Evaluation of the Speech Output of a Smart-Home System in a Car Environment

Paula M.T. Smeele¹, Sebastian Möller², Jan Krebber²

¹TNO Human Factors, Department of Perception
P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
²Institute of Communication Acoustics (IKA), Ruhr University Bochum, Germany
smeele@tm.tno.nl

Abstract
This paper reports on the evaluation of the speech output component of a dialogue system that enables the control of different home appliances via speech. It was simulated that participants called the system from the car and listened to feedback messages from the system. Their task was to judge the quality of the system’s output while performing a driving task in the TNO driving simulator. The following effects on the quality judgements were studied: type of system prompt (natural, synthesized), transmission degradations (circuit noise, speech codec, packet loss), and the driving task. The results are presented in terms of these aspects.

1. Introduction
The dialogue system developed in the FP5 EC-funded IST project INSPIRE (INfotainment management with SPeech Interaction via RE mote microphones and telephone interfaces, IST-2001-32746) enables the control of different home appliances via speech. It consists of speech and speaker recognition, speech understanding, dialogue management, and speech output components. The system can be addressed in a home environment via a microphone array or a portable microphone, and from remote locations, such as the office or the car, via the telephone network.

The present study is part of a series of system component assessments and focusses on the quality of the speech output of the dialogue system. The speech output component forms a very important part of a spoken dialogue system. It has been shown that the speech generated by the system helps to make up the “personality” of the system, and therefore may strongly affect the quality of the whole system perceived by its users [1].

2. Speech Output Component
An important feature of the INSPIRE dialogue system is that it can be used in different physical environments. In order to test their influence, three different experimental environments have been chosen in which different aspects of the system are tested:

(1) A living-room environment. Tested aspects: the system’s metaphor and the effect of room acoustics.

(2) An office environment. Tested aspects: different transmission channel degradations, user interfaces, and background noise effects.

(3) A car environment. Tested aspects: different transmission channel degradations and the effect of the driving task.

The system’s output is generated from pre-defined templates. These templates define the concatenation of full sentences, of phrases, or of individual expressions or words. Two types of system prompts have to be distinguished: static prompts are pre-defined for the application, and do not change according to the contents of the database. They are produced using natural pre-recorded speech. Dynamic prompts have to be generated online during system operation, because they contain dynamically changing contents (e.g. names of films, messages from the answering machine). These prompts always involve synthesized speech generated by a Text-To-Speech (TTS) system.

The evaluations of the speech output component in the living-room and office environments are carried out at the Institute of Communication Acoustics at the Ruhr University of Bochum and are not described here. This paper describes the evaluation of the speech output component in a car environment, carried out in the driving simulator at TNO Human Factors.

3. Evaluation in the Car Environment
In this evaluation, the focus is on the impact of channel degradations due to the telephone network. In addition, the participants’ preference for the system’s voice (natural vs. synthesized), and the effect of the driving task on the quality judgements of the participants are studied.

3.1. Participants
Sixteen native Germans participated in the experiment. Six participants were male and ten female. Fourteen participants were between 21 and 39 years old; two were older, they were 62 and 64 (mean age: 32.7). They were paid for their participation.

3.2. Test Conditions
Forty-eight different test conditions have been judged by each participant. Each test condition is characterized by the voice (natural or synthesized), the type of prompt (static or dynamic), the template sentence, and the transmission characteristics of the simulated telephone network. The system prompts were presented via the loudspeaker on the dashboard of the car simulator, at a comfortable sound pressure level of 68 dBA (Brüel & Kjaer 2236, fast setting). In all test conditions the background noise was engine noise, coming from loudspeakers positioned outside the car and generated by
the sound system. The level of the engine noise was determined relative to an average speech level calculated over all samples and was set at 62 dBA, in order to achieve a SNR of approx. 6 dB measured at the location of the driver’s head. The system prompts are grouped into two sessions (1 and 2) which correspond to different voice options:

Session 1: static information with naturally produced voice (male1), dynamic information with synthesized voice.
Session 2: both static and synthesized information with synthesized voice.

The system prompts have been processed by a transmission simulation tool in order to obtain realistic and still controllable characteristics of the assumed telephone channel between the car and the INSPIRE system. This tool is described in detail in [2]. The transmission characteristics vary with respect to the level of continuous narrow-band circuit noise level \( N_c \), the speech codec (G.726, GSM-EFR, G.729), and the probability of transmission errors (for G.729 codec: 0-15% random packet loss). The following 12 degradations are considered, for both static and dynamic prompts, resulting in 24 test conditions:

- 1 condition without telephone transmission
- 1 condition with default characteristics (G.711 with \( N_c = -70 \text{ dBm}0\)p)
- 3 conditions with different circuit noise levels: \( N_c = -50, -40, -30 \text{ dBm}0\)p
- 3 conditions with codecs with transmission errors: G.726 (0%), GSM-EFR (0%), G.729 (0%)
- 4 conditions with codecs with transmission errors: G.729 3, 5, 10, 15%.

### 3.3. Driving Task

A road scenario with a driving task has been specifically designed for the evaluation of the system prompts in the TNO driving simulator (see Fig. 1). The road was placed in a suburban setting and contained the following special events: sharp curves, a lead car driving with variable speed, a variable number of cars driving by in the opposite lane, road signs, trucks parked along the road, and a lead car driving off the road.

![Figure 1: The driving simulator at TNO.](image)

The road events were chosen to represent a varying degree of cognitive load for the driver. The driving task has been divided into 24 driving units of ca. 1-min duration each. Twelve of these units contained a special road event, the remaining twelve contained no special event (here the participants just drove the car on a more or less straight road with minor curves). During a driving unit one of the 24 system prompts was presented. The driving task, the suburban road and the order of the road events were identical in Sessions 1 and 2. However, the order of the sessions as well as the order of the 24 test conditions within a session have been randomized over the participants (resulting in different event-test condition pairs for each participant).

### 3.4. Instructions and Procedure

The participants were given some general information about the INSPIRE system. It was simulated that they would call the system from their car and receive feedback from the system. The functionality of the system would be restricted to programming the answering machine and TV/VCR, and to controlling the lights and blinds. No direct visible feedback from the home devices was given in the car simulator. Then, the participants answered some general questions related to their background. After the experiment, the participants were asked a set of additional quality-related questions (overall impression of the system, enjoyment, perceived helpfulness, pleasantness of listening to the system, perceived relaxation, strong and weak points of the system, proposals for improvement).

In order to stimulate the participants to pay attention to the contents of the speech prompts and not just to the surface form, they were asked to indicate the device the system prompt refers to (multiple-choice task). Subsequently, the participants rated four speech quality dimensions (i.e., overall quality, ease of listening, voice pleasantness and voice adequacy).

Regarding the driving task, the participants were instructed to drive in the car simulator as they would normally do in their own car. Their task was to listen to the speech prompts and judge the perceived intelligibility and quality while driving on a suburban road with a maximum speed of 80 km/h enforced by a speed limit sign (the maximum speed of the car simulator was actually fixed at 100 km/h). After each driving unit in which a speech prompt was presented, the participants had to stop driving and give their ratings. The stop was insured by introducing a red traffic light. Gradually reducing the speed of the car and then stopping the driving task was preferred over abruptly freezing the road scenario, which could cause the participants to suffer from motion sickness. The questions were presented, each question separately, on a touch screen mounted on the dashboard of the car located to the right of the driver. They rated the perceived intelligibility and speech quality using a continuous 5-point scale and entered their ratings by touching a slider on the screen. As soon as the participants completed the questions, the driving task automatically continued with the next driving unit. This procedure was repeated for all 24 speech prompts within a session. One session took about 45 min. After a short pause the participants continued with the second session. The test in the driving simulator was preceded by a 10-min intro session in order to familiarize the participants with the range of conditions to be encountered in the test, as well as with driving in a simulator to prevent motion sickness.

### 4. Data Analysis and Discussion

After the experiment in the driving simulator, the participants were given a final questionnaire with questions related to overall impression of the system, enjoyment, perceived
helpfulness, pleasantness of listening to the system, and perceived relaxation. The mean judgements (including Sessions 1 and 2, scale 0-6) are presented in Table 1.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Car environment</th>
<th>mean</th>
<th>std.</th>
</tr>
</thead>
<tbody>
<tr>
<td>overall quality</td>
<td>3.45</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>fun</td>
<td>4.33</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>perceived helpfulness</td>
<td>3.79</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>pleasantness</td>
<td>3.71</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>perceived stress/</td>
<td>3.69</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>relaxation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As can be seen in Table 1, the overall judgement of the speech output component is on the positive side of the scale (i.e. greater than 3). All aspects received a mean rating on the positive side of the scale, which shows that the system voices used in the speech output test seem to be generally acceptable for the given purpose.

Next, the ratings for “overall quality”, “ease of listening”, “voice pleasantness” and “voice adequacy” have been checked for outliers. The analysis reveals that the elderly participants rated considerably different than the younger participants. This in itself is an interesting finding, suggesting an age-dependent factor in the quality judgements of the speech prompts, but it also means that the ratings of the elderly participants had to be excluded from the data pool. The subsequent analyses included the ratings of the 14 younger participants only.

An analysis of variance (ANOVA) reveals that the ratings significantly differ between test conditions, and that this effect of test condition is significant for each of the judgements (overall quality, ease of listening, voice pleasantness, voice adequacy) on a p<0.001-level, for both Session 1 and 2. To analyze participants’ preference for one of the two voice options, an ANOVA is carried out on the two groups. The analysis reveals that the effect of voice is significant, indicating that participants generally prefer the naturally produced speech (p=0.01). With natural voice, participants gave higher ratings for overall quality (p<0.001), ease of listening (p=0.001), voice pleasantness (p=0.05), but not for voice adequacy (p=0.084). This indicates that participants have a general preference for a natural voice, but also that a synthesized voice can be as adequate as a natural voice.

### 4.1. Effect of Circuit Noise

Averaged over all quality judgements, the level of circuit noise is almost significant at the p<0.1-level. For each judgement separately, a higher level of circuit noise received lower ratings (p<0.001). Participants rated the conditions with the default settings (G.711, -70 dBm0p) and with a circuit noise level of -50 dBm0p significantly higher than the conditions with higher circuit noise levels. Fig. 2 shows the influence of circuit noise on the judgements of overall quality and ease of listening in Session 1 (the ratings for voice pleasantness and voice adequacy are not presented here and in the following figures to limit space).

For synthesized speech in Session 2, the effect of circuit noise is not significant averaged over all quality judgements (p=0.299). However, subsequent analyses show that the effect is significant for each judgement separately (p<0.01): a higher circuit noise level received lower ratings. Although the difference between the circuit noise conditions does not reach significance, the conditions with the default settings (G.711 with −70 dBm0p) and with a noise level of −50 dBm0p seem to have somewhat higher ratings than the conditions with higher noise levels.

### 4.2. Effect of Speech Codecs

The effect of speech codec is significant for all judgements for both natural and synthesized voice (p<0.04). The effect of distortions on the judgements of Session 1 resulting from low bit-rate coding (64 to 8 kbit/s) can be seen in Fig. 3. Post-hoc analyses (Tukey HSD) indicate that the condition with the default settings (speech codec G.711) is rated higher than the speech codecs G.726, GSM-EFR, and G.729, with no significant differences among these three codecs.

In Session 2, only for the judgement of ease of listening a significant effect of the speech codec is found (p<0.001). Here, the ratings for the test condition with the default settings (speech codec G.711) are higher than for the speech codecs G.726 and GSM-EFR, but similar to those for codec G.729.

### 4.3. Effect of Packet Loss

For all judgements, there is a significant effect of packet loss (p<0.01). Post-hoc analyses (Tukey HSD) reveal that this effect is mainly due to the test condition with 15% packet loss, for which the ratings are significantly lower than for the four conditions with lower packet loss percentages. Other
differences between the test conditions are not significant. Fig. 4 shows the mean ratings as a function of packet loss for Session 1. The picture is slightly different for Session 2, where a significant effect of packet loss is found for the judgement of ease of listening only (p<0.01). For this judgement, the test conditions with the lowest percentage of packet loss (0% and 3%) received higher ratings than the other three test conditions with higher percentages of packet loss.

Figure 4: Effect of packet loss on the quality judgments of Session 1.

4.4. Effect of Road Event in the Driving Task

The participants rated the quality of the system prompts while driving in the car simulator. The road was divided into 24 units each containing a road event which could be either a specific event or a non-event. During each unit a system prompt was presented. In Fig. 5 mean ratings of Sessions 1 and 2 as a function of event number are depicted. For each judgement separately, an ANOVA is carried out to study whether participants’ ratings depend on the type of road event. The analysis reveals that the effect of event is not significant for all judgements (p>0.62). Although the road events have been chosen to represent a varying degree of cognitive load for the driver, this result may indicate that the cognitive load of the events did not reach a critical threshold for influencing the judgements on speech quality.

Figure 5: Effect of road events on the judgements of Session 1 and 2.

4.5. Effect of the Driving Task

The question whether the driving task as an additional task to the speech quality judgement task affects participants’ ratings, can only be answered when the ratings of the car simulator experiment are compared to ratings obtained in other environments where the participants did not have a driving task. As mentioned earlier, the speech output component of the INSPIRE dialogue system has also been tested in a home and office environment. For the comparison, the office environment was chosen, because it included test conditions with similar channel degradations and the implementation of the same system metaphor as in the car environment. An ANOVA reveals no significant differences between participants’ ratings in the car and office environments (p=0.42). This may indicate that the cognitive load of the driving task as well as the level of engine noise, that is inevitably associated with driving a car, do not critically affect the speech quality judgements. Please note that it concerns a quality judgement task and not an intelligibility task!

5. Conclusions

In the car environment, several aspects have been studied that may influence the overall quality of the speech output component of the INSPIRE dialogue system significantly:

1. **Natural vs. synthesized speech:** The participants generally prefer naturally produced over synthesized speech. The results also indicate that a synthesized voice can be as adequate as a natural voice.

2. **Transmission characteristics:** The effects of circuit noise, speech codecs, and packet loss are clearly visible, and similar for natural and synthesized speech. Two minor differences are observed: synthesized speech is slightly less affected by the speech codec and slightly more by packet loss.

3. **Road events:** No significant effects of road events are observed on the quality judgements. This result may indicate that the cognitive load of the events did not reach a critical threshold for influencing the judgements on speech quality.

4. **Driving task:** No significant differences between ratings in the car and office environments are observed. This may indicate that the cognitive load of the driving task as well as the level of car engine noise do not critically affect the speech quality judgements.

6. Acknowledgements

The present work was enabled by the European IST project INSPIRE. For details see website [http://www.inspire-project.org](http://www.inspire-project.org)

7. References
