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

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

Department of Cybernetics, Faculty of Applied Sciences,
University of West Bohemia, Czech Republic
{gruber,legat,ircing,rompi,psutka}@kky.zcu.cz
http://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 [9], [6].

In the research being done within the COMPANIONS project [10] (`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 the process of the TTS corpus annotation with communicative functions.

2 Data collection process

We have decided to employ the Wizard of Oz (WoZ) approach [9] 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") [8], 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 utter-

ances 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.

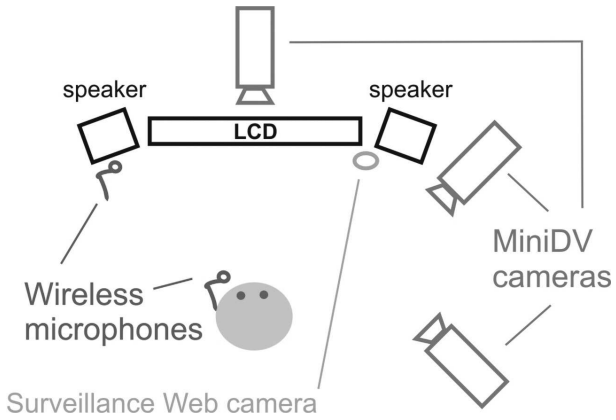

Fig. 1. Recording room setup

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.

Fig. 2. Snapshot of the WoZ system interface - user side

The speech was captured by two wireless microphones. 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 finding 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 [11].

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 [2] 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 [5] 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 in-

structed 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 annotated by communicative functions (CF) described bellow.

5 Annotation using communicative functions

The set of CFs was partial inspired by [7] and is listed in Table 2. The expressive speech corpus was annotated using communicative functions by means of a listening test. An additional label for "other" communicative function was introduced for the test purposes only — this label is not listed in the table. The test was aimed to determine objective annotation on the basis of several subjective annotations as the perception of expressivity is always subjective and may vary depending on particular listener.

A special web application working on the client-server basis was developed for the listening test purposes. This way listeners were able to work on the test from their homes without any contact with the test organizers. Various measures were undertaken to detect possible cheating, carelessness or misunderstandings.

The test participants have been instructed to listen to the recordings very carefully and subsequently mark communicative function(s) that are expressed within the given sentence. The number of CFs that should be assigned to a single sentence was not prescribed, this decision was left to listeners' discretion. In order to facilitate their job, listeners also had a few sample sentences labelled with communicative functions available during the whole course of the test.

That way we obtained subjective annotations that of course somehow vary across the listeners. A proper combination of those subjective annotations was needed in order to objectively annotate the expressive recordings. Therefore an evaluation of the listening test was made.

We utilized two approaches to the inference of the objective annotation:

- The first way is a simple majority method. Using this easy and intuitive approach, each sentence is assigned a communicative function that was selected by the majority of the listeners. If this majority accounts for less then 50% of all listeners, the classification of the sentence is considered to be unreliable.

comunicative function	*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

– The second approach is based on maximum likelihood method. Maximum likelihood estimation is a statistical method used for fitting a statistical model to data and providing estimates for the model's parameters. The maximum likelihood estimator is consistent. It means that having a sufficiently large number of observations (annotations in our case), it is possible to find the value of statistical model parameters with arbitrary precision. The parameter calculation is implemented using the EM algorithm [1]. Knowing the model parameters we are able to infer the objective annotation. Precision of the estimate is one of the outputs of this model. A sentence labelled with a communicative function with low precision can be eliminated from the expressive corpus.

Comparing these two approaches, 35 out of 7287 classifications were marked as untrustworthy using maximum likelihood method and 571 using simple majority method. The average ratio of listeners who marked the same communicative function for particular sentence using simple majority approach was 81%, when untrustworthy classifications were excluded. Similar measure for maximum likelihood approach cannot be easily computed as the model parameters and the estimate precision depend on number of iteration in the EM algorithm.

Finally, we decided to use the objective annotation obtained by maximum likelihood method. We have also successfully used this approach in recent works regarding speech synthesis research, see [4].

Further, we need to confirm that the listeners marked the sentences with communicative functions consistently and achieved some measure of agreement.

Otherwise the subjective annotations could be considered as accidental or the communicative functions inappropriately defined and thus the acquired objective annotation would be false. For this purpose, we make use of two statistical measures for assessing the reliability of agreement among listeners.

One of the measures used for such evaluation is Fleiss' kappa. It is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. We calculated this measure among all listeners separately for each communicative function. Computation of overall Fleiss' kappa is impossible because the listeners were allowed to mark more than one communicative function for each sentence. However, the overall value can be evaluated as the mean of Fleiss' kappas of all communicative functions.

Another measure used here is Cohen's kappa. It is a statistical measure of inter-rater agreement for categorical items and takes into account the agreement occurring by chance as well as Fleiss' kappa. However, Cohen's kappa measures the agreement only between two listeners. We decided to measure the agreement between each listener and the objective annotation obtained by maximum likelihood method. Again, calculation of Cohen's kappa was made for each communicative function separately. Thus we can find out whether particular listener was in agreement with the objective annotation for certain communicative function. Finally, the mean of Cohen's kappas of all communicative functions was calculated.

Results of agreement measures are presented in Table 3. Value of Fleiss' and Cohen's kappa vary between 0 and 1, the higher value the better agreement. More detailed interpretation of measure of agreement is in [3].

The Fleiss' kappa mean value of 0.5434 means that the measure of inter-listeners agreement is moderate. The Cohen's kappa mean value of 0.6632 means that the measure of agreement between listeners and objective annotation is substantial. As it is obvious from Table 3, communicative functions *OTHER* and *NOT-SPECIFIED* should be considered as poorly recognizable. It is understandable when taking into consideration their definitions.

Additionally, in Table 3, there are also shown probabilities of the particular communicative functions occurrence when maximum likelihood method was used for the objective annotation obtaining. It is obvious that communicative functions *SHOW-INTEREST* and *ENCOURAGE* are the most frequent.

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.

Since 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

communication function	Fleiss's kappa	Measure of agreement	Cohen's kappa	Cohen's kappa SD	Measure of agreement	Occurr. probab.
DIRECTIVE	0.7282	Substantial	0.8457	0.1308	Almost perfect	0.0236
REQUEST	0.5719	Moderate	0.7280	0.1638	Substantial	0.0436
WAIT	0.5304	Moderate	0.7015	0.4190	Substantial	0.0073
APOLOGY	0.6047	Substantial	0.7128	0.2321	Substantial	0.0059
GREETING	0.7835	Substantial	0.8675	0.1287	Almost perfect	0.0137
GOODBYE	0.7408	Substantial	0.7254	0.1365	Substantial	0.0164
THANKS	0.8285	Almost perfect	0.8941	0.1352	Almost perfect	0.0073
SURPRISE	0.2477	Fair	0.4064	0.1518	Moderate	0.0419
SAD-EMPATHY	0.6746	Substantial	0.7663	0.0590	Substantial	0.0344
HAPPY-EMPATHY	0.6525	Substantial	0.7416	0.1637	Substantial	0.0862
SHOW-INTEREST	0.4485	Moderate	0.6315	0.3656	Substantial	0.3488
CONFIRM	0.8444	Almost perfect	0.9148	0.0969	Almost perfect	0.1319
DISCONFIRM	0.4928	Moderate	0.7153	0.1660	Substantial	0.0023
ENCOURAGE	0.3739	Fair	0.5914	0.3670	Moderate	0.2936
NOT-SPECIFIED	0.1495	Slight	0.3295	0.2292	Fair	0.0736
OTHER	0.0220	Slight	0.0391	0.0595	Slight	0.0001
mean	*0.5434*	*Moderate*	*0.6632*		*Substantial*	

Table 3. Fleiss' and Cohen's kappa and occurrence probability for various communicative functions and for the "consecutive CFs" label. For Cohen's kappa, mean value and standard deviation is presented, since Cohen kappa is measured between annotation of each listener and the reference annotation.

speech expressing various communicative functions with implicit acoustic emotional cues.

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.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39(1), 1–38 (1977), with discussion
2. Ekman, P.: Basic emotions. In: Dalgleish, T., Power, M.J. (eds.) The Handbook of Cognition and Emotion, pp. 45–60. John Wiley & Sons Ltd, New York (1999)
3. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (March 1977)

4. Romportl, J.: Prosodic phrases and semantic accents in speech corpus for Czech TTS synthesis. In: Text, Speech and Dialogue, proceedings of the 11th International Conference TSD 2008. Lecture Notes in Artificial Intelligence, vol. 5246, pp. 493–500. Springer, Berlin–Heidelberg, Germany (2008)
5. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)
6. Strauss, P.M., Hoffmann, H., Scherer, S.: Evaluation and User Acceptance of a Dialogue System Using Wizard-of-Oz Recordings. In: IE 07. pp. 521–524. Ulm, Germany (2007)
7. Syrdal, A., Kim, Y.J.: Dialog speech acts and prosody: Considerations for TTS. In: Speech Prosody 2008. Campinas, Brazil (2008)
8. Železný, M., Krňoul, Z., Císař, P., Matoušek, J.: Design, implementation and evaluation of the czech realistic audio-visual speech synthesis. Signal Processing 12, 3657–3673 (2006)
9. Whittaker, S., Walker, M., Moore, J.: Fish or Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain. In: LREC 2002. Gran Canaria, Spain (2002)
10. Wilks, Y.: Artificial companions. Interdisciplinary Science Reviews 30, 145–152 (2005)
11. Zajíc, Z., Machlica, L., Müller, L.: Refinement approach for adaptation based on combination of MAP and fMLLR. Lecture Notes in Computer Science pp. 274–281 (2009)