Generating Embodied Descriptions Tailored to User Preferences

Mary Ellen Foster

Informatik VI: Echtzeitsysteme und Robotik
Fakultät für Informatik, Technische Universität München
Boltzmannstraße 3, 85748 Garching bei München, Germany
foster@in.tum.de

Abstract. We describe two user studies designed to measure the impact of using the characteristic displays of a speaker expressing different user-preference evaluations to select the head and eye behaviour of an animated talking head. In the first study, human judges were reliably able to identify positive and negative evaluations based only on the motions of the talking head. In the second study, subjects generally preferred positive displays to accompany positive sentences and negative displays to accompany negative ones, and showed a particular dislike for negative facial displays accompanying positive sentences.

1 Introduction

The facial displays and body language that accompany embodied speech can have a significant effect on all aspects of how the content is perceived. A number of recent studies have demonstrated that the non-verbal behaviour of an embodied agent can affect users' perceptions and reactions at a number of levels, ranging from prosodic stress to affective content. For example, Swerts and Krahmer [1] have demonstrated in a series of studies that nodding and raising the eyebrows on prosodically stressed syllables can enhance users' ability to perceive stress in speech. Rehm and André [2] found that users judged an embodied agent that used deceptive facial displays to be less trustworthy and less certain about what it was saying than an agent that did not use such displays. Marsi and van Rooden [3] implemented selected facial displays of certainty and uncertainty on a talking head in the context of a multimodal question-answering system. In an experiment, users generally rated the videos intended to show certainty as being more certain than those intended to show uncertainty, even when the spoken content was identical.

The above studies all measured how facial displays can add dimensions of meaning that are not present in the speech signal. A related experimental strategy is to generate combinations of speech and body language where the message expressed on the two channels is either consistent or inconsistent. Such studies can be used to measure the relative impact of the two communication channels. For example, in some experiments, Swerts and Krahmer [1] showed judges stimuli with visual stress on one syllable and spoken stress on another and asked them

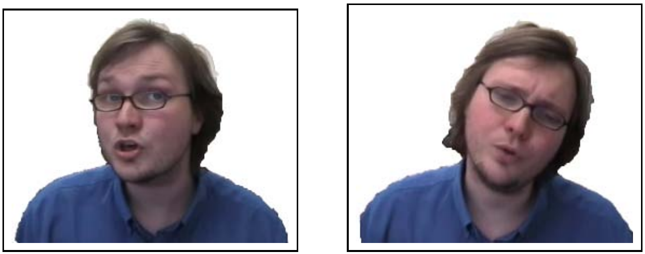
C. Pelachaud et al. (Eds.): IVA 2007, LNAI 4722, pp. 264–271, 2007.

to select the stressed syllable; they found that the visual channel has greater impact for this task. Studies of this type can also compare users' reactions to stimuli with consistent or inconsistent cues, by measuring such factors as task performance or subjective preferences. A study of this latter type is that of Berry et al. [4], who compared several versions of an embodied agent in the context of a recall task and found that subjects performed best when the affective content was consistent between the language and the facial displays.

In this paper, we examine the impact of using corpus-derived rules to select the non-verbal behaviour of a talking head in a multimodal dialogue system. We first recorded and annotated the head and eye movements of a speaker reading a number of sentences in the domain of the target dialogue system. We then implemented a simple rule-based method to select displays for the embodied agent based on this corpus data and performed two studies to test the effectiveness of the rules, using both of the experimental strategies mentioned above. In the first study (Section 3), subjects were asked to identify the intended evaluation based only on the facial displays and were generally able to do so. In the second study (Section 4), users were asked to give subjective preferences between consistent and inconsistent combinations of speech and facial displays, and tended to prefer the consistent combinations.

2 A Corpus of Conversational Facial Displays

The COMIC multimodal dialogue system[1] adds an embodied dialogue interface to a CAD-like application used in sales situations to help clients redesign their bathrooms. When COMIC generates a description of a tile design, it uses information from a model of user preferences to help make choices at all levels.



(a) Positive (b) Negative

Fig. 1. Characteristic facial displays from the corpus

To help select appropriate facial displays for the COMIC talking head to use when presenting the descriptions, we recorded the face of a speaker reading approximately 450 sentences generated by the system and annotated his facial displays; details of the recording and annotation process are given in [5]. The single biggest influence on the speaker's facial displays was the user-preference

[1] http://www.hcrc.ed.ac.uk/comic/

evaluation—that is, whether the user was expected to like or dislike the object being described. When he described a feature with a positive evaluation, he was more likely to turn to the right and to raise his eyebrows (Fig. 1(a)); on the other hand, on a feature with a negative evaluation, he was more likely to lean to the left, lower his eyebrows, and narrow his eyes (Fig. 1(b)).

3 Experiment 1: Recognisability

To test the recognisability of the facial displays of the recorded speaker, we performed a user study designed to measure whether subjects were able to identify the evaluation expressed in a sentence based only on the accompanying facial displays, where the displays were selected using a simple rule based on the findings from the corpus.

3.1 Subjects

This experiment was run over the web; subjects were recruited through the Language Experiments Portal,[2] a website dedicated to online psycholinguistic experiments. There were a total of 26 subjects: 14 females and 12 males. 20 of the subjects were between 20 and 29 years old, 4 were over 30, and 2 were under 20. 11 described themselves as expert computer users, 14 as intermediate users, and one as a beginner. 11 were native speakers of English, while the others had a range of other native languages.

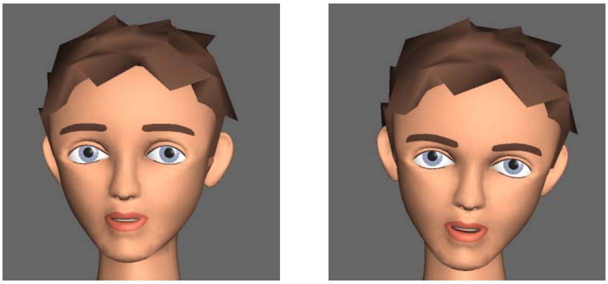
3.2 Methodology

Each subject was shown 16 videos in which a talking head described a tile design. There were four different types of videos: with no facial displays, using only nodding, and using the positive and negative displays from the corpus. After viewing each video, the subject was asked to indicate whether they thought the speaker believed that the user liked or disliked the design being described; they could also choose *don't know* if they were unable to determine the intended evaluation. All subjects saw videos for the same 16 sentences, in an individually randomly-chosen order. The schedule types were randomly allocated to the videos so that each subject saw four schedules of each type.

3.3 Materials

To create the materials for this study, we used the COMIC text planner to generate 16 neutral sentences. For each sentence, we then created a set of facial-display schedules, using a simple rule that added a display to every mention of a tile-design feature (style, colour, decoration, and manufacturer). There were four schedule types: one using the characteristic positive displays from the corpus (Fig. 1(a)), one using the negative displays (Fig. 1(b)), one using only nodding, and one that did not assign any motions at all. The following are the facial-display schedules created for one of the sentences in this study, where *tn=r*

[2] `http://www.language-experiments.org/`



(a) Positive (b) Negative

Fig. 2. Synthesised version of characteristic facial displays

indicates a right turn, *ln=l* a left lean, *bw=u* and *d* upward and downward brow motions, *nd=d* a downward nod, and *sq* narrowing the eyes:

	This	*design*	*features*	*orange*	*in*	*the*	*colour scheme.*
Positive				tn=r, bw=u			
Negative				ln=l, bw=d, sq			
Nodding				nd=d			
Nothing							

We then created a video for each generated schedule using the RUTH talking head [6] and the Festival speech synthesiser [7]. The start and end of each display were coordinated with the relevant words: for example, in the above sentence, all displays would start and finish at the same time as the word *orange*. Fig. 2 shows how the facial displays from Fig. 1 were reproduced on the RUTH head.

3.4 Results

The overall results of this study are presented in Fig. 3. The x axis in this graph shows the actual facial displays that were used; for each display type, the bars indicate the number of times that the subjects believed that the speaker was being positive or negative, or whether they could not tell, respectively. For example, the videos with negative facial displays were identified as positive on 28 trials and as negative on 69 trials, while on the remaining 7 trials the judges were unable to make a decision.

We can assess the significance of these results using a binomial test, which provides an exact measure of the statistical significance of deviations from a theoretically expected classification into several categories (in this case, three). The results of this test indicate that subjects were successful at identifying both the positive and negative facial displays (66% and 63%, respectively; both $p < 0.0001$), and also tended to identify the schedules with nodding as positive (64%, also $p < 0.0001$). There was a weaker but significant tendency to identify the schedules with only lip-synchronisation as negative (45%, $p < 0.05$). The difference between the response patterns for the different video types is statistically significant: $\chi^2 = 87.4$, $df = 6$, $p < 0.0001$.

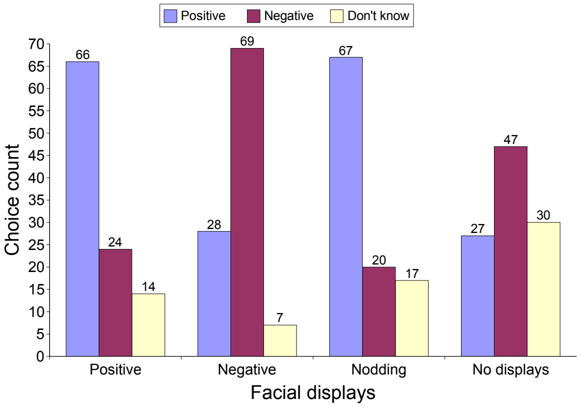


Fig. 3. Response counts for the recognition study

3.5 Discussion

The results of this experiment indicate the subjects were able to identify the intended evaluation when the talking head used the characteristic positive and negative displays of the corpus speaker. The tendency of subjects to identify a sentence with only a nod as positive could be due to several factors. First, while nodding is common across the entire corpus, particularly on words with predicted pitch accents, its frequency is relatively higher in positive contexts. The subjects could have been interpreting the nodding as a positive signal even though it was not intended as such. Another possibility is that the default interpretation of descriptions in COMIC—which operates in a sales domain—is that a design being described is assumed to be something that the user will like unless it is explicitly marked otherwise.

4 Experiment 2: Consistency

In this second experiment, we tested whether subjects' ability observed in the preceding study to identify the polarity of facial displays translates into any preferences regarding the consistency between the evaluation expressed on the verbal and non-verbal channels.

4.1 Subjects

Like the previous study, this one was also run over the web, again using the Language Experiments Portal to recruit subjects. In this case, there were 18 subjects: 8 females and 10 males. 14 of the subjects were between 20 and 29 years old, 3 were over 30, and one was under 20. 11 described themselves as expert computer users, 5 as intermediate users, and 2 as beginners. 8 of the subjects were native English speakers.

4.2 Methodology

Each subject was shown 12 pairs of videos. Both videos in a pair had identical speech content, but the face-display schedules differed: each trial included two of the three possible types of facial displays (positive, negative, or nodding only). For each trial, the subject was asked to indicate which video they preferred; subjects were encouraged to go with their first instincts and were not otherwise instructed on the selection criteria. All subjects saw videos of the same 12 sentences, in an individually-chosen random order. Six of the sentences suggested a positive evaluation, while the other six indicated a negative evaluation. The trials were balanced so that a subject made each pairwise comparison between schedules twice per sentence type (positive or negative), once in each order, while the allocation of comparisons to items was made randomly for each subject.

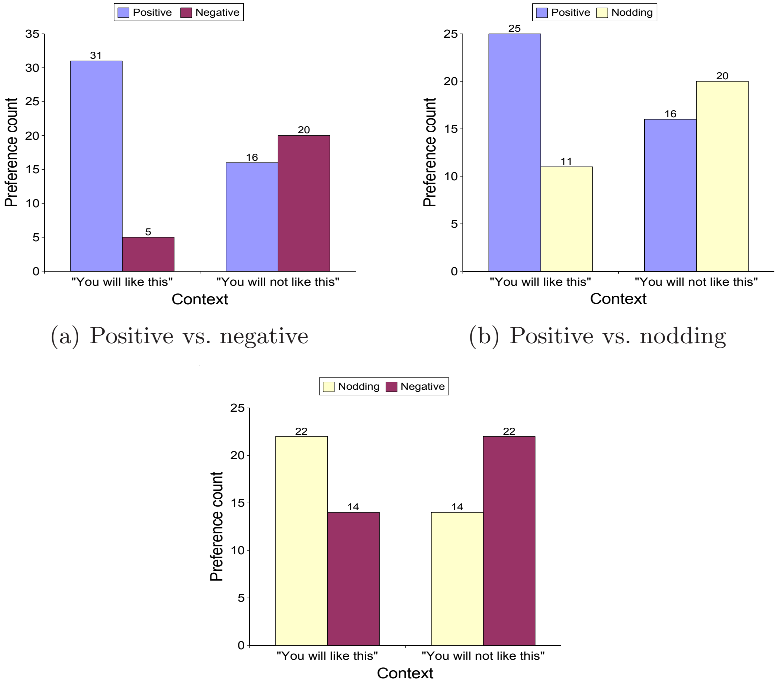
4.3 Materials

To create the materials for this experiment, we used the COMIC text planner to generate a further 12 sentences, again based on neutral user preferences. We then created a positive version of six of the sentences and a negative version of the other six by prepending either *You will like this* or *You will not like this.* For each sentence, we then used the rule-based method described in Section 3.3 to create three face-display schedules (positive, negative, and nodding); we also added a nod on the initial *this* in all cases. We then created videos of each version of each sentence, using the RUTH head and the Festival synthesiser as in the previous experiment. The following are the three schedules generated for one of the positive sentences in this study:

	You	*will*	*like*	*this:*	*it*	*is*	*classic.*
Positive				nd=d			tn=r, bw=u
Negative				nd=d			ln=l, bw=d, sq
Neutral				nd=d			nd=d

4.4 Results

The results of this study are shown in Fig. 4. On each graph, the bars on the left indicate the choices made in a positive context, while the right-hand bars show the choices made in a negative context. For example, when comparing positive facial displays against negative displays (Fig. 4(a)), subjects preferred the positive displays in a positive context 31 out of 36 times, while in a negative context they preferred the negative displays 20 times out of 36. There is a clear pattern: in a negative context, subjects generally preferred the less positive facial displays (negative over positive; negative over nodding; nodding over positive), while in a positive context these preferences were all reversed. The only preferences that were individually significant on a binomial test were those for positive displays over each of the others in a positive context (both $p < 0.05$). Using a χ^2 test, the preference between schedules using positive and negative displays (Fig. 4(a)) was



(a) Positive vs. negative

(b) Positive vs. nodding

(c) Nodding vs. negative

Fig. 4. Choice counts for the consistency study

found to be significantly different in the two contexts ($\chi^2 = 12.0$, $p < 0.001$); for the positive-nodding choice (Fig. 4(b)), the context had a marginally significant impact ($\chi^2 = 3.63$, $p \approx 0.06$); while for the nodding-negative choice (Fig. 4(c)), the effect was further from being significant ($\chi^2 = 2.72$, $p \approx 0.1$).

4.5 Discussion

The subjects showed a marginally significant difference between their responses to the videos with positive displays and those with nodding; the difference for the neutral vs. negative choice showed a similar trend but was not statistically significant. This suggests that, although videos with nodding may appear to be positive when presented in isolation, when they are contrasted with displays that are explicitly positive, the difference becomes apparent. As well, there appears to be a general preference across all of the trials for the positive facial displays: the strongest preferences are those for the positive displays over the others, while adding a negative context only reduces the preference to just under 50%.

5 Conclusion

The speaker recorded for our corpus showed characteristically different facial displays in positive and negative contexts. To test whether these displays are recognisable, we performed two user studies using talking-head videos generated using a simple corpus-derived rule. In the first experiment, subjects were reliably able to identify positive and negative evaluations based on the facial displays; they identified displays with only nodding (a common display in all contexts) as positive, and also had a weaker tendency to identify videos with no motion as negative. In the second study, subjects generally preferred consistent combinations of spoken and visual content.

The results of these studies extend findings from other studies of the influence of non-verbal behaviour on the interpretation of embodied speech (e.g., [1,2,3,4]) to show that varying the body language based on the user-model evaluation has a similar effect. They also confirm that the most marked behaviours of the speaker recorded for the corpus—namely, the facial displays he used in positive and negative contexts—are identifiable to human judges, and that people also prefer embodied descriptions where the facial displays are consistent with the polarity expressed in the speech. Knowing that the displays in the corpus are identifiable when resynthesised on the talking head indicates that this corpus is suitable for use in tasks such as the data-driven generation of non-verbal behaviour, which was the main motivation for creating the corpus.

References

1. Swerts, M., Krahmer, E.: On the perception of audiovisual cues to prominence (In press)
2. Rehm, M., André, E.: Catch me if you can – exploring lying agents in social settings. In: Proc. AAMAS '05 (2005)
3. Marsi, E., van Rooden, F.: Expressing uncertainty with a talking head. In: Proc. MOG 2007 (2007)
4. Berry, D.C., Butler, L., de Rosis, F., Laaksolathi, J., Pelachaud, C., Steedman, M.: Final evaluation report. Deliverable 4.6, MagiCster project (2004)
5. Foster, M.E.: Associating facial displays with syntactic constituents for generation. In: Proc. ACL 2007 Linguistic Annotation Workshop (2007)
6. DeCarlo, D., Stone, M., Revilla, C., Venditti, J.: Specifying and animating facial signals for discourse in embodied conversational agents. Computer Animation and Virtual Worlds 15(1), 27–38 (2004)
7. Clark, R.A.J., Richmond, K., King, S.: Festival 2 – build your own general purpose unit selection speech synthesiser. In: Proc. ISCA Speech Synthesis Workshop (2004)