Evaluating Talking Heads for Smart Home Systems

Christine Kühnel
Quality and Usability Lab,
Berlin Institute of Technology
Berlin, Germany
christine.kuehnel@telekom.de

Benjamin Weiss
Quality and Usability Lab,
Berlin Institute of Technology
Berlin, Germany

Ina Wechsung
Quality and Usability Lab,
Berlin Institute of Technology
Berlin, Germany

Sascha Fagel
Communication Science,
Berlin Institute of Technology
Berlin, Germany

Sebastian Möller
Quality and Usability Lab,
Berlin Institute of Technology
Berlin, Germany

ABSTRACT

In this paper we report the results of a user study evaluating talking heads in the smart home domain. Three non-commercial talking head components are linked to two freely available speech synthesis systems, resulting in six different combinations. The influence of head and voice components on overall quality is analyzed as well as the correlation between them. Three different ways to assess overall quality are presented. It is shown that these three are consistent in their results. Another important result is that in this design speech and visual quality are independent of each other. Furthermore, a linear combination of both quality aspects models overall quality of talking heads to a good degree.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Evaluation/methodology

General Terms

Experimentation

Keywords

Multimodal UI, talking heads, smart home environments

1. INTRODUCTION

Giving a system a face or even a whole body is supposed to improve the user satisfaction of this system [6]. This embodiment of a system has become a huge research field with studies for example focusing on mimic, gestures or emotions of Embodied Conversational Agents (ECAs) [7]. While more and more ECAs enter the first commercial applications, evaluation of this special form of output modalities moves into the field of interest (see [10] for an overview).

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

In this paper we present the results of a user study where different kinds of talking heads form the embodiment of a smart home system that allows the control of household devices via speech.

This study is embedded in a series of four user studies focusing on the evaluation of talking heads. Questionnaires and test designs to assess the influence of a talking head on user satisfaction and perceived quality are developed and thoroughly tested in this and subsequent user studies.

2. EXPERIMENTAL APPROACH

In this study three different talking heads, each combined with two different speech synthesis systems, were compared using a 2x3 within design, with the factors VOICE and HEAD being manipulated.

The three heads are of different appearance: The first head originates from the Thinking Head Project [2] and will be referred to by TH in the following. This head is based on a 3D model with the texture made from pictures of the Australian artist STELARC. It moves, smiles and winks when speaking. The remaining two heads were developed at the TU Berlin, one being the **M**odular **A**udiovisual **S**peech **SY**nthesizer (MASSY) [5], the other is a German Text-To-Audiovisual-Speech synthesis system based on speaker cloning (Clone) using motion capture [4]. Both heads are immobile apart from lower face movements, and the show no facial expressions.

The speech synthesis systems used are the **M**odular **A**rchitecture for **R**esearch on speech s**Y**nthesis (MARY) [11] based on HMM-synthesis, and the Mbrola system (Mbrola) [1] based on diphone synthesis. For both speech synthesis systems a male German voice ('hmm-bits3' for MARY and 'de2' for Mbrola) was used.

An informal listening test conducted before our experiment suggested that the speech produced by MARY is very natural and human-like and in general notably better than the Mbrola speech, which sounds more like a 'computer voice'. But MARY has one drawback: questions are not correctly intonated. After informal viewings of the heads, the following was concluded: while MASSY is of aesthetic but artificial appearance, both, Clone and TH look human-

like. Only TH exhibits eye and rigid head movements, the other two heads seem to be looking at a fixed point. And although Clone is compounded of smoothly synthesized video recordings, the eyes are clouded and do not look at the dialog partner. Furthermore, hair and ears are missing. This results in a rather unusual impression.

The analysis of this user study allows to tackle the following important questions arising from the considerations above.

- How important are appearance, animation and voice for overall quality?
- Which voice is preferred?
- How important is intonation for overall quality?
- Is a human-like head preferred in a smart home setting over an artificial head?
- Does the combination of voice and face have an influence on overall quality?

Ten sentences were recorded offline as videos for all 2x3 voice-head combinations. One example is:

'The following devices can be turned on or off: the TV, the lamps and the fan.'

Those sentences are of variable phrase length, contain both questions and statements and originate from the smart home domain. The 60 resulting videos will be referred to as 'stimuli' below.

Seven female and seven male participants aged between 20 and 32 (M= 27, SD=4.21) were paid to rate the six voice-head combinations in a two hour experiment. The experiment comprised two blocks, divided by a short break. The participants first received a short introduction and were asked four questions concerning their experience with talking heads and spoken dialog systems in general. They were seated in front of a screen on which the videos were displayed. The sound was played back over head-phones. Thus, the experiment can be described as watching-and-listening-only. Before the first block started the participants were shown six anchor stimuli, consisting of each voice-head combination speaking one sentence not contained in the abovementioned 10 sentences. Thus, every participant had seen the whole range of talking heads analyzed in this study before being asked for his rating.

In the first block all 60 stimuli were presented in randomised order. After every stimulus the participants were asked to answer four questions (per-sentence-questionnaire). One question concerning the content of the sentence was only included to focus their attention not only on the appearance but on understanding as well and was excluded from further analysis. With the remaining three questions the participants were asked to rate the *speech quality* sq ('How do you rate the speech quality?'), *visual quality* vq ('How do you rate the visual quality of the head?') and *overall quality* oq_1 ('How do you rate the overall quality?') of each stimulus. The answer format used was a five-point rating scale, with the descriptions 'very good', 'good', 'undecided', 'bad', 'very bad'.

In the second block a set of six stimuli followed by a questionnaire (per-set-questionnaire) was presented for every voice-head combination. This questionnaire assessed the *overall quality* oq_2 of the voice-head combination ('How do you rate the overall quality of the animated head?') and their overall impression sd ('Please use the following antonym pairs to rate your impression of the animated head.') using 25 semantic-differential items. Every item was rated on a five-point Likert scale with the poles described by antonyms. These items derive from a questionnaire currently being developed at our lab based on [3].

After the last per-set-questionnaire the anchor stimuli of all six conditions were displayed as freeze images simultaneously on one screen in random order. The participants could replay the stimuli by clicking on the images and where then asked to order them after their preference, giving ranks from '1 – least liked' to '6 – most liked'. Each number could only be given once, thus resulting in six ranks fr.

On paper they where finally asked to answer two open questions. The first question asking, why they had ranked the six conditions the way they had, the second question asking which condition they would prefer in a smart-home setting.

3. RESULTS AND ANALYSIS

The results of two participants were excluded from further analysis, one due to remarks he made during the test, implying he was not taking the test seriously and the other participant because he appeared as an outlier in most analyses. Thus the reported results originate from the ratings of 12 participants.

The results will be presented in four parts. The first part comprises the answers to the per-sentence-question, the second part reports the results of the per-set-questionnaire concerning the question assessing overall quality. The results of the 25 semantic differential items will not be discussed in this paper. The forced ranking task will be analyzed as a third part. The final open questionnaire is not reported in detail. In the last part, a comparison of the different questionnaires is presented.

3.1 Per-sentence-questionnaire

In the first block of the experiment the participants were ask to rate every stimulus on *speech quality*, *visual quality* and *overall quality*. Thus, the resulting data ('1 – worst' to '5 – best') can be analyzed concerning the influence of the text material the SENTENCES (1–10) are made of, the three different HEADS (TH, MASSY, Clone) and two VOICES (MARY, Mbrola). ANOVAs show, that *overall quality* is significantly dependent on all three variables, as is *speech quality*, whereas for *visual quality*, only the HEAD has a significant effect (cf. Table 1).

An inspection of Figure 1 confirms that *visual quality* is not influenced by VOICE or SENTENCE. Apparently, the test participants could clearly separate *visual quality* judgments from *speech quality*. A major result is the lacking influence of SENTENCE: The animation quality is not dependent on linguistic differences between the sentences. A Tukey posthoc test confirmed, that all three heads are rated significantly different ($\alpha = .05$), TH being rated better than MASSY than Clone.

Speech quality is significantly higher for MARY than for Mbrola. The effect of HEAD on *speech quality* is unexpected. The marginally significant influence shows, that the sub-

Table 1: Results of the ANOVA for overall, speech and visual quality

		F(df)	p	part. eta^2
Overall quality	SENTENCE	(9,659)=3.53	.000	0.03
	VOICE	(1,659)=96.71	.000	0.10
	HEAD	(2,659)=74.96	.000	0.15
Speech quality	SENTENCE	(9,659)=7.19	.000	.06
	VOICE	(1,659)=249.87	.000	.24
	HEAD	(2,659)=3.03	.049	.01
Visual quality	SENTENCE	(9,659)=0.46	.899	.00
	VOICE	(1,659)=0.17	.679	.00
	HEAD	(2,659)=389.09	.000	.5

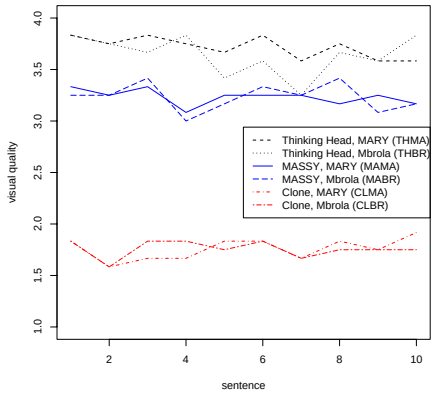

Figure 1: Visual quality

jects could not exclude the influence of the different heads entirely when concentrating on *speech quality*. One explanation could be that vision and audition are sensorily integrated with HEAD as a dominating factor with respect to the judgement of talking heads.

A Tukey posthoc test reveals, that stimuli with MASSY have been rated better than those with Clone ($\alpha = .05$). Concluding from Figure 2 and the effect size, this is not a great effect. The significant influence of SENTENCE was expected, as speech synthesis is known to perform differently for varying linguistic material.

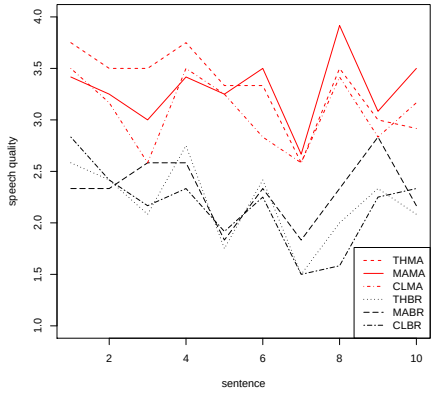

Figure 2: Speech quality

Overall quality seems to be the result of both, *speech* and *visual quality* (c.f. Figure 3). In its systematic, the influence of SENTENCE on *speech quality* is also present in *overall quality*, as are the differences between the levels of HEAD and VOICE. A regression analysis of *overall quality* based on data averaged over participants confirms this assumptions. Both, a linear model with *speech quality* and a linear model with *visual quality* yield $R^2 = .54$. However, the linear combination of both (cf. Eq. (1)) explains $R^2 = 96.7\%$ of the variance of *overall quality*.

$$oq = 1 + .65 \cdot sq + .46 \cdot vq \quad (1)$$

As the MARY synthesis can not produce appropriate intonation for questions, an additional analysis was conducted to test for a difference between questions and statements. There is a significant influence (F(1,707)=4.57; 0.033). However, there are two reasons why this result should not be interpreted as a difference between questions and statements. Firstly, there is no interaction effect with VOICE, therefore this effect is also valid for Mbrola. Secondly, the effect size is small compared to the SENTENCE effect and seems to originate from the latter.

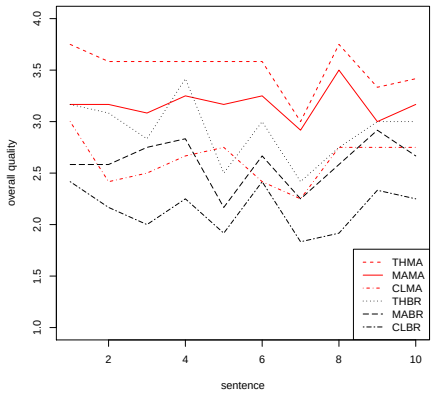

Figure 3: Overall Quality

3.2 Per-set-questionnaire

The per-set-questionnaire contained a question asking the participants to rate the *overall quality* oq_2 of each voice-head combination (1 – worst, 5 – best). The analysis of the effects of the two variables HEAD and VOICE showed a significant influence of HEAD (F(2,22)= 24.284, p=.000, part. eta^2=.688) but only a marginally significant impact of VOICE (F(1,11)=4.231, p=.064, eta^2=.278).

A t-test showed a clear, albeit not always significant difference in ranking for HEAD, with TH (M=3.46, SD=.50) better than MASSY (M=3.21, SD=.96) better than Clone (M=1.67, SD=.58) (cf. oq_2 in Table 2). The difference for VOICE, with MARY (M=2.92, SD=.52) better than Mbrola (M=2.64, SD=.48) is only marginally significant (t(11)=2.06, p=.064).

3.3 Forced Ranking

For this questionnaire the participants were forced to rank the six conditions according to their preference fr ('1 – worst', '6 – best'). A non-parametric two-way analysis of

variance [8] showed a significant effect of HEAD ($\chi^2(2)$=31, p=.001) but not of VOICE ($\chi^2(1)$=2.98, p=.08).

A Wilcoxon test yielded significantly different ranks for HEAD, with TH (M=5.08, SD=.7) better than MASSY (M=3.63, SD=.91) better than Clone (M=1.79, SD=.58) (cf. fr in Table 2), while the VOICE ranks reveal that MARY (M=3.72, SD=.66) is not significantly better judged than Mbrola (M=3.28, SD=.66) (Wilcoxon: Z=-1.07, p=.159).

Table 2: Ranking of Head for oq_2 and fr

	t-test (oq_2)		Wilcoxon (fr)	
	t(11)	p	Z	p
TH:MASSY	0.92	.337	-2.45	.007
TH:Clone	7.40	.000	-3.10	.000
MASSY:Clone	0.48	.000	-2.85	.001

3.4 Comparison of questionnaires

The seperate analyses for the three questionnaires – assessing *overall quality* in three different ways using two different scales – showed similar results. For the per-sentence-questionnaire both HEAD and VOICE have a significant influence on overall quality, while both, the per-set-questionnaire and the forced ranking yield a significant impact of HEAD but only a marginally significant impact of VOICE.

All questionnaires yield the same ranking for HEAD: TH > MASSY > Clone, and for VOICE: MARY > Mbrola. The rank order of the 2x3 voice-head combinations in *overall quality* achieved by a comparison of means for oq_1, oq_2 and fr show Thinking Head with MARY as the best rated option and Clone as the worst, regardless of the voice.

4. CONCLUSIONS

The reported study analyzed the perception of the quality of talking heads in the smart home domain in a watching-and-listening-only test. Three different questionnaires were used and the results compared. All questionnaires showed consistent results: In the smart home domain a human-like talking head system with a naturally sounding speech synthesis is preferred, followed by an aesthetic but artificially looking system.

Overall quality seems to be a result of visual quality (HEAD) and speech quality (VOICE), the participants were able to distinctly discern between these factors. The head variable has a significantly higher impact on overall quality than voice. This could be explained by the variation in quality, as the three different heads cover a broader range in quality than the two voices.

There was no interaction between HEAD and VOICE. According to [9] a consistent combination of head and voice would be better rated than the best possible combination. This is not confirmed by our findings; provided it can be assumed that the human-like TH is more consistent with the human-like MARY voice, while the animated MASSY head is in keeping with the computer-like Mbrola voice. Even though according to the open question a few participants did prefer a computer-like voice for MASSY, this effect is not reflected by the other data.

This could indicate that it might be sufficient to seperatly judge head and voice and combine the best rated ones, provided that a few fundamental conditions suggested by common sense (e.g. male voice for male head) are complied. This finding may not necessarily be generalized to all types of talking heads. But, we can expect a higher rating on overall quality if either one of the three heads or one of the two voices is improved.

The stimuli were composed of statements and questions. According to our analysis, this has no relevant effect for MARY voice on overall quality, contrary to our expectation. This does not imply that intonation is not important for voice synthesis systems. But we can assume that the intonation of questions in the smart home domain is less relevant.

In the subsequent studies the test design will be reused and the questionnaires will be validated further. In the next study the users will be interacting with the talking heads and the third study will be conducted in a fully equipped living room. Thus, influences of the setup and design on user perception of system quality can be analyzed.

5. ACKNOWLEDGMENTS

The authors would like to thank Rob Looijmans and David van der Pol for their help in setting up, conducting and analyzing the experiments. Thanks as well to the reviewers for helpful and encouraging comments on this work.

6. REFERENCES

[1] http://tcts.fpms.ac.be/synthesis/mbrola.html, last accessed 2008/05/22.

[2] http://thinkinghead.edu.au/, last accessed 2008/05/22.

[3] A. Adcock and R. Van Eck. Reliability and factor structure of the attitude toward tutoring agent scale (attas). *JILR*, 16(2):195–217, 2005.

[4] S. Fagel, G. Bailly, and F. Elisei. Intelligibility of natural and 3d-cloned German speech. In *Proc. AVSP*, 2007.

[5] S. Fagel and C. Clemens. An articulation model for audiovisual speech synthesis – determination, adjustment, evaluation. *Speech Comm*, 44, 2004.

[6] M. E. Foster. Enhancing human-computer interaction with embodied conversational agents. In *Proc. HCI International*, Beijing, July 2007.

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

[8] K. Kubinger. A note on non-parametric tests for the interaction in two-way layouts. *Biometric Journal*, 28:67–72, 1986.

[9] C. Nass and L. Gong. Maximized modality or constrained consistency? In *Proc. AVSP*, 1999.

[10] Z. Ruttkay and C. Pelachaud. *From Brows to Trust: Evaluating Embodied Conversational Agents.* Springer-Verlag, New York, USA, 2004.

[11] M. Schroeder and J. Trouvain. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *Int J Speech Tech*, 6:365–377, 2003.