Evaluating System Metaphors via the Speech Output of a Smart Home System

Sebastian Möller¹, Jan Krebber¹ and Paula Smeets²

¹Institut für Kommunikationsakustik (IKA), Ruhr-Universität, 44780 Bochum, Germany
²TNO Human Factors, Dept. of Perception, 3769 ZG Soesterberg, The Netherlands
sebastian.moeller@ruhr-uni-bochum.de

Abstract

This paper presents an evaluation of the speech output component of a smart home system developed under the European INSPIRE project. In particular, it is investigated how the “personality” of the system which is conveyed by the output speech is accepted by the user. Three different metaphors are compared: An assistant visualized by a talking head, an invisible assistant, and multiple intelligent devices. The results show that a personal assistant metaphor leads to more homogenous judgments for the individual voices. Potential reasons for this finding are discussed.

1. Motivation and introduction

The speech output component forms a very important part of a spoken or multimodal dialogue system. In fact, it has been shown that the speech generated by the system helps to make up the “personality” of the machine interaction partner, and thus has a strong influence on the quality of the entire system [1]. In a multimodal dialogue system, different options exist for speech output, e.g. speech-only vs. talking-head output, naturally produced vs. synthesized speech, or directed vs. undirected sound presentation. Decisions on these options should be taken in view of an optimum overall quality for the user.

In the EC-funded IST project INSPIRE (“INfotainment management with SPeech Interaction via REmote-microphones and telephone interfaces”), a dialogue system has been set up for controlling home appliances via speech (TV, VCR, lights, blinds, fan, answering machine). It combines speech input with speech and visual output, and the device feedback can directly be experienced by the user. For example, lights are switched on and off, the TV shows program listings, or the answering machine plays a message. The mentioned modalities can be used to support an optimum “personality” of the system. This personality is reflected by the so-called “metaphor”, i.e. the transfer of meaning to the machine interaction partner by the human interaction partner, due to the similarity to a human partner in its apparent shape, in its function, and in its use. It is the aim of the present paper to evaluate the effect of different “metaphors” on the overall quality of the dialogue system.

Three different metaphors have been implemented which differ with respect to the output modality, the system voice, as well as the sound direction:

- Multiple intelligent devices: Each of the addressed devices is “intelligent” in that it is able to maintain a spoken interaction with the user, using a different voice. The sound is reproduced near the location of the device.

- A single visible “assistant” or “servant” which operates the devices on behalf of the user. This assistant is visible in terms of a talking head (avatar) on a screen facing the user. The sound is reproduced by a loudspeaker near that screen.

- An invisible “assistant”, similar to the visible one, but immaterial and invisible, acting like a ghost somewhere in the room. In this case, the sound is presented from a number of loudspeakers at different locations of the room, generating a more-or-less diffuse sound field.

Because an interactive experiment would strongly limit the number of system options which can be tested, a listening-only test has been designed. It has been carried out in a realistic environment where the users get an impression of the device feedback. The experimental set-up and test design are described in Section 2. Section 3 presents subjective judgments and highlights the differences between the metaphors and the voices which have been used in the system. The results are discussed in Section 4, and show that the metaphor carries an influence on the quality judgments obtained for the different voices.

2. Experimental set-up

In order to get valid judgments on how users perceive the overall speech output quality of the INSPIRE dialogue system, an application-specific test design has been chosen which reflects the physical environment the system is used in, and also the system’s personal and functional characteristics. The latter are expected to be particularly important for the system metaphor.
2.1. Speech samples

In the INSPIRE system, speech is concatenated from pre-defined templates (full sentences, phrases, expressions or words). Because of this template-based approach, the quality of the resulting speech will not be stationary, and in particular it will depend on the contents conveyed by the message (what is said). Consequently, the prompts used in the test should closely reflect the sentences used in the later application. A number of application-near text prompts have been defined which fulfill this requirement. Whenever a prompt refers to a specific device, it is linked to a direct feedback from that device. For example, the system prompt “I will open the blinds. Please say stop when it is bright enough!” is linked to an opening action of the blinds during the test.

The prompts have been recorded with six different “voices”. Five of them are natural voices, including two male (m1 and m2) and three female ones (f1, f2 and f3). All of them originate from non-professional speakers. The sixth “voice” consists of speech prompts generated with a commercial TTS system, concatenating units of different length in a unit-selection approach (“voice” s).

2.2. Test site

The test site at IKA consists of a room which is decorated in a way typical for a living room (with a couch, armchairs, a low table, shelves), and which is equipped with the house appliances controllable by the current dialogue system. A number of loudspeakers are used for speech prompt output and device feedback. In addition, a TFT display is mounted on the wall for displaying the avatar in the talking head metaphor. In this environment, speech output can be provided in different ways: Either from one or several of the loudspeakers located in the four corners of the room, hanging under the ceiling (JBL Control 1c); or via device-related loudspeakers (Boss MA-12 Vocal) located close to the TFT display and to the answering machine. Background noise (pop synthesizer music with spectral characteristics close to white noise) can be inserted from all 4 ceiling loudspeakers.

2.3. Test design

Three test sessions have been carried out, each corresponding to a different metaphor. In the intelligent devices metaphor, the user can interact with the individual devices, which show specific voice and (sound) location characteristics. Namely, each device is characterized by a specific voice, and the sound of the device is always generated at the position of the device (i.e. via the closest loudspeaker). In the talking head metaphor, the user interacts with an assistant which is visible on the TFT display mounted on the wall. This assistant (and consequently the system) always has the same voice (m1) which is combined with a cartoon-like talking face (“Zach” from the CSLU toolkit v2.0 [2]). Speech and video sequences are synchronized, but because the synchronization procedure provided with the CSLU toolkit is optimized for English, and because of the slow response time of the talking face, the synchronization is not very good. The voice is reproduced by the loudspeaker mounted close to the TFT display. In the ghost metaphor, the system is characterized by one single voice (m1) which is reproduced by all 4 ceiling loudspeakers in a relatively diffuse way, without any visual support.

2.4. Rating procedure

Test subjects had to rate four characteristics of what they heard, on continuous rating scales as described in [3]. Each of the scales is introduced by a question and labelled with two or five describing adjectives (translated from German):

- What is your overall impression of what you just heard? (excellent; good; fair; poor; bad)
- Which effort was necessary to understand the meaning of the utterance? (complete relaxation possible, no effort required; attention necessary, no appreciable effort required; moderate effort required; considerable effort required; no meaning understood with any feasible effort)
- How pleasant was the voice you just heard? (pleasant; unpleasant)
- How well does the voice fit to the described system? (excellent; good; fair; poor; bad)

The ratings are calculated on a scale from 0 to 6, 5 corresponding to the most positive label of the inner (thick) part of the scale, and 1 corresponding to the most negative label of that part. The mean ratings will be abbreviated with the terms “overall quality”, “listening-effort”, “voice pleasantness” and “voice adequacy” in the following discussion.

2.5. Test subjects

26 subjects participated in the test. They were between 20 to 76 years old (mean: 34.0 years). 15 of them (58%) had experience with spoken dialogue systems, 5 (19%) with dictation systems, and 10 (38%) with synthesized speech. Subjects were paid for their participation.

3. Experimental results

In order to get an impression of the differences between the voices used in the metaphors, test results will first be analyzed for the intelligent devices metaphor alone (which includes all voices). Then, a comparison between the different metaphors will be performed, for the two voices (m1 and s) which are common to all metaphors.
3.1. Differences between voices

The mean ratings for each voice in the intelligent devices metaphor are depicted in Figure 1. An analysis of variance (ANOVA) shows a statistically significant influence of the voice on all ratings \( (p < 0.001) \). The synthesized voice is rated significantly worse in its overall quality than m1, f1, f2 (all \( p < 0.001 \)) and f3 (\( p = 0.005 \)), but not significantly worse than m2. For listening-effort, best ratings are obtained for m1 and f2. The synthesized voice seems to require significantly higher listening-effort than all natural voices except m2 \( (p < 0.001) \). For voice pleasantness, there are no significant differences between the synthesized and the m2, f2 and f3 voices. Here, the synthetic voice is rated even better than m2.

In general, m1 receives the best ratings for all judgments (except for listening-effort where it is similar to f2), and m2 receives the worst judgments (except for listening-effort where it is slightly better than the synthesized voice). The female voices range in the middle between the two quality extremities. The synthesized voice does not seem to be generally inferior in quality than the natural voices; instead, the characteristics of the particular voice seem to be of high importance. The main problem of the chosen synthesized voice is the required listening-effort. The judgments on adequacy, still, show that the voice can in principle be used in the addressed INSPIRE system; they remain on the positive side of the scale (>3.0).

3.2. Comparison between metaphors

Because the test conditions differ with respect to the voice, a comparison should be made for the individual voices separately. Two of the voices (m1 and s) have been judged upon in all three sessions, and can therefore be used for this cross-metaphor comparison. Figures 2 and 3 show the overall quality ratings for the noise-free case, for the male (m1) and the synthesized (s) voices. The result is surprising: Whereas the naturally produced voice is judged more positive in the intelligent devices metaphor compared to the other two metaphors (talking head and ghost), the opposite seems to be the case for the synthesized voice. The same tendency is shown in the ratings on listening-effort, voice pleasantness and voice adequacy, which are not reproduced here to limit space. It can also be observed in the case of background noise.

In principle, the observed differences between the metaphors may result from the choice of test conditions: In all three test sessions, the number of noisy and noise-free conditions was balanced (50%), the number of naturally produced vs. synthesized speech samples was nearly balanced, and the complexity of the prompts seems to be similar. The diversity of voices is larger in the session addressing the intelligent devices metaphor, but the range of quality judgments which is covered in all sessions seems to be comparable. The differences

![Figure 1: Mean ratings for different voice options, intelligent devices metaphor.](image1)

![Figure 2: Mean values and 95% confidence intervals for overall quality ratings on voice m1, comparison across metaphors. Numbers represent different test conditions.](image2)

![Figure 3: Mean values and 95% confidence intervals for overall quality ratings on voice s, comparison across metaphors. Numbers represent different test conditions.](image3)

4. Discussion

In principle, the observed differences between the metaphors may result from the choice of test conditions: In all three test sessions, the number of noisy and noise-free conditions was balanced (50%), the number of naturally produced vs. synthesized speech samples was nearly balanced, and the complexity of the prompts seems to be similar. The diversity of voices is larger in the session addressing the intelligent devices metaphor, but the range of quality judgments which is covered in all sessions seems to be comparable. The differences
may therefore be due to the differences between the addressed metaphors. Several potential explanations may be thought of which reflect these differences:

1. When each device is characterized by its own voice, the judgement of the speech output quality may be linked to the suitability of the voice for the addressed device. In the intelligent devices metaphor, m1 was associated with a light and used for the general instructions, and the synthesized voice was associated with the TV. In the other two metaphors, the m1 voice was associated with the fan and used for the general instructions, and the synthesized voice addressed the blinds, lights and TV. It does not seem convincing that this association may have caused the observed judgment differences.

2. The fact that the speech output is displayed at different locations may have confused the test subjects. In the intelligent devices metaphor, the general instructions have been displayed in a non-directional way (all four ceiling loudspeakers), and the other prompts from the location of the individual devices. Such a confusion would however have caused more critical judgements in general, and not only for one specific voice. It is thus not expected that the sound direction may be responsible for the observed differences between the metaphors.

3. The modality of the presented speech prompts differs between the sessions dedicated to the talking head and the other two sessions (intelligent devices and ghost). As the judgments for the visible and the invisible assistant are very similar, this does not seem to have exercised a significant effect on the quality perceived by the test users.

This observation is somehow surprising: One could think of a higher quality (due to better intelligibility) of the audio-visual speech prompts compared to the audio-only ones. The fact that the video and audio sequences have not been fully synchronized (problems of the CSLU tool which is designed for English) may be responsible for the missing influence of the talking head.

4. The use of a single “personal assistant” seems to concentrate the judgments on the central part of the scale. Thus, the ratings may be less extreme because they are associated with a single interaction agent – the personal assistant – instead of a number of different “assistants” (the devices). The fact that the voice changed for this personal assistant does not diversify the judgments to the same extent as the different positioning and the large diversity of voices of the different devices do. Finally, explanation (4) seems to be the most reasonable one. It has to be noted that the ratings for the individual speech prompts generated by the synthesized voice scatter significantly. This may be due to the time-variant characteristics of the chosen synthesis approach. The scatter shows the necessity to use relatively long speech samples in order to get stable judgments. It underlines the importance of using text prompts which are representative for the application scenario.

5. Conclusions and outlook

The experimental data shows that the system metaphor, i.e. the transfer of meaning to the system by its human user, carries an influence on the quality judgments with respect to the speech output component of the INSPIRE system. It has to be noted that the effect of the metaphor may be quite restricted when only the speech output is evaluated. The situation may change in an interactive scenario. No final conclusion can be drawn with respect to a preference for a specific metaphor. Depending on the voice, either the intelligent devices metaphor or a “personal assistant” (be it visible as a talking head, or invisible as a ghost) may be preferred. The judgments seem to be more homogeneous (close to the center of the judgement scale) for the talking head and the ghost than for the intelligent devices metaphor.

6. Acknowledgement

The present work has been performed in the frame of the EC-funded IST project INSPIRE (IST-2001-32746). The authors would like to thank Martin Rajman (EPFL) and Jettie Hoonhout (Philips) for fruitful discussions on the metaphors and the evaluation strategies, and Alexander Raake, Antonio Guzman Avalos, Rosa Pegam and Anders Krosch (all IKA) for their support in the experiments.

7. References

