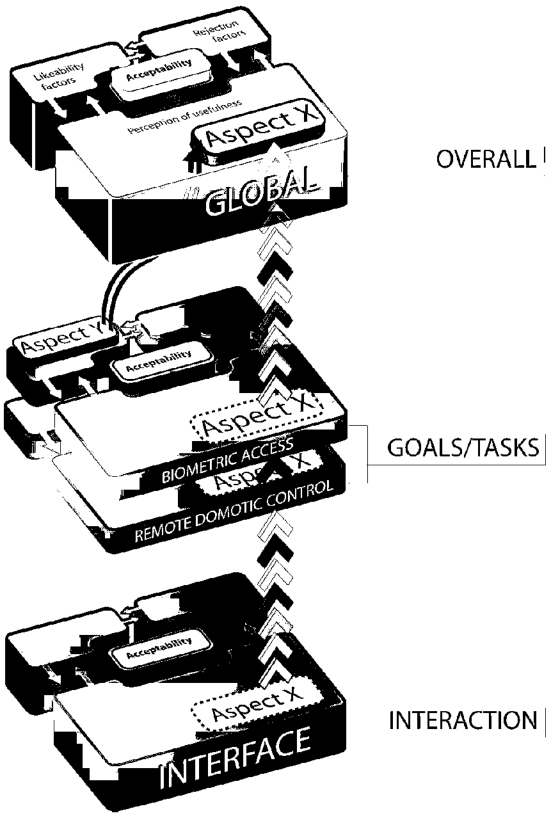

Fig. 5. Layered quality-factor frame structure, corresponding to the main system elements distinguishable to user perception.

4. Experimental set-up

4.1. *System implementation*

The architecture of the test environment is based on web technology, with which we simulate a mobile phone interface. Figure 6 shows the two different interaction scenarios we have compared: one (on the right) corresponding to what we have called the ECA interaction metaphor, and the other (on the left) with a still image (representing "home") that we call VOICE interaction metaphor (SLDS without ECA). Different users interact with these two different scenarios providing contrastive experimental data that will allow us to evaluate the ECA metaphor vs. the VOICE metaphor. The system is implemented on a web page that contains two frames. In the left frame there is a column of labels that show the test user what stage of testing he or she is (not to be confused with the dialogue stage which is not indicated). The main interface is displayed in the right frame and shows a mobile phone running a videotelephony application. Tactile interaction is not active at any stage.

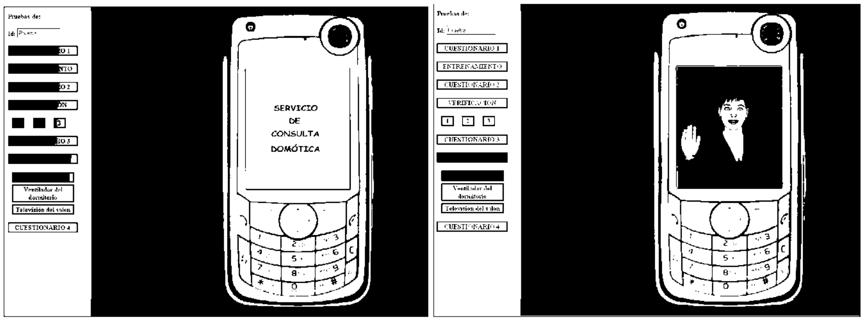

Fig. 6. Interface displays for ECA Metaphor and VOICE Metaphor.

All the contents of the evaluation are hosted on an Apache Tomcat web server. Throughout the test, users face a series of evaluation questionnaires and dialogue interactions. The questionnaires are implemented using HTML forms, and the information collected on them is transferred to JSP files and then stored in a database. Our test environment uses Nuance Communications' speech recognition technology.[e] The ECA character was created by Haptek.[f] The dialogues are implemented with Java Applet technology, and they are all packed and signed to guarantee fast download and access to the audio resources. Dialogue dynamics are programmed. Nuance's speech recognition engine provides a useful Java API that allows access to different grammars and adjusting a range of parameters depending on the characteristics of each application. Finally, interaction parameters (such as utterance durations, number of turns, number of recognition errors, etc.) are recorded automatically during the test interactions.

4.2. *Description of the experiment*

We tested the system with 16 undergraduate and graduate students (7 female and 9 male), aged 19 to 33, divided into two groups (8 users in each), one to test the system with the ECA interface (or interaction metaphor) and the other without ECA (VOICE metaphor).

Testing was carried out in a small meeting room. Users were seated at the head of a long table in front of a 15" screen. Two different views of the user interacting with the system were video-recorded to provide us with visual data to inspect and annotate the subject's behavior: A frontal view was taken from the top edge of the user's screen, and a lateral view was recorded from a wide-angle position to the right of the user. Both views were taken with Logitech Quickcam Pro 4000 webcams. Users spoke to system through headset microphone, and the system prompts were played through two small speakers.

All user-system dialogue was in Spanish. The entire test procedure was designed to take roughly 30 to 45 minutes, with minimal intervention on the part of experimenter.

The stages of the evaluation are as follows:

(1) *Brief explanation:* The user is told what the general purpose (to "evaluate automatic dialogue systems") and methodology of the evaluation are, as well as the tasks that lie ahead for him/her.

(2) *Opening questionnaire* to learn about the user's prior experience and expectations.

(3) *Training phase:* The user is asked to enroll in a secure access system, which requires interacting in guided dialogue with an application that registers his/her voice traits. (The system asks the user to repeat four four-digit sequences.)

(4) *Post-enrolment questionnaire* to capture the user's opinions on the form of access and related aspects such as privacy and security.

(5) *Verification phase (secure access task):* The user does three successive verification exercises. In each he/she is required to repeat a random four-digit sequence (up to three times, in the event of verification failures). The outcome of each exercise is predetermined (there is no real verification going on, but speech recognition *is* real). The idea is to let the user feel various situations that can arise during verification: In the first exercise the system reacts as if it had successfully recognized the user at the first attempt; in the second the user is rejected after three failed attempts; and in the third, the user is granted access at the second attempt.

(6) *Post-verification questionnaire:* Similar to the post-enrolment questionnaire, to see if users' opinions change after using the secure access system.

(7) *Domotic dialogue phase (information retrieval task):* Users are asked to find out the state (on/off) of three household devices ("the bathroom lights", "the fan in the bedroom", and "the living-room television set"). The automatic speech recognizer and the dialogue system function freely (i.e., they are not programmed to give certain answers; it is a real working system).

(8) *Final questionnaire:* To obtain the user's overall impression of the system, its main elements and the most important aspects of using it. Some questions are the same as in previous questionnaires, so that we may observe how user perceptions evolve throughout the various stages of using the system.

5. Experimental Results

We have obtained the results detailed in this section by a) comparing performance and questionnaire responses in the ECA metaphor group of users with those in the VOICE metaphor group; and b) observing how performance and responses to certain questions evolve throughout the test. We used two sample t-tests, setting the significance level at 5% (p=0.05). Questionnaire responses were collected on Likert-type 5-point response formats. User comments were also collected and compared to the findings in a) and b).

We now present the main findings obtained from these comparative analyses, focusing successively on each of the three quality evaluation categories introduced in Section 3: usefulness, likeability and rejection factors.

5.1. *Perceived usefulness*

In this section we look at how parameters related to the users' perception of system usefulness were affected by interaction features designed to make dialogue flow better and so gain in efficiency and clarity. We focus on three important aspects: perception of system errors, turn management and perception of dialogue capability.

5.1.1. *Sensitivity to errors and user frustration*

Average user awareness of system recognition errors is lower for ECA users. In spite of the fact that the minor difference we found in the actual average numbers of recognition errors between both of the tested interaction metaphors was not statistically significant, there was a striking, statistically significant, difference in the answers to the question *"Did the system make many mistakes?"* (1- very many ... 5 - none): a mean value of 3.8 for the ECA metaphor vs. 2.6 for the VOICE metaphor (t(12)=3.16; p =0.004). User frustration while interacting with the system was also markedly lower for the ECA group, as indicated by the mean values $\mu_{ECA} = 1.4$ vs. $\mu_{VOICE} = 2.6$ (t(9)=-2.52; p =0.016) of the responses to the question: *"Was the experience [of using the system] frustrating?"* (1 – no, not at all ... 5 – yes, very much so).

The measured differences in the two previous parameters between the ECA and VOICE scenarios possibly reflect relevant advantages, at least in terms of how it affects user perception, of the use of ECAs with appropriately designed gestures, both to deal with problematic dialog stages such as error recovery situations and to provide users with visual cues of how well the system is understanding her (i.e., with what level of confidence; see Section 2). We could be seeing here a variant of the persona effect [43] – a phenomenon widely reported in the literature according to which users tend to perceive tasks as easier when they interact with an ECA–, without there being any real improvement in performance (success in task execution and efficiency) when compared with users doing the same without an ECA. In our case no significant difference was found between the two test groups regarding perception of ease of use. However, believing the system made fewer mistakes could be a related effect.

There may be more to it, though, and user frustration and perception of performance quality may be linked to actual improvements in dialogue flow and in the users' knowledge of what is going on (what the system is doing and expecting the user to do). We now turn to exploring these possibilities briefly.

5.1.2. *Visual cues for turn switching*

The users' perception that "*Dialoguing with the system led quickly to solve the task proposed*" (1 - totally disagree ... 5 - totally agree) was on average greater in the ECA group (4.2) than in the VOICE group (3.2) (t(12)=3,16; p=0.004). This is not just a subjective impression induced by the presence of the ECA, which would make it an instance of the persona effect. A close examination of the ECA-supported dialogues shows that users easily learn when it is their turn to speak to the system. This helps prevent most of the typically observed failed barge-in attempts and time-outs, which we found occurred more often for our VOICE metaphor users. Some of these users said they had felt confused at certain stages of the dialogue (e.g., *"between tasks there were silences and I didn't know if I was supposed to say anything," "a couple of times I think I spoke too early and that's why the system didn't get what I said," "it would be better if some sort of visual sign told you when the system is ready to listen"*).

We also found consistent differences between the two groups of users in task duration and number of turns taken, which are, of course, two important efficiency indicators. However, none were statistically significant. This may be due to the small size of our test groups. Nevertheless, before we test the system with more users it is reasonable to explain our findings as a combination of a persona effect with the fact that ECA-metaphor users learn to interact with the system more easily, feel more in control, and actually experience a more coordinated dialogue than VOICE-metaphor users.

Thus, it seems our visual feedback channel featuring an ECA displaying contextual dialogue management cues may be providing supra-linguistic information that users are able to interpret correctly, leading to improved coordination, which in turn increases the users' impression of the dialogue being fast, efficient and under control.

But what are these visual cues that appear to be so useful? Our findings suggest that the visual information strategy for turn-switching that we have implemented –involving a combination of gestures and lighting and camera zoom effects– may be creating a "proxemic-code" that helps avoid the complicated, problem-laden interaction patterns reported in where user-ECA interaction suffers from rather severe coordination problems. Moreover, we observed no negative reactions, so users seem to accept proxemic shifts as a "natural" element of the interaction.

5.1.3. *User expectations and perception of dialogue capability*

The users' impression of how powerful the system's dialogue capabilities are, combined with the users' expectations regarding these capabilities, has an important impact on the users' overall assessment of the system Our experimental results show that the ECA-metaphor group was impressed with the system dialogue capability, although somewhat less than the VOICE-metaphor group, the former grading with an average of 3.9, and the latter 4.5 (3.0 being the neutral score), on the question: *"Were you positively or negatively surprised by the system's dialogue capability?"* (1 - very negatively surprised … 5 - very positively surprised) (t(13)=-2.12; p =0.027).

A plausible explanation has to do with the "humanizing" effect of the ECA discussed in Section 2 (in agreement with other reports such as and [44]). Since in fact we have the same dialogue engine behind both our ECA and VOICE-metaphor interfaces, users of the former tend to end up being less impressed with the system's conversational skills –having expected more but getting the same– than users of the latter.

This, of course, notwithstanding the fact that the users in the ECA group don't really "get the same," if we consider that, on average, they experience a smoother dialogue, as we saw previously. The following qualitative impressions expressed by our test users may add a little perspective to the analysis: *"In the beginning my main feeling was one of* *mistrust because it was a new experience, but afterwards it was pleasant and it was very easy to become accustomed to it." "I thought that the interaction with the system would be less comfortable, but the system understood me very well."*

Here we see that initial expectations might not be so positive after all, and that the experience of interacting with the system did in fact exceed at least some of the users' expectations. We clearly need to carry out further tests to shed light on the intricacies of user expectations and their evolution through system use.

5.2. *Likeability*

It is important that users feel comfortable during the interaction. We now look at factors related with pleasantness, amusement, and emotions. We also report users' opinions regarding the expressiveness of the ECA.

5.2.1. *User amusement and emotions*

ECA metaphor user enjoyment increased throughout the test: answers to the question *"Compared to other ways of interacting with a system (e.g., pressing buttons to choose options from menus), is spoken dialogue more fun or more tedious?"* (1 - much more tedious … 5 - much more fun) averaged 2.9 in the first questionnaire and 3.6 in the last (t=-2.39; p=0.024). In contrast, the average *pleasantness* score for VOICE metaphor users fell from the first questionnaire (3.6) to the last (3.3) (t=2.05; p=0.040).

Apart from frustration (which we looked at earlier in relation with error perception), the only other feeling for which our data shows a statistically significant difference between the ECA and the VOICE group is happiness (users in both groups felt similarly relaxed, confident, bored, dejected, angry and clumsy, for instance). The ECA group averaged 4.0, against 3.1 for the VOICE group, in their replies to the question: *"While you were interacting with the system, did you feel happy?"* (1 - no, not at all … 5 - yes, very much so) (t(13)=1.99; p =0.034).

It is clear that the observed difference in emotional response between the two test groups, favoring as it may the use of an ECA, was only very slight. After all, the whole experimental procedure is short and fairly simple, and test users have very little at stake performing the test, so it seems unlikely that strong emotional responses might appear. However, in future experiments we plan to design longer, more complex tasks and, by increasing the sample size, we hope to be able to determine more precisely how our ECA affects user emotions, if at all, and how these might affect overall usability and user acceptance.

5.2.2. *ECA expressiveness*

We invited the test users to give us their views regarding the ECA's gestures and expressiveness. These are a few revealing samples:

"I very much liked the expressiveness of the animations."

"I found the agent and the agent's gestures surprising."

"The face gestures were well designed, but the hand gestures could distract you."
"I liked the ECAs very much. They're very funny."

These opinions are encouraging, especially as there are studies that point out that in order to improve the believability and naturalness of an ECA it is essential to give it a consistent personality and to make it expressive

Furthermore, in our study we have observed that the users' opinion of the ECA's expressiveness increases with use after first contact (which occurs in the identity verification phase of the test): the average score for *"Is the agent expressive?"* (1 - no, absolutely not … 5 - yes, very much so) increased from 3.5 after first contact to 4.1 at the end of the test (t-value=-3.42; p-value=0.006). Similarly, users' impression of ECA friendliness (another relevant factor connected to user expectations; also increases slightly with use, from 4.1 to 4.5 (t-value=-2.05; p-value=0.040).

Expressiveness and friendliness may be "humanising" the ECA [47], but in a way that, rather than leading ultimately to disappointment, keeps users in a positive attitude and raises their interest in a natural-feeling interaction. This happens though the course of time (the little time our test lasts), which may be yet another piece of evidence that our ECA doesn't trigger unrealistic expectations upon first appearance, but gradually "wins users over."

Finally, we mention that in the present work we have not focused on specific gesture design (which gestures were preferred, which were perceived as being the clearest, and so on). However, prior to the present experiment we carried out a successful gesture validation test on the repertoire displayed by our ECA The comparative experiment discussed in this paper also serves as *implicit* overall gestural validation thanks to the interaction improvements we have observed. By analysing the video recordings of the user tests (which we will do shortly) we hope to obtain deeper insights on the effects of specific gestures –especially those we have designed with a view to improving dialogue robustness in difficult situations– and on how we might refine them.

5.3. *Rejection factors*

A major concern in identity verification systems is privacy. Therefore, "personifying" with an ECA a system designed to capture sensitive information, as voice features are, requires special care. These are the findings in our study that bear on this issue:

Responses to the question *"Would you feel uncomfortable using the remote control system for home devices because you would feel your privacy was being encroached on?"* (1 - no, not at all … 5 - yes, very much so) evolved significantly in the ECA metaphor group, averaging 2.5 in the first questionnaire and 3.3 in the last (t =-2.05; p=0.040). Similarly, for the question *"Would you have security concerns using the system, perhaps because you fear that unauthorized people might manage to remotely control your home devices?"* (same response format): replies averaged 2.5 in the first questionnaire and 3.5 in the last (t=-3.06; p=0.009).

These results are in accordance with previous work of ours in which we studied the effect an ECA could have on users interacting with a biometric authentication application. We found that the mere presence of an ECA (without any specifically designed gestures and with little expressiveness) can negatively affect users' perception of loss of privacy. However, our new findings seem clearer, suggesting that a more active ECA has a greater negative impact on the users' perception of security and privacy. This could be either because the user feels observed or because an animated figure makes the system look less serious and therefore less trustworthy. We need to continue testing to clarify this point.

6. Conclusions

In this paper we have presented a research scheme in which we have considered the main problematic situations that typically arise in automatic dialogue generation. In order to improve the robustness and the ease-of-flow of the dialogue we have implemented a gesture repertoire for an ECA. The gestures are designed to convey to the users meaningful supra-linguistic information regarding the state of the dialogue throughout the interaction, and to try to keep user in a positive frame of mind. We have proposed evaluating how well these strategies work by setting up an experiment to compare interaction with two different interfaces: one featuring our ECA and another with speech as the sole system output.

We found that the ECA contributed to keeping user frustration low, especially when recognition errors occurred (which is the most delicate scenario). This result suggests that our error management strategies are working, particularly: a) implicit confirmation with no ECA reaction when confidence in recognition is intermediate; b) performing a "*What was that you said?*"-type gesture to show the user the system isn't sure it has understood but is making an effort to (when confidence in recognition is low); and c) acknowledging misunderstandings with an apology and an accompanying gesture sequence to reassure the user that the system knows what has happened and is trying to put things right.

An encouraging result is that adding specific ECA gestures and 'camera movements' to mark turn changes seems to improve dialogue flow and prevent barge-in attempts and related problems. Users seem to be able to learn our proxemic code and accept it rather naturally.

On the negative side, the ECA's human-like appearance could potentially cause users to ultimately be somewhat disappointed with the system's dialogue capability, probably because of the false expectations such an appearance gives rise to, as has already been reported in the literature. Our results cannot confirm nor disprove this effect. However, we have seen indications that our ECA doesn't generate expectations in users that are too far off the mark. In fact, users seem to appreciate the ECA more after interacting with it for a while. This is an area we must examine more closely in future work.

Our findings also suggest that certain likeability and rejection factors might cancel each other out in terms of the effect they have on user acceptance. We have observed that interaction with our ECA is more enjoyable but increases privacy concerns, while, overall, no noticeable difference in acceptance was observed between the two test groups. However, with our data we cannot determine a precise relationship.

Many questions open up before us. For instance, why are ECA users more concerned about privacy? Is it because of the way the ECA behaves? Because it seems more natural, as if there were a real person in the interface, so users feel observed? Or does this effect depend primarily on whether the ECA is present or not (and not on its expressiveness)?

We plan to perform further user tests with this experimental set-up shortly, after which we will analyze all the gathered information, including the video recordings (what we have presented here is a first batch of results that do not fully exploit the possibilities of our dialogue and gesture strategies, or our acceptability evaluation frame). We expect the videos will help us study the reactions of users to the 'emotional' cues of the ECA.

We hope our work, while far from settling the debates reflected in the introduction of this paper, might help to show ways in which ECA technology can make a positive contribution to the quest for natural dialogue interfaces.

Acknowledgments

The work of Telefónica I+D was carried out with the support of the European Union IST FP6 program, project COMPANIONS, IST-34434. UPM activities were funded by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01. Our gratitude goes also to Bénédicte Fréon for her wonderful help with the graphics.

References

A.M. Noll, Natural language interaction with machines: a passing fad? or the way of the future?, *in Proceedings of the 18th annual meeting on Association for Computational Linguistics*, p.137, 1980.

N. Weinstein, Thinking about emotional machines, *Technological Review*, January Issue, 1998.

B. Schneiderman, Human values and the future of technology: A declaration of empowerment, *Proceedings of the conference on Computers and the quality of life*, ACM Special Interest Group on Computers and Society, 1990.

V. Gómez Pin, Entre lobos y autómatas. La causa del hombre, *Espasa Calpe*, Madrid, 2006.

S.M. Ali, "The End of The (Dreyfus) Affair": (Post)Heideggerian Meditations on Man, Machine, and Meaning*, in Proc. Cognitive Technology: Instruments of Mind : 4th International Conference, CI 2001,* Warwick, UK, pp. 149-156, 2001.

J.L. González Quirós, El porvenir de la razón en la era digital, *Síntesis*, Madrid, 1998.

COMPANIONS, European Commission Sixth Framework Programme Information Society Technologies Integrated Project IST-34434, http://www.companions-project.org/.

J. Cassell,, T. Bickmore, H. Vilhjálmsson, and H. Yan, More than just a pretty face: affordances of embodiment*, in Proceedings of the 5th international conference on Intelligent user interfaces*, pp. 52-59, ACM Press, 2000.

D. McNeill, Hand and Mind: What Gestures Reveal about Thought. *The University of Chicago Press*, Chicago, 1992.

P. Ekman, Facial Expression And Emotion, *American Psychologist*, 48(4), 384-392, 1993.

M. Montepare, S.B. Goldstein, and A. Clausen, The iden-tification of emotions from gait information, *J. Nonverbal Behavior*, vol. 11, no. 1, pp. 33–42, Spring 1987.

I. Poggi, C. Pelachaud and E.M. Caldognetto, Gestural Mind Markers in ECAs, *Gesture Workshop 2003*, pp 338-349, 2003.

N. Leßmann, A. Kranstedt, and I. Wachsmuth, Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max, *Proceedings Workshop W12, AAMAS* 2004, New York, 57 - 65.

S. J. Boyce, Spoken natural language dialogue systems: user interface issues for the future, *in Human Factors and Voice Interactive Systems,* D. Gardner-Bonneau Ed. Norwell, Massachusetts, Kluwer Academic Publishers: 37-62, 1999.

T. Rist, E. André, and S. Baldes, A flexible platform for building application with life-like characters, *in Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 158-165, ACM Press, 2003.

W.L. Johnson, J.W. Rickel, and J.C. Lester, Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments, *The International Journal of Artificial Intelligence in Education* (11), 47-78, 2000.

T. Bickmore, J. Cassell, J. Van Kuppevelt, L. Dybkjaer, and N. Bernsen, (eds.), *Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*, chapter Social Dialogue with Embodied Conversational Agents. Kluwer Academic, 2004.

S. Brave, C. Nass, and K. Hutchinson, Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent, *Int. J. Human-Computer Studies*, Nr. 62, Issue 2, pp. 161-178, 2005.

L. Bell, and J. Gustafson, Child and Adult Speaker Adaptation during Error Resolution in a Publicly Available Spoken Dialogue System, *Proceedings of Eurospeech 03*, Geneve, Schweiz, 2003.

R. Catrambone, Anthropomorphic agents as a user interface paradigm: Experimental findings and a framework for research, *in Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 166-171, Fairfax, VA, 2002.

J. Xiao, Empirical Studies on Embodied Conversational Agents, *Ph.D. Dissertation*, Georgia Institute of Technology, Atlanta, GA, December 2006.

A. Whitten, and D. Tygar, Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0, *Proceedings of the 8th USENIX Security Symposium*, Washington DC, 1999.

M. McTear, Spoken Dialogue Technology: Towards the Conversational User Interface, *Springer*, 2004.

ITU-T P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken. Dialogue Systems, *International Telecommunication Union (ITU)*, Geneva, 2003.

D.W. Massaro, M.M. Cohen, J. Beskow, and R.A. Cole, Developing and evaluating conversational agents. In *Embodied Conversational Agents* MIT Press, Cambridge, MA, 287-318, (2000).

M. White, M. E. Foster, J. Oberlander, and A. Brown, Using facial feedback to enhance turn-taking in a multimodal dialogue system, *Proceedings of HCI International 2005*, Las Vegas, July 2005.

S. Oviatt, and R. VanGent, Error resolution during multimodal humancomputer interaction, *Proc. International Conference on Spoken Language Processing*, 1 204-207, (1996).

S. Oviatt, M. MacEachern, and G. Levow, Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication*, vol.24, 2, 1-23, (1998).

K. Hone, Animated Agents to reduce user frustration, in *The 19th British HCI Group Annual Conference*, Edinburgh, UK, 2005.

S. Oviatt, Interface techniques for minimizing disfluent input to spoken language systems, *in Proc. CHI'94*, pp. 205-210, Boston, ACM Press, 1994.

I. Poggi, From a Typology of Gestures to a Procedure for Gesture Production, *Gesture Workshop 2001*, pp 158-168, 2001.

J. Cassell, Y.I. Nakano, T.W. Bickmore, C.L. Sidner, and C. Rich, Non-verbal cues for discourse structure, *in Proceedings of the 39th Annual Meeting on Association For Computational Linguistics*, 2001.

N. Chovil, Discourse-Oriented Facial Displays in Conversation, *Research on Language and Social Interaction*, 25, 163-194, 1992.

A. Kendon, Conducting interaction: patterns of behaviour in focused encounters*, Cambridge Univer-sity Press*, 1990.

J. Cassell and K.R. Thorisson, The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, vol.13, pp.519-538, (1999).

R. San-Segundo, J.M. Montero, J. Ferreiros, R. Córdoba, and J.M. Pardo, Designing Confirmation Mechanisms and Error Recover Techniques in a Railway Information System for Spanish, *SIGDIAL*, Denmark, 2001.

S. Oviatt, and B. Adams, Designing and evaluating conversational interfaces with animated characters, *Embodied conversational agents*, MIT Press: 319-345, 2000.

H. Schaumburg, Computers as tools or as social actors: the users' perspective on anthropomorphic agents, *International Journal of Cooperative Information Systems*, pp 217-234, 2001.

R. Catrambone, J. Stasko, and J. Xiao, ECA as user interface paradigm, *From brows to trust: evaluating embodied conversational agents*, Kluwer Academic Publishers, Norwell, MA, 2004.

I. Bulyko, K. Kirchhoff, M. Ostendorf, and J. Goldberg, Error correction detection and response generation in a spoken dialogue system, *Speech Communication* 45, pp 271-288, 2005.

A. Sasse, *Usability and trust in information systems.* Cyber Trust & Crime Prevention Project, 2004.

S. Möller, P. Smeele, H. Boland, and J. Krebber, *Evaluating spoken dialogue systems according to de-facto standards: A case study.* Computer Speech & Language 21 (2007) 26-53.

J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone and R. S. Bhogal, The persona effect: affective impact of animated pedagogical agents, in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press New York, NY, USA, pp. 359-366, 1997.

J.H. Walker, L. Sproull, and R. Subramani, Using a human face in an interface, in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 85-91, ACM Press, 1994.

A.B. Loyall, and J. Bates, Personality-rich believable agents that use language. In Johnson, W.L., Hayes-Roth, B., eds.: Proceedings of the First International Conference on Autonomous Agents (Agents'97), Marina del Rey, CA, USA, ACM Press (1997) 106–113

N.C. Krämer, G. Bente, and J. Piesk, *The ghost in the machine. The influence of Embodied Conversational Agents on user expectations and user behaviour in a TV/VCR application.* IMC Workshop (2003) 121-128.

B. Reeves and C. Nass. *The media equation: How people treat computers, television and new media like real people and places.* CSLI Publications, Stanford,CA, 1996.

B. López, Á. Hernández, D. Díaz, R. Fernández, L. Hernández, and D. Torre, Design and validation of ECA gestures to improve dialogue system robustness, Workshop on Embodied Language Processing, in the *45th Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 67-74, Prague, 2007.

Á. Hernández, B. López, D. Díaz, R. Fernández, L. Hernández, and J. Caminero, A "person" in the interface: effects on user perceptions of multibiometrics, Workshop on Embodied Language Processing, in the *45th Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 33-40, Prague, 2007.