INTERSPEECH 2009 BRIGHTON

Talking Heads for Interacting with Spoken Dialog Smart-Home Systems

Christine Kühnel, Benjamin Weiss, Sebastian Möller

Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin

christine.kuehnel@telekom.de

Abstract

In this paper the relation between the quality of a talking head as an output component of a spoken dialog system and the quality of the system itself is investigated. Results show that the quality of the talking head has indeed an important impact on system quality. The quality of the talking head itself is found to be influenced by visual and speech quality and the synchronization of voice and lip movement.

Index Terms: evaluation, embodied conversational agents, spoken dialog system, smart-home

1. Introduction

Developers of multimodal systems are frequently confronted with the question of whether or not a certain input or output component would positively contribute to system quality. At our lab a multimodal smart-home system is set up to investigate evaluation of multimodal systems. The system is based on the spoken dialog system INSPIRE developed in the frame of the EU-funded IST-project 2001-32746 [1]. The focus of the experiment reported here is on system output. Two options, voice-only or talking head as output component are compared. Voice-only, being the traditional output for spoken dialog systems used for example for train information services [2] and call-routing [3] is used as a reference condition. Two different talking heads are chosen as an alternative output as the impact of an embodied conversational agent (ECA) is found to enhance human-computer interaction and improve user satisfaction in general [4]. In the smart-home domain only a few studies have been reported so far, mainly on the application of ECAs for subsets, e.g. for giving advice via the TV (e.g. [5] and [6]). In those studies the focus has been on the acceptance of ECAs, on their entertainment and the trust an ECA can evoke. Furthermore, the impact on an ECA on effectiveness and efficiency has been analyzed. The results in general showed a positive influence of the ECA.

We are interested in the impact of a talking head on system quality as perceived by the user. Does smart-home system quality profit from a talking head in general? If this is the case, how does the talking head quality contribute to overall system quality? Are those findings supported by performance measures as well? These are the questions to which the present paper will give answers, thus contributing to the work mentioned above. In the following section working hypotheses are motivated based on findings reported in the literature. The smart-home system INSPIRE and the output components are described in Section 3, as well as our methodology – including the measures we used to answer the questions stated above. In Section 4 results are presented and discussed in Section 5.

2. Working Hypotheses

Studying the literature we find that positive ratings obtained for ECA-enhanced systems are explained with the so called 'persona effect': the positive effect on user's interaction induced by a life-like interface agent [7]. Furthermore, findings reported in [8] indicate that in the smart-home domain the perceived quality of the output component has a high impact on systems' overall quality as measured by user ratings. This is explained with the fact that it is part of the directly experienced interface and not hidden from the user like other system components – for example the dialog management. We hypothesize therefore that for the smart-home domain

H1: A talking head is preferred to voice-only output.

H2: The perceived quality of the talking head has an impact on system quality.

An evaluation of audio-visual quality reported in [9] showed that both, audio and visual quality have an impact on perceived overall quality. In the case of talking heads – presented on a screen with the voice played back via loudspeakers – the user perceives the talking head much like he would perceive a video played back via a computer: as a compound of audio and visual stimuli. We presume therefore, (1) that in the case of talking heads, both audio and visual quality have an effect on overall quality as well. Furthermore, findings of [10] indicate that (2) consistency of voice and face are important for talking heads.

In [1] user preferences concerning different system metaphors (a talking head, voice-only and several voices) in the smart-home domain are analyzed. It is found that users do not prefer the talking head over voice-only output. It is assumed that a lack of synchronization between lip movement and the acoustic signal is responsible for the indifference of the ratings. We conclude that (3) synchronization is important for talking head quality. Based on (1–3) we assume that

H3: The quality of talking heads is influenced by visual quality, speech quality, fit of voice and head and synchronization of voice and lip movements.

For testing this hypotheses an experiment was carried out during which audio was recorded, system behavior was logged and subjective ratings were obtained using questionnaires described in the next section.

3. Methodology

3.1. The smart-home system

It is argued in [11] that the setting of an experiment should be as realistic as possible to obtain valid judgements and gain insights transferable to real applications. This requirement is met by conducting the experiment inside a fully functional living room. The room is equipped with a TV set, several lamps, blinds, and a phone as well as a sofa. All the devices can be

(a) Thinking Head (b) MASSY

Figure 1: The two different head components.

controlled by the user via the spoken dialog smart-home system INSPIRE.The users' speech is recorded by a portable wireless lapel microphone. The focus is not on speech input and a performance test of the speech recognizer yielded a low recognition rate. Therefore, the recognizer was replaced by a transcribing wizard. The resulting delay is supposed to be constant during the time of the experiment. The transcriptions are interpreted by a key-matching module. The dialog is managed via a set of generic dialog nodes connected through a local and a global branching logic. Initiative is mixed between user and system until a specific device and an action have been specified; then, the system takes the initiative to guide the user through further steps. System output is predefined as a gloze filled with information depending on dialog state and context. In case the user gave incoherent or incomplete information the user utterance is repeated by the system and clarification is asked for. The output is played back via one of six possible output components – or 'metaphors' – which will be described in the next section.

3.2. The output components

Four talking head and two voice-only metaphors are used as output components of the system. The talking heads are combinations of two animated heads and two speech synthesis systems. The first head (TH) originates from the Thinking Head Project [12]. This head is based on a 3D model of the Australian artist STELARC. In addition to having a human-like texture build from pictures of STELARC, it exhibits random head movements and extra-linguistic facial expressions like smiling and winking. The second head was developed at TU Berlin: MASSY, the **M**odular **A**udiovisual **S**peech **SY**nthesizer providing an accurate audio-visual speech synchronization, but being immobile otherwise [13]. See Figure 1 for pictures of the two heads.

The speech synthesis systems producing the respective voices include the **M**odular **a**rchitecture for **r**esearch on speech s**y**nthesis (Mary) [14] and the Mbrola system (Mbrola) [15]. For both systems a male German voice was selected, namely 'hmm-bits3' for Mary and 'de2' for Mbrola. Both speech synthesis systems are also used for the voice-only output.

3.3. Test design

The experiment took about 45 minutes. In total 49 participants, aged between 20 and 61 years (Median=26, SD=8.34) were paid for their attendance. For half of the participants the Mary voice was used for system output (voice-only or talking head), the other half listened to the Mbrola voice when interacting with the system. This factor (`Voice`) was varied between-

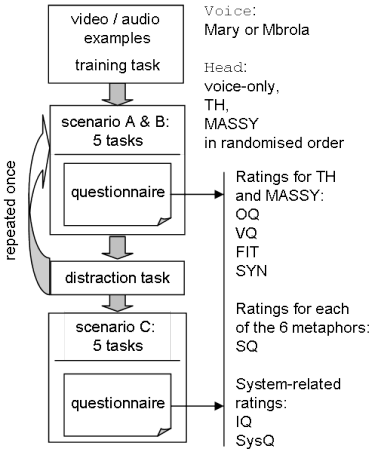

Figure 2: Experimental procedure.

How do you rate the overall quality?

◯ very good ◯ good ◯ undecided ◯ bad ◯ very bad

Figure 3: Example of one question to collect quality ratings.

subjects, while `Head` (= voice-only, TH, and MASSY) was varied within-subjects. Thus, every participant interacted with three of the six metaphors.

In Figure 2 the procedure of the experiment and the ratings obtained (abbreviations explained in Section 4) are depicted. The experiment is divided into multiple parts. The participant is seated on a sofa from where the whole room is visible. First, a short example audio or audio-visual file of the three system outputs is played on a screen in front of the participant. After this a training task is solved by the participants interacting with the system via the first metaphor. The main body of the experiment consists of three scenarios A, B, and C. During each of the three scenarios the participants interact with INSPIRE via one of the metaphors to solve 5 tasks. Task are paraphrased such as not to prime the language used by the participants. An example is:

> *'It is too bright outside. You'd like to shade the room.'*

The sequence of metaphors is altered for every participant. After each of the three scenarios the participant is asked to fill out a questionnaire, depending on the metaphor. Please confer to Fig. 2 for the experimental structure and the ratings obtained.

Quality aspects of the talking head metaphor were assessed in terms of OVERALL QUALITY (*'How do you rate the overall quality?'*), VISUAL QUALITY (*'How do you rate the visual quality of the head?'*), and SPEECH QUALITY (*'How do you rate the speech quality?'*). Additionally, participants were asked to rate the goodness of the COMPONENTS' FIT (*'How well does the voice fit with the head?'*) as well as the quality of SYNCHRONIZATION (*'How do you rate the synchronization of voice and lip movements?'*). The answer format used was a five-point rating scale, with the descriptions ranging from 'very good' (= 2) to 'very bad' (= −2) (cf. Fig. 3). In the case where the metaphor is voice-only these questions are reduced to the question assessing speech quality.

For the smart-home system we considered two aspects as important: overall SYSTEM QUALITY (*'How do you rate the overall quality of the system?'*) and INTERACTION QUALITY

(*'How do you rate the quality of the interaction?'*). Again, we used the scale described above.

After each of the first two runs – between answering the questionnaire and starting the next dialog – a distraction task is given. This task consists of building a given device with wooden bites following illustrated instructions. We choose two different tasks of similarly low difficulty. Participants were told, that the task was meant as a creative break only and that neither time taken nor task success was of any importance.

4. Results

In this section results for talking head quality, system quality and their relation are presented. For all scales assessed, it is verified whether they are measuring the intended parameter by comparing the results with results obtained in a previous non-interactive experiment [16]. For this purpose a few of the results of that experiment are summarized here:

`Head`: Thinking Head (TH) better than MASSY

`Voice`: Mary better Mbrola

For each analysis the data has been tested for normality of distribution. If a normal distribution could not be confirmed equivalent nonparametric tests have been used.

4.1. Talking head quality

The scales OVERALL QUALITY, VISUAL QUALITY, COMPONENTS' FIT, and SYNCHRONIZATION have been assessed only for the four talking heads while SPEECH QUALITY has been assessed for all six methapors. ANOVAs show that – as expected – OVERALL QUALITY (OQ) and VISUAL QUALITY (VQ) are dependent on `Head` (cf. Table 1). Effects of `Voice` on the scale SPEECH QUALITY would have been expected, as well as interaction effects of `Voice` and `Head` for COMPONENTS' FIT, and SYNCHRONIZATION. That no such effects are found is probably due to the fact that `Voice` was not varied within-subject.

Table 1: Results of the ANOVA for OVERALL and VISUAL QUALITY.

		F(1,46)	p	part. η^2
OQ	`Head`	9.05	.004	0.164
VQ	`Head`	10.37	.002	0.184

Given that the ranking of `Head` is similar for both voice groups, ratings are compared across the groups. A Wilcoxon signed rank test yields a higher OVERALL QUALITY for the TH (M=0.76, SD=0.56) than for MASSY (M=0.31, SD=0.83) (Z=-2.822, p=.002). The same results are obtained on the scale VISUAL QUALITY: TH (M=0.49, SD=0.869) is better than MASSY (M=-0.06, SD=0.84) (Z=-2.829, p=.002).

Comparing the two voices on the scale SPEECH QUALITY we find that Mary (M=0.82, SD=0.83) is preferred over Mbrola (M=0.46, SD=0.81) (Mann-Withney-U: U=2034, p=.006). Both findings, the ranking of `Head` and of `Voice`, are in line with the results found previously. Based on this we are confident that the scales measure the intended quality aspects, namely overall quality of the metaphor, visual quality and speech quality. In the following the hypotheses stated in the beginning of the paper will be tested.

Table 2: Impact of output component on dialog duration (dd) and SYSTEM QUALITY ($SysQ$), VO: voice-only.

	dd		$SysQ$	
	Z	p	Z	p
VO : TH	-1.80	.036	-0.21	ns
TH : MASSY	-4.85	.000	-2.20	.016
VO : MASSY	-5.53	.000	-1.74	.049

H1 is tested based on results gathered with the scale OVERALL QUALITY. In the case of voice-only output OVERALL QUALITY is assumed to be represented by SPEECH QUALITY. No differences between TH (M=0.76, SD=0.56) and voice-only (M=0.77, SD=0.86) output is found on this scale. Both, TH and voice-only are considered better than MASSY (M=0.31, SD=0.83) (Z=-2.822, p=.002 and Z=-2.737, p=.003 respectively). Thus, H1 is neither confirmed nor rejected.

H3 can be confirmed, as indicated by Spearman's correlations between OVERALL QUALITY and the scales VISUAL QUALITY (ρ=.47, p=.000), SPEECH QUALITY (ρ=.34, p=.001), COMPONENTS' FIT (ρ=.26, p=.010), and SYNCHRONIZATION (ρ=.45, p=.000). Furthermore, a linear combination of VISUAL QUALITY, SPEECH QUALITY (SQ) and SYNCHRONIZATION (SYN) explains $r^2 = 43\%$ of the variance of OVERALL QUALITY:

$$OQ = 1 + .40 \cdot VQ + .28 \cdot SQ + .28 \cdot SYN$$

COMPONENTS' FIT shows the smallest correlation with OVERALL QUALITY, and does not contribute significantly to the model.

Selected parameters extracted from log data are analyzed, namely *dialog duration* (dd) and *number of system turns* ($\#st$). These are traditionally used to measure inefficiency. While significant differences between the three metaphors are found for *dialog duration* (cf. Tab. 2), this is not the case for *system turns*. With voice-only, participants were fastest (M=160.49, SD=51.30), followed by TH (M=188.24, SD=100.55) and MASSY (M=290.82, SD=159.61). This can be explained with the processing time needed for animating the heads, which was even longer for MASSY than for TH. The dialog flow was not influenced by this, therefore no results are found for *system turns*. In any case, no positive influence of the talking head – the before mentioned 'persona effect' – can be found on these parameters.

4.2. Smart-home system quality

The rating of the smart-home system is assessed with INTERACTION QUALITY (IQ) and SYSTEM QUALITY ($SysQ$). To test **H2** the impact of the output component is analyzed. The results of a Wilcoxon signed rank test show that SYSTEM QUALITY is indeed rated better, when the output component with the higher OVERALL QUALITY is used (namely voice-only and TH). This confirms the hypothesis. The difference in smart-home SYSTEM QUALITY is significant only for the comparison of voice-only (M=0.80, SD=0.69) and TH (M=0.79, SD=0.65) to MASSY (M=0.53, SD=0.69) (cf. Tab. 2). This replicates the findings for metaphors' OVERALL QUALITY.

Inefficiency measured by *dialog duration* and *system turns* is – as would be expected – negatively correlated with INTERACTION QUALITY (Spearman's ρ = -0.33, p=.000) and SYSTEM QUALITY (Spearman's ρ = -0.29, p=.000). To find

out, whether the reduced efficiency found for interactions with MASSY could explain the metaphors' OVERALL QUALITY ratings, correlations of OVERALL QUALITY with *dd* are computed. No significant correlations are found, indicating that something else but the interaction quality is responsible for the differences in ratings.

5. Discussion

In this paper results from an experiment on the quality of talking heads, smart-home system quality and the interrelations between them have been reported. The hypotheses stated at the beginning have been tested based on subjective ratings and efficiency measures and the results are summarized below.

H1: We cannot confirm our first hypothesis, namely a preference of a talking head as output component for a smart-home system. Especially the missing impact on interaction performance has been found before (cf. [17]). In our case, an explanation could be that the interaction with the smart-home does not depend on a sophisticated dialog – which might profit from a talking head – but rather on a series of commands possibly followed by system questions. Besides, for the TV and EPG tasks important information was displayed on the TV screen, thus diverting users' attention from the talking head.

H2: Considering the impact of output component quality on system quality we can confirm our second hypothesis: the better the output component – as measured by user ratings – the better the perceived system quality.

H3: We were also able to confirm that visual quality, speech quality and synchronization between voice and lip movements are important for metaphors' overall quality. The fit of voice and head does not seem to be as important, given that commonsense rules – no female face with male voice – are followed.

By analyzing efficiency measures we were also able to examine the impact of interaction quality on perceived output component quality. The dialog when interacting with the talking heads took longer than for the voice-only output. This is due to a longer processing time required for the animated heads. But this has no effect on output components' overall quality while an effect on interaction and systems overall quality is found. Participants were obviously able to distinguish between these aspects.

6. Conclusions

We will further analyze our data, especially the audio recordings in order to better understand the interrelations between the quality of the talking head and the system. But, based on our findings so far, we cannot recommend using a talking head as an output component for a smart-home system. We could not confirm a positive 'persona effect' on interaction or user ratings for the rather restricted dialog found in this domain. This might be different for other applications also conceivable in a smart home like information or companionship services; applications that require a more human-like dialog and do not offer additional distraction on the visual channel. Another possibility – not analyzed in this study – is that privacy concerns (who wants someone in the living room constantly looking at him) outplay potential advantages in the dialog offered for example by facial expressions.

7. Acknowledgments

We would like to thank Matthias Siebke, Fabian Brinkmann and Raffael Tönges for their help in setting up and conducting the experiment and analyzing the data gathered. The project was financially supported by the Deutsche Forschungsgemeinschaft DFG (German Research Community), grant MO 1038/6-1.

8. References

[1] S. Möller, J. Krebber, A. Raake, P. Smeele, M. Rajman, M. Melichar, V. Pallotta, G. Tsakou, B. Kladis, A. Vovos, J. Hoonhout, D. Schuchardt, N. Fakotakis, T. Ganchev, and I. Potamitis, "INSPIRE: Evaluation of a Smart-Home System for Infotainment Management and Device Control," in *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC)*, vol. 5, May 2004, pp. 1603–1606.

[2] P. Baggia, G. Castagneri, and M. Danieli, "Field trials of the italian arise train timetable system," in *Proc. Interactive Voice Technology for Telecommunications Applications (IVTTA)*, 1998, pp. 97–102.

[3] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may I help you?" *Speech Communication*, vol. 23, no. 1-2, pp. 113–127, 1997.

[4] M. E. Foster, "Enhancing human-computer interaction with embodied conversational agents," in *Proc. Human Computer Interaction International (HCII)*, Beijing, July 2007.

[5] N. C. Krämer, *Soziale Wirkungen virtueller Helfer*, ser. Medienpsychologie. Stuttgart: Kohlhammer, 2008.

[6] D. C. Berry, L. T. Butler, and F. de Rosis, "Evaluating a realistic agent in an advice-giving task," *International Journal of Man-Machine Studies*, vol. 63, no. 3, pp. 304–327, 2005.

[7] D. M. Dehn and S. Van Mulken, "The impact of animated interface agents: a review of empirical research," *International Journal of Human-Computer Studies*, vol. 52, no. 1, pp. 1–22, 2000.

[8] S. Möller and J. Skowronek, "Quantifying the impact of system characteristics on perceived quality dimensions of a spoken dialogue service," in *Proc. European Conference on Speech Communication and Technology*, vol. 3, Geneva, 2003, pp. 1953–1956.

[9] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman, "Experienced quality factors: qualitative evaluation approach to audiovisual quality," in *Multimedia on Mobile Devices*, February 2007.

[10] C. Nass and L. Gong, "Maximized modality or constrained consistency?" in *Proc. International Conference on Auditory-Visual Speech Processing (AVSP)*, 1999, pp. 1–5.

[11] S. Möller, J. Krebber, and P. Smeele, "Evaluating the speech output component of a smart-home system," *Speech Communication*, vol. 46, no. 1, pp. 1–27, 2006.

[12] D. Burnham, A. Abrahamyan, L. Cavedon, C. Davis, A. Hodgins, J. Kim, C. Kroos, T. Kuratate, T. Lewis, M. Luerssen, G. Paine, D. Powers, M. Riley, Stelarc, and K. Stevens, "From talking to thinking heads: Report 2008," in *Proc. International Conference on Auditory-Visual Speech Processing (AVSP)*, 2008.

[13] S. Fagel and C. Clemens, "An articulation model for audiovisual speech synthesis – determination, adjustment, evaluation," *Speech Communication*, vol. 44, no. 1–4, pp. 141–154, 2004.

[14] M. Schroeder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[15] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vreken, "The MBROLA project: Towards s set of high-quality speech synthesizers free of use for non-commercial purposes," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 1393–1396.

[16] C. Kühnel, B. Weiss, I. Wechsung, S. Fagel, and S. Möller, "Evaluating talking heads for smart home systems," in *Proc. International Conference on Multimodal Interfaces (ICMI)*, 2008.

[17] H. Prendinger, J. Mori, S. Saeyor, K. Mori, K. Okazaki, Y. Juli, S. Mayer, H. Dohi, and M. Ishizuka, "Scripting and evaluating affective interactions with embodied conversational agents," *Zeitschrift Künstliche Intelligenz*, vol. 1, pp. 4–10, 2004.