

Test Series on the Influence of Talking Heads on the Quality of a Smart Home System – Experimental Outline and First Results

Christine Kühnel¹, Benjamin Weiss¹, Ina Wechsung¹, Sascha Fagel² & Sebastian Möller¹

¹Quality and Usability Lab, Berlin Institute of Technology, Berlin

²Communication Science, Berlin Institute of Technology, Berlin
Germany

{BWeiss;Christine.Kuehnel;Ina.Wechsung;Sebastian.Moeller}@telekom.de
; sascha.fagel@tu-berlin.de

ABSTRACT

In this paper a series of four user studies on the evaluation of talking heads in the smart home domain is outlined. Results of the first two experiments are reported. The findings of the first user study, a watching-and-listening-only test is verified through the second test, a web-based experiment. Both studies link three talking head components to two open source 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. Furthermore, three different ways to assess overall quality are presented, the questionnaires are validated and adjustments motivated.

SUMMARY

A growing research community is working on embodied conversational agents (ECAs), focusing for example on emotions (Krämer, 2008). With reporting the first results of a series of user studies this paper contributes to work done on evaluation (cf. Ruttkay & Pelachaud, 2004). In this series of four user studies three different talking heads, each combined with two different speech synthesis systems are compared using a 2x3 within design, with the factors VOICE and HEAD being manipulated.

The first head originates from the Thinking Head Project [4]. This head is based on a 3D model with the texture made from pictures of the Australian artist STELARC. The remaining two heads were developed at the TU Berlin [1]: The first one is the Modular Audiovisual Speech SYnthesizer (MASSY), the other is a German Text-To-audiovisual-Speech synthesis system based on cloned video recordings of a real speaker (Clone).

The speech synthesis systems used are the Modular Architecture for Research on speech sYnthesis (MARY) [2] based on unitselection, and the Mbrola system (Mbrola) [3] based on diphone synthesis. For both speech synthesis systems a male German voice ('hmm-bits3' for MARY and 'de2' for Mbrola) was used, which is congruent to the male heads reported above.

The heads are being evaluated in a series of four user experiments, differing with respect to the degree of interactivity. In the first two experiments the overall quality of the agent metaphor – voice and head decoupled of the smart home system (Erickson, 1997) – was investigated. Thus, effects of application system performance on the perceived quality of the ECA were avoided. The disjointed evaluation of the agent metaphor was done by recording 10 sentences offline as videos for all 2x3 voice-head combinations and judging each combination after their appearance in the videos in a watching-and-listening-only paradigm. One example of the recorded sentences 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. In the first two user studies the 60 resulting videos (respectively a selection of those) were displayed to the participants on a screen.

In the first experiment three different questionnaires and the experimental setup were tested for validity. The test was repeated with a reduced design as a web-based experiment, to further validate the – now partially adapted – questionnaires. The web-based experiment has the advantage of allowing access to much more test participants than usually possible in a lab-based experiment. A third and fourth experiment are planned to analyze possible changes in the quality ratings of the ECA when moving from purely watching/listening mode to a situation where the user interacts with the ECA. Thus, in the third experiment the user will be interacting with the smart home system through the talking head as an agent metaphor while he is still sitting in front of a screen. The fourth experiment will be set in a realistic living room, where the test participants can move around while interacting.

Advantages and disadvantages of a talking head as an output module of a smart home system will be focused on.

Analyzing the first user test separately, three interesting issues were tackled. (1) The question whether a human-like or an artificial talking head would be preferred in the smart home domain could be answered for our stimuli. (2) The influence of audio and visual quality on overall quality of the agent metaphor could be described by a simple linear model. (3) By analyzing the interaction between the factors VOICE and HEAD, we found out that there are no particularly well- and ill-fitting combinations of head and voice in our experiment.

The findings of the first experiment are summarized in the following paragraphs. All questionnaires showed consistent results: In the smart home domain a human-like talking head system (the Thinking Head) with a natural sounding speech synthesis system (MARY) is preferred.

Overall quality can be described by visual quality (HEAD) and speech quality (VOICE); a linear model is presented, explaining $R^2 = 96.7\%$ of the variance of overall quality. Furthermore, the participants were able to distinctly discern between these factors. In two of three tests the HEAD variable has a significant impact on overall quality whereas VOICE has not.

There was no interaction between HEAD and VOICE. This could indicate that it might be sufficient to separately judge head and voice and combine the best rated ones, provided that a few fundamental conditions suggested by common sense (such as male voice for male head) are complied. This finding can not be assumed for talking heads in general. 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. As the results might change for an interaction instead of a listening-and-watching-only situation, the subsequent test will also be analyzed to verify these findings.

The final paper presents the results and analysis of both, the first watching-and-listening-only experiment and the web experiment in comparison in more detail. The experimental approach is described, and changes in the questionnaires are motivated and validated.

REFERENCES

- Erickson, T., 1997. *Designing Agents as if People Mattered*. Intelligent Agents (ed. J Bradshaw). Menlo Park: AAI Press.
- Krämer, N.C., 2008. *Soziale Wirkungen virtueller Helfer*. Medienpsychologie. Kohlhammer, Stuttgart.
- Ruttkey, Z. & Pelachaud, C., 2004. *From Brows to Trust: Evaluating Embodied Conversational Agents*. Human Computer Interaction Series. Springer-Verlag, New York, USA.
- [1] <http://fourier.kgw.tu-berlin.de/>, last accessed 2008/05/22.
- [2] <http://mary.dfki.de/>, last accessed 2008/05/22.
- [3] <http://tcts.fpms.ac.be/synthesis/mbrola.html>, last accessed 2008/05/22.
- [4] <http://thinkinghead.edu.au/>, last accessed 2008/05/22.