THE RESTAURANT BOOKING CORPUS – CONTENT-IDENTICAL COMPARATIVE HUMAN-HUMAN AND HUMAN-COMPUTER SIMULATED TELEPHONE CONVERSATIONS

Ingo Siegert[1], *Jannik Nietzold*[2], *Ralph Heinemann*[2], *Andreas Wendemuth*[2]

[1]*Mobile Dialog Systems, Institute for Information and Communications Engineering, Otto-von-Guericke University*
[2]*Cognitive Systems Group, Institute for Information and Communications Engineering, Otto-von-Guericke University*
ingo.siegert@ovgu.de

Abstract: A new dataset, the Restaurant Booking Corpus (RBC) is introduced, comprising 90 telephone dialogs of 30 German speaking students (10 males, 20 females) interacting either with one out of two different technical dialogue systems or with a human conversational partner. The aim of the participants was to reserve a table each at three different restaurants for four persons under certain constraints (late dinner time for one day, sitting outside, reachable via public transport, availability of vegetarian food, getting the directions to the restaurant). The purpose of this constraints was to enable a longer and realistic conversation over all three calls. This dataset is explicitly designed to eliminate certain factors influencing the role model of the interlocutor: the effect of a visible counterpart, the speech content, and the dialog domain. Furthermore, AttrakDiff is used to evaluate the correct implementation of the conversational systems. A human annotation and an automatic recognition is pursued to verify that the speech characteristics are indistinguishable for the human-directed and the device-directed calls.

1 Introduction

During the last years, the market for commercial voice assistants has rapidly grown. Microsoft Cortana had 133 million active users in 2016, the Echo Dot was the best-selling product on all of Amazon in the 2017 holiday season. Furthermore, 72% of people who own a voice-activated speaker say their devices are often used as part of their daily routine. The attractiveness of today's voice assistants is especially based on their ease of use. Using nothing but speech commands, users can play music, search the web, create to-do and shopping lists, shop online, get instant weather reports, and control popular smart-home products.

Besides making the operation of technical systems as simple as possible, voice assistants should enable a natural interaction. Therefore, one aspect that still needs improvement is to automatically recognise device-directed utterances. Although, multiple solutions are implemented to detect if a system should react to an uttered speech command, in particular used are push-to-talk and keywords as wake-word, this interaction initiation is still very unnatural. Furthermore, the currently preferred wake-word method is error-prone. Therefore, more sophisticated addressee detection systems are needed, to determine when an utterance is directed towards the voice assistant system.

Various studies using different datasets have already been realised in this field of research. In these studies, the technical system is either a robot, a research system, or a Wizard-of-Oz (WOZ)-experiment. In [1], 150 multiparty interactions of two to three people playing a trivia

question game with a computer are utilised. In [2], data of 38 sessions of two people interacting in a more formal way with a "Conversational Browser" are recorded. The authors of [3] used two different experimental settings (standing and sitting) of a WOZ data collection with 10 times two speakers interacting with an animated character. The data used in [4] are recorded interactions of three to four people sitting around a computer display, answering questions from the users. The authors of [5] used 250 hours of human interactions with voice controlled far-field devices.

A disadvantage of all of these studies is that they condone differences between human-directed and system-directed speech. Especially the speech content and the speech situation is strongly different, due to the limited understanding of actual technical systems. The task the participants have to solve in the different presented datasets always puts the participants the human conversational partners and the technical systems in different roles. These roles influence the speaking behaviour [6].

To overcome this issues, the Restaurant Booking Corpus is recorded. This dataset lets the participants perform the same task with a human being or a technical system as conversational partner. RBC explicitly assigns the same role to the human conversational partner as well as to the technical system. To support this role model, certain influencing factors are eliminated: the effect of a visible counterpart, the speech content, and the dialog domain. Additionally, the statements given by the human conversational partner and the technical systems is identical.

2 Study Design

The recorded corpus can be used for various analyses. However, the main purpose was to design a dataset of naturalistic human-human and human-machine conversations with only less variations between these conversations. Therefore, the only difference between the interaction was the voice of the conversational partner, all other factors were hold identical. It is assumed that by this design the participants' speech is indistinguishable for human directed and system directed speech.

2.1 Task

As task for the participants, a scenario was chosen were the speech content could be designed in such a way that it can be naturally similar for both a human and a technical system. But in the same time allowing a certain degree of freedom to avoid mindless repetitions. Thus, as scenario restaurant bookings for a 3-day journey at three different restaurants was used.

As cover story the participants were told that they should test different technical telephone-based dialog systems in comparison to a human agent. They were furthermore informed that they will speak with a human being in addition to the two conversations with technical systems. The order of the two technical systems and the human interlocutor was not given to them. This was permuted for every participant based of the order of conversations. This should also permute over all restaurants so that each restaurant is used for the two technical systems and the human being.

As task for the conversation, they had to book restaurant tables for dinner during a planned 3-day weekend trip with some friends to a City called "Quedlinstätt". A list with three recommended restaurants was given to the participants. Furthermore they should consider specific requirements: a table for four persons, the date of the trip, the availability of vegan dishes, the dishes of the day, a table outside, availability of public transport or the walking distance to the accommodation. This task description was given to all participants in written form including the details of the journey.

To allow a certain variance in the three calls, a script sheet was designed to cover the specific implementation for each restaurant. The differences are mainly in the name, the availability of free tables on specific dates, the availability of vegan dishes, dishes of the day, possibility to sit outside, the location and direction, Table 1 gives a short overview.

	Alte Wassermühle	Restaurant Adler	Restaurant Krone
Possible reservation Date	Saturday	Sunday	Friday
Vegan Dishes	wide range of vegan and gluten-free dishes	always vegan alternative	vegan own creations
Dishes of the day	dependent of market	Young-peas risotto, haunch of venison	none
Outside tables	terrace to the river	garden tables	tables on market place
Location	Westringbrücke	between city and park	on the market
Public transport	tram	only car or taxi	waling distance

Table 1 – Individual differences of designed restaurants.

After the introduction, the participant had to read the task description aloud. The purpose was, to lower the first excitement at the beginning of the experiment and to capture the reading voice of the participants. This reading part was recorded as well. Then the participants could freely chose the order of their calls. After they had called all three restaurants, the participants had to answer two questionnaires. First AttrakDiff [7] was used to evaluate the three different systems and to maintain the cover story. Furthermore, a short form of a self-defined questionnaire used in [8] was utilized to obtain socio-demographic information as well as the participants' experience with technical systems.

2.2 Implementation

To implement the technical dialog systems, a WOZ technique was used. For each restaurant all possible answers were pre-generated using MaryTTS (voice: bits3-hsm male German) based on the corresponding script. Thereby, two versions of technical systems are developed. For the first version, denoted as TS1, it is assumed that the technical system only understands single-slot requests. For the second version, it is assumed to have a more elaborated system also able to understand multi-slot requests and back references. This system is denoted as TS2. During the experiment all system answers are played-back by a human operator. Furthermore, the human agent, denoted as H, directly works with the script to give the same answers as the two technical systems. The telephone system of the participant was implemented as a simulated SIP-Dialer, see Figure 1. The corresponding sounds were played back via headset.

3 Recording Setup

The recordings took place at the Institute of Information and Communication Engineering, Cognitive Systems Group, University Magdeburg. They were conducted in a living-room-like surrounding. The aim of this setting was to enable the participant to get into a natural communication atmosphere (in contrast to the distraction of laboratory surroundings).

To conduct the recordings, a headset (Sennheiser Chat 3) was used for the participant to simulate the SIP-Telephone. Additionally a high quality shotgun microphone (Sennheiser ME

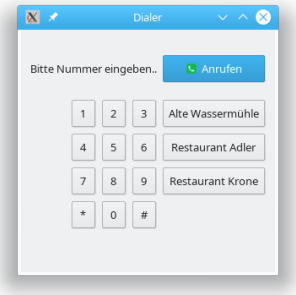

Figure 1 – Screenshot of the Mockup SIP-Dialer.

66) was utilized. The same type of microphone was also used to record the human agent. The technical system's utterances as well as the output of the SIP-Dialer are directly recorded from the operator's computer line-out. All recordings were stored uncompressed in WAV-format with 44.1 kHz sample rate and 16 bit resolution.

4 Dataset Characteristics

Subjects/Experiments	30
Language	German
Sex	Male 10 / Female 20
Age	Mean 23.97 (Std: 3.45) Min: 18; Max: 31
Total Recorded Data	5 h 37 min
Duration per Call	Mean: 193.6 $\pm$ 43.6 sec
Total Number of Utterances	4 835 (2 087 Agent)
Duration per Utterance (Caller) Duration per Utterance (TS1, TS2) Duration per Utterance (Human Agent)	2.96 sec 3.78 sec 4.29 sec
Annotation	Utterances, Transcription, context

Table 2 – Overview of RBC's dataset characteristics.

RBC contains recordings of 30 German speaking participants, all students at the Otto von Guericke University Magdeburg. The dataset comprises 20 female and 10 male participants, the age ranges from 18 to 31 years (23.97 $\pm$ 3.45 y). The participants came from different study courses including computer science, engineering science and humanities. Thus, this dataset is not biased towards technophilic students. The data collection took about 20 minutes (5 minutes introduction, 10 minutes recording and 5 minutes questionnaires) per participant. Via manual annotation, each utterance was annotated with its speaker, context and textual transcription. The contrext comprises the interlocutor, such as human or technical system, off-talk, laughter, and more. The textual transcription is obtained using Google Cloud Speech API automatic speech recognition service with an additional manual correction. Table 2 summarises the dataset characteristics.

The call duration for the different systems of RBC is depicted in Figure 2-left. It can be seen that the calls having a human partner (H) are significantly shorter than with a technical system. Furthermore the simple technical system (TS1) has the longest call duration. However, if the length of the individual utterances is considered (Figure 2-right), it can be seen that the duration

for the callers' utterances is nearly similar between H and TS2 and a bit shorter for TS1. This can be explained by the experimental design, the human partner is able to understand complex statements, TS2 at least combined requests. For TS1 the caller has to formulate a statement for each information. This led to shorter utterances of the participant. But this increases the number of utterances and thus also the call duration of TS1 calls.

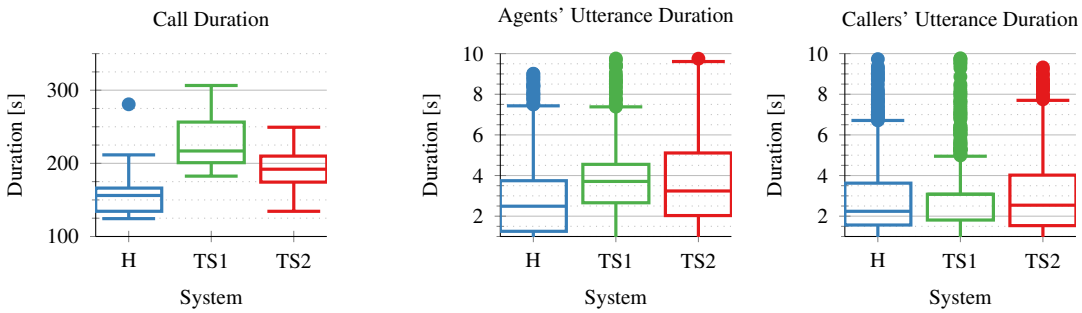

Figure 2 – Boxplots of durations for each call (left) and each utterance (right) for the different conditions. H denotes the human agent, TS1 is the single-slot system and TS2 the multi-slot system.

5 Validation of Experimental Design

In the following, some initial results are presented to validate the experimental design. By AttrakDiff the usability and design of the conversational partners are evaluated. Afterwards, it is tested whether humans are able to distinguish human-directed and device-directed utterances for this dataset. Furthermore, an addressee recognition system is employed to test the automatic recognition performance as well.

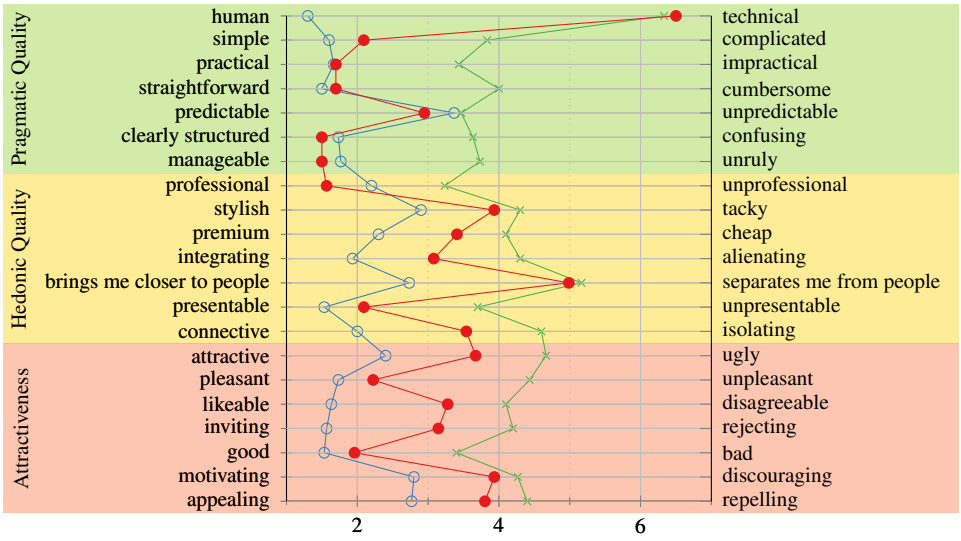

Figure 3 – Evaluation of the AttrakDiff questionnaire for the three conversational partners: Human (—○—), TS1 (—×—) and TS2 (—●—).

6 Perception of the three systems

AttrakDiff is generally employed to understand how participants evaluate the usability and design of interactive products [7]. It distinguishes three aspects, the pragmatic quality (PQ), the hedonic quality (HQ), and the attractiveness (ATT). The evaluation of these aspects for the

three conversational partners is given in Figure 3. It can be seen that in accordance with our experimental design, the human conversational partner is evaluated very positive, the single-slot system (TS1) is evaluated much worse for all items. The multi-slot system (TS2) is in between these two evaluations with a slight shift towards the human partner in case of the pragmatic quality.

	UAR	UAP
Human Evaluation (GER)	$60.54 \pm 3.25\%$	$60.14 \pm 2.33\%$
Human Evaluation (NON-GER)	$53.57 \pm 4.77\%$	$53.35 \pm 4.38\%$
Automatic Recognition (H vs. TS1 & TS2)	50.55 ± 0.22	50.62 ± 0.28
Automatic Recognition (R vs. TS1 & TS2)	80.55 ± 0.76	80.51 ± 0.60

Table 3 – Human Evaluation and Automatic Recognition Results on RBC.

6.1 Human Recognition Results

To evaluate the human recognition ability, a labelling experiment employing ten native German speaking and ten non-German speaking annotators was conducted. Each participant had to evaluate 450 randomly selected samples from RBC. The task was to decide whether the utterance is directed towards the human partner or to one of the technical systems. The selection of these samples followed a two-step selection process. First, all valid utterances are selected manually, skipping off-talk, laughter and all samples with a direct hint of the call partner. Afterwards from the remaining samples five utterances for each condition were randomly selected for each speaker. Utterances from the reading part are skipped. The labelling was conducted with ikannotate2 [10]. For the evaluation, TS1 and TS2 are combined into one class. Thus the overall UAR and UAP for this two-class problem was calculated.

The results (Table 3) reveal that neither the German speaking annotators nor the non-German speaking annotators could identify the addressee sufficiently. Although the speech content improves the labelling result by 7% absolute for the German annotators, this is still far below rates seen as acceptable. The result of this labelling were expected for our dataset and thus proves the experimental design.

6.2 Automatic Recognition Results

The automatic recognition experiments used the same two-class problem as for the human labelling: detecting whether an utterance is device-directed (TS1+TS2) or human-directed (H). Utterances from the reading part and off-talk, laughter utterances are skipped as before. The "emobase" feature set of OpenSMILE was utilized, as this set provides a good compromise between feature size and feature accuracy. Differences between the data samples of different speakers were then eliminated using standardisation [11]. As recognition system, a Support Vector Machine (SVM) with linear kernel and a cost factor of 1 was utilized with WEKA. The same system has already been successfully used for addressee detection archiving very good results ($>$ 86% UAR) [12]. A Leave-One-Speaker-Out (LOSO) validation was applied and the overall UAR and UAP as the average over all speakers was calculated. This strategy allows to revise the generalisation ability of the actual experiment and compare it with the human labelling.

The achieved results (Table 3) show that the characteristics between human-directed and device-directed speech (H vs. TS1 & TS2) are also hardly distinguishable for an automatic system. The recognition performance is similar to chance-level. This was expected based on the chosen experimental design. As validation that the recognition system is properly designed and suited for the data, additional experiments to distinguish read speech from device-directed speech (R vs. TS1 & TS2) were pursued. Thereby, the achieved recognition results are much better and comparable to similar experiments.

7 Conclusion

In this paper, a new dataset on natural human-human and human-computer conversation within a limited setting is presented. The focus of this dataset is on the retention of the speaker role.

Therefore, the conversation is designed in such a way that both the technical system and the human being have the same tasks and the same speech content when interacting with the participant. The conversations are realised via a simulated telephone and recorded as high quality speech. Additionally, the participants' socio-demographic characteristics were gathered. Furthermore, by using AttrakDiff as a quantifying measurement of the quality of the conversational partner the correct performance of the human agent and the correct implementation of the technical systems could be verified. The assumed result of this dataset, that the speech characteristics became indistinguishable could be verified by human annotations and the application of an automatic recognition system.

In total, 30 subjects took part in the experiment. The mean recording time per person is about 11 m, resulting in 5.5 hours of recorded material. The dataset is enriched with additional information gained from post-processing (off-talk, overlaps, laughter) and textual transkriptions of all utterances. As RBC aims to represent the same speaker role for human and technical systems, it allows to analyse these effect in a controlled natural conversation.

Finally, twelve participants (three male, 9 female) of this experiment took also part in the recordings for the Voice Assistant Conversation Corpus (VACC), see [9]. This corpus consists of conversations between one and two interaction partners with a commercial voice assistant system (Amazon's ALEXA) in two different conditions and scenarios. By using a formal and an informal scenario conducted either in the condition with a accompanying person or in the condition without a accompanying person, VACC expands the RBC dataset investigating different role models for participant, interaction partner and technical system by deliberately eliminated factors.

Availability

The Restaurant Booking Corpus is available for research purposes upon written request from the authors (ingo.siegert@ovgu.de).

References

[1] TSAI, T., A. STOLCKE, and M. SLANEY: *Multimodal addressee detection in multiparty dialogue systems*. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2314–2318. 2015.

[2] SHRIBERG, E., A. STOLCKE, D. HAKKANI-TÜR, and L. HECK: *Learning when to listen: Detecting system-addressed speech in human-human-computer dialog*. In *Proc. of the INTERSPEECH'12*, pp. 334–337. Portland, USA, 2012.

[3] BABA, N., H.-H. HUANG, and Y. I. NAKANO: *Addressee identification for human-human-agent multiparty conversations in different proxemics*. In *Proc. of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, pp. 6:1–6:6. 2012.

[4] LUNSFORD, R. and S. OVIATT: *Human perception of intended addressee during computer-assisted meetings*. In *Proc. of the 8th ACM ICMO*, pp. 20–27. Banff, Alberta, Canada, 2006.

[5] MALLIDI, S. H., R. MAAS, K. GOEHNER, A. RASTROW, S. MATSOUKAS, and B. HOFFMEISTER: *Device-directed utterance detection*. In *Proc. of the INTERSPEECH'18*, pp. 1225–1228. 2018.

[6] BIGOT, B., J. PINQUIER, I. FERRANÉ, and R. ANDRÉ-OBRECHT: *Looking for relevant features for speaker role recognition*. In *Proc. of the INTERSPEECH'10*, pp. 1057–1060. 2018.

[7] HASSENZAHL, M., M. BURMESTER, and F. KOLLER: *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität*. In G. SZWILLUS and J. ZIEGLER (eds.), *Mensch & Computer 2003*, vol. 57 of *Berichte des German Chapter of the ACM*, pp. 187–196. Vieweg+Teubner, Wiesbaden, Germany, 2003.

[8] RÖSNER, D., J. FROMMER, R. FRIESEN, M. HAASE, J. LANGE, and M. OTTO: *LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions*. In *Proc. of the 8th LREC*, pp. 96–103. Istanbul, Turkey, 2012.

[9] SIEGERT, I., J. KRÜGER, O. EGOROW, J. NIETZOLD, R. HEINEMANN, and A. LOTZ: *Voice assistant conversation corpus (vacc): A multi-scenario dataset for addressee detection in human-computer-interaction using amazon's alexa*. In H. KOISO and P. PAGGIO (eds.), *Prof of the 11th REC 2018*. Miyazaki Japan, 2018.

[10] SIEGERT, I. and A. WENDEMUTH: *ikannotate2 – a tool supporting annotation of emotions in audio-visual data*. In B. M. JÜRGEN TROUVAIN, INGMAR STEINER (ed.), *Elektronische Sprachsignalverarbeitung 2017. Tagungsband der 28. Konferenz*, vol. 86 of *Studientexte zur Sprachkommunikation*, pp. 17–24. TUDpress, Saarbrücken, Germany, 2017.

[11] BÖCK, R., O. EGOROW, I. SIEGERT, and A. WENDEMUTH: *Comparative study on normalisation in emotion recognition from speech*. In P. HORAIN, C. ACHARD, and M. MALLEM (eds.), *Intelligent Human Computer Interaction: Proceedings of the 9th International Conference, IHCI 2017, Evry, France, December 11-13, 2017*, pp. 189–201. Springer International Publishing, Cham, 2017.

[12] SIEGERT, I., T. SHURAN, and A. F. LOTZ: *Acoustic addressee-detection – analysing the impact of age, gender and technical knowledge*. In W. M. ANDRE BERTON, UDO HAIBER (ed.), *Elektronische Sprachsignalverarbeitung 2017. Tagungsband der 28. Konferenz*, vol. 90 of *Studientexte zur Sprachkommunikation*, pp. 113–120. TUDpress, Ulm, Germany, 2018.