J Multimodal User Interfaces (2010) 3: 285–297
DOI 10.1007/s12193-010-0052-2

ORIGINAL PAPER

Non-verbal communication strategies to improve robustness in dialogue systems: a comparative study

David Pardo · Beatriz L. Mencia · Álvaro H. Trapote · Luis Hernández

Received: 11 November 2009 / Accepted: 28 August 2010 / Published online: 17 September 2010

Abstract This paper explores the use of embodied conversational agents (ECAs) and their visual communicative ability to improve interaction with spoken language dialogue systems (SLDSs) through an experimental case study in the application context of secure access by speaker verification followed by remote home automation control. After identifying a set of typical interaction problems with SLDSs and associated with each of them a particular ECA gesture or behaviour, we conducted a comparative evaluation based on ITU recommendations for the evaluation of spoken dialogue systems. User tests were carried out dividing the test users into two groups, each facing a different interface setup: one with an ECA, and the other only with voice output. The ECA group encountered fewer interaction problems. Users' impressions, however, were similar in both groups, with a slight advantage observed for the ECA group. In particular, the ECA seems to help users to better understand the flow of the dialogue and reduce confusion. Results also suggest that rejection (based on privacy and security concerns) is a dimension in its own right that may influence subjective evaluation parameters closely related to user acceptance.

Keywords Embodied conversational agents · Non-verbal communication · Robustness · Comparative evaluation · Spoken dialogue system · Voice authentication · Speaker verification · Gesture design

D. Pardo · B.L. Mencia (✉) · Á.H. Trapote · L. Hernández
Signal, Systems and Radio Communications Department, Universidad Politécnica de Madrid (UPM), Madrid, 28040, Spain
e-mail: beatriz@gaps.ssr.upm.es

D. Pardo
e-mail: david@gaps.ssr.upm.es

Á.H. Trapote
e-mail: alvaro@gaps.ssr.upm.es

1 Introduction

Justine Cassell describes Embodied Conversational Agents (ECAs) as 'virtual humans', featuring a visual representation, i.e., an animated figure—a face and sometimes also a body—, capable of proactively engaging in conversation with humans using speech, movement, gesture and facial expression [8]. Conversational systems are meant for engaging in 'natural' or 'humanlike' communication with users (see, e.g., [9] and [3], and ECAs aspire to emulate face-to-face conversation between humans.

Most research on ECAs has been directed towards the technical aspects of incorporating animation to a dialogue system output, towards endowing the ECA system with 'intelligence' and a greater sensitivity to contextual elements to decide the course of action to take (e.g., [26], and, finally, towards manipulating high-level communication or relational aspects of the interaction; in other words, these are linguistic and social interaction approaches (e.g. [29]). However, relatively little attention has been given to studying the role ECAs may play in reducing communication errors and improving the flow of the dialogue [1], in particular regarding low-level miscommunication problems related to the limitations of current speech recognition technology.

Typical spoken language dialogue systems (SLDSs) rely on automatic speech recognition (ASR), spoken language understanding, natural language generation, speech synthesis and on managing dialogue states—which basically involves deciding what to say or do next. Decisions are based on considerations such as which information items needed to accomplish the interaction goals are known, or how well the system understood the utterance from the user. The most common problems with SLDSs are rooted in the limitations of speech recognition, but they may affect other stages of the process toward the generation of the output. For instance, a

non-understanding is, first of all, an ASR problem (the ASR module fails to render any interpretation of the user's utterance), but it also creates a problem at the dialogue management level since it has to decide how to deal with not having understood what the user has said; it has to decide what to do next. There may be a variety of strategies available for a particular situation (e.g., asking the user to repeat or rephrase what she said, or repeating the system's previous utterance [5]), and the decision should be taken in accordance with the goal of maximising robustness to the problem encountered; i.e., detecting it and recovering from it, preventing further complications, and doing so with minimal disruption of the flow of the dialogue. Now, a critical problem with SLDSs is precisely their lack of robustness, as described profusely in the literature (see, e.g., [30] and [4]). Problems include miscommunication, turn-handling problems, and failure to convey clearly and intelligibly to the user the stage of the dialogue and how it is progressing.

There are different approaches to deal with these communication problems in dialogue systems. For instance, [25] focus their efforts in improving the natural language generation module with the Limsi Arise system, and [40] propose adapting the generated output to the users' preferences. The Mask Kiosk project [24] seeks to confer naturalness to a fairly complex dialogue system featuring touch and voice with NLG by incorporating a very simple avatar into a multimodal interface, but the avatar does not play an important or significant role in the interaction.

We have also worked on the idea of adding a virtual human to the interface, but instead of using a complex system and a simple avatar, we have designed a relatively simple dialogue system with a virtual human that plays an active role engaging with the user with specifically designed behavioural strategies for each type in a set of typical spoken interaction problems. Adding a visual channel that incorporates a more or less human-looking animated character—an embodied conversational agent, or ECA [11])—that personifies the user's interlocutor makes it possible, at least in theory, to convey 'natural' cues through gesture and general demeanour which may help the user follow the progress of interaction more easily. For instance, the ECA might display meta-cognitive gestures suggesting what it is 'thinking', reinforcing or complementing (adding to) the information conveyed verbally concerning the state of the dialogue. Indeed users have been found to prefer visual cues over verbal ones when they face uncertainty in dialogue. ECAs have also been found to have psychological effects (generally positive) merely by appearing in the interface, without it having to do anything in particular to help the users; a phenomenon known as the 'persona effect' [27].

We have sought to design gestural and verbal behaviour for our ECA for specific problem-prone dialogue situations, and to compare user interaction performance and user opinions—their subjective experience using the system—with and without an ECA in the interface (different users for each of these interfaces). It therefore seems reasonable to begin by adopting a user-centred approach to evaluation that goes beyond merely observing operational parameters that are close to the underlying technology, such as recognition error rates and number of turns needed to complete a specific unitary task. The focus needs to shift more toward how users perceive the interaction and how satisfied they are with the experience, one that, it is hoped, is more socially engaging.

Along these lines, the PARADISE evaluation scheme [41] proposes predicting users' subjective impressions from 'objective' metrics covering task success rates and a set of interaction 'costs'; Moeller et al. [31] have thought out a taxonomy relating quality factors that affect users' acceptance of a spoken dialogue system; and two ITU recommendations, on questionnaire design for subjective assessments and for filling a collection of objective parameters in order to quantify system and interaction performance [21, 22], put together, suggest combining system and interaction performance measures, as well as event data that is registered automatically, with users' responses to questionnaires. We should stress, however, that the approaches and standards we have just mentioned concern the evaluation of spoken dialogue systems. The conceptual categories on which they are based do not cover the nonverbal interaction elements that an ECA introduces.

The literature on evaluation of systems with ECAs usually focuses on measuring the users' subjective experience during the interaction. For instance, Cerrato and Ekeklint [14] propose new parameters and metrics to evaluate multimodal dialogue systems endowed with ECAs which are focused on the user rather than on the system, such as prosodic variation and communicative nonverbal behaviour. In [32] Noor Christoph proposes an empirical evaluation methodology using typical usability procedures, including questionnaires, interviews, logs files and biological measures, to evaluate systems with ECAs. Buisine et al. [6] take a more direct approach to evaluate ECAs comparatively, asking users about its technical and behavioural features (e.g., which ECA is more expressive? Which is the most likeable?). Also worthy of mention are Catrambone [13] efforts, which combine personal information about users (e.g., gender and personality) and their opinions regarding an ECA (e.g., 'The agent was friendly', 'annoying' or 'cold') with interaction performance measures. An interesting observation from their study was that the perception of the ECA was strongly influenced by the nature of the task). In fact, the variety of the experimental and evaluative approaches is mirrored in the variability of the results reported, which may have to do not only with varying features of the ECAs employed in each study but also with the differing nature of the tasks that motivate each interaction scenario [15].

We can conclude that there are different ways to evaluate interactive systems with ECAs but no standard evaluation method. What we have done is to take as close a guidance as possible from the abovementioned ITU-T recommendations [21, 22], adding a few ECA-specific items to our questionnaires in the same spirit as the dialogue-related categories in [21]. Within this methodological framework we have conducted, as mentioned above, a comparative user study with the same SLDS system with and without an ECA in the interface.

The rest of the article is organised as follows: Section 2 explains the dialogue strategies we have designed to increase robustness in a face-to-face communication system, and the ECA behaviour we have associated with them. Section 3 describes the experimental setup of our comparative study. Section 4 presents our data analysis and the main results of the experiment. Finally, conclusions form Sect. 5.

2 Dialogue and gesture strategies

Consider the following scenario: A multimodal dialogue researcher is away from home attending a conference. He'll be back home in the evening, but having been away for a week he fears he'll find the house too cold. So, he calls his remote home automation control system to turn the heating on so that the house will be warm when he arrives. When the system answers, an ECA, appears on the screen of his phone and asks him to undergo a biometric authentication procedure, specifically, speaker (or voice) recognition. After providing a few voice samples the system lets him through. The user then asks the ECA to turn the heating on, and when this is done the ECA provides the appropriate feedback.

Our test system was designed according to this straightforward two-phase scenario, which speaker verification followed by a simulated remote home automation control application which the test users were asked to query.

Before speaker verification phase a user enrols into the system. Enrolment is implemented as a dialogue in which the system asks the user to repeat four random four-digit sequences, one after the other. Speech recognition problems may affect the course of the dialogue since each sequence uttered by the user must match the requested number sequence before the speaker verification unit produces the speaker's voiceprint. Speaker verification is similar to enrolment, the goal being to grant or deny access to the main application.

The main application is a service with which users call 'home' using mobile phones (simulated on a computer screen) to check the state of various home appliances. At the beginning of each 'call' the system speaks a welcome message and reminds the user of what the application does: provide information about various home devices—lights, TV, fan, heater, etc.—and/or carry out actions with them—e.g., switch them on or off. The communicational goal in each call is for the system to collect the three information items that form the user's request: a device, a location (a room) and an action. Once the system is confident it has correctly understood of which device in which room the user wants to do which action, it answers appropriately.

Table 1 shows an example of a complete interaction through both phases: the speaker verification and the home automation dialogues. We highlight some of the dialogue situations for which we have designed a behavioural strategy, which we describe in the next sub-section.

Dialogue difficulties related to speech recognition include non-understandings (the system fails to put together an intelligible reckoning of the user's utterance), misunderstandings (what the system understands is not what the user has said), no-inputs (the user doesn't speak when the system expects her to, or she does and the system doesn't realise) and dealing with low confidence in recognition (what to do when the system is unsure that its interpretation of the user's utterance is correct).

When recognition difficulties are detected it is advisable to initiate a recovery process to find out what the correct communicative intention of the user is. Unfortunately, this may produce frustration in the user [17]. In fact, once an error occurs it is common to enter an error spiral, because as the user becomes increasingly frustrated the system will typically find it more and more difficult to understand her, so further errors are made [34, 35]. A system's robustness is a matter of how it deals with all of these situations.

But in order to improve the robustness of a spoken dialogue interface is necessary to take into account not only the accuracy and reliability of the speech recognition engine but also the capabilities of the dialog manager and the design of the interaction flow. In particular, turn-taking and interruption handling provide another family of problems that fit this slightly broader notion of robustness. For the dialogue to flow smoothly turn possession must be clearly understood and turn changes seamlessly negotiated. Uncertainty in this area is connected with confusion in the user's mind regarding the state of the dialogue, i.e., what information the system has already for the purpose of reaching the objectives of the interaction, what it expects the user to do (provide a piece of new information? confirm information that was already given? repeat or rephrase an utterance? wait until the system says or does something?), or whether the dialogue is progressing normally or some sort of failure has occurred.

These are the kinds of problems we have focused on for the purpose of observing whether an ECA in the interface, displaying behaviour designed for each particular situation, can make a difference in dialogue performance and/or in the way users perceive the system and their interaction with it.

Broadly speaking, we can gather the previous problem situations in the following three categories which will be useful for conceptualising the ECA's responses in each case:

Table 1 Complete Dialogue Example

Dialogue	Situation
	User calls the remote home automation control system.
SYSTEM: *In order to grant you access you this system I need to verify your identity. Please repeat the following sequence of numbers: five, four, six, eight.*	Initiation of the speaker verification task. The ECA reads out the number sequence marking the regular tempo with its hands.
USER: *Five, four, six, eight.* SYSTEM: *I'm sorry. I've heard the wrong number sequence. Could you repeat, please?*	Wrong number sequence recognised (the system 'heard' '*five, four, three, eight*').
USER: *Five, four, six, eight*. SYSTEM: *Now please repeat the following number sequence: two, three nine, four*.	The acoustic features extracted from the user's utterance do not match the template recorded at the time of the user's enrolment closely enough. Response: the system asks the user to produce another spoken sequence (but does not explain the reason so as not to worry or frustrate the user).
USER: *Two, three nine, four*. SYSTEM: *Positive authentication*	There is a good enough match with the registered acoustic features. The task is successful.
SYSTEM: *Welcome to the remote home automation control system, Guillermo. What can I do for you?*	Initiation of the remote home automation control task.
USER: *I want to know if the lights in the bathroom are switched off*.	Requesting information regarding the status (on/off) of a device.
SYSTEM: *In which room do you want to check whether the lights are switched off?*	Failure to understand one of the information slots, and implicit confirmation of the rest of the user's request.
USER: *In the bathroom.* SYSTEM: *Could you say that again, please?*	Non-understanding.
USER: *In the bathroom*. SYSTEM: *Did you ask about the bedroom lights?*	Low confidence in the recognised utterance. Response: explicit confirmation.
USER: *No*. SYSTEM: *I'm sorry. I misunderstood. Could you tell me again what you want?*	Acknowledgement of misunderstanding.
USER: *Are the bathroom lights off?* SYSTEM: *The lights in the bathroom are switched on*.	High confidence in the complete recognised utterance. Response: answer the query.

– Turn management: Here the body language and expressiveness of agents could be exploited to help regulate the flow of the dialogue [2, 43].
– Recognition error recovery and low confidence in the recognition result: ECAs may help limit such feelings of frustration and make error recovery more effective [19].
– User confusion: The expressive capacity of ECAs could be used to help the user keep track of what the system is doing and expecting from the user.

We have designed 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. Table 2 shows each dialogue stage, what prompts it, and the associated ECA behaviour. The gesture repertoire of our ECA is partially based on relevant gestures described in [2] and [11], and on recommendations in [11, 12, 23], and [37], to which we have added a few suggestions of our own.

2.1 Dialogue strategy and ECA gesture scheme

We designed a set of ECA behaviours and animation effects to respond to the problem situations in dialogue briefly described in the previous section. In [28] we have already reported user validation tests for each of the ECA gesture sequences we have designed (following [18]). We first asked the users to watch a system simulation (a video recording of a user interacting with the system), so that they could see the ECA performing the gestures in the context of a real dialogue. After watching the simulation the users were asked to fill out a questionnaire. The questionnaire allowed users to view isolated clips of each of the dialogue gestures. To each gesture clip were associated questions basically covering the aspects of gesture interpretation, gesture design and user expectations. We found that the gestures worked acceptably well with the test subjects: overall, they were correctly interpreted, regarded as natural, and they did not generate undesirable or unexpected effects. Finally, we gave due consideration to useful comments such as, '*I would like it better if the ECA would move faster when switching dialogue turns*'), which led only to some minor changes. The results can be viewed in a video sample in.[1]

We now describe the main behavioural strategies that were devised and validated. The ECA's verbal and gestural behaviour for each stage is summarised in Table 2.

[1] http://www.gaps.ssr.upm.es/es/investigacion/interaccion-multimodal.

Table 2 Critical dialogue situations and associated ECA's behaviour strategies

Dialogue stage	Description	Gestures design references	ECA behaviour
Initiation	At the beginning of the dialogue [13, 33, 38]	[23] Kendon	Look straight at the camera, smile, wave hand. Zoom in for task explanation. Start speaking. Zoom out, lights dim. Stop speaking.
Turn management	Take Turn: when the system starts to speak [7]	[2, 11] Cassell and Own suggestion	Take Turn: look straight at the camera, raise hand into gesture space. Camera zooms in. Start speaking. Light gets brighter.
	Give Turn: when the system prepares to listen to the user [7]	[2, 11] Cassell and Own suggestion	Give Turn: look straight at the camera, raise eyebrows. Camera zooms out. Stop speaking. Lights dim.
Wait	When a timeout occurs	Own suggestion (inspired by [16])	Slight leaning back, one arm crossed and the other touching the cheek. Shift of body weight.
Non-understanding	When the system cannot understand something the user has said [37]	Own suggestion (inspired by Pelachaud [36])	Slight leaning of the head to one side, stop smiling, mildly squint. Start speaking
Confirmation (high confidence)	The system has recognised the user utterance with a high level of certainty [37]	Cassell [10] (show willingness to chat)	Start speaking. Nod gesture, smile, eyes fully open. Stop speaking
Acknowledgement misunderstanding	After user informs the system that it has misunderstood what he or she has said.	Own suggestion (inspired by Pelachaud [36])	Apology: Head aside, raise inner eyebrow central, head down, eyebrow of sadness (to show remorse). Start speaking
	Speech: a) apology; b) repetition or rephrase request		Request: Show expression of interest by opening eyes, and smiling slightly. Stop speaking
Error recovery with correction (see 'Confidence in the accuracy recognition')	When the user has corrected a recognition error and the system confirms the correction	Own suggestion (inspired by Cassell [12] and Pelachaud [36])	Start speaking. Lean towards the camera, beat gesture. Stop speaking
Speaker Verification Failure	When the user's identity could not be verified.	Own suggestion	Sad and remorseful expression for not having been able to verify the user's identity, accompanying a verbal message informing the user of the failure.
Wrong sequence of numbers recognised	The system *believes* to have *understood* a sequence of numbers uttered by the user, but it is not the one requested	Own suggestion (inspired by Pelachaud [36]—remorse)	Same behavioural sequence as for *Acknowledgement of misunderstanding*.

Initiation. The inclusion of an ECA at this stage can 'humanise' the system [33] to the users. This can make users overly optimistic about the conversational capabilities of the system, leading ultimately to disappointment with the interaction. The ECA might also be distracting and hinder the users' concentration (new users especially) on the goals of the interaction [13, 38]. The purpose of a behaviour sequence at dialogue initiation should therefore be to present a human-like interface that captures the attention of the users, focuses it on the interaction goal and directs users straight to its pursuit. We have designed a welcome gesture based on the recommendations in [23]: a smile and a wave of the hand.

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.

The gestural behaviour we have designed for our ECAs 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. When it's the user's turn the camera zooms out slightly, lights dim and ECA raises an eyebrow. With these sequences we aim to help the user associate different gestures, camera shots and levels of light intensity with each of the turns.

Confidence in the accuracy of recognition. 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 interpre-

Fig. 1 Behavioural routine for explicit confirmation

tation of the user's utterance [37]. If the system's confidence in having understood the user correctly is low, on the other hand, the recovery strategy is one of explicit confirmation (e.g., '*Did you ask about the bathroom lights?*').

A common problem with explicit confirmation in these cases is that users will often reply with a short utterance like '*yes*' or '*no*'. Such utterances are hard for the system to detect, and easy to get wrong if it does detect them. In order to prevent the occurrence of this sort of problem we added a second part to the ECAs behaviour: the ECA stresses, '*Am I correct?*', and accompanies this utterance with a gesture designed both to show interest and to engage optimistically with the user: it leans forward slightly to catch the user's attention and to enhance the social engagement and enter a space of mutual trust. The gesture combines elements from the I propose intention gesture in—'head forward, raised eyebrow, look at [the user]'—and the hope affective state gesture—'raised eyebrow, large eye aperture'—suggested in [36]. The object of this optimistic stressing of the utterance is to elevate the user's motivation and induce him to increase the energy, and perhaps also the length, of her response.

If the user points out that recognition errors have taken place and gives the correct information at the same time (e.g., '*No, I mean the bathroom lights.*'), the ECA repeats the corrected information emphasising 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): '*So, do you want to know if the [ECA beat gesture begins] lights in the bathroom [ECA beat gesture ends] are switched off?*'

Non-understanding. A non-understanding occurs when the system fails to obtain a representation of the user's utterance that is intelligible within the scope of its grammar [5]. When this happens it is important to make sure that the user realises what has happened, and also to get him to try again. But we also want to avoid a build up of anxiety, frustration or disappointment in the user.

The ECA stops smiling, brings its head forward very slightly, and turns it to one side while mildly squinting. The goal of this behaviour is to show the user that the system is making an effort to understand and to recover from the situation—which is taken seriously, hence the losing of the smile—, sending the encouraging message that the system probably expects the problem will be easily solved. Our gesture takes some elements from the wh-question intention gesture and the sorry-for and sadness affective state gestures suggested in [36]. The verbal response that accompanies the gesture sequence is chosen randomly from one of the following: a. '*Could you say that again, please?*' b. '*I didn't catch that. Please tell me what you want again.*'

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. (The system realises immediately in the verification task if it 'hears' the wrong number sequence. See the dialogue example in Table 1). The ECA 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 ('*I'm sorry, I didn't catch that. Please could you tell me what you want again?*'). A two-part behaviour is synchronised with these utterances: first the ECA lowers its head and puts on a 'eyebrow of sadness'—elements that [36] has associated with remorse; then it lifts its head up, opens its eyes and smiles while it repeats the request.

Waiting. It sometimes happens that the user doesn't realise it is her turn to speak. This is one possible cause of a *no-input* situation, i.e., one in which the system is waiting for the user to say something for a lapse of time long enough to indicate that there may be a problem. No-inputs can occur also for other reasons, for instance when the user speaks while the ASR module is inactive.

The system decides there is a no-input situation when there has been a timeout. At this point the ECA does nothing. To help the user realise the system is waiting for her to say something, after a second timeout the ECA performs the following gesture sequence: it leans back slightly, crosses its arms and brings its left hand to its face, and it shifts its 'weight' slightly onto its left leg. (This gesture sequence is largely our own design, but we drew inspiration from [16]).

Speaker verification failure. If the system is unable to verify the identity of the user after one or two valid attempts (i.e., when the recognised utterance corresponds to the number sequence the system requested), the ECA kindly informs the user ('*I wasn't able to verify your identity*') with a sad expression on its face to show empathy with the user, aiming to control her frustration and keep her in a calm mood. Again, our goal is to see whether an empathetic ECA generates a better subjective experience even when a task ends in failure.

3 Comparative study: interaction with and without an ECA

We now present an evaluation of the performance and the users' subjective experience with the dialogue system outlined in the scenario description in the previous section. To reiterate, the two functions offered by the system are identity verification through voice and remote home automation control. Our approach is comparative: we set up two versions of the same system, one with an ECA in the interface (i.e., visually present on screen) and the other with speech output only, divided our test users between them and contrasted the interaction performance results and questionnaire responses for both groups.

We should emphasise that both interaction set ups (with and without ECA) have the same underlying SLDS (speech recogniser, speech output, etc.) and the same spoken dialogue engine. The only difference is the visual communication channel with the ECA in one of the setups.

We have relied on Nuance Communications' technology[2] to implement the speech recognition. The animated character was designed using tools created by Haptek.[3]

We tested the system with 40 users divided into two groups. Due to system errors happened during the interaction two users (1 from each group) were excluded because we considered the results were conditioned by the bad performing of the system. So we obtained results from 38 people (19 users in each group). One of the groups has an ECA integrated in the interface (ECA metaphor) and the other was implemented just only with voice (VOICE methaphor). Users were mainly students (28 out of the 38–73.7%) and under 24 years of age (25 out of the 38–68.5%). 16 were female and 22 male. Regarding their prior experience with virtual agents, about 60% of the users have ever used a system where an agent is included. Finally, a high percentage of users (81.6%) indicated they had previously used traditional spoken dialogue systems. However, only 52.6% of our test users had some prior use experience with SLDSs that had natural language capabilities. Regarding user satisfaction with SLDS technology, only 13.2% of the users describe their previous interaction experiences as having been 'good'.

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 behaviour: 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.

As the example in Sect. 2 shows, we regard the speaker verification and the home automation as two phases of one conceptual dialogue whole. However, in testing we divided this 'whole' in order to capture the users' subjective impressions at each stage of the dialogue.

The evaluation procedure was as follows: Test users were given a brief explanation of the general purpose (to 'evaluate automatic dialogue systems') and methodology of the evaluation, and of the tasks that lay ahead for him/her. An opening questionnaire registered the users' prior experience and expectations. The test proper began with the enrolment phase in which the test users enrolled in the speaker verification system. This was followed by a verification phase in which users were asked to perform three successive speaker verifications. Finally, a remote home automation control phase required users to find out, in succession, the state (on-off) of three household devices ('the bathroom lights', 'the fan in the bedroom', and 'the living-room television set'). At the end of the entire interaction procedure divided into the three main phases we have just seen, the test users completed a questionnaire that registered their impressions regarding the system, its main elements and the most important aspects of using it.

4 Analysis and discussion of results

In this section we present some of the more interesting results we have obtained from comparing the performance and the users' subjective experience of user–system interaction with ECAs and with a voice-only interface, both scenarios and dialogue systems otherwise being identical.

First we present results related with the performance of the interactions in both interface setup groups. Secondly we present analyses of the users' subjective impressions as collected in the questionnaires. Finally we draw some relationships (correlation analyses) between these two kinds of parameters: 'objective' and 'subjective'.

4.1 Interaction performance parameters

ITU-T Suppl. 24 to P-Series Recommendation [22] lists the most common parameters for describing the interaction with a spoken dialogue system. We have selected parameters from this recommendation and adapted them to carry out a comparative assessment of interaction performance. Table 3 summarises our results. To obtain the figures in the table we

[2] http://www.nuance.com.

[3] http://www.haptek.com.

Table 3 Interaction parameters for the ECA and voice-only setups

Dialogue- and communication-related interaction parameters			
Abbrev.	Description	With ECA	Voice-only
DD	Average dialogue duration (all sub-dialogues counted) (sec.)	343.15	368.45
STD	Average system turn duration (ms.)	7173	6571
UTD	Aveage user turn duration (ms.)	3573	3873
# System turns	Accumulated number of system turns throughout all interactions (all users)	517	556
# User turns	Accumulated number of user turns throughout all interactions (all users)	416	499
# Turns	Accumulated number of turns throughout all interactions (all users)	933	1055
Meta-communication-related interaction parameters			
Abbrev.	Description	With ECA	Voice-only
#Time-out	Number of time-out prompts (after 3 successive 7 second internal time-out periods)	7	15
#ASR Rejection	Number of speech recogniser rejections	22	32
Task-related interaction parameters			
Abbrev.	Description	With ECA	Voice-only
TS	Task success rate (%)	98.25	98.25
Speech-input-related interaction parameters			
Abbrev.	Description	With ECA	Voice-only
WER	Word Error Rate (%)	35.73	34.00
%PA:CO	Percentage of correctly parsed user utterances	76.29	68.38
%PA:PA	Percentage of partially parsed user utterances	5.67	5.38
%PA:IC	Percentage of failed-to-be-parsed user utterances	18.04	26.24

pooled the data from the three stages of the test dialogue: enrolment, verification (three successive dialogued verification tasks) and home automation control (three successive device queries). All dialogue parameters in the table refer to this overall dialogue (composed of the sub-dialogues we have just mentioned). The figures are averages or totals (depending on the parameter) for all users in the respective test groups, and include all of the interactions accumulated from both phases in the test scenario (biometric access and home automation).

Table 3 shows that interactions with the ECA were, on average, slightly shorter in duration than with the voice-only interface, and they also yielded fewer turns. We conducted Mann-Whitney U tests and found that the differences in user turns ($Median_{ECA} = 21$; $Median_{voice\text{-}only} = 26$) and overall number of turns ($Mdn_{ECA} = 47$; $Mdn_{voice\text{-}only} = 54$) were statistically significant ($U_{user\text{-}turns} = 94$, $p < .01$, $r = -.41$; $U_{turns\text{-}total} = 112$, $p < .05$, $r = -.33$).

The total number of time-out prompts (after the system has been waiting for the user to say something) with the voice-only interface was twice that with the ECA, and the number of user utterances for which the speech recogniser was unable to provide any interpretation was 50% higher in the voice-only group than in the ECA group. These differences were not statistically significant. More data with instances of timeouts and ASR rejections would be needed to confirm an effect.

Word error rates were very similar in both test groups. However, statistically significant differences were found in the dialogue system's understanding of the users' utterances. Number sequences (in the secure access phase), queries (in the home automation control phase), confirmations and corrections given by the user could either be correctly or incorrectly understood (adding to the PA:CO and PA:IC figures, respectively). In addition, in the home automation sub-dialogues queries could also be partially correctly understood (when only one of the two information slots specified by the user—device and room—was correctly understood by the system): the relative number of times this happens gives the %PA:PA figure. The number incorrectly parsed user utterances (we may call this #PA:IC) was 70 for the ECA interface group (%PA:IC = 18.04% of all utterances in the group) and 112 for the voice-only interface group (%PA:IC = 26.23% of all utterances in the group). Both the difference in number of incorrectly parsed utterances (#PA:IC: $Mdn_{ECA} = 3$; $Mdn_{voice\text{-}only} = 5$) and the difference in the proportion of the same relative to the total number of user utterances detected by the system (%PA:IC:

$Mdn_{ECA} = 0.15$; $Mdn_{voice\text{-}only} = 0.22$) were statistically significant ($U_{\%PA:IC} = 101$, $p < .05$, $r = -.38$; $U_{\%PA:IC} = 98.5$, $p < .05$, $r = -.39$).

Inspired in the smoothness parameter proposed in [42], we computed a parameter we called *roughness* derived from several performance parameters related to complications in the interaction, to give an overall measure of these. Specifically,

$$\text{Roughness} = \#\text{ASR rejection} + \#\text{PA:PA} + \#\text{PA:IC} + \#\text{Time-out}$$

The median value of roughness for the interactions in the ECA group was 5 vs. 8 for the voice-only group. A Mann-Whitney test revealed that the difference is statistically significant ($U = 110$; p (exact, two-tailed) $< .05$; $r = -.34$).

We may infer from these general observations that the ECA is helping users interact more efficiently with the system, as a result of their having, generally, fewer problems and recovering more easily from them—i.e., the system is more robust. The reason may be that the information provided by the ECA throughout the interaction reduces user confusion regarding the state of the dialogue at any particular moment.

Effectiveness in reaching the interaction goals was identical in both groups: almost all of the tasks with the home automation system (the queries) were completed successfully.

We recorded another parameter to measure the users' visual attention. This is important in studies with visual elements on system interfaces. In the case of ECAs performing gestures, the point of these may be lost if the user isn't looking, which reduces the confidence in interpretations of differences in comparative studies based on such visual information. It is also interesting to consider visual attention as a measure of cognitive attention and engagement with the task.

Inspecting the video recordings of the test interactions we computed the number of times users turned their gaze away from the screen for longer than three seconds, and we assumed that when users in the ECA group were directing their gaze towards the screen they were looking at the ECA. Such visual attention loss was observed on 22 occasions in the ECA group vs. 37 in the voice-only group. These are low numbers considering that each of the nineteen users in each group was required to carry out seven tasks (seven sub-dialogues), and the difference was not statistically significant. Nevertheless, the difference between the groups may be an indication that the ECA draws the users' attention and helps her focus on the task, which could be a factor behind the improved interaction performance. More data is needed to confirm these effects, however.

In any case, the figures show that users were looking at the screen most of the time, an observation which allows us to consider (though obviously not, by itself, confirm) the visual behaviour of the ECA as a possible factor in the explanation of interaction performance differences as well as differences in the users' expressed subjective experience. To the latter we turn our attention in the following sub-section.

4.2 Subjective parameters

The questionnaires were designed following recommendations by the ITU-T and also drawing items from Hone and Graham's SASSI tool [20]. In this subsection, first we present general observations with relation to the data organised in the categories suggested by the former, and then we present a factor analysis inspired in the methods of the latter.

4.2.1 Approach #1: categories recommended by the ITU-T

ITU-T Suppl. 24 to P-Series Rec [21] provides guidelines regarding the sort of aspects that should be covered when studying users' impressions of spoken dialogue systems. Following this recommendation we grouped our questionnaire items in seven blocks (or categories) as follows (with one or two of the questions in each block given, in brackets, as an example):

- G1: Overall impression ('*Rate your overall impression of the system.*')
- G2: Personal impression ('*I felt bored while using the system*')
- G3: Information obtained from the system ('The system always gave me the information I wanted.')
- G4: Behaviour of the system ('*The system made many mistakes.*')
- G5: Dialogue ('*The dialogue was too long.*' '*It was easy to understand what the system was saying.*')
- G6: Perceived system personality ('*The agent was polite.*' '*The agent expressive.*')
- G7: Expected future use of the system and perceived usability ('*I would use this system/service.*')

We paid special attention to the effect an ECA might have in blocks G2 and G6, in which 'humanising' effects could be reflected. Specifically, in block G2 we included a set of items on the feelings and basic emotions evoked in the user through the interaction. In block G6 there are items covering the personality attributed to the conversational agent and the quality and naturalness of the gestural behaviour. Questionnaire responses were collected on Likert-type 5-point response formats.

We compared the average scores for each of these blocks of questionnaire items across the two experimental conditions (interface with and without an ECA). Table 4 shows the means per question block and user group. On average, scores did not differ greatly. In fact, no statistically significant differences were found with the per-block averages.

Table 4 Mean values for each block and interface group

Block	Interface	Mean (stand. dev.)
G1	ECA	3.89 (0.74)
	Voice-only	3.79 (1.08)
G2	ECA	4.08 (0.36)
	Voice-only	3.59 (0.51)
G3	ECA	4.43 (0.53)
	Voice-only	4.38 (0.55)
G4	ECA	3.53 (0.52)
	Voice-only	3.37 (0.59)
G5	ECA	4.04 (0.53)
	Voice-only	3.93 (0.57)
G6	ECA	3.69 (0.54)
	Voice-only	–
G7	ECA	4.20 (0.51)
	Voice-only	4.20 (0.50)

One item addressed whether the system had surprised users negatively or positively (response categories: 1 = 'very negatively' ... 5 = 'very positively'). The responses show that voice-only interface users were very slightly more positively surprised by the system's dialoguing capability than ECA users ($\mu_{ECA} = 3.9$, $\mu_{voice} = 4.3$; medians were 4 for both groups). The differences were not quite statistically significant at the $p = .05$ level ($U = 121.5$, two-tailed exact $p = .073$). Still, the observation is interesting, especially in view of the fact that in terms of actual dialogue performance, the interactions of the voice-only group fared objectively worse than those of the ECA group, as we observed above.

This observation suggest that ECAs may generate greater expectations which lead them subsequently to being less impressed with the system's performance compared to voice-only users, even if on average they had fewer interaction problems.

4.2.2 Approach #2: factor analysis on user responses

To better understand the dimensions involved in the subjective evaluations of our test users, we performed an exploratory factor analysis—specifically, Principal Component Analysis (PCA)—on the questionnaire filled out by our users at the end of their interactions with the system (as explained in Sect. 3). The questionnaire contained a fairly large number of items (58 for the voice-only user group and 68 for the ECA group) addressing a wide variety of aspects.

For a sample as small as ours (38 users), it has been suggested that only factor loadings greater than 0.7 can be considered significant [39]. In some cases we 9 have, however, taken into account questions loading only 0.6 when interpreting the meaning of factors. We found that the interpretation of the factors was easier and more coherent when we did include such items, and our choice is further justified by the fact that we obtained increased alpha scores of internal consistency for the corresponding scales. Partly because of the limited number of user cases we had to work with, before reaching stable factor solutions, questionnaire items had to be screened out either because they were found not to correlate significantly with the other items or because they did not load significantly on any factor resulting from the analysis, or in order to obtain an acceptable value for the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy.

The stable solution we report here is shown in Table 5. (KMO: .71; data-set variance accounted for: 72%; Varimax rotation applied.) We interpreted the five factors extracted as follows: C1. Dialogue robustness and ease of use. C2. Perceived overall performance quality (note that it is also related to the perceived 'naturalness' of the interaction). C3. Habitability. We borrowed the term from the corresponding factor extracted by Hone and Graham with their SASSI tool [20]. They explain: ' 'Habitability' refers to the extent to which the user knows what to do and knows what the system is doing. It can be understood in terms of the adequacy of the user's conceptual model of the system as a dialogue partner.' C4. Dialogue speed and flow. C5. Perceived usefulness and inclination to use the system. Interestingly, this component is dominated by an item on privacy concerns, which suggests that such concerns plays a prominent role in the users' practical acceptance of the system (in terms of how interested they may be in using it). Factors 1 and 2 combined can be interpreted as 'system response accuracy', to use Hone and Graham's [20] terminology. They relate to how well the system is perceived by users to do what it is supposed to do, bearing in mind that our users seem to regard the system primarily as a speech recognition system. We have already seen that habitability is a factor also identified in Hone and Graham's study, as is 'speed'. This is encouraging, and it suggests that our dialogue system, with a visual communication channel for half of the users, is perceived along general dimensions already found for a speech-only system.

4.3 Correlations between objective and subjective data

Correlation analyses involving performance data and questionnaire responses enable us to gain useful insights regarding the connections between interaction particulars (e.g. system comprehension, smoothness, etc.), users' impressions regarding such particulars and the users' overall valuation of the system.

In our tests we found that the users' overall impression regarding the system (the data from both user groups considered together) was correlated with the roughness parameter we computed (see Sect. 4.1) ($r_S = -.33$; $p < .05$). This general (albeit weak) trend makes sense: the better the interaction goes we should expect a better overall opinion of a dialogue system. Furthermore, robustness was also correlated

Table 5 Exploratory Factor Analysis results for the final questionnaire

Questions*	C1[a]	C2[b]	C3[c]	C4[d]	C5[e]
Whenever the system misunderstood you, you found it easy to solve the problem and get the system to finally understand you	.813				
It was easy to learn how to speak to the system	.781				
It is easy to use the system once you have learnt how to handle it	.718				
The system make many mistakes	.614				
The interaction felt natural		.876			
The system worked well		.719			
It is comfortable to remotely control home devices with this system		(.541)			
You always knew what to say to the system			.855		
You sometimes lost control of the dialogue and didn't know what was going on			−.814		
It is easy to learn how to use the system			.705		
The pace of interaction was fast				.790	
The dialogue was too long				−.762	
The interaction with the system flowed nicely				.624	
You would feel uncomfortable using the remote control system for home devices because it may compromise your privacy					−.844
The pros you see in the system outweigh the cons					.639
A remote domotic control system such as this one is useful					.614
You would use a system/service like the one we have described					(.540)

*Five-point Likert response formats, with anchors 1 = strongly disagree; 5 = strongly agree. *Only loadings higher than 0.6 are shown. (Two items load lower than 0.6 on all factors. For these we show the highest loading in brackets)

[a]Dialogue robustness and easy of use

[b]Perceived overall performance quality and naturalness of the interaction

[c]Habitability

[d]Dialogue speed and flow

[e]Perceived usefulness and inclination to use the system (future use)

(overall), with the composite variable computed for component C1 from the PCA for the final questionnaire ('Robustness and ease of use'), which we may plausibly, if somewhat loosely, regard as the subjective counterpart of our robustness parameter ($r_S = -.41$; $p < .05$).

More interestingly, the comparative aspect of the evaluation is offered in Figs. 2(a) and (b), which show, for the voice-only and the ECA groups of users respectively, the diagram of correlations between robustness (R), the 'Robustness/Ease of use' PCA component (RE) and overall impression (OI). Roughness is now only correlated with RE in the voice group while it is only correlated with OI in the ECA group. Thus, the overall impression of voice-only users seems to be directly related with their subjective impression of roughness, rather than with actual roughness (i.e., with an objective measure of it). The overall impression of ECA users is directly related to both actual and perceived roughness. Interestingly, for the ECA group actual and perceived roughness are not correlated. This may be due to the fact that there was less variability in the roughness of the interactions within the ECA group to relate with the variability in RE (again, a sort of 'perceived roughness' factor) which is roughly the same as in the voice-only group. Indeed, the test users responded to the questions related to the RE component rather similarly, even though there was a significant difference in the roughness of the interactions of both groups. This, in turn, may have to do with the fact that the test users only got to interact with one of the interface setups, and their similar evaluations reflect the lack of points of reference against which to rate their answers.

These notes of caution notwithstanding, it is plausible that smoother interaction was only correlated significantly to the overall subjective experience in the ECA group—the group that actually fared better with the dialogue system—because while the ECA may tends to help the interaction flow better, it also makes users more sensitive to interaction problems when they do occur (note that this would contradict a previous interpretation in Sect. 4.2.2 that the ECA

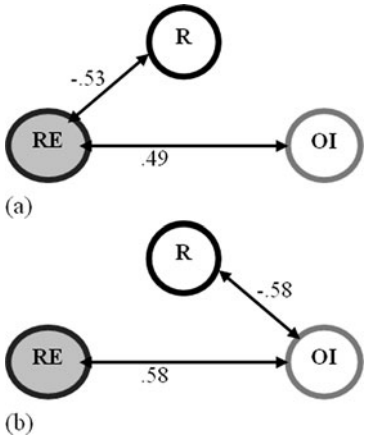

Fig. 2 Correlation diagram for robustness (R), component 'Robustness/Ease of use' and overall impression, for the voice-only group (**a**) and the ECA group (**b**). *Arrows* indicate correlations; figures by the *arrows* are the corresponding Spearman correlation coefficient

would be instilling patience in users). This effect could also be connected to the greater expectations users have of dialogue systems with ECAs, which are betrayed when there are errors and problems recovering from them.

5 Conclusions and future work

In this paper we have presented a comparative study of a speech-output-only dialogue system and the same enriched with visual communication—an embodied conversational agent that performs context-appropriate gestures. We have focused specifically on empathic gesture sequences performed when interaction problems (mainly ASR-related) occur in the dialogue.

Standard objective interaction parameters reveal the dialogue works better with the ECA, especially with regard to turn management and error recovery, which are basic indicators of robustness. This suggests that the ECA may be helping users understand what is going on and what they are supposed to do (say) throughout the dialogue.

Following the ITU P-851 recommendation we collected the impressions of the test users in questionnaires, dividing the questions into seven categories. Results do not show a clear difference in measures of the subjective experience of the users, despite the differences observed in the important objective parameters just mentioned. ECA users were slightly more positive about the experience, but this does not translate into a stronger inclination to use a similar system in the future (which is commonly regarded as an important indicator of user acceptance).

We then tried a factor analysis approach to explore our questionnaires, which gave us extra insight into the structure of experiential factors that are related ultimately with user acceptance and the dynamics of the users' experience. Our analyses reveal similar factor structures to those identified by other authors (especially Hone and Graham [20]): system response accuracy, habitability and speed are good examples of such factors. Privacy concerns were strongly coupled with perception of usefulness and inclination to use the system, which suggests paying special attention to such concerns.

Finally, correlation analyses revealed that while in the voice-only interface user group 'objective' dialogue roughness was correlated with a composite variable based on subjective valuations (questionnaire items) of interaction problems but not with the users' overall impression regarding the system, for the ECA user group the opposite is true. This may be a consequence of the slightly smaller number of interaction problems in the ECA group, or perhaps another indication that our ECA induces heightened expectations in users that the system is subsequently unable to live up to.

Clearly, further testing is needed to confirm the tendencies and, especially, our interpretations of the results outlined in this paper. Nevertheless, we believe the insights discussed provide a sound foundation and directions for further inquiry.

Acknowledgements The activities described in this paper were funded by the Spanish Ministry of Science and Technology as part of the TEC2009-14719-C02-02 project.

References

1. Bell L, Gustafson J (2003) Child and adult speaker adaptation during error resolution in a publicly available spoken dialogue system. In: 8th European conference on speech communication and technology-EUROSPEECH 2003, ISCA, Geneva, Switzerland, pp 613–616
2. Bickmore T, Cassell J (2005) Social dialogue with embodied conversational agents. In: Jan van Kuppevelt LD, Bernsen NO (eds) Advances in natural multimodal dialogue systems. Text, speech and language technology, vol 30. Springer, Berlin, pp 23–54
3. Bickmore TW (2004) Unspoken rules of spoken interaction. Commun ACM 47(4):38–44. Special issue: Human-computer etiquette: managing expectations with intentional agents
4. Bohus D, Rudnicky AI (2005) Sorry, I didn't catch that!-an investigation of non-understanding errors and recovery strategies. In: 6th SIGdial workshop on discourse and dialogue, ISCA
5. Bohus D, Rudnicky AI (2008) Sorry, I didn't catch that! an investigation of non-understanding errors and recovery strategies. In: Dybkjær L, Minker W (eds) Recent trends in discourse and dialogue. Text, speech and language technology, vol 39. Springer, Amsterdam, pp 123–154
6. Buisine S, Abrilian S, Martin J (2004) Evaluation of multimodal behaviour of embodied agents. In: Ruttkay Z, Pelachaud C (eds) From brows to trust: evaluating embodied conversational agents. Springer, Berlin, pp 217–238
7. Bulyko I, Kirchhoff K, Ostendorf M, Goldberg J (2005) Error-correction detection and response generation in a spoken dialogue system. Speech Commun 45(3):271–288
8. Cassell J (2000) Embodied conversational agents. MIT Press, Cambridge
9. Cassell J (2000) Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents. In: Embodied conversational agents. MIT Press, Cambridge, pp 1–27

10. Cassell J, Thórisson K (1999) The power of a nod and a glance envelope vs. emotional feedback in animated conversational agents. Appl Artif Intell 13:519–538
11. Cassell J, Bickmore T, Campbell L, Vilhjalmsson H, Yan H (2001) More than just a pretty face: Conversational protocols and the affordances of embodiment. Knowl-Based Syst 14(1–2):55–64
12. Cassell J, Nakano Y, Bickmore T, Sidner C, Rich C (2001) Non-verbal cues for discourse structure. In: Proceedings of the 39th annual meeting on association for computational linguistics, association for computational linguistics. Morgan Kaufmann, Toulouse, pp 114–123
13. Catrambone R, Stasko J, Xiao J (2004) ECA as user interface paradigm. In: From brows to trust: evaluating embodied conversational agents. Springer, Berlin, pp 239–267
14. Cerrato L, Ekeklint S (2004) Evaluating users' reactions to human-like interfaces: prosodic and paralinguistic features as measures of user satisfaction. In: From brows to trust: evaluating embodied conversational agents. Springer, Berlin, pp 101–124
15. Dehn DM, Van Mulken S (2000) The impact of animated interface agents: a review of empirical research. Int J Hum-Comput Stud 52(1):1–22
16. Fagerberg P, Stahl A, Höök K (2003) Designing gestures for affective input: an analysis of shape, effort and valence. In: Ollila M, Rantzer M (eds) Proceedings of mobile ubiquitous and multimedia, MUM 2003. Linköping University Electronic Press, Norrköping
17. Goldberg J, Ostendorf M, Kirchhoff K (2003) The impact of response wording in error correction subdialogs. In: ISCA tutorial and research workshop on error handling in spoken dialogue systems. ISCA, Citeseer, pp 101–106
18. Hartmann B, Mancini M, Buisine S, Pelachaud C (2005) Design and evaluation of expressive gesture synthesis for embodied conversational agents. In: Proceedings of the 4th international joint conference on autonomous agents and multiagent systems. ACM, Association for Computational Linguistics, The Netherlands, pp 1095–1096
19. Hone K (2005) Animated agents to reduce user frustration. In: The 19th British HCI group annual conference, Edinburgh
20. Hone KS, Graham R (2001) Towards a tool for the subjective assessment of speech system interfaces (SASSI). Nat Lang Eng 6(3–4):287–303
21. Rec ITU-T P851 (2003) Subjective quality evaluation of telephone services based on spoken dialogue systems. International recommendation, International Telecommunication Union
22. Suppl ITU-T 24 to P-Series Rec (2005) Parameters describing the interaction with spoken dialogue systems. International recommendation, International Telecommunication Union
23. Kendon A (1990) Conducting interaction: patterns of behaviour in focused encounters. Cambridge University Press, Cambridge
24. Lamel L, Bennacef S, Gauvain J, Dartigues H, Temem J (1998) User evaluation of the MASK kiosk. In: Fifth international conference on spoken language processing, Citeseer
25. Lamel L, Rosset S, Gauvain J, Bennacef S, Garnier-Rizet M, Prouts B (2000) The LIMSI ARISE system. Speech Commun 31(4):339–354
26. Lee J, DeVault D, Marsella S, Traum D (2008) Thoughts on FML: Behavior generation in the virtual human communication architecture. In: Why conversational agents do what they do. Functional Representations for Generating Conversational Agent Behavior. AAMAS 2008, Estoril, Portugal
27. Lester JC, Converse SA, Kahler SE, Barlow BA, Stone ST, Bhogal RS (1997) The persona effect: affective impact of animated pedagogical agents. In: Pemberton S (ed) Proceedings of the SIGCHI conference on human factors in computing systems, 1997, Atlanta, Georgia, pp 359–366
28. López-Mencía B, Hernández-Trapote A, Díaz-Pardo D, Fernández-Pozo R, Hernández-Gómez L, Torre Toledano D (2007) Design and validation of eca gestures to improve dialogue system robustness. In: Proceedings of the ACL 2007 workshop on embodied language processing. Association for Computational Linguistics, Prague, Czech Republic, pp 67–74
29. ter Maat M, Heylen D (2009) Turn management or impression management? In: Proceedings of 9th international conference on intelligent virtual agents, IVA 2009. Springer, Berlin, pp 467–473
30. McTear M (2008) Handling miscommunication: Why bother? In: Dybkjær L, Minker W (eds) Recent trends in discourse and dialogue. Springer, Amsterdam, pp 101–122.
31. Möller S, Smeele P, Boland H, Krebber J (2007) Evaluating spoken dialogue systems according to de-facto standards: A case study. Comput Speech Lang 21(1):26–53
32. Noor C (2004) Empirical evaluation methodology for embodied conversational agents. In: From brows to trust: evaluating embodied conversational agents. Kluwer, Dordrecht, pp 67–99
33. Oviatt S, Adams B (2000) Designing and evaluating conversational interfaces with animated characters. In: Cassell J, Sullivan J, Churchill EF (eds) Embodied conversational agents. MIT Press, Cambridge, pp 319–345
34. Oviatt S, VanGent R (1996) Error resolution during multimodal human-computer interaction. In: Proceedings of the fourth international conference on spoken language processing, vol 1. Institute of Electrical & Electronics Engineers, pp 204–207
35. Oviatt S, MacEachern M, Levow G (1998) Predicting hyperarticulate speech during human-computer error resolution. Speech Commun 24(2):87–110
36. Pelachaud C (2003) Overview of representation languages for ECAs. Project reports, Paris VIII, IUT Montreal
37. San-Segundo R, Montero J, Ferreiros J, Córdoba R, Pardo J (2001) Designing confirmation mechanisms and error recover techniques in a railway information system for Spanish. In: Proceedings of the second SIGdial workshop on discourse and dialogue, vol 16. Association for Computational Linguistics, Association for Computational Linguistics, Aalborg, pp 136–139
38. Schaumburg H (2001) Computers as tools or as social actors?—the users' perspective on anthropomorphic agents. Int J Coop Inf Syst 10(1–2):217–234
39. Stevens JP (1992) Applied multivariate statistics for the social sciences. Lawrence Erlbaum, Las Vegas
40. Walker M, Whittaker S, Stent A, Maloor P, Moore J, Johnston M, Vasireddy G (2004) Generation and evaluation of user tailored responses in multimodal dialogue. Cogn Sci 28(5):811–840
41. Walker MA, Litman DJ, Kamm CA, Abella A (1997) Paradise: a framework for evaluating spoken dialogue agents. In: Proceedings of the 35th annual meeting of the association for computational linguistics (ACL-97). Association for Computational Linguistics, Madrid, pp 271–280
42. Weiss B, Kühnel C, Wechsung I, Fagel S, Möller S (2010) Quality of talking heads in different interaction and media contexts. Speech Comm 52(6):481–492
43. White M, Foster M, Oberlander J, Brown A (2005) Using facial feedback to enhance turn-taking in a multimodal dialogue system. In: Proceedings of HCI international, vol 2. Lawrence Erlbaum Associates, Inc, Las Vegas