Wizard of Oz Data Collection for the Czech Senior Companion Dialogue System

Milan Legát[1], Martin Grůber[1] and Pavel Ircing[1]

Abstract. In this paper, we present the setup of a Wizard of Oz environment used for collection of data for the implementation of the Czech Senior Companion dialogue system. We also discuss some aspects of using WoZ method for collection of emotional data and summarize some statistics about data set recorded. The domain of the collected data is limited to reminiscing about photographs. In each session a dialogue between elderly person and (WoZ) experimenter was recorded. Both audio and video data were collected.

1 INTRODUCTION

The dialogue systems are currently becoming a very active research area of many scientists. It is due to the fact that ASR and speech synthesis systems have made considerable progress in recent years which allowed their utilization in various areas. It must be, however, noted at this point that there are still many issues that are remaining to be solved. One of them is, unquestionably, incorporation of expressivity, i.e. developing of speech synthesizers producing expressive speech on the one hand and making ASR systems robust enough to handle utterances of speakers expressing their emotional states on the other hand. This issue is even more important when speaking about dialogue systems.

The human dialogue or discourse as such are very complex activities and require knowledge and reasoning capabilities of all participants. Thus, when developing a dialogue system we first need to restrict its domain to make the problem solvable. Ideally, the computer should be able to act in the same way, human would do in a particular domain. One can encounter simple dialogue systems when calling to a centre providing information about train schedules or services offered by a telecommunication company, etc. More advanced dialogue systems operating in the restaurant domain were presented in [1, 2].

In this paper we deal with the collection of audio and video data intended for the development of the Czech Senior Companion dialogue system. As this domain is still not limited enough it was restricted to the reminiscing about photographs. Our task is to make a system that would play the role of a partner to elderly people in the dialogue about their photographs. As mentioned above, we have used WoZ technique for data acquisition. This method is based on simulation of the dialogue system by a human, so-called "wizard". Ideally, the users do not notice the simulation and behave as if they were interacting with an automatic system rather than a human [3].

[1] Department of Cybernetics, University of West Bohemia in Pilsen, Univerzitni 8, 306 14, Pilsen, Czech Republic. Email: {**legatm, gruber,ircing**}**@kky.zcu.cz**.

This research is done within the COMPANIONS project [4] (www.companions-project.org). Another paper which is closely related to this one is [5] where the reader can find some additional information on the COMPANIONS project as well as some problem specifications and requirements posed on the data set being recorded.

The rest of this paper is organised as follows. In Section 2, we describe the data collection scenario as a whole. Section 3 serves to present the setup of the recording room including positions of the cameras. And also an example of screen shots taken by each of cameras is shown in that section. Section 4 is intended for brief description of WoZ software being used. In Section 5 we present the summary of the recorded data set. We discuss some pros and cons of the proposed approach to data collection in Section 6. Finally, we draw some conclusions and outline our future work in Section 7.

2 DATA COLLECTION SCENARIO

In each session a dialogue lasting approximately 55 minutes was recorded. The talking points of these dialogues were the sets of user's photos. The maximum number of pictures in each session was 12. This amount has shown to be sufficient as the users, we have recorded, really enjoyed reminiscing about their photographs and were inclined to spend a lot of time on a single photo. This is not very surprising as the elderly people like narrating and every single photo has its own story.

The computer (WoZ more precisely) acted as a dialogue partner the role of which was to stimulate the conversation and to give the user the feeling of being listened to by someone. This task was managed by using the set of typical questions, backchannel utterances and also pre-recorded non-speech dialogue acts expressing comprehension, amusement, hesitation, etc. We used our TTS system to generate the speech [6] output.

To be able to keep the interaction smooth with no unnatural silent pauses, the crucial thing was to have an appropriate set of pre-prepared sentences for the recording session so that the wizards can use them very quickly at a dialogue runtime simply by clicking on them. In Fig. 1 there is shown one of the users' photographs accompanied by the set of pre-prepared sentences. For this dialogue scenario preparation we used the information gained from the users along with their pictures. We visited all the users in their homes before recording session to speak with them about the chosen pictures and learn more about the topics related to them.

During the recording of a few initial dialogues we noticed that the users tended to be nervous at the beginning of the recording session. It was due to their unfamiliarity with talking to the computer as well as the environment of the recording room. That is why we have decided to spend some time before the session chatting with them to make them feel more relaxed. It is worth

mentioning here that the visit to user's home is also recommendable because it enables to gather important information for running a sophisticated dialogue with the user.

- Where was this picture taken?
- Was it a vacation?
- Did you go there regurarly?
- This can be which year?
- Who are the people on the picture?
- Your daughter?
- What is her name?
- Do you also remember the name of her friend?
- Which one is Veronica? *(the daughter)*
- How old was she at that time?
- Did she like the mountains? *(We already know that the picture was taken in the mountains.)*
- Did you also do some skiing there?
- Cross-country or down-hill?
- That must have been wonderful.
- You surely have nice reminiscences.

Figure 1: An example of a set of pre-prepared set of sentences related to a picture.

In addition there is also space to explain everything to the user in advance and give him/her positive feeling of the project. Nothing is as stressful as being uncertain about something and we need the users to be absolutely relaxed for the recording session.

3 SETUP OF RECORDING ROOM

A dedicated room has been established for the recording purposes. Speech from both the person and the avatar[2] were captured by a high-quality head microphones and recorded to a computer placed in another room (in order to minimize noise), sampled at 22kHz. The user was also recorded by three miniDV cameras simultaneously (see Fig. 3). The video data are currently just being archived and are intended for future use in audiovisual speech recognition, emotion detection, gesture recognition, etc.

In Fig. 2, there is shown the setup of the recording room. In addition to the three miniDV cameras, there was also a surveillance web camera placed in the recording room to check the status of the speaker and provide the wizards with a visual feedback. The only contact between a user and the computer was through speech, there was no keyboard nor mouse on the table.

The subjects were recorded from front, side and back view to provide data usable in various ways. The front view can be used for lips tracking which is beneficial for visual speech synthesis and recognition or for training of emotion detection algorithms. Combined with side view, it can be also utilized for 3D head modelling.

Since in the side view there was captured not only face but also the whole upper part of a body, it can be used for hands gesture and body movement tracking. The back view shows what was displayed on the LCD screen and in some cases what the speaker point at on the photograph. This information can be useful for example for tagging people on the picture when they are pointed at by the user while talking about them. This could be helpful for computer vision while seeking for the objects pictured on the photo.

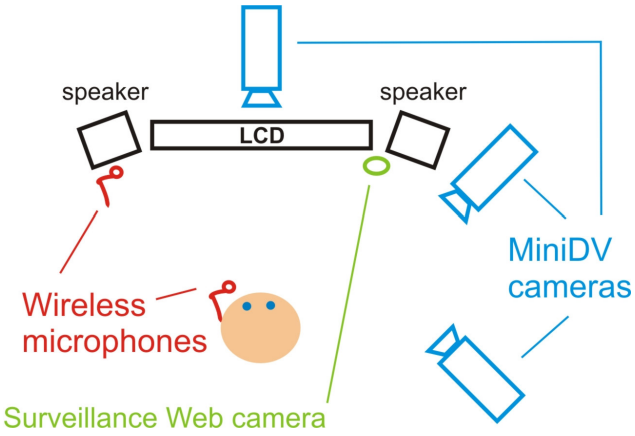

Figure 2: Recording room - cameras and microphones position

Figure 3: Screenshots from MiniDV cameras – right, front and back view

4 WoZ SOFTWARE USED

We have implemented a new network application equipped with the Czech 3D talking head avatar developed at University of West Bohemia in Pilsen [6]. On one side there is the "Presenter" which is the interface presented to the user during the recording session. On the other side there is the "Wizard" which is the interface used for managing the dialogue. It allows cooperation of two human wizards which is very useful for maintaining the

[2] Naturally, only the speech from the interviewed person will be used for ASR training the speech from the avatar is nevertheless important for dialogue tagging.

dialogue smooth even if it develops unpredictably and no pre-arranged sentences are available.

Figure 4: The "Presenter" interface

The "Presenter" interface is shown in Fig. 4. The users were presented with their photos on the right of the screen and 3D talking head avatar on the left. The bottom part served for displaying subtitles which were used during a couple of initial recording sessions to provide better understanding to the user. However, all subjects reported after recording sessions that the synthesized speech was clearly intelligible and it was decided to hide the subtitles for the rest of recording sessions.

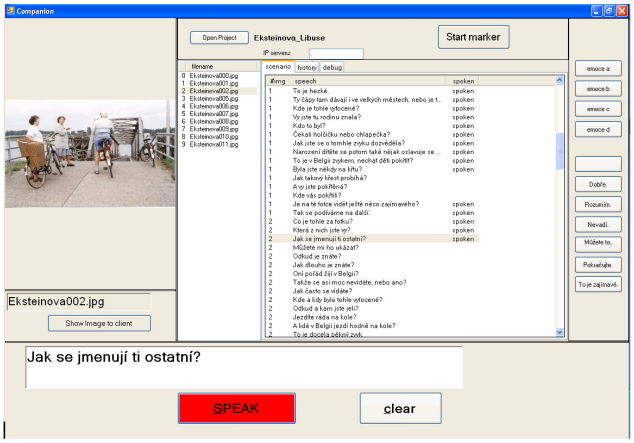

Figure 5: The "Wizard" interface

In Fig. 5, the "Wizard" interface is depicted. The middle part of the screen serves to display the pre-prepared scenario for a dialogue. Note that the wizards could select the sentences from the scenario, the assumption on how the dialogue could develop, by clicking on them. Each sentence of the scenario was given a number related to the picture displayed on the left. This enabled the orientation in large pre-prepared scenarios. Under the picture there is a button for displaying the picture on the "Presenter" screen. Once a sentence is selected by clicking on the list, it appears in the bottom edit box just above the buttons "SPEAK" and "clear". The displayed sentence can be modified before pressing "SPEAK" button and also an arbitrary text can be typed into the edit box. The right part of the screen is intended for displaying buttons bearing non-speech acts (smile, laughter, assentation, hesitation) and quick phrases (Yes. No. It's nice. Alright. Doesn't matter. Go on; etc.).

5 THE SENIOR COMPANION DATA SET

Almost all audio recordings are stored using 22kHz sample rate and 16-bit resolution. The first six dialogues were recorded using 48kHz sample rate, later it was reduced to the current level according to requirements of the ASR team.

The total number of dialogues recorded so far is 65, however, recording of even more dialogues is planned. Based on gender, the set of speakers can be divided into 37 females and 28 males. Mean age of the speakers is 69.3 years; this number is almost the same for both male and female speakers. The oldest person was a female, 86 years old. The youngest one was also a female, 54 years old.

All the recorded subjects were native Czech speakers; two of them (1 male and 1 female) spoke a regional Moravian dialect. This dialect differs from regular Czech language in pronunciation and, also, a little in vocabulary.

Approximately one half of the subjects stated in the after recording form that they have a computer at home. Nevertheless, most of them do not use it very often. Almost all the dialogues were rated as friendly and smooth. And even more, the users were really enjoying reminiscing on their photos, no matter that the partner in the dialog was an avatar.

Duration of each dialogue was limited to 1 hour, as this was the capacity of tapes used in miniDV cameras, resulting in average duration 56 minutes per dialogue. During the conversation, 8 photographs were discussed in average (maximum was 12, minimum 3).

To briefly outline how the dialogues develop we present the following figure:

AVATAR: *What about this photo?*
USER: Well, this is my son with his dog, his name is Cindy.
AVATAR: *What is your son's name?*
USER: And the son's name is Roman.
AVATAR: *How old is your son?*
USER: He is 28 years old.
AVATAR: *Do you have other children?*
USER: Yes, I have one more son.
AVATAR: *What is your son doing?*
USER: This son, Roman, has his own company and the other is named Jiří and he works in a warehouse in Plzeň.
AVATAR: *Tell me something about them.*
USER: Well, both of them are rather good sons, one of them is single, the other is divorced because just his wife with that Lucinka left for Norway and he stayed alone whereas the other son Roman has a girlfriend that he is only probably going to marry.

Figure 6: Initial phase of a dialogue

Thus, we have gathered more than 60 hours of speech data but the most important thing is that we have knowledge now how such dialogues develop and what is crucial for senior companion dialogue system development. Moreover, we have a

set of sentences to design a speech corpus for limited domain speech synthesis.

6 DISCUSSION ON THE CHOSEN METHOD

There are several issues related to the topic of collection of data for purposes of the development of this kind of dialogue system that merit a discussion. Regarding the task of the prospective system, which is to act as a companion of the elderly people, we should be aware that this is a highly emotional area in many aspects. The question is whether we are able to collect rich enough data from this point of view using WoZ method. The problem is that the avatar even if driven by a skilled wizard does not have the capabilities of expressing all emotional states which we would like to be simulated by the final system. This would suggest that recording of human-human dialogues would be a better idea.

On the other hand, we probably can not expect the users to behave in the same way in interaction with computer as they would do in human-human dialogue. Thus, human-human dialogue recording can be an unreliable source of data for some important aspects of dialogue system design, in particular the style and complexity of interaction [7].

This is obviously an open issue and a special study would be needed to support these assumptions.

7 CONCLUSIONS & FUTURE WORK

Up to now, we have recorded more than 60 hours of a unique audiovisual corpus for Czech language which is currently being annotated. The recording was made using high quality technical equipment – external sound card, pre-amplifier, two wireless head microphones (separately one for the subject and one for the avatar) and three miniDV cameras.

Speakers' audio tracks are supposed to be used for statistical model training in the field of automatic speech recognition, in the future it could be used e.g. for the recognition of emotions.

Avatar's audio tracks are being analyzed and the sentences uttered by the avatar are supposed to be recorded by a professional speaker for the purposes of the speech synthesis. The corpus for the speech synthesis will be recorded as a dialogue, where parts of real existing dialogues will be used. The speaker will be supposed to listen to the dialogue and in appropriate moments his/her turns will come on. His/her task will be to respond according to the avatar's track and to convey some expressivity in the speech. Text-to-speech system making use of a corpus containing such sentences is planned to be used in the conversational agent.

Video recordings are assumed to be used in the audiovisual speech recognition. They could also be used for emotion recognition, in conjunction with the speakers' audio tracks.

The main goal of this work is to develop a front-end for a conversational agent, which will be able to communicate with elder people in the limited domain of reminiscing about photographs. Nevertheless, the collected data is also supposed to be used for the development of the core of the Czech dialogue system.

ACKNOWLEDGEMENT

This work was funded by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

REFERENCES

[1] S. Whittaker, M. Walker, and J. Moore. Fish and Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain. In: *Language Resources and Evaluation Conference.* (2002)

[2] P. M. Strauβ, H. Hoffmann, and S. Scherer. Evaluation and User Acceptance of a Dialogue System Using Wizard-of-Oz Recordings. Intelligent Environments. In: *Proc. 3rd IET International Conference on Intelligent Environments IE 07,* pp. 521-524. (2007)

[3] P. Strauβ, H. Hoffman, W. Minker, H. Neumann, G. Palm, and S. Scherer et al. Wizard-of-Oz Data Collection for Perception and Interaction in Multi-User Environments. In *International Conference on Language Resources and Evaluation.* (2006)

[4] Y. Wilks. Artificial companions. Interdisciplinary Science Reviews, 30:145-152(8). (2005)

[5] Y. Wilks, D. Benyon, Ch. Brewster, P. Ircing, and O. Mival. Dialogue, Speech and Images: The Companions Project Data Set. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC08), Marrakech, Morocco. (2008)

[6] J. Matoušek, J. Romportl, D. Tihelka, and Z. Tychtl. Recent improvements on ARTIC: Czech text-to-speech system. In Proc. INTERSPEECH, Jeju, Korea, pp. 1933-1936. (2004)

[7] M. Železný, Z. Krňoul, P. Císař, and J. Matoušek. Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. In Signal Processing, 86:12, pp. 3657-3673. (2006)

[8] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of Oz Studies – Why and How. Knowledge-Based Systems, 6: 4. (1993)