Wizard-of-Oz tests for a dialog system in smart homes

Jan Krebber[1], Sebastian Möller[1], Rosa Pegam[1], Ute Jekosch[1], Miroslav Melichar[2], Martin Rajman[2]

[1] *Institut of Communication Acoustics, Ruhr-University Bochum, D-44780 Bochum, Germany,*
Email: (jan.krebber sebastian.moeller rosa.pegam ute.jekosch)@ruhr-uni-bochum.de

[2] *Artificial Intelligence Laboratory, Ecole Polytechnique Fédéral de Lausanne, CH-1015 Lausanne, Switzerland,*
Email: (miroslav.melichar martin.rajman)@epfl.ch

Introduction

Within the European IST-project INSPIRE (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces) [1] a spoken dialog system for accessing different kinds of home appliances (TV, VCR, lamps, blinds etc.) is currently being developed. The purpose of the system is to provide an intelligent and uniform speech user interface. The interface is capable of supporting the user in controlling different complex devices, and it is hoped to increase the acceptance and quality of the offered services. Speech input is performed via microphone arrays (in the home) or via telephone or internet access (for remote use). The system is evaluated with the help of a human experimenter (the so-called Wizard-of-Oz, WoZ) who replaces the automatic speech recognition; in this way, the word accuracy (WA) can be adjusted. The set-up has been used to test different system metaphors. The article focuses on the architecture of the WoZ tool and on the way it was used to test the quality of the dialog system.

Purpose of the INSPIRE System

One of the main interests of the INSPIRE project is to control different devices via speech, in a uniform way for all devices. The devices which are available in the home show a different level of complexity; e.g., a fan can be switched on or off, whereas complex devices like a Video Cassette Recorder (VCR) or a Electronic Program Guide (EPG) need a large number of information to perform operations on a specific TV program. The use of a single dialog manager makes the system appear uniform for all devices.

System Evaluation Set-Up

The system evaluation focuses on several components like the overall quality of the speech output, the robustness of the automatic speech recogniser (ASR), the influence of the transmission channel in case of remote access, the influence of the dialog structure, vocabulary and others. As the INSPIRE project is within a tight time schedule, the system evaluation has to be performed in parallel to the system implementation. This means that the ASR is not available for the first evaluation experiments. In these experiments, the human operator replaces missing parts of the system. The operator has to act in a pre-defined way, in order to reflect the behaviour of the later system components. For this purpose, the operator is supported by a graphical user interface (GUI).

The INSPIRE spoken dialog system contains several modules (Fig. 1). The signal pre-processing and the ASR provide a transcription of the user utterances. This transcription forms the input to the speech understanding module, which fills a certain number of slots with information contained in the user utterance. The number of slots to be filled depends on the chosen device, and on the task which has to be carried out. The mapping between the surface forms (user utterance) and the canonical values allowed for each slot is described in the keyword list.

The dialog manager gets information from the speech understanding module, the device interface, the dialog flow table and the solution table (indicating the operations which are possible on each device). It takes the decision on how to react, and provides a response via the speech output module (either from pre-recorded speech files or by a text-to-speech synthesizer). The device interface drives the devices and keeps track of their state, e.g. to give the information to the dialog manager that a certain lamp is already at its maximum position and brightness cannot be increased. For evaluating

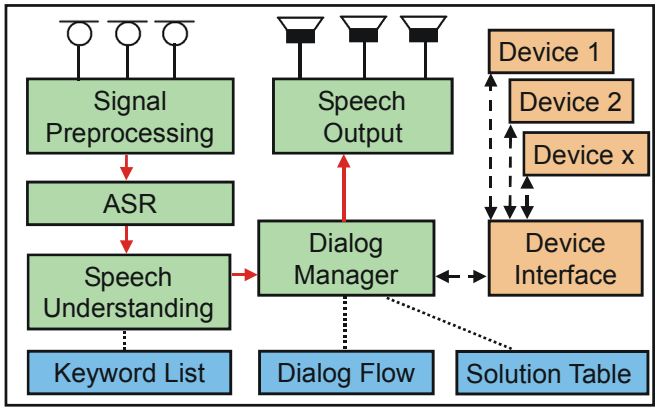

Figure 1 : Modules of the INSPIRE spoken dialog system.

the dialog manager, it is necessary to provide a defined and reliable speech input. Due to the fact that the ASR is not capable of a WA of 100%, the signal pre-processing, the ASR and some parts of speech understanding are replaced by the wizard (Fig. 2). In this way, the dialog manager can be tested with nearly perfect speech recognition. In real life, this will not be a realistic situation. Consequently, it is also possible to set a lower pre-defined WA, by using a noise generator to exchange certain words by phonemically close ones. The Wizard also has the possibility to intervene in speech understanding and in the dialog flow. He or she may provide help and repetition prompts even if they are not foreseen at the certain point in the dialog. The WoZ interface allows the automatic logging of the interaction during the

test sessions. This information can be annotated with the help of a second tool, as described in [2].

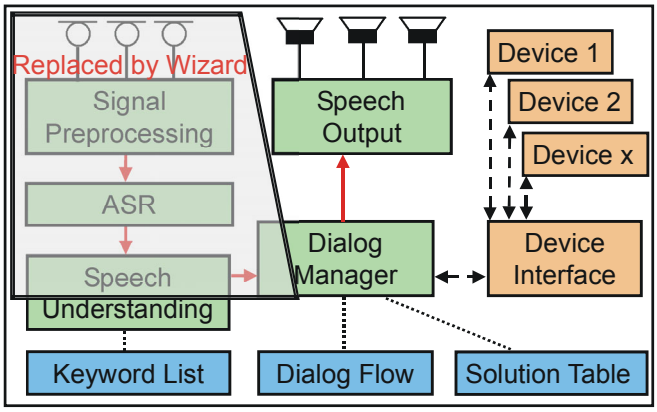

Figure 2: Evaluation set-up with signal pre-processing, ASR and parts of the speech understanding module and the dialog manager replaced by a wizard.

Discussion

The advantages of the INSPIRE WoZ tool are that

- it is possible to reach a WA of nearly 100 %,
- it is possible to test the user satisfaction related to the WA, and to evaluate the dialog manager independently of the WA, and that
- the wizard is able to control the flow of the dialog system by overruling the dialog manager, if necessary.

Because the dialog manager is part of the wizard tool, no changes will be necessary for performing the step from the evaluation environment to the final system.

The drawbacks of the INSPIRE WoZ tool are that

- the wizard has to be trained to operate the system; the wizard has to be familiar with the internal characteristics of several components like the keyword list, the dialog flow table and the solution table; beside of that, the wizard has to know what system reactions are caused by his or her actions;
- the WoZ tool requires a high cognitive load of the wizard; this restricts the experiments to a maximum of four test subjects (up to four scenarios, ten minutes length within two hours per test subject) with one wizard per day.

The influence of the "human factor" in the experimental set-up is still more than a simple transcription, as the wizard has to take decisions of the dialog manager due to the fact that the dialog manager still lacks some functionality.

Tests with the INSPIRE WoZ tool

In an initial experiment, subjective quality judgements and interaction parameters have been collected with the WoZ tool. The purpose of the test was to investigate the influence of different system metaphors, i.e. *the transfer of meaning to the machine interaction partner by the human interaction partner, due to similarity to a human partner in its apparent shape, in its function and in its use*. Three different metaphors have been tested:

- A "talking head" (avatar) metaphor, i.e. a visible assistant shown on a screen, providing audio output through a loudspeaker placed next to the screen.
- An "intelligent devices" metaphor, where each device was equipped with a loudspeaker. Besides the audio output of the devices, there was no additional visual information for the test subjects.
- A "ghost" metaphor, in which loudspeakers were mounted to the ceiling to provide a "homogenous" sound field in the test room ("homogenous" compared to the other metaphors). Again, there was no additional visual information for the test subjects.

First results with 28 participants showed that preference was mainly given to the "intelligent devices" metaphor. The users were highly in favour of the direct interaction with the device, even though they could not access the devices from other rooms than the one the device was located in. The ghost metaphor was rated worse, mainly because the participants felt uncomfortable in the "spooky" environment. Still, they liked that they could access all devices from every room. The avatar metaphor was ranked lowest. The participants didn't like that they had to go to the servant and not vice versa, even though the avatar offered a direct and uniform access to the system.

Conclusion

The WoZ tool for the INSPIRE system is capable of replacing the ASR and some parts of the missing functionality of additional modules. It is possible to run tests with a defined WA. Firsts tests showed that the participants preferred the "intelligent devices" metaphor.

Acknowledgment

The study was carried out at the IKA, Ruhr-University Bochum (PD U. Jekosch, Prof. R. Martin). It was supported by the EC-funded project INSPIRE (IST-2001-32746).

References

[1] INSPIRE. www.inspire-project.org .

[2] Skowronek, J. (2002). *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstequalität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem*, Diploma thesis, IKA, Ruhr-Universität, D-Bochum.

[3] Fraser, N.M., Gilbert, G.N. (1991). *Simulating Speech Systems*, Computer Speech and Language **5**, 81-99.

[4] Smeele, P., Krebber, J., Möller, S., Ganchev, T., Vovos, A., Kladis, B. (2004). *System Component Assessment Report*, Deliverable 6.1, IST project INSPIRE (IST-2001-32746), Institut für Kommunikationsakustik, Ruhr-Universität, D-Bochum.