# Czech Senior COMPANION: Wizard of Oz Data Collection and Expressive Speech Corpus Recording

**Martin Grůber, Milan Legát, Pavel Ircing, Jan Romportl, Josef Psutka**

University of West Bohemia
Univerzitní 8, 30614, Plzeň, Czech Republic
{gruber, legat, ircing, rompi, psutka}@kky.zcu.cz

**Abstract**

This paper presents part of the data collection efforts undergone within the project COMPANIONS whose aim is to develop a set of dialogue systems that will be able to act as an artificial "companions" for human users. One of these systems, being developed in Czech language, is designed to be a partner of elderly people which will be able to talk with them about the photographs that capture mostly their family memories. The paper describes in detail the collection of natural dialogues using the Wizard of Oz scenario and also the re-use of the collected data for the creation of the expressive speech corpus that is planned for the development of the limited-domain Czech expressive TTS system.

**Keywords**: data collection, corpus recording, expressive speech synthesis, dialogue system

## 1. Introduction

The research area of the automatic dialogue systems is recently receiving a considerable surge of attention from the scientific teams dealing with speech technologies and natural language processing. It is largely due to the fact that automatic speech recognition (ASR) and speech synthesis (TTS) systems have made considerable progress in recent years which allowed their utilization in various areas. However, one should bear in mind that those two components still constitute only a "front-end" and "back-end" of a system that would be able to engage in a natural dialogue with a human user. What still needs a lot of research effort are the "central" modules dealing with natural language understanding (NLU), information extraction (IE) and dialogue management (DM). Since human dialogues are very complex and require both specific and background knowledge and reasoning capabilities of all participants, the development of a general-purpose, unrestricted computer dialogue system is currently unfeasible.

Thus, when designing 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 at least in a given domain. For example, rather simple dialogue systems are nowadays often encountered 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 (Whittaker et al., 2002), (Strauss et al., 2007).

In the research being done within the COMPANIONS project (Wilks, 2005) (www.companions-project.org), it was decided to develop a computer system that would be able to conduct a natural dialogue with elderly users, mostly to keep the company and letting them to stay mentally active. As this restriction is still not sufficient enough, it was decided to narrow the task further to the reminiscing about family photographs. The system was named "Senior Companion" and was originally planned to be developed in two languages - Czech and English.

No dialogue system can be designed without prior knowledge about the specifics of the conversations that such system is going to deal with. Therefore at least a small sample of representative dialogues needs to be gathered, even when the developers plan to use rule-based techniques in the NLU, IE and DM modules. When (as is the case of the COMPANIONS project) there is an intention to employ machine-learning algorithms in all those modules, the amount of representative data that are necessary to gather is even more crucial.

Therefore this paper deals mostly with the data gathering efforts undergone in the preparation of the development of the Czech Senior Companion. The paper is organized as follows - Chapter 2 describes the basic premises of the data collection method and technical measures taken to ensure representative and high-quality corpus. Chapter 3 contains a brief description of the gathered corpus of natural dialogues, both quantitative and qualitative. Chapter 4 explains how the data from the dialogue corpus were re-used for the recording of the expressive speech corpus that will be used for the development of the new limited-domain TTS system and Chapter 5 presents a setup for the the annotation of the TTS corpus with communicative functions.

## 2. Data collection process

We have decided to employ the Wizard of Oz (WoZ) approach (Whittaker et al., 2002) in order to gather a corpus of human-computer dialogues. It means that human subjects were placed in front of the computer screen and were told that the program they are interacting with is fully autonomous, i.e. using "artificial intelligence" techniques to conduct a natural dialogue. In reality, the automatic speech recognition, understanding and response generation was simulated by a human operator (the "wizard"). Only the speech produced by a computer was genuinely generated using a TTS system coupled with 3D avatar

("talking head") (Železný et al., 2006), which further reinforced the subjects' belief that they are truly interacting with a computer only.
The wizard 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. To keep the dialogue smooth and natural, the crucial thing was to have pre-prepared sentences and questions that will be used. These sentences were saved in a so-called scenario. However, sometimes the dialogue does not follow the prepared scenario exactly, so the task of the wizards was to type the appropriate sentences on-line. This could have caused unnatural pauses in some cases but in general this problem was not so serious.
The recording of natural dialogues consists of separate sessions. In each session, one elder person (subject) was left alone in a recording room where necessary recording equipment was placed in. The setup of the recording room is depicted in Figure 1.

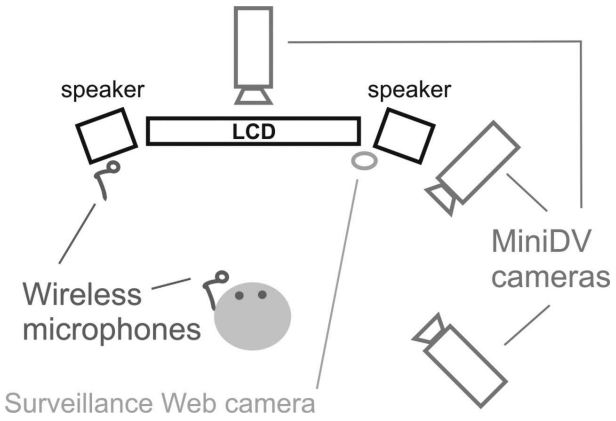


Figure 1: Recording room setup

In the recording room, the subject faces an LCD screen and two speakers, the speech is recorded by two wireless microphones, and the video is captured by three miniDV cameras. There is also one surveillance web-camera, just to monitor the situation in the recording room. The only contact between a user and the computer was through speech, there was no keyboard nor mouse on the table.
A snapshot of the screen presented to human subjects is shown in Figure 2. On the left upper part of the LCD screen, there is visualized 3D model of a talking head. This model is used as the avatar, the impersonate companion that should play a role of the partner in the dialogue. Additionally, on the right upper part, there is shown a photograph which is currently being discussed. On the lower half of the screen, there is a place used for displaying subtitles (just in case the synthesized speech is not intelligible sufficiently). The subtitles were used only during the first few dialogues. Later, the subtitles were not displayed because the generated speech was understandable enough and subjects did not have any problems to comprehend it. The speech was captured by two wireless microphones.


Figure 2: Snapshot of the WoZ system interface - user side

One microphone was used for the speech of the subject, the second one recorded the speech of the avatar. For high quality recording, an external preamplifier and an external Creative Sound Blaster Extigy sound card were used. 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 video of the session was also recorded, using three miniDV cameras. The subjects were recorded from the front, side and back view to provide data usable in various ways. The front view could be later used for the audio-visual speech recognition (where such viewing angle will be beneficial for lip-tracking) and also for the emotion detection algorithms. Along with the side view, it can be also used 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 subjects pictured on the photo.

## 3. Dialogue corpus characteristics

First some statistics - the current data set contains 65 dialogues. 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 spoke a regional Moravian dialect. This dialect differs from regular Czech language in pronunciation and, also, a little in vocabulary. 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).
Each participant of the experiment was asked to fill out a short questionnaire right after the recording session, asking question both about their computer literacy and the rating of the just finished dialogue with the computer. Approximately one half of the subjects stated in this 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 about their photos, no matter that the partner in the dialog was an avatar. This was confirmed not only by the questionnaires by also by wizards' observations during the sessions. A rather representative excerpt from one of the dialogues is presented in Table 1.

| | |
|---|---|
| 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. |

Table 1: Excerpt from a WoZ dialogue

To summarize, we have gathered more than 60 hours of speech data and, most importantly, we feel that we have a rather good knowledge about the way in which the conversation about the photographs usually develops and what kinds of "system" responses were the most appropriate for keeping the conversation rolling. Last, but not the least importantly, we have found out that the avatar operated by wizards, although equipped with a neutral voice only and a very limited set of facial expressions, is able to elicit quite a strong emotional response from the users. This is an important findings since the idea of an artificial companion being able to both detect and generate affective response is one of the hallmarks of the COMPANIONS project.
The resulting dialogue corpus can be of course also readily used for various machine-learning procedures, designed mainly to tailor the ASR system to the specific domain, such as re-training of the language models. Since we have quite a large amount of speech data for each individual user, we can also extensively test new speaker adaptation methods (Machlica et al., 2009).
Moreover, we have devised a way how the recorded data can be used to design and record a speech corpus for limited-domain expressive speech synthesis. The principle of this method is described in the following two chapters.

## 4. Design and recording of the expressive speech corpus

Development of the affective TTS system is a challenge that has still not been satisfactorily resolved. The main problem is that even just the classification of the affective (non-neutral) speech utterances is difficult.
Many methods of emotional (affective) state classification have been proposed. Very briefly and in simplicity - the basic distinction is whether a particular classification system is categorical, or dimensional. Among many we can name a categorical classification system (Ekman, 1999) which distinguishes emotional states such as anger, excitement, disgust, fear, relief, sadness, satisfaction, etc. In a dimensional model, emotions are defined as positions (or coordinates) in a multidimensional space where each dimension stands for one property of an emotional state. Various dimensions have been proposed out of which a widely accepted set is the one presented in (Russell, 1980) with two axes: valence (positive vs. negative) and arousal (high vs. low activation). Other models also consider a third dimension that is power or dominance and some even a fourth dimension: unpredictability. However, as was mentioned above, it is quite difficult to classify human speech according to either one of these models with the perspective of finding out acoustic correlates useful for generation purposes.
Therefore instead of labeling the emotions in the utterances (affective states) explicitly, we have settled for the assumption that a relevant affective state (of the conversational agent) goes implicitly together with a communicative function (CF) of a speech act (or utterance) which is more controllable than the affective state itself. It means that we do not need to think of modelling an emotion such as "guilt" per se - we expect it to be implicitly in an utterance like "I am so sorry about that" with a communicative function "apology".
Thus we have decided to proceed with the affective TTS corpus creation as follows. First, we hired a professional female speaker (stage-player) and instructed here not to express a specific emotions but just to put herself in the place of a Senior Companion. In order to facilitate such an empathy, a special software application was developed - it played back the parts of the WoZ dialogues where the subject was speaking (to provide the speaker with the relevant context) and at the time where the avatar have originally spoken, the dialogue was paused and the speaker was prompted to record the avatar's sentence herself. The text of the actual sentence was displayed on the screen even when the real (context) dialogue was being played so that the speaker had enough time to get acquainted with it before the recording. The recording equipment was again carefully selected and set-up in order to ensure the highest possible technical quality of the corpus - the speaker was placed in the anechoic room and the recording was done

using a professional mixing desk. The glottal signal was captured along with the speech.

That way we have recorded approximately 7,000 of (mostly short) sentences. Those were carefully transcribed and are ready to be annotated by a communicative functions (CF) described bellow.

## 5. Communicative functions

The set of CFs was partial inspired by (Syrdal and Kim, 2008) and their (draft) set is listed in Table 2. This list is most probably going to be further modified on the basis of the preliminary annotations of the recorded affective speech data.

| *dialogue act* | *example* |
|---|---|
| directive | Tell me that. Talk. |
| request | Let's get back to that later. |
| wait | Wait a minute. Just a moment. |
| apology | I'm sorry. Excuse me. |
| greeting | Hello. Good morning. |
| goodbye | Goodbye. See you later. |
| thanks | Thank you. Thanks. |
| surprise | Do you really have 10 siblings? |
| sad empathy | I'm sorry to hear that. |
| | It's really terrible. |
| happy empathy | It's nice. Great. |
| | It had to be wonderful. |
| showing interest | Can you tell me more about it? |
| confirmation | Yes. Yeah. I see. Well. Hmm. |
| disconfirmation | No. I don't understand. |
| encouragement | Well. For example? |
| | And what about you? |
| not specified | Do you hear me well? |
| | My name is Paul. |

Table 2: Set of communicative functions

Once the set of CFs is fixed, the entire TTS corpus will be annotated with the appropriate communicative functions by ten or more annotators simultaneously and the "winning" label of each of the utterances will be selected by the statistical algorithm that was already developed for similar task involving parallel annotations.

## 6. Conclusions and future work

This paper described data collection and annotation efforts needed for preparation of the corpora that were and/or are going to be used for the development of the Czech Senior Companion dialogue system.

Once the TTS corpora annotation is finished, the unit-selection algorithm in the Czech TTS system will be modified by changing the target-cost function so that the function will include a new feature for communicative function representation. The modified unit-selection algorithm will be hopefully able to generate speech expressing various communicative functions with implicit acoustic emotional cues.

## 7. Acknowledgements

This work was funded by the Ministry of Education of the Czech Republic, project No. 1M0567, and in part 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.

## 8. References

P. Ekman. 1999. Basic emotions. In Tim Dalgleish and Mick J. Power, editors, *The Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons Ltd, New York.

L. Machlica, Z. Zajíc, and A. Pražák. 2009. Methods of Unsupervised Adaptation in Online Speech Recognition. In *Specom 2009*, Saint Petersburg, Russia.

J. A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

P.-M. Strauss, H. Hoffmann, and S. Scherer. 2007. Evaluation and User Acceptance of a Dialogue System Using Wizard-of-Oz Recordings. In *IE 07*, pages 521–524, Ulm, Germany.

A. Syrdal and Y.-J. Kim. 2008. Dialog speech acts and prosody: Considerations for TTS. In *Speech Prosody 2008*, Campinas, Brazil.

M. Železný, Z. Krňoul, P. Císař, and J. Matoušek. 2006. Design, implementation and evaluation of the czech realistic audio-visual speech synthesis. *Signal Processing*, 12:3657–3673.

S. Whittaker, M. Walker, and J. Moore. 2002. Fish or Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain. In *LREC 2002*, Gran Canaria, Spain.

Y. Wilks. 2005. Artificial companions. *Interdisciplinary Science Reviews*, 30:145–152.