FAU IISAH Corpus — A German Speech Database Consisting of Human-Machine and Human-Human Interaction Acquired by Close-Talking and Far-Distance Microphones

Werner Spiegl, Korbinian Riedhammer, Stefan Steidl, Elmar Nöth

University of Erlangen-Nuremberg, Pattern Recognition Lab
Martensstr. 3
91058 Erlangen, Germany
werner.spiegl, korbinian.riedhammer, stefan.steidl, noeth@informatik.uni-erlangen.de

Abstract

In this paper the FAU IISAH corpus and its recording conditions are described: a new speech database consisting of human-machine and human-human interaction recordings. Beside close-talking microphones for the best possible audio quality of the recorded speech, far-distance microphones were used to acquire the interaction and communication. The recordings took place during a Wizard-of-Oz experiment in the intelligent, senior-adapted house (ISA-House). That is a living room with a speech controlled home assistance system for elderly people, based on a dialogue system, which is able to process spontaneous speech. During the studies in the ISA-House more than eight hours of interaction data were recorded including 3 hours and 27 minutes of spontaneous speech. The data were annotated under the aspect of human-human (off-talk) and human-machine (on-talk) interaction.

1. Introduction

The here presented data were recorded during usability studies at the Pattern Recognition Lab of the University of Erlangen-Nuremberg. The studies took place in the intelligent, senior-adapted house (ISA-House, German: ISA-Haus), which is on the one hand the title of a project in the research association FitForAge of the Bavarian Research Foundation, and on the other hand an existing and fully furnished demonstration and living room with a dialogue based home assistance system adapted to the situation and needs of elderly people. Following (McTear, 2004) and (Bernsen et al., 1997) one part of these usability studies were Wizard-of-Oz (WOZ) experiments to find appropriate, real-life dialogues to configure and model the speech interface part of the home assistance system.

1.1. The Intelligent Senior-Adapted House (ISA-House)

The ISA-House project works towards an assistance system featuring an intuitive speech interface which is specifically designed for elderly people (Ott et al., 2009). The system is equipped with highly adapted functionality to fit the needs and situation of seniors (Soutschek et al., 2008). To get an idea of what elderly people expect from such an assistance system, we interviewed a group of 46 seniors asking how, in terms of technologies, assistance would be appreciated. From that we found simple tasks like switching the light but also more complex ones like telephony or scheduling assistance. Beside this functionality, we integrated in our system on the one hand a home server for controlling the heater, ventilation, electric socket and on the other hand a media server for controlling TV and radio. The communication and controlling of all these devices is resolved by using the UPnP (Universal Plug and Play) interface technology.

The general system design is now as follows. The system listens to the user to automatically determine when it receives the focus of attention. If so, the utterance is recognized and interpreted, and the requested action is performed, e.g., to switch on the light. In case of confusing orders or missing information, e. g., it is unclear which particular light, the system asks back. All the above comes along with strong challenges:

- Appropriate recording technology: Though headsets produce the best signal quality, they are quite uncomfortable if worn all day. Distant (room) microphones provide the most comfort but have inferior signal quality due to noise and reverberation.
- Age-adapted speech recognizer: Standard recognition systems are often designed and trained on speech and language of people aged around 30 and thus yield more errors on the aging voice.
- Focus of attention: The system has to automatically recognize if it is in the focus of the user. Keywords or push-to-talk solutions are the simplest and safest approaches to attract the attention of the system. However, they are counterintuitive to spontaneous interaction.
- Dialogue handling: Building a dialogue system with suitably modeled dialogues and an appropriate dialogue strategy can be a tricky task for someone not part of the target user group. It is not obvious which words people use to interact with their environment and what system responses they expect.

1.2. Solution Strategy and Concepts

To get to a solution that actually satisfies the target group, a group of seniors citizens is constantly involved in the development process. These are members of SEN-PRO, an German advisory board for product development for elderly people, organized by the Institute of Psychogerontology of the University of Erlangen-Nuremberg.

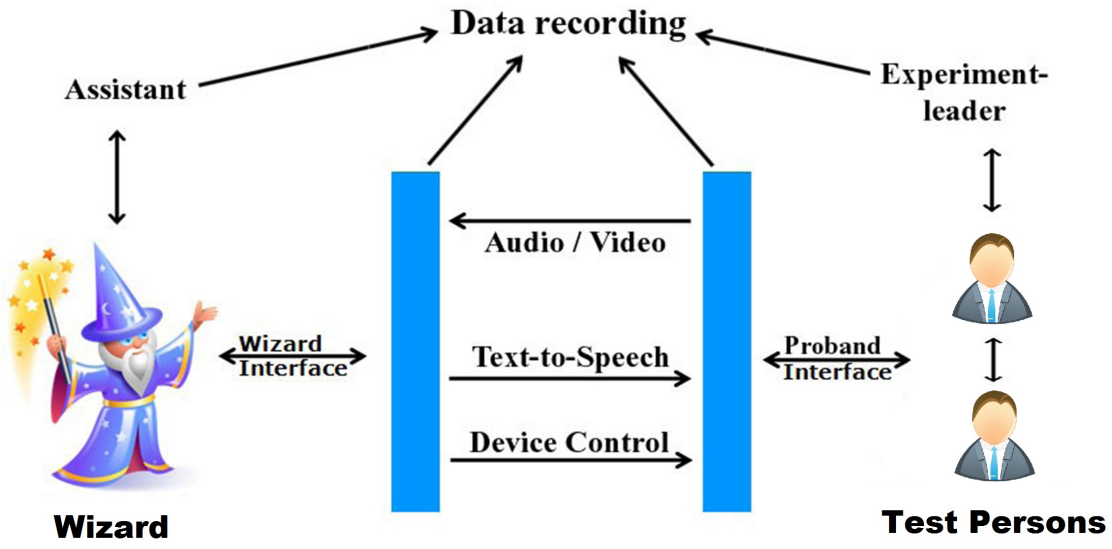

Figure 1: Scheme of our Wizard-of-Oz setup, after (Bernsen et al., 1997). Additionally, it is illustrated that two test persons are involved in our experiments to analyze on- and off-talk: the test persons interact with the system or they talk to each other.

As mentioned above, a first step towards the system design was to interview the senior citizens to find out in what way an automatic home assistance system could actually help them. The second step is now to see how seniors would actually communicate with and use the system. That, however, would require a fully functional system – which leads to some sort of chicken-or-the-egg problem: We can only build a satisfactory system if we know how the target group uses it, but we can only see how the users interact with the system if it is already working as expected. This is where we need a little magic to break the circle: Instead of many iterations of re-designing the system and risking that test subjects adapt to it, a hidden human "wizard" replaces the inner working of the system, that is the recognition and interpretation of utterances, and the subsequent responses and actions. Doing so, the front-end, i.e., what the user sees and hears, remains untouched thus the user assumes a fully working system. That kind of setup is well-known as Wizard-of-Oz (WOZ) experiment and it is a standard approach to develop user interaction systems. An example for data acquisition for an interactive TV can be found in (Brutti et al., 2008).

1.3. Overview

The rest of the article is structured as follows: In section 2 the recording environment and setup is described, the way in which the recordings were conducted. After that in section 3 the corpus is described, including the description of the audio data, the further processing and the transcription of the data.

2. Recording Environment and Setup

As pointed out above, the WOZ studies were performed in the ISA-House. In the following the planning and processing of the studies are presented including the recording of the data.

2.1. Wizard-of-Oz Scenario

The sessions of our WOZ study were planned and supervised in collaboration with a psychologist of the Institute of Psychogerontology. In general we arranged sessions of 90 minutes length for groups of two test persons, to study human-machine as well as human-human interaction, particularly on-/off-talk:

On-talk: A test person interacts with the system, i. e., the system has the focus of attention.

Off-talk: A test person does not interact with the system, e. g., communicates with the other test person

In total, 31 seniors (19m, 12f), native German speakers, aged 61 to 78 (68 $\pm$5) years participated in the WOZ experiment. In three cases, the second test person of a session was not of the target age group due to scheduling issues, so we got 17 sessions.

The agenda of a session included several steps to guide the test persons through the experiment. In the first step the participants got a short introduction to the ISA House. Then we explained the procedure of the test and the tasks. After that each of the two test persons was equipped with a close-talking microphone. Before the actual WOZ experiment, the test persons were told to read out loud a text in order to allow the system to adapt to their voice. They read the German version of "The North Wind and the Sun", a phonetically rich text by Aesop consisting of 108 (71 disjoint) words. After that, the system would greet the test persons, signaling the begin of the experiment. From that point, all uttered speech is considered to be spontaneous, as there were explicitly no instructions on how to interact with the system. In addition to utterances addressed to the system, there is (spontaneous) human-human interaction as the test persons were encouraged to also talk to each other. This WOZ experiment took approximately 30 minutes. Afterwards the test persons were interviewed about their ex-

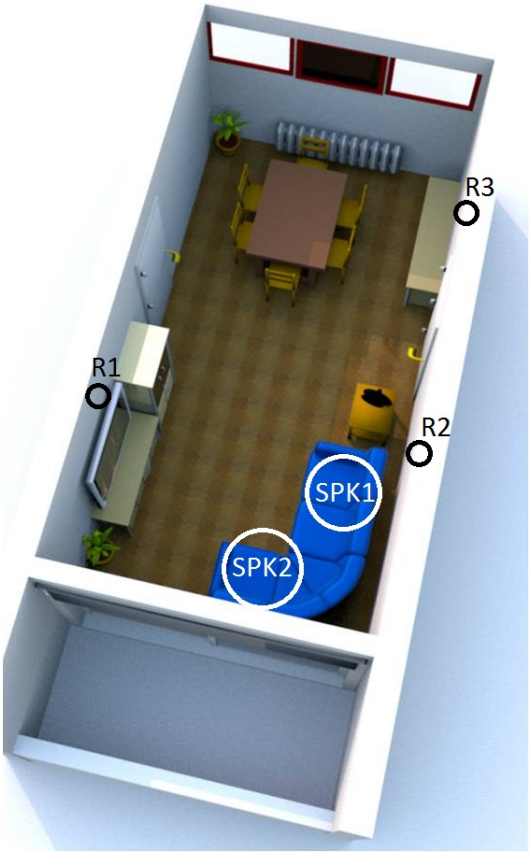

Figure 2: 3D-Model of the ISA-House: the recording room with the positions of the two test persons (SPK1, SPK2) and the pressure zone microphones (R1-R3)

perience concerning the tested system. The interviews were not part of the recordings.

2.2. Technical Setup

The setup of the WOZ sessions has to be split into two parts: On the one hand there is the test person interface with the recording-microphones and on the other hand the wizard interface for responding to the test person inquiries: device control and audio channel for further queries. The scheme for the recording setup is illustrated in figure 1.

2.2.1. Test person interface

The sessions were recorded using several microphones that were jointly captured by a mixing desk, the Phonic Helix Board 12 Firewire MKII, using the built-in Firewire port and the studio software Steinberg Cubase. Besides the head mounted SHURE WH20XLR of each test person, three T.Bone GZ400 pressure zone microphones (PZMs) were used as far-field microphones. PZMs are designed to cut off sound interferences from reflections with a smooth response of frequencies. Next to this an additional advantage over standard far-field microphones is the plain and unobtrusive attachment of the PZMs on the wall in the living room. In the ISA-House the microphones were mounted to the wall behind, opposite and to the far right of the speakers in approx. 1.5m height. For positioning of the speakers and the far-filed microphones refer to figure 2.

2.2.2. Wizard interface

The wizard in the side room was provided the headset recordings, each test person on one channel. Two software tools were developed to allow for wizard responses and to assure a fast, comfortable and non-exhausting environment for the colleague, who gives the wizard: an application for controlling all the available devices and functionalities in the ISA-House (e. g., the light, TV, radio, heater, telephone) and a software tool, which triggers a text-to-speech (TTS) system with predefined and parameterizable utterances, so the wizard doesn't have to type all the queries (e. g., "Which radio station do you want to hear?") and the answers (e. g., "Now it is 7 pm!"). As underlying TTS the software MaryTTS (http://mary.dfki.de/) was used.

3. Description of the FAU IISAH Corpus

The name FAU IISAH Corpus is an acronym for *Friedrich-Alexander-University - Interaction in the Intelligent, Senior-Adapted House*.

3.1. Recordings

The audio was acquired at a sampling rate of 48 kHz using a quantization of 24 bit. Afterwards, the recordings were normalized and converted to 16 kHz and 16 bit quantization as it is the most common format for speech recognition task. All in all more than eight hours of human-human and human-machine interaction data were recorded. Next to this spontaneous speech 31 recordings of read speech are acquired with about 22 minutes in total read by the test persons (see above, the test persons read the text "The North wind and the Sun").

3.2. Transcription

The recordings of spontaneous speech are processed and transcribed under the following objectives:

- Modeling dialogues for the ISA-House dialogue system
- Training a speech recognizer, adapted to the situation and condition of elderly people
- Building an on-/off-talk classifier for the ISA-House system

Therefore next to the usual transcription procedure (annotating spoken words, noise, accent, laughing etc.) the data is partitioned into turns: a turn describes a clear advice or information unit (e. g., "Turn on the light!", "At home we have a Panasonic TV.", "Yes!", "Thank you!"). Additionally, the utterances of the test persons have to be divided under the aspect of on- and off-talk, which denotes the decision whether the test person talks to the system or not.
From processing and dividing the interaction data into turns 3 hours and 27 minutes of speech data spoken by the test persons resulted. Divided by on-/off-Talk the test persons used 2891 turns for communicating with the other test person (off-talk) and 2752 turns of interaction with the system (on-talk). On the whole 5643 turns were uttered with 1751 different words. In table 1 the numbers are summarized. The linguistic and statistical analysis of the data is still ongoing.

	Turns	Words
On-talk	2752	964
Off-talk	2891	1188
On-/Off-talk	5643	1751

Table 1: Turns and different words used by the test persons during the WOZ studies (The word sets for on-talk and for off-talk are not disjunct, so on-/off-talk word number is not the sum, but the word number of the union of the sets.)

4. Conclusion

In this article we describe a speech database covering human-human and human-machine interaction in the context of an ambient assistance system for elderly people.

4.1. Areas of application for the FAU IISAH Corpus

The acquired and labeled data can greatly contribute to the following fields of speech-related research.

- The relatively high age of the speakers allows to investigate problems in understanding and recognizing elderly (spontaneous) speech and language.
- The synchronous signal acquisition using headsets and three distant room microphones at different typical speaker positions (opposite, behind, on the side) allows to study the effects of reverberation and speaker position on automatic speech, speaker and age recognition.
- The alternation of human-human and human-machine interaction provides realistic data to study on-/off-Talk in an ambient living context which is required for any good human-machine interaction.
- As there were no constraints how to address the system or how to form system requests, the resulting spontaneous dialogues give a good insight in how senior citizens, often not very tech savvy, interact with a virtual assistant.

4.2. Future work

At first the results of the evaluated data will be integrated into the ISA-House and its dialogue system, speech recognition and on-/off-talk component. Additionally, the data will be used to analyze the characteristics of the aging voice under various aspects, e. g., see (Mwangi et al., 2009). In conjunction with other data of test persons of different age groups the FAU IISAH Corpus can be used to improve age recognition applications like in (Spiegl et al., 2009) and (Bocklet et al., 2008).
In the second half of 2010 additional usability studies are planned for the ISA-House to evaluate the whole system by users of the target group: seniors. In this studies the wizard will be replaced by the real working dialogue system.

5. Acknowledgment

In the research association FitForAge of the Bavarian Research Foundation a team of scientists and engineers of 13 chairs of the four Bavarian Universities Erlangen-Nuremberg, Munich, Regensburg and Würzburg with 25 industrial partners is cooperating to develop products and services for an aging society. The aim of the research association is the development of technology-based solutions for aging people in their future life, to assist them at home, in their professional life, in the communication with their surroundings and in their participating of road traffic. Finally, not only elderly people should profit from this solutions, but all groups of society.

6. References

N. O. Bernsen, H. Dybkjaer, and L. Dybkjaer. 1997. *Designing Interactive Speech Systems: From First Ideas to User Testing*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, and E. Nöth. 2008. Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines. In IEEE Computer Society Press, editor, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 1605–1608.

A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo. 2008. WOZ Acoustic Data Collection for Interactive TV. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

M. F. McTear. 2004. *Spoken Dialogue Technology: Towards the Conversational User Interface*. Springer, London.

S. Mwangi, W. Spiegl, F. Hönig, T. Haderlein, A. Maier, and E. Nöth. 2009. Effects of Vocal Aging on Fundamental Frequency and Formants. In Acoustical Society of the Netherlands (NAG) and German Acoustical Society (DEGA), editors, *Proceedings of the International Conference on Acoustics NAG/DAGA 2009*, pages 1761–1764.

St. Ott, W. Spiegl, St. Soutschek, A. Maier, St. Steidl, and E. Nöth. 2009. Home Assistance System for Elderly People. In Russian Bavarian Conference on Bio-Medical Engineering Communication, editor, *Proceedings of the 5th Russian-Bavarian Conference on Biomedical Engineering*, page 173f.

St. Soutschek, W. Spiegl, St. Steidl, J. Hornegger, H. Erzigkeit, and J. Kornhuber. 2008. Technology Integration in the Daily Activities of the Elderly. *KI - Künstliche Intelligenz*, 4/2008(4):49–54.

W. Spiegl, G. Stemmer, E. Lasarcyk, V. Kolhatkar, A. Cassidy, B. Potard, St. Shum, Y. Ch. Song, P. Xu, P. Beyerlein, J. Harnsberger, and E. Nöth. 2009. Analyzing Features for Automatic Age Estimation on Cross-Sectional Data. In Interspeech 2009 10th Annual Conference of the International Speech Communication Association, editor, *Proceedings of Interspeech 2009*, pages 2923–2926.