DEVELOPMENT OF EXPRESSIVE SPEECH SYNTHESIS FOR CZECH LIMITED DOMAIN DIALOGUE SYSTEM

M. Grůber, M. Legát,

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia,
Univerzitní 8, 306 14 Pilsen, Czech Republic
gruber@kky.zcu.cz, legat@kky.zcu.cz

Abstract
The paper describes several stages of the development of Czech expressive speech synthesis for limited domain dialogue system. The first phase includes the creation of a corpus containing natural human-machine dialogues, gathered using a WoZ technique. The recording setup and the process itself are described in the first part of the paper. In the second phase, an expressive corpus for speech synthesis is created. The theoretical background of the expressive TTS corpus design, the technical measures taken to ensure the corpus appropriateness and the corpus recording process are presented. All this research is being done within COMPANIONS project scenario, where one of the aims is to develop a Czech dialogue system allowing elderly people to reminiscence about their photographs.

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 remain 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.

Human dialogues as such are very complex 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 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] and [2].

In this paper, we mostly deal with the collection of audio and video data intended for the development of the Czech senior companion dialogue system. To be more restrictive regarding the content of the dialogues between a senior and a machine, the domain is limited 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. For required data acquisition a Wizard of Oz technique (WoZ) has been used [3]. 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 [2].

Currently, we have a concatenative speech synthesis system [4], which produces high-quality neutral speech. To incorporate some expressivity into the system, the recording of an expressive speech corpus is needed. The corpus was designed, recorded, annotated using so-

called communication functions and it is planned to be included into the existing neutral speech corpus.

A set of communication functions, partially inspired by dialogue acts described in [5], was created on the basis of the collected data. Each unit in the enriched corpus will be given an extra feature related to the communication function from which it originates. This extra feature will then be taken into account in the unit selection process at synthesis run time.

This research is being done within the COMPANIONS project [6] (www.companions-project.org). Another paper, which deals with similar data set collecting used in this dialogue system, is [7] 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 organized as follows: In Section 2, we describe the natural data collection process. Section 3 serves to present the set of the communication functions which will be used within the expressive speech synthesis. Section 4 is intended for a brief description of preparation works and recording of the expressive corpus. Finally, we draw some conclusions and outline our future work in Section 5.

2 Data Collection

To record the natural dialogues, the WoZ technique was used. This means that the dialogue between a human subject and a dialogue system was simulated. 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 events expressing comprehension, amusement, hesitation, etc. We used our TTS system [4] to generate the speech output.

To keep the dialogue smooth and natural, the crucial thing was to have a set of pre-prepared sentences (henceforth referred to as scenario) which the wizards could use very quickly at a dialogue runtime simply by clicking on them. A part of one of these scenarios along with the corresponding photograph is shown in Figure 1. Obviously, the dialogues did not always follow exactly the prepared scenario, 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.

- 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 sentences related to a picture*

The recording of the natural dialogues was done in hourly sessions. In each session, one elder person (object) was recorded being alone in a recording room. The setup of the recording room is depicted in Figure 2.

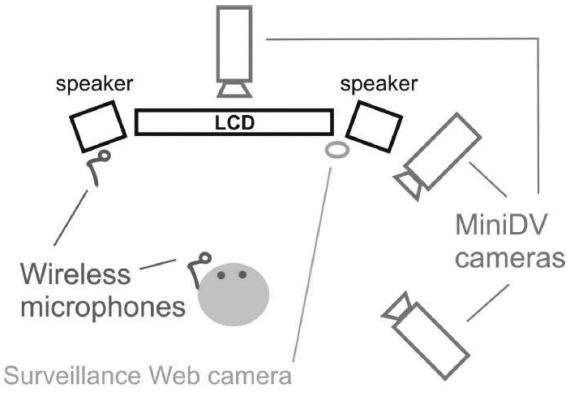

Figure 2 *Setup of the recording room*

In the recording room, the object was sitting at the table where only an LCD screen and two speakers were placed on. The only contact between the subject and the computer was through speech, there was neither keyboard nor mouse on the table. Speech was recorded by two wireless head microphones and the video was captured by three miniDV cameras. There was also one surveillance web-camera, just to check the status of the object and provide the wizards with a visual feedback.

In Figure 3, there is shown a snapshot of the screen which was presented to the object as a visual communication interface of the dialogue system. On the left upper part of the screen, there is visualized 3D model of a talking head. This model is used as an avatar, the impersonate companion that should play a role of the partner in the dialogue. In addition, on the right upper part, there is shown a photograph 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). However, after some recording sessions, all subjects reported that the synthesized speech was clearly intelligible and therefore it was decided to hide the subtitles for the rest of the recording sessions.

Figure 3 *Snapshot of the presenter*

Speech was captured by two wireless microphones. One microphone was used for speech of the object, the second one recorded 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 were stored using 22 kHz sample rate and 16-bit resolution. The first six dialogues were recorded using 48 kHz sample rate, later it was reduced to the final level according to the requirements of the ASR team.

The video was captured by 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 used for the visual speech synthesis, taking advantage of algorithms for lips tracking. 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 pointed 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.

The 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 was 69.3 years; this number was 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 objects 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.

Approximately one half of the objects 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 assessed 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 it in the Figure 4.

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 4 ***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 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. It is clear that many real future dialogues might, and very likely will, develop in a different way than the recorded ones. However, the coverage of the most widely used and the most relevant phrases should be assured. Besides, in these dialogues the human plays an active role whereas the machine only expresses an interest by various ways and stimulates the user to speak more by asking questions, and thus acquires more information about the photograph and the user.

3 Communication function

As mentioned above, in these dialogues we will not try to express an emotion explicitly, but we will focus on a set of affective states instead. However, these affective states will of course include emotions implicitly.

Each sentence in a dialogue is classified in two steps. In the first step, there is taken a decision, whether the sentence is a question or an indicative sentence. In addition, the type of the question can be more precisely specified using the following groups:

- question - multiple choice
- question - wh
- question - yes/no

In the second step, the communication function (real feeling) is determined. The communication is a term for a part of a dialogue that is uttered in other than neutral style, i.e. expressively. For this purpose a set of communication functions was proposed, see Table 1. However, to keep the consistency in labelling there is also a communication function marked *not specified*.

CF	**example**
Directive	Tell me that.
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	It's really terrible.
Happy empathy	It's nice. Great.
Showing interest	Can you tell me more about it?
Confirmation	Yes. Yeah. Well.
Disconfirmation	No. I don't understand.
Encouragement	Well. For example? And what about you?
Not specified	Do you hear me well?

Table 1 *Communication functions*

This set of communication functions is partially inspired by dialogue acts described in [5]. For our future application, this set can be still slightly modified; some communication functions can be removed, added or merged. Exact set of communication functions will be determined after the analysis of annotations (see below).

In addition to the communication function, some non-speech events like cough, smile, laughter, hesitation etc. are required to be included to the corpus. These events are very important parts of a dialogue when we are talking about a "natural" dialogue system. They have to be implemented into the speech synthesis and have to sound as naturally as possible. These events are very significant features that can show that the system is still listening to the speaker and understands him/her.

To illustrate a link between communication functions and emotions, let us consider following situation. When the subject thinks back on a certain event in the past, for example a wedding, he may speak emotionally (let us say expressively) and the emotion for this part of the dialogue could be marked as *joy* or *happiness*. When using communication functions, each sentence uttered during this part of the dialogue can be marked with different label, e.g. *directive* ("Tell me something about this wedding."), *happy empathy* ("It had to be really nice wedding."), *confirmation* ("Well, I understand."), *disconfirmation* ("I don't know that place.") or *encouragement* ("What was the wedding presents?").

It is supposed that these various styles differ in acoustic parameters, e.g. in intonation, F0 contour, intensity etc. as well as various emotional styles do [8]. Thus, the concatenation of speech units coming from sentences marked with the same communication function should be acoustically smoother. Moreover, considering this limited domain synthesis – reminiscing about photographs – it is very likely to have, during the synthesis time, the appropriate sentence pre-recorded and already stored in the speech corpus as a whole. Then the required sentence could be just played back. Using this approach, we can achieve better coverage of various parts of the dialogues in this domain and finer resolution of various speaking styles.

4 Preparation works and recording of the expressive speech corpus

Since the concatenative speech synthesis is planned to be used, we firstly needed to record expressive speech corpus. This corpus will be added to the neutral corpus we already have. The newly recorded expressive corpus was properly annotated with respect to the communication functions, which will than become one of the features used for the target cost computation.

The data used for the speech synthesis purpose has to be high-quality and should be recorded by professional speaker with ability to show the feelings and express them in speech. Therefore, the recording took place in an anechoic room and was performed by a stage player (who recorded also our neutral speech corpus). In addition, high-quality equipment like external mixing desk including sound card were used. Speech and the glottal signal were captured.

The expressive corpus for the speech synthesis was recorded as a dialogue, based on the previously recorded WoZ dialogues (described above). The speaker was listening to the natural dialogue (only the channel with the subject's speech) and in the appropriate moments the dialogue was paused and the speaker was prompted to record the required sentence (the original avatar speech). Her task was to respond in the same sense as the avatar did in the original dialogue, but different words could have been possibly used (it was up to the speaker and her feelings). She was also supposed to convey a suitable expressivity in speech.

A special application dedicated to this type of recording was developed. It allows the playback of the natural dialogue; in the appropriate time, before the talking head starts speaking in the original recording, the dialogue is paused and the speaker has an opportunity to record a desired sentence. The text of the sentence is displayed on the screen in advance so that the speaker has enough time to get acquainted with it before the recording.

Almost all the natural dialogues were re-recorded this way. We have recorded approximately 7,300 of (mostly short) sentences. Those were transcribed and annotated by communication functions described above.

5 Conclusion and future work

More than 60 hours of a unique audiovisual corpus for Czech language has been recorded and manually transcribed. The recordings were made using high-quality technical equipment – external sound card, pre-amplifier, two wireless head microphones (separately one for the senior and one for the avatar) and three miniDV cameras.

Speakers' audio tracks were mainly used during the expressive speech corpus recording. However, the tracks are also 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 have been analyzed and the sentences uttered by the avatar were re-recorded by a professional speaker for the purposes of the expressive speech synthesis.

According to the recorded expressive speech corpus, a set of communication functions was proposed to cover our limited domain task. The corpus was later transcribed and afterwards annotated using proposed communication functions. During annotations, the annotators were listening to the utterances and marked each sentence with one or more communication functions (we assume that one sentence can represent more than one expressive style).

The annotations will then be used in the concatenative speech synthesis system, where the relevant communication function will be taken as one of the features of a speech unit.

We suppose that this approach should change the style of the synthesized speech. Now, the style is only neutral. When we put the neutral and the expressive corpora together, a huge database with acoustically various speech units will be created. Hopefully, the synthesized speech will be natural and will express feelings suitable in any specific situations within this limited domain.

6 Acknowledgement

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

References

[1] S. Whittaker, M. Walker, and J. Moore, "Fish or fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain." in *Language Resources and Evaluation Conference*, Gran Canaria, Spain, 2002.

[2] P.-M. Strauss, H. Hoffmannt, and S. Schererl, "Evaluation and user acceptance of a dialogue system using Wizard-of-Oz recordings," in *Proceedings of the 3rd IET International Conference on Inteligent Environments IE 07*, Ulm, Germany, 2007, pp. 521–524.

[3] M. Legát, M. Grůber, and P. Ircing, "Wizard of Oz data collection for the Czech senior companion dialogue system," in *Fourth International Workshop on Human-Computer Conversation*. Bellagio, Italy: University of Sheffield, 2008, pp. 1–4.

[4] J. Matoušek, J. Romportl, D. Tihelka, and Z. Tychtl, "Recent improvements on ARTIC: Czech text-to-speech system," in *Proceedings of Interspeech, 8th International Conference on Spoken Language Processing – ICSLP*, vol. 3. Jeju, Korea: Sunjin Printing Co., 2004, pp. 1933–1936.

[5] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," in *Proceedings of Speech Prosody*, Campinas, Brazil, 2008.

[6] Y. Wilks, "Artificial companions," *Interdisciplinary Science Reviews*, vol. 30, pp. 145–152, June 2005.

[7] Y. Wilks, D. Benyon, C. Brewster, P. Ircing, and O. Mival, "Dialogue, speech and images: The companions project data set," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, E. L. R. A. (ELRA), Ed., Marrakech, Morocco, May 2008.

[8] M. Grůber and M. Legát, "Single speaker acoustic analysis of Czech speech for purposes of emotional speech synthesis," in *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, vol. 2. Aberdeen, UK: The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2008, pp. 84–87.