the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

Quality of Talking Heads in Different Interaction and Media Contexts

Benjamin Weiss[a,*], Christine Kühnel[a], Ina Wechsung[a], Sascha Fagel[b], Sebastian Möller[a]

[a] *Quality and Usability Lab, Deutsche Telekom Labs, TU Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany*
[b] *Institut für Sprache und Kommunikation, TU Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany*

Abstract

We investigate the impact of three different factors on the quality of talking heads as metaphors of a spoken dialogue system in the smart home domain. The main focus lies on the effect of voice and head characteristics on audio and video quality, as well as overall quality. Furthermore, the influence of interactivity and of media context on user perception is analysed. For this purpose two subsequent experiments were conducted: the first was designed as a non-interactive rating test of videos of talking heads, while the second experiment was interactive. Here, the participants had to solve a number of tasks in dialogue with a talking head. To assess the impact of the media context, redundant information was provided via an additional visual output channel to half of the participants. As a secondary effect, the importance of participants' gender is examined. It is shown that perceived quality differences observed in the non-interactive setting are blurred when the interactivity and media contexts provide distraction from the talking head. Furthermore, a simple additional feed-back screen improves the perceived quality of the talking heads. Gender effects are negligible concerning the ratings in interaction, but female and male participants exhibit different behaviour in the experiment. This advocates for more realistic evaluation settings in order to increase the external validity of the obtained quality judgements.

Keywords: embodied conversational agent, smart home, talking head, usability, WOZ

1. Introduction

A growing research community is working on Embodied Conversational Agents (ECAs). Human-human dialogues are studied and insights concerning, for example, emotions (Krämer, 2008) and gestures (Kipp, 2004) are transferred to the human-computer interaction (HCI). Evaluation of ECAs in HCI addresses e.g. the impact of human-like expressions on system efficiency (Ruttkay and Pelachaud, 2004) or on the intelligibility of the audio-visual speech (Massaro et al., 2000). With the LIPS-Challenge a unified approach was started in 2008 to annually evaluate different synthesis systems and techniques with the same training data, subjects, test material and metrics, including intelligibility and naturalness of the produced sentences (cf. Theobald et al., 2008, for the specification).

It is assumed that the supposed similarity to human

*Corresponding author. Tel.: +49 30 8353 58526.
Email addresses: BWeiss@telekom.de (Benjamin Weiss), Christine.Kuehnel@telekom.de (Christine Kühnel), Ina.Wechsung@telekom.de (Ina Wechsung), Sascha.Fagel@tu-berlin.de (Sascha Fagel), Sebastian.Moeller@telekom.de (Sebastian Möller)

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

face-to-face communication (Krämer and Bente, 2002) eliminates the need of learning special strategies for HCI (Xiao et al., 2002). It is argued that one of the main advantages of ECAs is a facilitated interaction compared to typical user interfaces by intuitively using (subconscious) social cues (Cassell et al., 2000). Such cues can be both verbal and non-verbal such as signalling disagreement with gestures.

Users might be more motivated to interact with an ECA than a traditional uni-modal interface (Takeuchi and Naito, 1995). In Pandzic et al. (1999) participants rated the waiting time of a theatre information system more satisfying with a talking head than with speech only or text. Apparently, they were more distracted and thus accepted the comparable waiting time more easily. More positive ratings of dialogue systems that include ECAs are explained with the so called 'persona effect': the positive effect on user's interaction induced by a life-like interface agent (Dehn and Van Mulken, 2000; Lester et al., 1997; Van Mulken et al., 1998). Consequently, studies are conducted to examine the persona effect on measures of efficiency and effectiveness: log-data in the case of the dialogue itself or for example knowledge acquisition test results as a correlate for effectiveness in tutoring systems for teaching. However, results are ambiguous (cf. Yee et al., 2007), as sometimes positive effects are observed (e.g. Gong and Nass, 2007; Cowell and Stanney, 2005), and sometimes no effects are found (e.g. Prendinger et al., 2004; Xiao, 2006).

Due to differences in test design, results of evaluation studies are difficult to compare and interpret. For instance, ECA features which are varied in a test include appearance and gender (Buisine et al., 2004; Zimmerman et al., 2005), but also the level of anthropomorphism (cf. Gong, 2008; King and Ohya, 1996; Nowak, 2004), or the usage of non-verbal gestures (Buisine et al., 2004). Using either text (e.g. Sproull et al., 1996) or speech (e.g. Berry et al., 2001) as a reference (anchor) condition is another important methodological difference, that complicates comparability. In particular, the persona effect is highly dependent on task domain: Zimmerman et al. (2005) could confirm a positive effect of human-like ECAs on perceived usability for a financial task domain, but not for entertainment or tutoring. In the studies of Xiao et al. (2002) and André et al. (1998) no persona effect was observed, but the ratings assessed were dependent on type of task – e.g. three different avatars were rated more positively after a text-editor tutorial than after the task of selecting items for travelling in Xiao et al. (2002).

Not in all studies – especially not in tutoring application – did test participants actually interact with the ECA (e.g. Breitfuss et al., 2008; Buisine et al., 2004; McBreen and Jack, 2000; Nowak and Rauh, 2005). This is a relevant distinction, as for example the more human-like ECA is rated more intelligent in a non-interactive condition, but not after an interaction (Koda and Maes, 1996).

To sum up, the persona effect has been observed in several conditions. However, results differ for the context factors mentioned above and may therefore not be generalized over different ECAs, tasks or domains. Additionally, the effect is more obvious in subjective ratings than in interaction parameters. In a comparison of three different facial models, a standard cartoon-like head speaking one welcome message was rated significantly more appealing than a synthesis based on real video samples (Pandzic et al., 1999). The standard face but with additional texture was rated worst. Apart from

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

this interesting ranking of the varying degrees of naturalness in timing and appearance (the sample-based head displayed the most natural face but least natural movement) the ratings were remarkably low in general. A similar selection of facial models is used in our experiments.

As a complement to these studies, our work focuses on factors influencing perceived quality in the smart home domain. This is analysed based on data obtained in two subsequent experiments: a non-interactive rating experiment and an interactive task-solving experiment.

One of the main questions is the impact of talking head components (TTS, animated head) on the talking head quality: *How do participants perceive talking heads? What is the influence of the TTS and head component?* The impact of interactivity is another focus: *Are characteristics of talking heads still perceived as different when entering a real dialogue?* Changes in perception and judgement due to an additional output channel are analysed as well: *Does additional information offered on a screen impede the evaluation of talking heads? Does it enhance the interaction quality? Are participants able to distinguish between talking head, system, and interaction quality?*

The application is a smart home environment, where several devices, for example a TV and an answering machine can be controlled via spoken input. The ECA – in our case a talking head – acts as the interface and thus as the metaphor of the system. We chose male ECAs as appropriate for this purpose, as a male voice was rated more competent than a female voice for a technical domain in Nass et al. (1997).

Participants prefer ECAs with an obvious gender over androgynous ones, and they prefer ECAs of their own gender (Nowak and Rauh, 2005; McBreen and Jack, 2000, only for women). Therefore, dependence of ratings on participants' gender is analysed as well.

The remainder of this article is structured as follows. In Section 2 the methodology of both experiments is described and the talking heads and their components are characterized. In Section 3 set-up, procedure, and results of the first experiment are explained. The second experiment is presented in Section 4, including analysis of assessed ratings and extracted log-data. The results cover the impact of degree of interactivity and media context on the ratings of the different metaphors. Gender effects are presented in Section 5. The results are discussed in Section 6, and conclusions are drawn in Section 7.

2. Methodology

The goal of these two experiments is to assess the impact of different text-to-speech and text-to-visual-speech components on the perceived quality of both the talking heads as metaphors of a dialogue system and the system itself. It is of special interest how context factors like the degree of interactivity and amount of available media feedback interact with the perception of quality.

In a first watching-and-listening-only experiment (E1) six different talking heads, combinations of three head components and two German speech synthesis systems, were compared using a 3x2 within-subject design. The aim of this experiment was to evaluate the influence of the two components (head and voice) on the speech, video and overall quality ratings of the talking heads in a non-interactive setting. The second Wizard-of-Oz experiment (E2) compared the four best-rated head-voice combinations in an interactive setting, addressing the validity of the quality scores obtained in

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

E1 for future system usage.

The first head (TH) originates from the Thinking Head Project (Burnham et al., 2008). This head is based on a 3D model of a human being, in this case 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. As the control of the visual articulation was built for English and does not define separate phonetic articulators (like lip-spreading or jaw opening), but target visemes, a German synthesis was made by hand using the most appropriate English visemes and applying basic co-articulation rules from Massy (see below). The original English visemes were created from motion-capture data. The two following head components do not exhibit facial expressions or movements apart from visual articulation. The second head was developed at TU Berlin: Massy (MS), the **M**odular **A**udiovisual **S**peech **SY**nthesizer is a parametric 3D head model and provides accurate audio-visual speech synchronization and includes articulators like the velum and tongue body which are not always visible. MS also accounts for co-articulation with rules based on empiric data of German (Fagel and Clemens, 2004). The third head is a 2D German Text-To-Audiovisual-Speech synthesis system based on speaker cloning (CL) using motion capture. The co-articulation behaviour was extracted from the videos. CL was developed by a cooperation between TU Berlin and GIPSA-lab Grenoble (Fagel et al., 2007). Pictures of the three head components are displayed in Fig. 1. Because of the low quality scores obtained in the first experiment, Clone was not used in E2.

The speech synthesis systems producing the respective voices include the **M**odular **A**rchitecture for

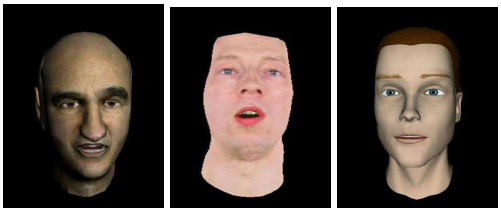


Figure 1: Three head components from left to right: Thinking Head, Clone, and Massy.

Research on speech s**Y**nthesis (Mary) (Schroeder and Trouvain, 2003) and the Mbrola system (Mbrola) (Dutoit et al., 1996). A male German voice was selected for both systems, namely 'hmm-bits3' for Mary and 'de2' for Mbrola. Both were considered best from a selection of the two TTS-systems in an earlier informal listening test.

3. Experiment E1

3.1. E1 – Procedure

In the first experiment (E1), videos of the talking heads speaking short, meaningful sentences (approx. 2 seconds) related to a smart-home domain were presented to 14 participants (aged 20 to 32, $M = 27$, $SD = 4.21$, gender balanced, paid for attendance). Thus, the quality of the agent metaphor – i.e. voice and head decoupled of the system (Erickson, 1997) – could be analysed. This approach allows the assessment of the metaphor quality in a strictly controlled setting, and without any interfering interaction.

60 videos were pre-recorded presenting the talking heads uttering 10 sentences for all 2x3 voice-head combinations. The sentences are of variable phrase length, and contain both questions and statements. One example is:

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

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

The participants first received a short introduction and were asked four questions concerning their experience with talking heads and spoken dialogue systems in general.

The body of experiment E1 was divided into two parts, one per-sentence part and one per-set part. The per-sentence part consisted of single stimuli presented in randomized order. After every stimulus the participants were asked to answer four questions (per-sentence-questionnaire). One question concerning the content of the sentence – included only to focus their attention not exclusively on the appearance but on understanding as well – 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 of the talking head (*'How do you rate the overall quality?'*) for each stimulus (cf. Fig. 2). In the per-set

How do you rate the overall quality?

○ very good ○ good ○ undecided ○ bad ○ very bad

Figure 2: Example of one question to collect quality ratings.

part a set of six stimuli was presented for every voice-head combination followed by a questionnaire (per-set-questionnaire). The order of the six sets was randomized for each participant. This questionnaire assessed the participants' detailed impression of the talking head (*'Please use the following antonym pairs to rate your impression of the animated head.'*) using 25 semantic-differential items (cf. Fig. 3). Every item was rated on a five-point scale with the poles described by antonyms. These items derive from a questionnaire currently being developed at our lab based on Adcock and Van Eck (2005). For the analysis, the ratings of the quality scales

pleasant	○	○	○	○	○	unpleasant
reserved	○	○	○	○	○	enthusiastic

Figure 3: Example of one item of the semantic differential.

are transformed to -2 (very bad) to 2 (very good). The same is done for the results of the semantic differential. For those antonyms not intrinsically positive or negative, the more human-like and dynamic (e.g. enthusiastic) are chosen as positive ends of the sub-scales.

3.2. E1 – Results for the different head and voice components

The non-interactive experiment E1 yielded three major results: Overall Quality of the metaphors can be described as a linear combination of the Visual Quality related to the appearance of the talking head and the Speech Quality related to the synthesis system (data averaged over all 10 sentences for each participant, Pearson's $r = .83$, $E_p = .49$, $p < .001$).

$$OQ = .47 + .51 \cdot SQ + .33 \cdot VQ$$

Furthermore, in E1 the participants were able to distinctly discern between these two aspects, insofar as Visual Quality is only dependent on the head component and Speech Quality much more on the speech than the head component. No interaction effects between these two factors could be found. Finally, the participants prefer the metaphor which receives the best head and voice ratings – the more human-like talking head (TH), and the more natural speech synthesis (Mary). The two parts

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

of the experiment – single sentences randomized and a set of sentences for each metaphor – are consistent in their results. Considering all data, there is a clear ranking for the six different combinations, i.e. TH better than MA better than CL, with Mary rated higher than Mbrola for each head component. For more information and detailed statistics, please confer to Kühnel et al. (2008).

Analysing the semantic differential revealed three factors (*naturalness*, *friendliness*, *attractiveness*), which all differed systematically for the six metaphors (cf. Weiss et al., 2009). The more friendly, natural and attractive a metaphor was rated, the higher the Overall Quality was.

TH is considered as friendly as MS, but more natural and more attractive. It is interesting to see that friendliness does not correspond to human-like texture, or natural extra-linguistic movements. We can only speculate that friendliness might depend on other features representing a different personality (such as head shape or more constant slight smiling) rather than on the degree of artificiality.

4. Experiment E2

4.1. E2 – Procedure

In E2, the talking heads were presented to 46 participants (22 men, 24 women) as metaphors of a spoken-dialogue system. They had not taken part in E1 and were thus unfamiliar with the metaphors. The age of the participants ranged between 20 and 60 years (M=28.92, SD=7.65) and they were paid for their attendance. The participants were seated in front of a table inside a laboratory room which is designed for audio and video experiments. The metaphor was displayed on a screen (21") in front of the participants. When not articulating the talking heads remained static.

The participants interacted via headphones with the metaphor using free speech. They were asked to complete 7 different tasks originating from the smart-home domain once with each of the four metaphors (head and voice combinations). These tasks were grouped in an *answering machine scenario* (AM) consisting of 3 tasks and an *electronic program guide scenario* (EPG) consisting of 4 tasks. A sample dialogue for each scenario can be found in Figure 4.

The focus of E2 is on talking head quality. Therefore, the interaction should be comparable between participants in E2. To achieve this, the dialogue flow was controlled: the tasks were written on separate cards and offered to the participants in a predefined order. Every participant had to carry out both scenarios once with each metaphor. To avoid boredom the tasks were altered slightly in expression and content while the level of difficulty of each task remained constant. The order of scenarios (AM → EPG or EPG → AM) was varied between participants as depicted in Fig. 5.

4.1.1. Degree of interactivity

Log-data was recorded in terms of system output and time stamps. The system output consisted of pre-recorded films of the talking heads, played by the wizard once he received input from the participants. Because of the controlled interaction, the possible prompts and their order as played by the wizard are of limited variety. Thus, system output and order of system output is basically the same for each participant. To monitor the success in controlling the dialogue flow parameters are extracted from the log-data and analysed. These parameters are

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

AM: 'You like to call back Andrea immediately. If she does not answer, try again in two hours.'

P: *I would like to call back the caller.*

S: Trying to establish a connection.

S: The line is busy. Would you like to try again later? To stop trying, please say 'abort'.

P: *I would like to call back later.*

S: When would you like to try again?

P: *In two hours.*

S: The reminder is set to two hours.

EPG: 'You decide to watch Bleak House tonight. Find out when the movie starts. Ask the system to remind you at the beginning of the movie.'

P: *When does Bleak House start?*

S: Bleak House starts at 8:15 pm. Would you like to record the movie or watch it?

P: *I would like to watch it tonight. Please alert me when the movie starts.*

S: The beginning of the movie will be indicated to you.

Figure 4: Sample interaction for AM and EPG tasks. Statements of the participant (P) and system (S).

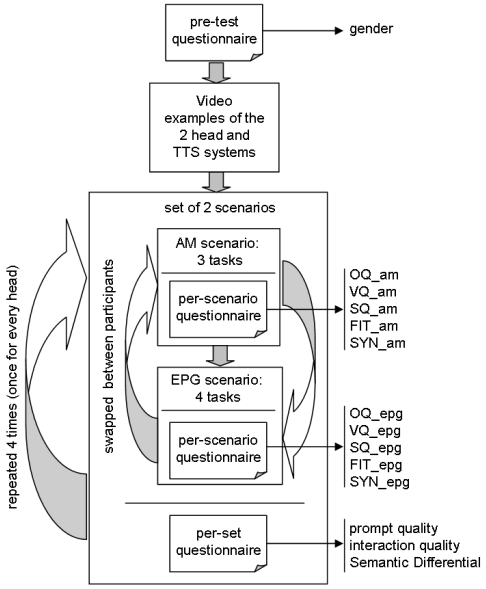


Figure 5: Structure of experiment E2 with associated measures.

- dialogue duration (*dd*): beginning of interaction (welcome message of head component) to end of interaction of each set,
- number of system turns (*#turns*): number of videos played per set,
- number of help- (*#help*), 'no input'- (*#noInput*) and 'no match'-messages (*#noMatch*): number of additional videos played per set,
- number of times the participant departed from the predefined task order (*#back*).

and will be explained in the following.

If the participants deviated from the required dialogue flow by changing the order of tasks, this was logged as a parameter (*#back*). If the participant skipped a task, the wizard had a few standard videos to bring him back on track. This also was logged (*#help*). If the participant said something which was unforeseen and not accounted for by the pre-recorded videos the wizard could

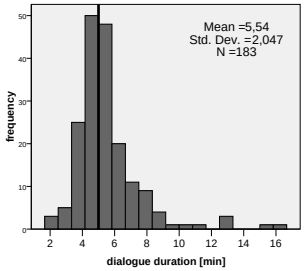
the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

play a 'no match' prompt (*#noMatch*). And if the participant remained quiet for a while, a 'no input' prompt was played and logged as *#noInput*.

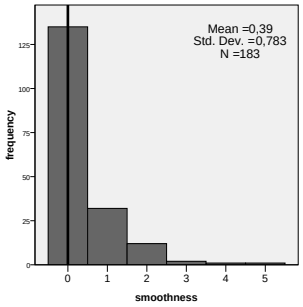
To measure smoothness of dialogue flow an additional parameter 'smoothness' (*sm*) is defined: the sum of *#help*, *#noInput*, *#noMatch* and *#back*. Task success is not measured as the task cards and the dialogue were designed in such a way that every participant solved every task.

In the case of a non-cooperative participant the above introduced interaction parameter would deviate from the expected values found by running the experiment with one expert (*#turns* = 26, *sm* = 0 and *dd* = 5 min). The number for *#turns* and *sm* represent optimal values while the number found for dialogue duration is an approximation.

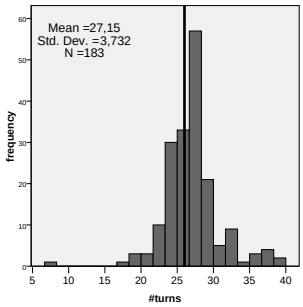
As wizard the observation was made that participants indeed tended to adhere to the predefined dialogue flow. To validate this finding participants behaviour as measured by the three parameters is compared to the expected values. As shown in Fig. 6 the mean value does not vary much from the expected value indicated by the vertical line. Thus, the observation that most participants followed the predefined dialogue closely (*#turns* $\in [24, 28]$, *sm* = 0 and *dd* $\in [4, 6]$ min) could be confirmed. When comparing the ratings of those participants to those who did vary more no effect could be found. Although the dialogue appeared non-optimal to the wizard, the participants did not notice.



(a) dialogue duration (*dd*)



(b) smoothness (*sm*)



(c) number of turns (*#turns*)

Figure 6: Distribution of the parameters with mean and standard deviation, vertical line indicates expected values.

4.1.2. Assessing the impact of additional media

To analyse the impact of redundant information, half of the participants received visual information on an additional screen (cf. Fig. 7). This additional screen was used to simulate the feedback from an answering machine and an electronic program guide according to the task. In the case of the AM scenario an answering machine was displayed, indicating with a red light that new messages were available. This turned to green once all new messages were played, so that the first task was

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

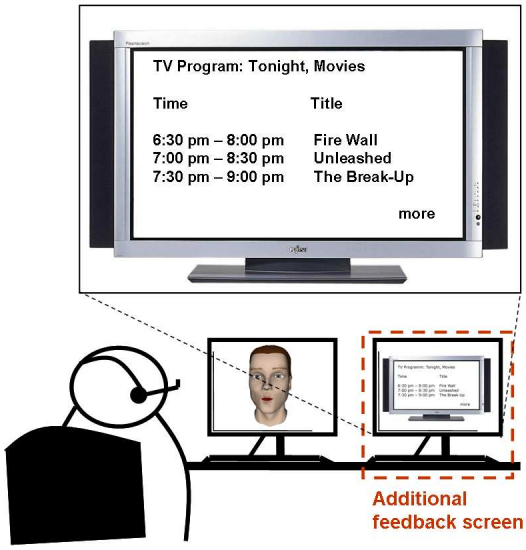


Figure 7: Feedback on the additional screen.

solved. In the EPG scenario, the TV program, lists of recorded films or an alarm clock was displayed, depending on the task (cf. Fig. 7).

4.1.3. Rating scales and questionnaires

Like in E1, quality aspects of the metaphor were assessed in terms of Overall Quality, Visual Quality, and Speech Quality. These quality aspects were rated by the participants after completing each scenario (per-scenario-questionnaire). Additionally, participants were asked to rate the goodness of the components' fit Fit (*'How well does the voice fit with the head?'*) as well as the quality of synchronization Syn (*'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' to 'very bad' (cf. Fig. 2), identical with the one used in E1.

These five items are referred to either by scenario (in the case of the Overall Quality e.g.: OQ_{AM} for the answering machine scenario and OQ_{EPG} for the electronic program guide scenario), or by the average of both values (e.g. Overall Quality: $\overline{OQ} = \frac{1}{2} \cdot [OQ_{AM} + OQ_{EPG}]$).

After both scenarios (at the end of the interaction with each metaphor), Prompt Quality PQ (*'How do you rate the quality of the prompts of the talking head?'*) and Interaction Quality IQ (*'How do you rate the quality of the interaction?'*) were assessed with the scale described above on a per-set-questionnaire. Please refer to Fig. 5 for the experimental structure of E2 and the scales obtained. As in E1, this questionnaire also included a semantic differential, where every item was rated on a seven-point scale with the poles described by antonyms. These 50 items include the 25 items used in E1 to assess the detailed impression of the metaphor and further items related to prompt quality and interaction quality.

4.2. E2 – Results: degree of interactivity

To analyse the influence of the interaction, results of E2 regarding the different head and voice components will be presented in relation to the results of E1. In addition to the quality ratings, data from the semantic differential and interaction parameter from the log-data will be examined in this section.

4.2.1. E2 – Results: quality scales

In contrast to the results of E1, the metaphor ratings in the interaction experiment E2 are less clear. Despite two additional scales, the linear combination of the four quality ratings Speech Quality, Visual Quality, Fit, and Synchronization is not as suitable to describe average Overall Quality (Pearson's $r = .66, E_p = .61$, $p < .001$, apart from the two scales also used in E1, Syn does significantly contribute to this model, whereas Fit

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

does not).

$$\overline{OQ} = .28 + .35 \cdot \overline{SQ} + .21 \cdot \overline{VQ} + .16 \cdot \overline{\text{Syn}}$$

Comparing the experiments E1 and E2 the clear ranking of heads found in E1 is conspicuously missing in E2. The only significant result replicated in E2 is a higher Speech Quality for the Mary voice compared to Mbrola ($F(1, 181) = 10.35$, $p < .01$). However, no differences are found for Visual Quality, Overall Quality for each scenario, or any quality rating obtained at the end of the interaction with each metaphor (Prompt Quality or Interaction Quality).

Analysing the additional items present in E2 – synchronization of audible speech and lip movements and the fit of head and voice components, a better fit was ascribed to the TH with Mary voice than the others ($F(3, 178) = 6.44$, $p < .001$, Tukey posthoc tests) and MS was judged to have a better synchronization than TH ($F(1, 182) = 6.61$, $p < .05$).

Nonetheless, participants had been able to perceive differences between the metaphors. This could be observed not only for some of the quality ratings (Speech Quality, Fit, Syn), but also for the semantic differential. When analysing quality aspects separately based on the semantic differential, we find the same ranking already known from E1 for 2 of the resulting factors. The detailed results are presented in the following section.

4.2.2. E2 – Results: semantic differential

The semantic differential with its 50 antonyms allows to analyse the metaphors' perception in a more analytic way than explicit quality ratings. This will be used to explain the differences in quality ratings between E1 and E2. The factor analysis of the data from the semantic differential using Horn's parallel analysis (Horn, 1965) and oblique rotation – as recommended by Costello and Osborne (2005) – reveals 6 factors (52% explained variance, with a Kaiser-Meyer-Olkin measure of sampling adequacy (Hutcheson and Sofroniou, 1999), $KMO = .91$. They can be entitled as

1. "*quality of the prompts*" (Cronbach's $\alpha = .95$),
2. "*metaphor's attractiveness*" ($\alpha = .91$),
3. "*metaphor's naturalness*" ($\alpha = .91$),
4. "*cognitive demand during the interaction*" ($\alpha = .90$),
5. "*trust towards the system*" ($\alpha = .85$), and
6. "*entertaining value*" ($\alpha = .82$).

Only two pairs of factors correlate strongly with each other ($r > .5$), *quality of the prompts* with *cognitive demand* ($r > .62$) and *naturalness* with *entertaining value* ($r > .69$). The last pair is of special interest, as both factors are the only ones showing different results for the metaphors, using the means of the items for each factor ($F(3, 178) = 10.59$, $p < .001$; $F(3, 178) = 10.02$, $p < .001$, respectively). For both factors TH receives higher values than MS and Mary is rated better than Mbrola. This ranking was also obtained for quality scales in the non-interactive setting.

The similarity in the results for *naturalness* and *entertaining value* (cf. Fig. 8) can be explained due to the feature differences of the TH compared to MS: In contrast to MS, TH exhibits both, a natural texture most probably responsible for the result for factor 3 and random extra-linguistic movements resulting in the difference in factor 6.

However, there is no explanation why no other factor differs for the metaphors. In the non-interactive experiment (E1), all three factors found, including *attractiveness* did show results.

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

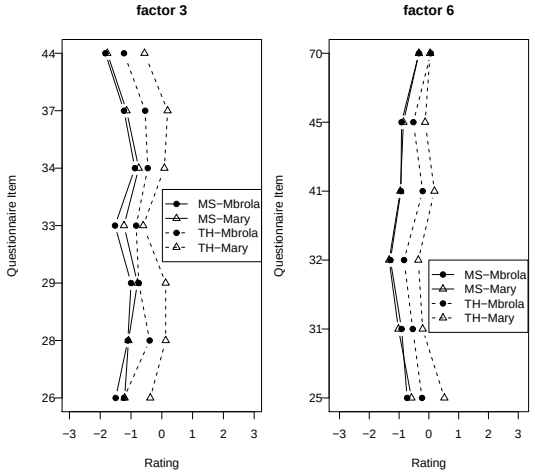


Figure 8: Results for factor 3 "metaphor's naturalness" and factor 6 "entertaining value". The Y-axis shows the number of the questionnaire items the factors consist of.

4.2.3. *E2 – Results of the correlation analysis between different scales*

It is assumed that some of the resulting factors correlate with the quality ratings. Especially in the case of Overall Quality and Visual Quality, which do not differ for the metaphors, correlations could give useful insight, whether quality ratings may be influenced by some unforeseen factor. Results of correlation analyses are presented in Table 1. There are only two strong correlations ($r > .6$) between factors and quality ratings: (1) *attractiveness* and Overall Quality and (2) *quality of prompts* and Prompt Quality. Whereas the latter result is quite obvious – of course there is a strong correlation between the Prompt Quality scale and the averaged ratings of the factor 1 comprising sub-scales related to the prompts – the former result indicates the high impact of this hedonic aspect on perceived overall quality.

In addition, there is a strong relation between Interaction Quality and Overall Quality.

4.2.4. *E2 – Results for the interaction parameters*

Analysing the interaction parameters, no differences between the four metaphors nor between the one- and the two-screen setting could be found. Due to the controlled dialogue flow, the missing impact on interaction of both the metaphors and the degree of media context could be expected.

Two different groups could be identified based on interaction parameters, namely a group which had a smooth dialogue and a group characterized by a problematic dialogue (see Section 4.1.1). But these groups do not differ in any of the ratings obtained.

Furthermore, there are no correlations to be found between the interaction parameters and any of the questionnaire items. It is not possible to build a meaningful model using the interaction parameters to explain any of the questionnaire items.

To sum up, the course of interaction has no perceivable influence on the metaphor ratings. Vice versa, neither metaphor characteristics nor media context have any impact on the interaction.

4.3. *Degree of interactivity – summary*

Our analysis shows that the degree of interactivity has an impact on the quality rating of the metaphors. In summary, the participants were able to distinguish between the different metaphors in both experiments. However, in E2 these differences are only visible for Speech Quality, Fit, Synchronization, *naturalness* and *entertaining value*. Both factors resulting from the semantic differential are independent from Overall Quality in the interactive setting.

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

Table 1: E2: Pearson's r for correlations between different scales (bold for results > .6), fx designates Factor x. See labels on the y-axis for explanation of the abbreviations.

	$\overline{OQ}$	$\overline{SQ}$	$\overline{VQ}$	$\overline{\text{Fit}}$	$\overline{\text{Syn}}$	f1	f2	f3	f4	f5	f6	PQ	IQ
Overall Quality $\overline{OQ}$	—												
Speech Quality $\overline{SQ}$	.57	—											
Visual Quality $\overline{VQ}$	.47	.32	—										
Fit $\overline{\text{Fit}}$	.40	.49	.29	—									
Synchronization $\overline{\text{Syn}}$	.49	.43	.53		—								
quality of prompts f1	.43	.28	.35	.14	.21	—							
attractiveness f2	**.61**	.47	.55	.41	.47	.42	—						
naturalness f3	.32	.38	.39	.47	.37	.19	.46	—					
cognitive demand f4	.44	.16	.32	.18	.36	**.62**	.34	.19	—				
trust f5	.27	.16	.28	.09	.14	.38	.41	.11	.32	—			
entertaining f6	.35	.37	.34	.32	.23	.17	.47	**.70**	.18	.08	—		
Prompt Quality PQ	.44	.26	.26	.12	.24	**.68**	.43	.14	.53	.26	.20	—	
Interaction Quality IQ	**.66**	.43	.49	.30	.50	.54	.55	.33	.55	.26	.33	.58	—

Correlations indicate that Overall Quality is mainly dependent on *attractiveness* and Interaction Quality. But no influence of the head and voice components on either scale is found. The influence of the respective interaction can be excluded.

4.4. E2 – Results for the degree of media context

The influence of additional media on perceived quality is another focus of our analysis of E2. When comparing the results for 23 participants with only the metaphor's screen and those 23 which had an additional screen simulating either of the two devices used, there is one major finding: an obviously reduced capability to judge the fit of the head and voice components of the metaphor once an additional feedback screen is provided. When the talking head is the only source of visual information, participants were able to judge Fit, resulting in different ratings for MS and TH, which is in line with the quality rankings found in E1 ($F(3,91) = 7.70$, $p < .001$. In the two-screen case, the metaphors were not rated differently concerning Fit ($F(3,91) = 1.90$, $p = .14$).

As main effect, participants with a second screen gave higher ratings for *entertaining value* ($F(1,178) = 6.14$, $p < .05$). The analysis of the impact of the degree of media context shows that a second screen had an impact only in the EPG-scenario: OQ_{EPG} increases with the additional screen ($F(1,178) = 3.87$, $p < .05$). No differences could be found for the AM scenario (OQ_{AM}), see Fig. 9 for the mean values.

5. E1 & E2 – Gender effects

Participants had been recruited to equally represent gender in both experiments. In E1, women rated consistently more positive than men for all scales assessed

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

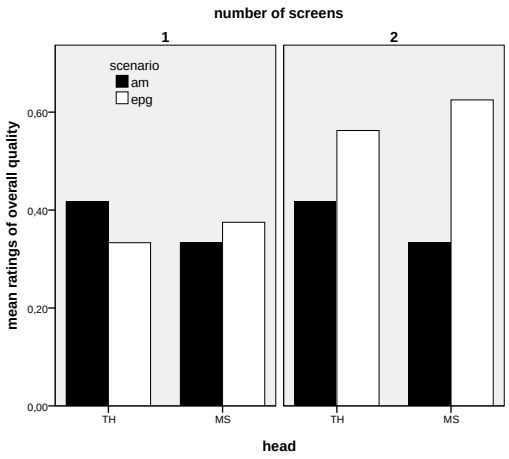


Figure 9: OVERALL QUALITY for the one- and the two-screen setting by scenario.

Table 2: E1 – Gender effects on overall quality (OQ), speech quality (SQ), visual quality (VQ) measured by group mean (M) and standard deviation (SD).

	women		men			
scale	M	SD	M	SD	F(1,706)	*p*
OQ	3.03	0.91	3.31	0.82	27.23	<.001
SQ	3.05	1.03	3.49	0.99	50.31	<.001
VQ	2.96	1.04	3.26	1.21	26.45	<.001

(see Table 2). However, controlling for this factor in E2 only small effects could be found for the questionnaire results: Male participants were more critical concerning SYN in the AM scenario ($F(1,178) = 7.48$, $p < 0.01$) and for *attractiveness* ($F(1,178) = 8.65$, $p < 0.01$), whereas women in general rated the *cognitive demand* higher ($F(1,178) = 5.14$, $p < 0.05$).

One additional observation were made: Women followed the design more diligently than men. This was tested by comparing the interaction parameters for men and women. While men and women needed about the same number of system turns to solve the given tasks, women kept more closely to the order of task cards than men as measured by smoothness. Women also solved the tasks more quickly than men (cf. Table 3).

Table 3: E2 – Gender effects on interaction measured by the group mean and standard deviation of dialogue duration (*dd*), number of turns (*#turns*) and smoothness (*sm*), with differences that are significant at p<.005 in bold. As the data is not distributed normally a Mann-Whitney-U test was used.

	women		men			
	M	SD	M	SD	U	*p*
dd	**5.15**	1.46	**5.98**	2.47	3176	**.002**
#turns	27.15	3.85	27.15	3.62	4198	.472
sm	**1.01**	1.7	**1.57**	1.84	3171	**.002**

6. Discussion

In this paper, quality differences for combinations of TTS and head components have been analysed with regard to the degree of interactivity and the media context. The main finding is a major influence of the **degree of interactivity**:

Without real interaction (E1), there is a clear ranking of the TH head module being better than the MS (and CL), and the Mary TTS module being better than Mbrola. While not all metaphors (combinations of the modules) differ significantly for all assessed scales, this ranking is significant and systematic for metaphor's OVERALL QUALITY, SPEECH QUALITY, and VISUAL QUALITY. Participants favoured the more natural voice (HMM-synthesis) and head component (realistic texture, extra-linguistic random movements). In contrast to the re-

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

sults in Nass and Gong (1999) concerning consistency, the more artificial head does not profit from the more computer-like voice.

However, this clear ranking could not be replicated in the interactive setting (E2). Considering only the previously used scales for assessing metaphor quality (Overall, Speech, and Visual Quality), only one yields significant results comparable to the non-interactive experiment, namely Speech Quality.

Additionally, the metaphors show differences on the scales synchronization (Syn) and fit of voice and head (Fit), which were not assessed in E1. Concerning fit, TH with Mary is rated best, which is in line with the ranking found in E1. However, MS is rated to have a better synchronization of voice and lip movements than TH. This result can be explained by the different mapping of audible speech and visual articulation: In the case of TH, for synthesizing German phonemes most adequate Australian-English visemes had to be used, whereas MS was originally built for German speech (Fagel et al., 2008). This is the only significant difference between MS and TH incoherent with the previously found ranking in E1. Nevertheless, the better synchronization of MS is not reflected in any of the quality scales, not even in the two significant factors out of six extracted from the semantic differential (*naturalness* and *entertainment value*). For both of these highly correlated factors, MS receives lower ratings than TH – in line with the ranking found in E1. This is can be interpreted as follows: synchronization does not influence any contrasting quality aspects of the metaphors assessed by the semantic differential or the scale Visual Quality.

The results show that participants had been able to distinguish the metaphors in E2 on at least some of the scales used. Wherever differences are found, the ranking does not differ from the one in E1. However, on scales like Overall Quality or the factor *attractiveness* the metaphors were not rated differently in E2. Equally, no ranking could be found on the Visual Quality scale. The latter might be due to the fact that participants were forced to look at their task cards regularly. Thus, their visual attention was diverted from the talking heads and they were not able to differentiate Visual Quality. As the spoken output of the metaphors was played via headphones, participants were able to distinguish the Speech Quality of the metaphors and rank them accordingly.

But this cannot explain the missing discrimination concerning the more general scales Overall Quality or *attractiveness*, that is present for more concrete scales (fit, synchronization, *naturalness*). Possible reasons for this are:

1. In an interactive setting, participants cannot distinguish between the quality of the talking heads and the system, as the metaphors are considered representative of the whole system. Therefore, they unconsciously rate the system quality instead of the talking head quality.
2. Participants are distracted by the interaction and they can perceive only some of the characteristics as varying between the talking heads, which are more salient (simple classification like *naturalness* and fit, but not *attractiveness*).

No impact of metaphor characteristics on Overall Quality could be found, while ratings on scales clearly assessing metaphor-related aspects (Fit, Synchronisation, *naturalness*) show differences for the metaphors. This supports hypothesis (1), namely that participants conflated system quality — which was constant by de-

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

sign – with the metaphors' Overall Quality. Möller and Skowronek (2003) already found a similar result for uni-modal spoken dialogue systems, namely that – in contrast to speech input – perceived speech output quality is conjoint with the quality of the whole system.

Following hypothesis (2), the impact of interactivity explains the poor results for Overall Quality. Based on this argumentation, differences in Fit and Synchronisation might just be more present than Overall Quality and *attractiveness*.

Strong correlations are found between Overall Quality and both *attractiveness* and Interaction Quality. Neither does *attractiveness* differ significantly between the metaphors, nor are there interaction parameters explaining Interaction Quality. However, it has been shown before that the perceived dialogue is not necessarily influenced by the de-facto course of dialogue (cf. Frøkjær et al., 2000). In our case the dialogue flow was controlled and thus only very few parameters could be used to model the subjective ratings as proposed with the PARADISE framework by Walker et al. (1998). This was done to force the participants to rate the head and not the system quality and we could interpret the results as having succeeded in this. Based on this point, we see the ratings on Overall Quality and *attractiveness* as neither influenced by the dialogue flow nor the metaphor characteristics.

The other factor varied in a controlled way is the **degree of media context**: providing redundant information on a screen and thus offering a distraction on the visual channel further hinders the participants' ability to perceive and judge the differences between the metaphors. This could be observed by the different ratings obtained with the Fit scale for the one- and the two-screen setting. At the same time this additional information leads to a higher metaphor quality and higher ratings for *entertaining value*. This is in line with the general finding that the metaphor ratings are obviously influenced by other factors. We interpret this finding as showing that for the subjects the metaphors represent the whole system and the metaphor's rating benefits from the increased system quality due to the additional screen (hypothesis (1)): As the second screen provides information of the electronic program guide in a more salient modality – film names, times and channels as text lists in addition to spoken lists – Overall Quality (of the metaphor) is higher in this case than without this second screen. We conclude therefore that participants are less able to distinguish between discrete quality aspects the more influencing or distracting factors are present. In other words, the more complex a stimulus, the more difficult to obtain analytic ratings.

Apart from the factors presented so far, the effects of gender have also been analysed. Concerning gender differences, some interesting results could also be found. Men and women exhibit different interaction behaviour as recorded by the interaction parameter. This is in line with findings reported in (Canada and Brusca, 1991). Women tend to follow the instructions much more closely while men stray from the given path to 'play' with the system. When considering the subjective data from the experiments, contradictory results are obtained. In the non-interactive experiment, female users rate the heads consistently better than male participants. In the interactive experiment, this could not be confirmed, as there are conflicting results. We consider the gender effect on ratings negligible.

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

7. Conclusion

Analysing the results of these two experiments leads to three important observations concerning the design of user studies including ECAs. (1) If an evaluation of singular system components or aspects is desirable, function tests should be carried out with as few distractions as possible. (2) Replacing one system component with a better one will not necessarily lead to a higher perceived quality of the talking head (and thus not improve the overall quality of the whole system) if this improvement may not be perceived by the users. This result shows the importance of the degree of interactivity and may not be valid for e.g. tutoring systems without interaction. (3) When choosing participants for user tests in the smart home domain, gender has to be taken into consideration not only for interpreting the results, but also concerning the expected behaviour during the study. It has to be decided whether a more playful or an obedient attitude regarding interaction behaviour is preferable for the given study.

Irrespective of these findings, we conclude that a more human-like head and voice is preferable for the output component of a smart-home system. However, other aspects of multi-modal systems may be more efficient, such as choosing the most appropriate medium for the information to be presented.

To answer the open questions, another experiment is currently being conducted. The four metaphors are evaluated again. This time participants interact with a fully functional smart home system. They are in a real living room and experience direct feedback from the system. Thus, this time an even higher distraction is provided. Furthermore, the interaction is not as strictly defined and more interaction parameters can be gathered. We hope to be able to answer the question of whether or not the metaphor quality is conflated with system quality / particular interaction and if the user is too distracted to actually evaluate the head. As a reference, the system with the metaphors is compared to a system without a head component to test for acceptability of the embodiment and the persona effect in the smart home domain.

At the same time the TH is enhanced with facial expressions, rendering the metaphor more affective.

Acknowledgment

The project was financially supported by the Deutsche Forschungsgemeinschaft DFG (German Research Community), grant MO 1038/6-1.

Adcock, A., Van Eck, R., 2005. Reliability and factor structure of the attitude toward tutoring agent scale (ATTAS). Journal of Interactive Learning Research 16 (2), 195–217.

André, E., Rist, T., Müller, J., 1998. Webpersona: a lifelike presentation agent for the world-wide web. Knowledge-Based Systems 11 (1), 25–36.

Berry, D. C., Butler, L. T., de Rosis, F., 2001. Evaluating a realistic agent in an advice-giving task. International Journal of Human-Computer Studies 63 (3), 304–327.

Breitfuss, W., Prendinger, H., Ishizuka, M., 2008. Automatic generation of gaze and gestures for dialogues between embodied conversational agents: System description and study on gaze behavior. In: Proc. AISB 2008 Symposium on Multimodal Output Generation (MOG 2008). pp. 18–25.

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. Springer-Verlag, New York, pp. 217–238.

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

Canada, K., Brusca, F., 1991. The technological gender gap: Evidence and recommendations for educators and computer-based in-

struction designers. Educational Technology Research and Development 39 (2), 43–51.

Cassell, J., Sullivan, J., Prevost, S., Churchill, E., 2000. Embodied conversational agents. MIT Press, Cambridge.

Costello, A., Osborne, J., 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment Research & Evaluation 10 (7).

Cowell, A. J., Stanney, K. M., 2005. Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. International Journal of Human-Computer Studies 62 (2), 281–306.

Dehn, D. M., Van Mulken, S., 2000. The impact of animated interface agents: a review of empirical research. International Journal of Human-Computer Studies 52 (1), 1–22.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vreken, O., 1996. 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). pp. 1393–1396.

Erickson, T., 1997. Designing Agents as if People Mattered. AAAI Press, Menlo Park.

Fagel, S., Bailly, G., Elisei, F., 2007. Intelligibility of natural and 3d-cloned German speech. In: Proc. International Conference on Auditory-Visual Speech Processing (AVSP). Paper L2-1.

Fagel, S., Clemens, C., 2004. An articulation model for audiovisual speech synthesis – determination, adjustment, evaluation. Speech Communication 44 (1–4), 141–154.

Fagel, S., Kühnel, C., Weiss, B., Wechsung, I., Möller, S., 2008. A comparison of German talking heads in a smart home environment. In: Proc. International Conference on Auditory-Visual Speech Processing (AVSP).

Frøkjær, E., Hertzum, M., Hornbæk, K., 2000. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In: Proc. Conference on Human Factors in Computing Systems (CHI). pp. 345–352.

Gong, L., 2008. The boundary of racial prejudice: Comparing preferences for computer-synthesized white, black, and robot characters. Computers in Human Behavior 24 (5), 2074–2093.

Gong, L., Nass, C., 2007. When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. Human Communication Research 33, 163–193.

Horn, J., 1965. A rationale and a test for the number of factors in factor analysis. Psychometrika 30 (2), 179–185.

Hutcheson, G., Sofroniou, N., 1999. The multivariate social scientist. Sage Publications, Thousand Oaks.

Kühnel, C., Weiss, B., Wechsung, I., Fagel, S., Möller, S., 2008. Evaluating talking heads for smart home systems. In: Proc. International Conference on Multimodal Interfaces (ICMI).

King, W. J., Ohya, J., 1996. The representation of agents: anthropomorphism, agency, and intelligence. In: Proc. Conference on Human Factors in Computing Systems (CHI).

Kipp, M., 2004. Gesture Generation by Imitation – From Human Behavior to Computer Character Animation. Dissertation.com, Boca Raton.

Koda, T., Maes, P., 1996. Agents with faces: The effects of personification of agents. In: Proc. IEEE International Workshop on Robot and Human Communication. pp. 189–194.

Krämer, N., 2008. Soziale Wirkungen virtueller Helfer. Medienpsychologie. Kohlhammer, Stuttgart.

Krämer, N. C., Bente, G., 2002. Virtuelle Helfer: Embodied Conversational Agents in der Mensch-Computer-Interaktion. In: Virtuelle Realitäten. Hogrefe, Göttingen, pp. 203–225.

Lester, J. C., Stone, B. A., Converse, S. A., Kahler, S. E., Barlow, S. T., 1997. Animated pedagogical agents and problem-solving effectiveness: a large-scale empirical evaluation. In: Proc. World Conference on Artificial Intelligence in Education. pp. 23–30.

Massaro, D., Cohen, M., Beskow, J., Cole, R., 2000. Developing and evaluating conversational agents. In: Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.), Embodied conversational agents. MIT Press, pp. 286–318.

McBreen, H. M., Jack, M., 2000. Empirical evaluation of animated agents in a multi-modal e-retail application. In: Proc. AAAI Fall Symposium on Socially Intelligent Agents. pp. 122–126.

Möller, S., Skowronek, J., 2003. Quantifying the impact of system characteristics on perceived quality dimensions of a spoken dialogue service. In: Proc. European Conference on Speech Communication and Technology, Geneva. Vol. 3. pp. 1953–1956.

Nass, C., Gong, L., 1999. Maximized modality or constrained consistency? In: Proc. International Conference on Auditory-Visual Speech Processing (AVSP).

Nass, C., Moon, Y., Green, N., 1997. Are computers gender-neutral? gender stereotypic responses to computers. Journal of Applied Social Psychology 27 (10), 864–876.

Nowak, K. L., 2004. The influence of anthropomorphism and agency on social judgment in virtual environments. Journal of Computer-Mediated Communication 9 (2).

the published article can be found here: DOI: 10.1016/j.specom.2010.02.011

Nowak, K. L., Rauh, C., 2005. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. Journal of Computer-Mediated Communication 11 (1).

Pandzic, I. S., Ostermann, J., Millen, D. R., 1999. User evaluation: Synthetic talking faces for interactive services. The Visual Computer 15 (7/8), 330–340.

Prendinger, H., Mori, J., Saeyor, S., Mori, K., Okazaki, K., Juli, Y., Mayer, S., Dohi, H., Ishizuka, M., 2004. Scripting and evaluating affective interactions with embodied conversational agents. KI Zeitschrift 1, 4–10.

Ruttkay, Z., Pelachaud, C. (Eds.), 2004. From Brows to Trust: Evaluating Embodied Conversational Agents. Springer-Verlag, New York.

Schroeder, M., Trouvain, J., 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology 6 (4), 365–377.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., Waters, K., 1996. When the interface is a face. Human Computer Interaction 11 (2), 97–124.

Takeuchi, A., Naito, T., 1995. Situated facial displays: towards social interaction. In: Katz, I., Mack, R., Marks, L. (Eds.), Human Factors in Computing Systems: CHI'95 Conference Proceedings. ACM Press, New York, pp. 450–455.

Theobald, B.-J., Fagel, S., Bailly, G., Elisei, F., 2008. Lips2008: Visual speech synthesis challenge. In: Proc. INTERSPEECH, Brisbane. pp. 2310–2313.

Van Mulken, S., André, E., Muller, J., 1998. The persona effect: How substantial is it? In: Proc. HCI on People and Computers.

Walker, M. A., Litman, D. J., Kamm, C. A., Abella, A., 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. Computer Speech and Language 12 (4), 317–347.

Weiss, B., Kühnel, C., Wechsung, I., Möller, S., Fagel, S., 2009. Comparison of different talking heads in non-interactive settings. In: Proc. Human Computer Interaction International (HCII), San Diego. pp. 349–357.

Xiao, J., 2006. Empirical studies on embodied conversational agents. Ph.D. thesis, Georgia Institute of Technology.

Xiao, J., Stasko, J., Catrambone, R., 2002. Embodied conversational agents as a UI paradigm: A framework for evaluation. In: Embodied conversational agents – let's specify and evaluate them! Workshop in conjunction with International Conference on Autonomous Agents and Multiagent Systems (AAMAS).

Yee, N., Bailenson, J. N., Rickertsen, K., 2007. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In: Proc. Conference on Human Factors in Computing Systems (CHI). pp. 1–10.

Zimmerman, J., Ayoob, E., Forlizzi, J., McQuaid, M., 2005. Putting a face on embodied interface agents. In: Proc. Conference on Designing Pleasurable Products and Interfaces (DPPI). pp. 233–248.