Available online at www.sciencedirect.com
ScienceDirect
Interacting with Computers 19 (2007) 484–493
Interacting with Computers
www.elsevier.com/locate/intcom

The effects of speech–gesture cooperation in animated agents' behavior in multimedia presentations

Stéphanie Buisine [a,*], Jean-Claude Martin [b,c]

[a] Ecole Nationale Supérieure d'Arts et Métiers, 151 boulevard de l'Hôpital, 75013 Paris, France
[b] LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
[c] LINC-IUT de Montreuil, 140 rue de la Nouvelle France, 93100 Montreuil, France

Received 31 July 2006; received in revised form 13 April 2007; accepted 18 April 2007
Available online 18 May 2007

Abstract

Until now, research on arrangement of verbal and non-verbal information in multimedia presentations has not considered multimodal behavior of animated agents. In this paper, we will present an experiment exploring the effects of different types of speech–gesture cooperation in agents' behavior: redundancy (gestures duplicate pieces of information conveyed by speech), complementarity (distribution of information across speech and gestures) and a control condition in which gesture does not convey semantic information. Using a Latin-square design, these strategies were attributed to agents of different appearances to present different objects. Fifty-four male and 54 female users attended three short presentations performed by the agents, recalled the content of presentations and evaluated both the presentations and the agents. Although speech–gesture cooperation was not consciously perceived, it proved to influence users' recall performance and subjective evaluations: redundancy increased verbal information recall, ratings of the quality of explanation, and expressiveness of agents. Redundancy also resulted in higher likeability scores for the agents and a more positive perception of their personality. Users' gender had no influence on this set of results.

Keywords: Embodied conversational agents; Multimodal behavior; Redundancy; Experimental evaluation

1. Introduction

Multimedia education is one of the primary application fields for embodied conversational agents. These virtual characters are used to present the educational material, answer users' questions and give feedback about their progression. They are also expected to increase entertainment and motivation in the learning process (Johnson et al., 2000, 2003; Lester et al., 1999a; Stone and Lester, 1996) and a recent research topic especially focuses on social mechanisms, such as politeness, arising from human-agent interaction (Johnson et al., 2005; Krämer, 2005). Embodied conversational agents in pedagogical-like applications were actually shown to increase perceived easiness and entertainment (Van Mulken et al., 1998), to increase learning transfer and interest ratings (Moreno et al., 2001), and sometimes to increase memorization (Beun et al., 2003) in comparison with equivalent systems with no agent.

Because they are visually embodied and use speech synthesis, animated agents can partly behave like a teacher in the classroom, i.e. they can support or illustrate their verbal explanations with hand gestures. For example, they can point to the educational material or depict particular properties of objects or ideas, like shapes, sizes, or spatial relationships. McNeill (1992) has identified four types of gestures that speakers routinely use when they talk: (1) deictic or pointing gestures indicating entities in the conversational space; (2) iconic gestures that capture concrete aspects of the semantic content of speech (e.g. shape, size); (3) metaphoric gestures capturing abstract aspects of the

* Corresponding author. Tel.: +33 1 44 24 63 77; fax: +33 1 44 24 63 59.
E-mail addresses: stephanie.buisine@paris.ensam.fr (S. Buisine). martin@limsi.fr (J.-C. Martin).

0953-5438/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.intcom.2007.04.002

semantic content (e.g. uncertainty); and (4) beat gestures that accompany the rhythm of speech independently of the semantic content. In tutors' behavior some of these spontaneous gestures can have an educative function: in a study investigating human teachers' multimodal behavior, Goldin-Meadow et al. (1999) showed that children understand a math lesson better when the teacher produces hand gestures matching the speech content than in conditions with no hand gesture.

From such a result we can assume that efficient pedagogical agents should display matching speech–gesture combinations. However, this recommendation is not sufficient to specify an agent's multimodal behavior, since speech and gestures can cooperate in different ways (see for example the types of cooperation listed by Knapp, 2002). In this paper, we present an in-depth study of the effects of two types of multimodal combinations in embodied agents' behavior. Our goal is to contribute to the field of pedagogical or presentation agents by providing insights for the design of multimodal behavior, and also to the field of multimodal output systems, by isolating the effects of different multimodal strategies on users.

We focus on two types of speech–gesture cooperation called redundancy and complementarity (cooperation also studied by Cassell and Prevost, 1996; Cassell et al., 2000). We define redundancy as a duplication of information in several modalities (e.g. verbal/pictorial, visual/auditory), and complementarity as the distribution of information across several modalities (the integration of modalities being necessary to understand the information). To illustrate these two types of speech–gesture cooperation, consider a history lesson about Hannibal's route of invasion: if the teacher says "Hannibal went from North Africa to Italy" and produces two deictic gestures on a map (one indicating North Africa, the other one indicating Italy), the gestures are considered *redundant* to the speech content (they duplicate the identification of countries mentioned by speech). Conversely, if the teacher says "Hannibal went from North Africa to there" and completes the utterance with a pointing gesture to Italy, the gesture is considered *complementary*. The listener has to integrate both modalities to get the full message. The same reasoning applies to iconic gestures: if the initial utterance is accompanied by an iconic gesture showing the land route from North Africa to Italy, the gesture is considered *complementary* to the speech content because it conveys a new piece of information (Hannibal went by land and not by sea). If the teacher details the route verbally ("through Spain...") and uses the same iconic gesture, the latter becomes *redundant* to the speech content. With such examples in mind, our initial research question was the following: in a pedagogical context, which one of the two strategies, the redundant or the complementary, would be the more efficient? Which one would be preferred by tutees? Although the study by Goldin-Meadow et al. (1999) provides empirical evidence of the role of gestures in education, it did not investigate the respective effects of redundant and complementary gestures. Indeed, Goldin-Meadow's concept of matching gestures (Goldin-Meadow, 1999; Goldin-Meadow et al., 1999) includes both redundant and complementary gestures, while mismatches are gestures conflicting with speech. Likewise, Cassell et al. (1999) compared the effects of matching and mismatching speech–gesture combinations on the listener's memorization but without examining the differences between redundant and complementary matching combinations.

Some embodied pedagogical systems such as AutoTutor (Graesser et al., 2005) include an agent who is not embedded in the learning environment and comments the educational material from a separate window. In a similar situation (agent unable to move about and producing deictic gestures from a distance), Craig et al. (2002) showed that the agent's redundant pointing gestures had no effect on the learning performance in comparison to a condition with no agent. Therefore the gestures might have an influence on the learning process only when the agent is embedded in the learning environment and can designate the illustrative items unambiguously. Some of the existing systems in this category are typically implemented to produce complementary speech–gesture combinations. For example, the Cosmo agent (Lester et al., 1999b), who teaches Internet packet routing, is capable of pointing unambiguously to an item in his environment while giving a verbal explanation about it. In this case, a module called the deictic planner manages the use of the appropriate demonstrative (this, these, that, those) in order to optimize the speech content. As a result, Cosmo produces multimodal utterances such as "this router has more traffic" with a coordinated deictic gesture. Here, speech and gesture cooperate by complementarity because each modality conveys a specific piece of information and the listener has to integrate the modalities to understand the message (the router concerned can be identified only by gesture). Cooperation by complementarity allows the amount of information given by each modality to be reduced: the Rea agent (Cassell, 2001) is another example of an implementation optimizing the distribution of meaning across speech and gestures. When Rea talks about an object, she can describe some of its features by hand gestures (e.g. shape or size) without mentioning them by speech. Conversely, the Steve agent (Rickel and Johnson, 1999), who teaches procedural tasks with complicated machinery, tends to use speech–gesture redundancy. One typical example of Steve's multimodal utterance, is "open cut-out valve 3" accompanied by a pointing gesture to this particular valve. In this example, the valve can be identified by speech alone and by gesture alone. Therefore we can consider that the gesture is redundant to the speech. Finally, other systems such as Max (Kopp et al., 2005) generate the gestures in accordance with the availability of each modality, the postural context, and a part of random choice. We assume that such a strategy results in a mix of complementary and redundant gestures. All these systems were repeatedly user-tested and proved to be efficient, but the effects of redundant and

complementary speech–gesture cooperation were never tested.

In short, the effects of speech–gesture cooperation in a learning context seem to have never been investigated, either with human tutors or with embodied conversational agents. Yet the effects of other kinds of redundancy in multimedia learning have been previously discussed, for example redundancy between text and image. Tutors' gestures are not analogous to pictures for multiple reasons: in our previous examples gestures do not replace images since pictorial or visual material is always used to support the verbal discourse. Gestures provide a new intermediate communicative modality between speech and image: they can be used to integrate speech and image (e.g. deictic gestures towards the image synchronized to relevant verbal information) or they can provide visual information closely integrated to speech (because they come from the same source – the agent – and they are temporally synchronized with the speech content). However, without confusing image and gestures, we can nonetheless examine previous results on multimedia redundancy to see what they suggest about the effects of speech–gesture cooperation. The first step is to identify possible media combinations, since the effects of redundancy depend on the media involved (for reviews, see Le Bohec and Jamet, 2005; Moreno and Mayer, 2002). Verbal redundancy, which involves presenting simultaneously written and auditory forms of the same text, is known to enhance memorization. Redundancy between an auditory text (auditory-verbal material) and an image (visual-non-verbal material) also facilitates learning. However, according to the cognitive load theory (Kalyuga et al., 1999), redundancy between written text (visual-verbal material) and image (visual-non-verbal material) leads to split attention and thus disturbs learning (see also Dubois et al., 2003). As speech conveys auditory-verbal material and gesture conveys visual-non-verbal material, the previous set of results suggests that, compared to a control condition with no gestures, speech–gesture redundancy facilitates learning. However, neither the cognitive load theory (Kalyuga et al., 1999) nor the dual-processing model of working memory (Moreno and Mayer, 2002) enable the effects of complementarity (speech and gesture both bring part of the message) to be predicted: this strategy relates speech and graphic material better (by means of deictic and iconic gestures) than the control condition; it also reduces the total amount of information compared to redundancy (no duplication); however, it may require an additional effort to integrate auditory-verbal and visual-non-verbal material into a single mental representation.

The following experiment was designed to study the effects of speech–gesture cooperation of animated agents in a learning context supported by images. To this end, we will test the following strategies: redundancy between speech and gesture, complementarity, and a control condition in which gesture does not convey semantic information. We did not implement a control condition with no agent (replacing him with e.g. an arrow pointing to the image, synchronized to the verbal discourse) because similar situations were previously tested (Beun et al., 2003; Moreno et al., 2001; Van Mulken et al., 1998) and showed that even with limited or no functionality, animated agents are useful in pedagogical applications (at least they improve subjective experience). We chose to focus on our research goal which is the comparison of agent's multimodal strategies. We will investigate the effects of these strategies on the memorization of the verbal content of presentations (cued written recall of agents' discourse), and on the memorization of the visual material (graphic recall of the images presented). As male and female subjects sometimes use different cognitive strategies, with visual-spatial vs. auditory-verbal proneness for males and females, respectively (Kimura, 1999), we will also explore the effects of users' gender on the results. In addition, we are interested in evaluating the effects of speech–gesture cooperation on subjective perception of the tutees: quality of presentation, likeability, expressiveness and perceived personality of the animated agent. Personality being a collection of emotional, thought and behavioral patterns unique to a person, it appears necessary to involve several agents in the experiment in order to test whether speech–gesture cooperation has a consistent effect on perceived personality, whatever the agent's appearance (a Latin-square design can be used to cross agents' appearance and speech–gesture cooperation). In the event of such a phenomenon appearing, we also wish to determine whether it relies on a conscious or an unconscious process, i.e. whether users consciously perceive the differences in speech–gesture strategies and base their judgments on them.

The present experiment was designed on the basis of a preliminary test with 18 users (Buisine et al., 2004) which enabled us to adjust agents' behavior (e.g. avoid some gestures such as crossing the arms which were negatively perceived), develop more accurate performance indices (cued written recall and graphic recall) and additional subjective indices (perception of agents' personality).

2. Method

2.1. Participants

One hundred and eight students from an undergraduate psychology institute at the University of Paris V participated in the experiment. There were 54 male students (mean age = 26.7 years, SD = 9.2, 18- to 53-years-old) and 54 female students (mean age = 23.1 years, SD = 5.6, 18- to 51-years-old).

2.2. Materials

To enable a within-user design, the three types of cooperation (redundancy, complementarity, control condition) given to agents of varying appearance were applied to the

presentation of different objects. We used 2D cartoon-like Limsi Embodied Agents (Abrilian et al., 2002): one female agent and two male agents, namely Lea, Marco and Jules. As we needed to control the parameters of their behavior fully, the agents were not interactive for this experiment – in this respect they can be called presentation agents as defined by André et al. (1999). They appeared in front of a whiteboard and made short technical presentations associated with an image displayed on the whiteboard.

The objects presented by the agents were a video-editing software program, a video-projector remote control and a photocopier. The main difficulties were ambiguities related to the position, the color and the shape of keys and/or menu items of the three objects. Hence these objects were particularly relevant to studying multimodal spatial references. They also have similar functional behaviors, and the preliminary test (Buisine et al., 2004) suggested that they were equivalent in complexity. The explanations concerned the identification of 10 buttons or menu items of each object, and a description of their function. They were equivalent in duration for the three objects (75 s for redundant and control conditions, 60 s for the complementary condition).

Multimodal agents' behavior was manually specified using a low-level XML language. The same scripts were used for the three appearances in order to ensure independence between agents' behavior and their appearance. The three types of speech–gesture cooperation were generated as follows:

- *Redundancy:* The agent described or referred to every button/menu item both by speech and arm gesture (see Fig. 1 upper window). In speech, absolute localization of items (e.g. "on the top left side") was used whenever possible; otherwise the agent used relative localization (e.g. "just below, you will find..."). The agent also verbalized shape, color and size of items whenever it was a discriminating feature. Regarding hand and arm gestures, the agent displayed shape and size via iconic gestures (with both hands) when possible. A deictic gesture was used for every object. Finger or palm hand shape was selected according to the precision required (size of the item to be designated). When necessary, preceding a deictic gesture, the agent moved closer to the target item. S/he also glanced at target items for 0.4 s at the beginning of every deictic gesture. Non-semantic gestures (i.e. not related to any object of the lesson) were inserted in order to obtain natural-looking animation: beat gestures (which have a syntactic rather than a semantic function), self-centered gestures, etc. In total, redundant scenarios included 14 semantic gestures and 23 non-semantic arm gestures. Strokes of all gestures were placed manually during agents' speech.
- *Complementarity:* Half of the semantic gestures from redundant scenarios (deictic gestures towards the image or iconic gestures) were selected to create complementary scenarios. The information they conveyed (identification of items, shape, or size) was removed from speech. Non-verbal behavior of agents was completed by non-semantic gestures. We thus ensured that information conveyed by gesture was not duplicated in speech and information conveyed by speech was not duplicated in gesture (see Fig. 1 middle window). The agent moved closer to the target item when necessary and glanced at it for 0.4 s at the beginning of every deictic gesture. Complementary scenarios included 7 semantic gestures and 30 non-semantic gestures.

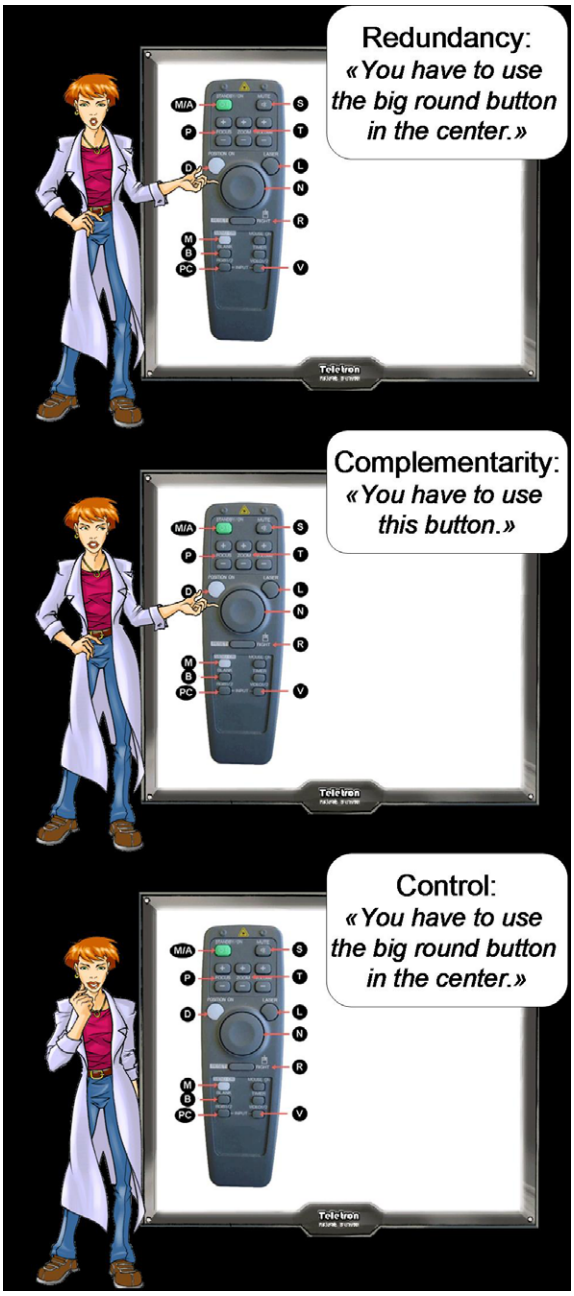

Fig. 1. Each agent (the female agent Lea in this screenshot) was tested with the three types of speech–gesture cooperation: redundant (upper window), complementary (middle window) and control (lower window).

– *Control condition:* The speech content was the same as in redundant scenarios (describing localization, shape, color, size of items), and non-semantic gestures were used throughout the presentation (see Fig. 1 lower window).

The rate of semantic gestures (deictic or iconic) among arm/hand movements was maximal in redundant scenarios (14/37), intermediate in complementary scenarios (7/37), and non-existent in control scenarios (0/37), but the total number of gestures was the same in the three conditions. Animation features that were common to all scenarios included lip movements, periodic eye blinks, and eyebrow movements appropriately inserted for the animation to be perceived as natural. We used IBM ViaVoice for speech synthesis with voice intonation set to neutral. The experiment was conducted in French.

2.3. Design

Combinations between agents' appearance, speech–gesture cooperation and content of presentation were determined by means of a repeated-measurement Latin-square design (Myers, 1979): such a design enables the three variables to be investigated with less expenditure of time (each user saw three presentations, see Table 1) than complete factorial designs would involve (27 presentations). It also removes some sources of variance such as repetition effects. Male and female users were paired across these combinations.

2.4. Procedure and data collection

Users were instructed to watch three short multimedia presentations carefully and were informed that they would have to recall the content of the three presentations afterwards. The presentations were displayed on a 17 in. computer screen, 1024 * 768 resolution, with loudspeakers for speech synthesis.

After the presentations, the data collection consisted of:

– *Graphic recall:* Users had to draw the three objects from memory. Although rarely used, this method of measuring performance seemed interesting to assess the memorization of visual material.
– *Cued written recall:* Users were provided with the images used for the presentations and had to recall the verbal explanation given by the agents.
– A questionnaire in which users had to evaluate the presentations and the agents according to several criteria: the quality of presentations (ranking of the three presentations), the likeability of agents (ranking of the three agents) and their expressiveness (ranking of the agents). We also included in the questionnaire an open question about agents' personality in order to test whether speech–gesture cooperation and/or agents' appearance influenced the perception of agents' personality. In all the questions users were invited to explain their judgment criteria (e.g. what feature they based their ranking of agents' likeability on) and were particularly prompted to make explicit their observations about the way each agent gave explanations.

2.5. Data analysis

Graphic recall was initially evaluated on a 15-point grid for each object: 3 points for the representation of global features such as general shape and the most meaningful components of the object; 10 points for the representation (not necessarily the exact position) of specific items commented on during the explanation; 2 points for the representation of additional items not commented on in the explanation. The cued written recall was evaluated on a 30-point grid: for each one of the 10 specific items commented on, the user was attributed 1 point if s/he mentioned it, 2 points if s/he mentioned it and approximately recalled its function, 3

Table 1
The Latin-square design used for the experiment

	Lea		*Marco*		*Jules*	
A	Redundancy	[RC]	Complementarity	[VS]	Control	[P]
B	Complementarity	[P]	Control	[RC]	Redundancy	[VS]
C	Control	[VS]	Redundancy	[P]	Complementarity	[RC]
	Marco		*Jules*		*Lea*	
D	Redundancy	[RC]	Complementarity	[VS]	Control	[P]
E	Complementarity	[P]	Control	[RC]	Redundancy	[VS]
F	Control	[VS]	Redundancy	[P]	Complementarity	[RC]
	Jules		*Lea*		*Marco*	
G	Redundancy	[RC]	Complementarity	[VS]	Control	[P]
H	Complementarity	[P]	Control	[RC]	Redundancy	[VS]
I	Control	[VS]	Redundancy	[P]	Complementarity	[RC]

Each user was allocated to a group (A–I) and followed the three experimental conditions of the corresponding row (in this order). The agent performing each condition is indicated in italics as column title (Lea, Marco, Jules); the speech–gesture cooperation and the object presented (in square brackets: RC for remote control, P for photocopier, VS for video software) are indicated in each cell. The user's gender was balanced in each group (A–I).

points if s/he used the same wording as in the agent's explanation. Finally, these two measures of performance (graphic and written recall) were expressed as percentages.

Rankings of presentations and agents according to the subjective variables were converted into scores (from 0 to 2; e.g. the first rank in likeability became a 2-point score in likeability). This data (graphic recall, cued written recall, quality of presentation, likeability of agents and expressiveness) were submitted to analysis of variance with user's gender as the between-user factor. For each dependent variable, the analysis was successively performed using speech–gesture cooperation and agents' appearance as within-user factors. By way of control, the effects of the objects were also tested. Post-hoc comparisons were performed by means of Fisher's LSD. We also examined relations between dependent variables by means of a linear correlation analysis. Words used to describe personality were merely classified as positive (e.g. nice, competent, serious, open, enthusiastic, clever, cool, funny), negative (e.g. cold, inexpressive, strict, unconcerned) or neutral (e.g. standard, technical, discreet). The distribution of these three categories as a function of speech–gesture cooperation and agent's appearance was studied using a Chi-square analysis. All the analyses were performed with SPSS software.

Finally, qualitative data about judgment criteria was categorized into nine ad-hoc dimensions and were analyzed descriptively.

3. Results

Table 2 summarizes the mean scores and standard deviations of all numerical dependent variables. Speech–gesture cooperation was proved to influence the cued written recall significantly ($F(2,212) = 12.04$, $p < .001$), with redundancy leading to a better recall than complementarity ($p < .001$), and control condition ($p < .001$). The difference between complementarity and control condition is not significant. Speech–gesture cooperation had no effect on graphic recall, but its main effect on subjective ratings of quality of explanation was significant ($F(2,212) = 12.01$, $p < .001$), with redundancy yielding a better evaluation than complementarity ($p = .001$) and control condition ($p < .001$), and no significant difference between complementarity and control condition. Speech–gesture cooperation also influenced the likeability ratings of agents ($F(2,212) = 6.34$, $p = .002$), with once again the same pattern: redundancy made agents more likeable than complementarity ($p = .001$) and control condition ($p = .014$), with no significant difference between complementarity and control condition. Finally, the effect of speech–gesture cooperation on the evaluation of expressiveness was also significant ($F(2,212) = 6.49$, $p = .002$). Redundant agents were judged as more expressive than complementary ($p = .052$) and control ones ($p < .001$), complementary and control agents being not significantly different. The influence of the user's gender was tested in each of the previous calculations, and no significant effect appeared in any case.

Regarding the influence of agents' appearance, the only significant effect arose on ratings of agents' likeability ($F(2,212) = 3.17$, $p = .044$). Marco appeared to be more likeable than Lea ($p = .024$) and Jules ($p = .035$). Likeability score of Lea and Jules did not significantly differ, and once again, the user's gender had no significant effect.

The object is the only variable that influenced graphic recall ($F(2,212) = 42.13$, $p < .001$): the remote control was better recalled than the software ($p < .001$) and the photocopier ($p < .001$), with no difference between the software and the photocopier. There was also a main effect of object on cued written recall ($F(2,212) = 4.04$, $p = .019$): the remote control was better recalled than the software ($p = .044$) and the photocopier ($p = .002$). Likewise, the object influenced quality of explanation ratings ($F(2,212) = 11.39$, $p < .001$): explanations concerning the remote control obtained better evaluations than those concerning the software ($p < .001$) or the photocopier ($p < .001$), with no significant difference between the

Table 2
Means and standard deviations for each speech–gesture cooperation, agent-appearance and object condition for graphic recall, written cued recall, ratings of quality of explanation, ratings of agents' likeability and ratings of agents' expressiveness

Condition	Graphic recall		Cued written recall		Quality of explanation		Agent's likeability		Agent's expressiveness	
	M	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD
Speech–gesture cooperation										
Redundancy	50.3	20.1	**48.7****	19.7	**1.4****	0.7	**1.3***	0.7	**1.3***	0.8
Complementarity	50.9	23.1	41.2	18.8	0.9	0.8	0.8	0.8	1	0.8
Control condition	48.3	23.3	40.7	20.7	0.7	0.8	0.9	0.8	0.8	0.8
Agent appearance										
Marco	51.2	22.7	44.4	20.9	1.1	0.8	**1.2***	0.8	1.1	0.8
Lea	49.2	21.3	43.6	19.3	1	0.8	0.9	0.8	0.9	0.8
Jules	49	22.6	42.6	20	1	0.9	0.9	0.9	1	0.9
Object										
Remote control	**60.8****	16.4	**46.6***	17.6	**1.4****	0.8	1.2	0.8	1	0.8
Video software	45.4	24	42.6	23.1	0.9	0.8	1.1	0.8	1	0.8
Photocopier	43.2	21.3	41.5	18.7	0.8	0.8	**0.8***	0.8	1	0.8

The values in bold font are significantly different from those in the same column, $^{*}p < .05$; $^{**}p < .01$.

software and the photocopier. Finally, there was a significant effect of objects on likeability scores of agents ($F(2,212) = 4.08$, $p = .018$): agents presenting the photocopier were less likeable than those who presented the remote control ($p = .005$) or the software ($p = .045$), with no significant difference between agents presenting the remote control and the software.

Table 3 presents the results of the linear correlation analysis between the five numerical dependent variables (graphic recall, cued written recall, quality of presentation, likeability of agents and expressiveness). Overall the correlation coefficients appear to be rather weak, but they nonetheless show a correlation between the graphic and the cued written recall ($r = 0.581$, $p < .01$), and a correlation between the quality of presentation and the likeability of agents ($r = 0.405$, $p < .01$).

Regarding the user's perception of the agents' personality, 56.8% of descriptive words fell into the positive category, 29.3% into the negative category, and 8.3% into the neutral category (5.6% of personality questions were not answered). Table 4 presents the distribution of categories as a function of speech–gesture cooperation and agents' appearance. Speech–gesture cooperation was proved to influence personality perception significantly ($\chi^2(6) = 13.46$; $p = .036$): Table 4 shows that redundant agents were judged more positively than complementary and control agents. Conversely, agents' appearance did not significantly influence the distribution of words used to describe personality ($\chi^2(6) = 5.52$; NS).

For the study of conscious judgment criteria, we established nine categories from the data elicited from the users: for example, users said they were influenced by the content of presentation (object, verbal discourse), agents' gestures, look (clothes, hair dressing, accessories such as glasses...), facial expressions (in particular smiles), agents' voice, etc. Table 5 details the judgment criteria that users put forward for the evaluation of the quality of presentation, the likeability of agents, their expressivity and personality.

Finally, we analyzed the answers to the question: "Did you notice any difference in the way the agents made their presentation?" Many users responded by emphasizing the content of presentations (object, vocabulary...). Non-verbal behavior was also widely discussed: 15% of users mentioned that some agents moved to the whiteboard to point to the picture; 31% of users said they noticed a difference in the gestures made by the agents. Finally, 2% of users expressed the notion of cooperation between speech and gesture, even if they did not use the words redundancy and complementarity.

4. Discussion

The primary purpose of this experiment was to test whether speech–gesture cooperation (redundancy, complementarity) influences learning and subjective evaluations of users. In this respect, our results clearly show the advantages of the redundant strategy in the context we set up. Multimodal redundancy improved recall of the verbal content

Table 3
Bivariate correlation coefficients between graphic recall, cued written recall, quality of explanation ratings, agents' likeability ratings and agents' expressiveness ratings

Dependent variables	Graphic recall	Cued written recall	Quality of explanation	Agents' likeability
Graphic recall				
Cued written recall	**0.581****			
Quality of explanation	**0.197****	**0.205****		
Agents' likeability	0.051	**0.118***	**0.405****	
Agents' expressiveness	0.021	0.100	**0.361****	**0.395****

*$p < .05$; **$p < .01$.

Table 4
Number of positive, neutral and negative words used to describe personality for each speech–gesture cooperation and agent-appearance condition

Condition	Positive	Neutral	Negative	No answer	Total
Speech–gesture cooperation					
Redundancy	**75***	7	**19***	7	108
Complementarity	54	9	39	6	108
Control condition	55	11	37	5	108
Total	184	27	95	18	324
Agent appearance					
Marco	68	11	24	5	108
Lea	57	7	38	6	108
Jules	59	9	33	7	108
Total	184	27	95	18	324

The values in bold font are significantly different from those in the same column, *$p < .05$.

Table 5
Percentages of judgment criteria elicited from users for their assessment of quality of explanation, likeability of agents, expressivity and personality

Judgment criteria	Quality of explanation (%)	Likeability (%)	Expressivity (%)	Personality (%)
Content of presentation	62	20	11	24
Agents' appearance	0	15	10	12
Agents' gender	3	9	6	2
Agents' look	1	21	9	27
Agents' voice	8	12	10	12
Agents' facial expressions	0	15	15	13
Agents' gestures	21	6	33	5
Agents' locomotion	4	2	6	5
Speech–gesture cooperation	1	0	0	0

of presentations, evaluations of quality of presentation, likeability, expressiveness and personality of agents.

Redundancy influenced verbal but not graphic recall: we can thus hypothesize that users paid sufficient attention to the images on the whiteboard whatever the agents' strategy, and that redundancy helped users in encoding verbal information (identification of items and functionalities) and/or in relating the verbal discourse to the visual material (attribution of functionalities to the proper items). Overall, redundancy yielded a relative increase of 19% of verbal information recalled (49% of information recalled with redundancy vs. 41% on average for complementarity and control condition).

Our data showed no difference between complementary and control conditions. The absence of effect of complementarity was not so predictable because this strategy has the advantage of relating speech and graphic material better than the control condition and reducing the total amount of information – in our experiment we achieved a 20% decrease in the time needed to present a scenario with a complementary strategy (see Fig. 1 to illustrate this decrease). On the contrary, the literature on redundancy in education made predictable the benefit of multimodal redundancy on verbal recall, and perhaps also its benefit on subjective ratings of quality of explanation, in comparison to the control condition. However, our experiment also showed some original findings that previous literature could not have anticipated: multimodal redundancy may improve the social perception of animated agents, since agents with redundant behavior appeared more likeable and their personality more positive. One could hypothesize that redundant agents were rated as more likeable just because they enabled the users to increase their memorization (and not because of their speech–gesture cooperation strategy). The linear correlation analysis between our numerical dependent variables (Table 3) contradicts such a hypothesis since it showed that likeability was not related to written recall. However, likeability appeared to be correlated with quality of explanation: to investigate whether high ratings of likeability were due to multimodal redundancy or to perceived quality of explanation, it would be interesting to design a control condition in which quality of explanation would not be so important to the user (e.g. in a conversational context).

Multimodal redundancy was shown to increase the ratings of quality of explanation, likeability and expressiveness. However, this does not mean that users consciously perceived speech–gesture cooperation. Indeed, most users perceived differences in agents' gestural behavior, but nothing in their comments suggests that they perceived a difference between redundancy and complementarity. Such a result is consistent with the classic view that a speaker's non-verbal behavior usually remains at the periphery of the listener's attentional field (Rimé and Schiaratura, 1991). Only two users in our experiment explicitly verbalized the notion of speech–gesture cooperation, and only one of them said she based her evaluation of quality of explanation on this feature. This set of results has two implications: users are influenced by features they do not perceive and users think they are influenced by features which are actually neutralized (e.g. many users mentioned the influence of agents' voice, although Marco and Jules had the same voice and Lea's scores did not significantly differ from Jules' ones in any variable). Both of these kinds of variables must be taken into account in agent system design and carefully controlled: variables that were shown to modify users' performance and subjective attitude (e.g. speech–gesture cooperation), as well as variables claimed as important by users, even if they are not (e.g. agents' look and voice).

The only variable actually influenced by agents' appearance was likeability: Marco, whatever his speech–gesture strategy, was significantly preferred to Lea and Jules. It is important to understand why this agent had higher likeability scores in order to learn lessons for future agent design. The study of qualitative comments elicited from the users showed that a key feature for Marco's likeability was his wide smile. Fig. 2 presents the three agents with their maximum smiling face: we can see that Marco's smile was designed broader than those of Lea and Jules, and many users said they appreciated it.

Another important feature of agent's likeability is his/her look, as mentioned in previous empirical research (McBreen et al., 2001), in this respect, Lea's white coat yielded contradictory comments: some users found her more pleasant and more serious because of her coat; others found her too strict. Finally, Jules' glasses seemed to penalize him: they were perceived negatively by most of the users, maybe because his eyes were not so visible through the glasses.

Fig. 2. Marco (left), Lea (middle) and Jules (right) with their maximum smiling face.

Users' gender had no significant effect in any of the previous results (performance data and subjective evaluation). This is a positive finding which suggests that it may be possible to design a single agent system suitable for both male and female users. However, this absence of influence of users' gender (like all our results in general) has to be validated in other age groups, in particular with educational applications intended for children. Cultural influences should also be addressed, since they are likely to modify users' preferences for agents' appearance (Cowell and Stanney, 2003) and more generally their perception of speech and gestures (Johnson et al., 2005; Knapp, 2002). The present study, mainly conducted with Europeans, would have to be replicated with people from other ethnic origins to strengthen or complement the results.

Finally, contrary to the results of a preliminary test (Buisine et al., 2004), we observed important effects of the object in this experiment (on graphic and written recall, quality of presentation and likeability of agents). In this experiment our goal was to neutralize the object in order to study the effect of speech–gesture cooperation in an unbiased way. However, it should be pointed out that the strong influence of speech–gesture cooperation arose in spite of this bias: we observed the benefits of redundancy when the presentations were not equivalent.

5. Conclusion

To summarize, we obtained a consistent corpus of results in which speech–gesture redundancy proved to increase the recall of verbal information, the subjective ratings of quality of explanation, the expressiveness of agents, their likeability and their personality. Complementary and control conditions did not significantly differ in the data we collected: of course, the introduction of unambiguous pointing gestures and iconic gestures in animated agents' behavior remains an important technical improvement, but to transfer this improvement to the cognitive side of interaction, gestures have to support speech in a redundant manner. Complementarity enables the amount of information conveyed by each modality to be decreased, but in a learning context it may not improve information recall or subjective evaluation of the situation.

In an extension to the present study we should address the naturalness of agents' behavior when it is based on a single multimodal strategy. Human spontaneous behavior being normally composed of several strategies mixed together (Cassell et al., 2000), we could compare the effects of an optimized behavioral strategy (redundancy between speech and gesture) vs. a natural one (mix of redundant and complementary behaviors). Although we had no negative comments from users about speech–gesture redundancy, and further assume that it was not consciously perceived, a more natural strategy could appear to be preferred, especially in long-term interaction.

As a secondary result, our experiment provided a few indications for the graphic design of animated agents. For example our results showed that a cartoon-like wide smile, although unrealistic, is an important feature for likeability of animated agents. This result can be related to Kohar and Ginn's recommendations (1997) according to which dramatized characters, because of the emotions they display, make better interface agents than more realistic and human-like characters. This recommendation is also applicable to pedagogical agents, because engagement and entertainment facilitate the learning process (Lester et al., 1999a).

Our agent experimental platform enabled us to highlight the effects of alternate multimodal strategies on a pedagogical-like situation. Similar experiment could be conducted with videotaped people instead of agents, but this would represent a much more costly and complex procedure, since it would involve training a tutor or an actor to accurately and consistently control her speech–gesture cooperation (which is normally an automatic and unconscious process). In conclusion, we should underline that our users were students: our findings can thus be applied to the design of presentation agents for students, e.g. for e-learning systems like the Adele agent (Johnson et al., 2003) or the AutoTutor system (Graesser et al., 2005). Our pattern of results would need a validation to be used for systems dedicated to children, but it nonetheless provides a strong hypothesis in favor of speech–gesture redundancy. Our experiment also raises the interesting question of whether the same hypothesis applies to human tutors' behavior and whether the use of speech–gesture redundancy can be recommended in the classroom.

Acknowledgements

This work was partly supported by the EU/HLT funded project NICE (IST-2001-35293). The authors thank

Marianne Najm, Fabien Bajeot and Marion Wolff (Paris-5 University) as well as Sarkis Abrilian and Christophe Rendu (LIMSI-CNRS) for their contribution.

References

Abrilian, S., Buisine, S., Rendu, C., Martin, J.C., 2002. Specifying cooperation between modalities in lifelike animated agents. In: Proceedings of PRICAI'2002 Workshop on Lifelike Animated Agents Tools, Affective Functions, and Applications, pp. 3–8.

André, E., Rist, T., Müller, J., 1999. Employing AI methods to control the behavior of animated interface agents. Applied Artificial Intelligence 13, 415–448.

Beun, R.J., de Vos, E., Witteman, C., 2003. Embodied conversational agents: effects on memory performance and anthropomorphisation. In: Rist, T., Aylett, R., Ballin, D., Rickel, J. (Eds.), IVA'2003 International Conference on Intelligent Virtual Agents, LNCS, vol. 2792. Springer, Berlin, pp. 315–319.

Buisine, S., Abrilian, S., Martin, J.C., 2004. Evaluation of multimodal behaviour of embodied agents. In: Ruttkay, Z., Pelachaud, C. (Eds.), From Brows to Trust: Evaluating Embodied Conversational Agents. Kluwer Academic Publishers, pp. 217–238.

Cassell, J., 2001. Embodied conversational agents: representation and intelligence in user interface. AI Magazine 22, 67–83.

Cassell, J., McNeill, D., McCullough, K.E., 1999. Speech–gesture mismatches: evidence for one underlying representation of linguistic and non-linguistic information. Pragmatics and Cognition 7, 1–33.

Cassell, J., Prevost, S., 1996. Distribution of semantic features across speech and gesture by humans and computers. In: Proceedings of Workshop on the Integration of Gesture in Language and Speech, pp. 253–270.

Cassell, J., Stone, M., Yan, H., 2000. Coordination and context-dependence in the generation of embodied conversation. In: Proceedings of International Natural Language Generation Conference, pp. 171–178.

Cowell, A.J., Stanney, K.M., 2003. Embodiment and interaction guidelines for designing credible, trustworthy ECAs. In: Rist, T., Aylett, R., Ballin, D., Rickel, J. (Eds.), IVA'2003 International Conference on Intelligent Virtual Agents, LNCS, vol. 2792. Springer, Berlin, pp. 301–309.

Craig, S.D., Gholson, B., Driscoll, D., 2002. Animated pedagogical agents in multimedia educational environments: effects of agent properties, picture features, and redundancy. Journal of Educational Psychology 94, 428–434.

Dubois, V., Gyselinck, V., Choplin, H., 2003. Multimodalité et mémoire de travail [Multimodality and working memory]. In: Proceedings of EIAH'03 French-speaking conference on Environnements Informatiques pour l'Apprentissage Humain, pp. 187–198.

Goldin-Meadow, S., 1999. The role of gesture in communication and thinking. Trends in Cognitive Sciences 3, 419–429.

Goldin-Meadow, S., Kim, S., Singer, M., 1999. What the teacher's hands tell the student's mind about math. Journal of Educational Psychology 91, 720–730.

Graesser, A.C., Chipman, P., Haynes, B.C., Olney, A., 2005. AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions in Education 48, 612–618.

Johnson, W.L., Mayer, R.E., André, E., Rehm, M., 2005. Cross-cultural evaluation of politeness in tactics for pedagogical agents. In: Looi, C.K., McCalla, G., Bredeweg, B., Breuker, J. (Eds.), AIED'05 International Conference on Artificial Intelligence in Education. IOS Press, Amsterdam, pp. 298–305.

Johnson, W.L., Rickel, J., Lester, J., 2000. Animated pedagogical agents: face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education 11, 47–78.

Johnson, W.L., Shaw, E., Marshall, A., LaBore, C., 2003. Evolution of user interaction: the case of agent Adele. In: Proceedings of IUI'2003 International Conference on Intelligent User Interfaces. ACM Press, pp. 93–100.

Kalyuga, S., Chandler, P., Sweller, J., 1999. Managing split-attention and redundancy in multimedia instruction. Applied Cognitive Psychology 13, 351–371.

Kimura, D., 1999. Sex and Cognition. MIT Press, Cambridge.

Knapp, M.L., 2002. Nonverbal communication in human interaction. Thomson Learning, Florence Wadsworth.

Kohar, H., Ginn, I., 1997. Mediators: guides through online TV services. In: Proceedings of Demo Session in CHI'97 International Conference on Human Factors in Computing Systems. ACM Press, pp. 38–39.

Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I., 2005. A conversational agent as museum guide – design and evaluation of a real-world application. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T., (Eds.), IVA'2005 International Conference on Intelligent Virtual Agents, LNCS, vol. 3661. Springer, Berlin, pp. 329–343.

Krämer, N.C., 2005. Social communicative effects of a virtual program guide. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (Eds.), IVA'2005 International Conference on Intelligent Virtual Agents, LNCS, vol. 3661. Springer, Berlin, pp. 442–453.

Le Bohec, O., Jamet, E., 2005. Les effets de redondance dans l'apprentissage à partir de documents multimédia [Redundancy effect and the multimedia learning process]. Le Travail Humain 68, 97–124.

Lester, J., Towns, S., FitzGerald, P., 1999a. Achieving affective impact: visual emotive communication in lifelike pedagogical agents. International Journal of Artificial Intelligence in Education 10, 278–291.

Lester, J., Voerman, J., Towns, S., Callaway, C., 1999b. Deictic believability: coordinating gesture, locomotion, and speech in lifelike pedagogical agents. Applied Artificial Intelligence 13, 383–414.

McBreen, H., Anderson, J., Jack, M. 2001. Evaluating 3D embodied conversational agents in contrasting VRML retail applications. In: Proceedings of International Conference on Autonomous Agents Workshop on Multimodal Communication and Context in Embodied Agents, pp. 83–87.

McNeill, D., 1992. Hand and Mind. University of Chicago Press.

Moreno, R., Mayer, R.E., 2002. Verbal redundancy in multimedia learning: when reading helps listening. Journal of Educational Psychology 94, 156–163.

Moreno, R., Mayer, R.E., Spires, H., Lester, J., 2001. The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents?. Cognition and Instruction 19 177–213.

Myers, J.L., 1979. Fundamentals of Experimental Design, third ed. Allyn & Bacon, Inc., Boston.

Rickel, J., Johnson, W.L., 1999. Animated agents for procedural training in virtual reality: perception, cognition, and motor control. Applied Artificial Intelligence 13, 343–382.

Rimé, B., Schiaratura, L., 1991. Gesture and speech. In: Feldman, R.S., Rimé, B. (Eds.), Fundamentals of Nonverbal Behavior. Cambridge University Press, pp. 239–284.

Stone, B., Lester, J., 1996. Dynamically sequencing an animated pedagogical agent. In: Proceedings of National Conference on Artificial Intelligence, pp. 424–431.

Van Mulken, S., André, E., Müller, J., 1998. The persona effect: how substantial is it? In: Proceedings of HCI'98 International Conference on Human–Computer Interaction. Springer, pp. 53–66.