Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon's ALEXA

Ingo Siegert[1], Julia Krüger[2], Olga Egorow[1], Jannik Nietzold[1], Ralph Heinemann[1], Alicia Lotz[1]
[1]Institute for Information and Communications Engineering, Cognitive Systems Group,
Otto-von-Guericke University, 39106 Magdeburg, Germany,
[2]Department of Psychosomatic Medicine and Psychotherapy, Otto von Guericke University, 39120 Magdeburg, Germany

Abstract

A new conversation corpus in the area of human-computer interaction is introduced. It consists of conversations between one and two interaction partners with a commercial voice assistant system (Amazon's ALEXA) in two different settings. The fundamental aim for building up this corpus is to investigate how humans address technical systems. Thereby, two different scenarios, a formal and an informal one, are designed. The scenarios are conducted by the participants alone and with an accompanying person. Furthermore, questionnaires are used to get a self-evaluation of the participants in terms of their experience of the interaction and their conscious changes in voice and behaviour while addressing a technical system. Additionally, also their experience with technical systems and the evaluation of the utilized commercial voice assistant is retrieved via questionnaires. The corpus consists of high-quality microphone recordings of 27 German speaking subjects, all students at the University Magdeburg.

Keywords: Corpus, Addressee Detection, Speech Assistant, Multi-Scenario, Multi-User, Speaking-Style

1. Introduction

Human-computer interaction (HCI) still receives increased attention, the market for commercial voice assistants is rapidly growing. 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 the addressee of a user's utterance.

Today, 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. Besides this unnaturalness in the interaction initiation, especially the currently preferred keyword method is error-prone. It can result in users' confusion, e.g., when the keyword has been said but no interaction with the system was intended by the user. Therefore, technical systems should be able to perform an addressee detection. Various aspects have already been investigated in this field of research, however most of the studies dealing with speech-enabled technical systems utilize datasets either with one human and a technical system (Lee et al., 2013), groups of humans (mostly two) interacting with each other and a technical system (Shriberg et al., 2012; Vinyals et al., 2012) or teams of robots and teams of humans (Dowding et al., 2006). These studies are mostly done using one specific scenario (Shriberg et al., 2013), just a few researchers analyse how people interact with technical systems in different scenarios (Lee et al., 2013). In these studies, the technical system is either a robot (Dowding et al., 2006; Katzenmaier et al., 2004), a research system (Shriberg et al., 2012; Vinyals et al., 2012), or a Wizard-of-Oz (WOZ)-experiment (van Turnhout et al., 2005). To the best of our knowledge, a current commercial system has not been used so far to examine addressee detection in HCI. Furthermore, previous research concentrated on analysing observable users' speech characteristics in the recorded data. The question whether users themselves recognise differences or even perhaps deliberately change their speaking style when interacting with a technical system has not been evaluated. The fact that users can be aware of speaking differently with technical systems than with humans has been described in (Frommer et al., 2017).

To address these issues, we designed the Voice Assistant Conversation Corpus (VACC) based on the interaction with a commercial voice assistant (Amazon's ALEXA). VACC further includes users' self-reports on their experiences during the interaction with the system, especially regarding their speaking style.

2. Study Design

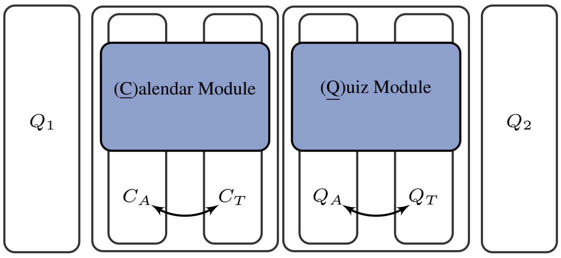

Figure 1: A sketch of the experimental procedure. Q_1 and Q_2 are the two questionnaire rounds. The order of the the scenarios (Calendar Module and Quiz Module) is fixed. A and T denote the experimental conditions (a)lone and (t)ogether with an confederate respectively. They can be interchanged.

The recorded corpus can be used for various analyses. However, we created it based on the following research questions: 1) How do humans speak with current speech-based assistant systems? 2) Which differences in the speaking style during the interaction with the technical system can be observed when they are alone or with an confederate? 3) Do they themselves recognise differences in the interaction? 4) Do the differences in the observed and/or reported interaction style differ between a formal and an informal interaction setting?

VACC consists of recordings of interaction experiments accompanied by various questionnaires presented before and after the experiment (see Sec. 4. and Fig. 1). The initial instruction of the experiment entailed information about the basic capabilities and the keyword-based addressing of ALEXA. Two experimental modules followed, arranged according to their complexity level. There were two conditions for each module, which were permuted for different participants. Thus, each experiment contained four "rounds". A round was finished when the aim was reached or broken up to avoid frustration if hardly any success could be realized. Although, the proscribed role of the confederate is distinct from that of ALEXA, we decided for such an attempt to gather natural interactions, as they would occur in daily life when using speech-enabled assistants.

Module 1 ("Calendar Module"): In this more formal interaction the participant had to make appointments with a the project partner. He/she was instructed that ALEXA could give information about the confederates' calendar for request including exemplary commands. In condition C_A ("without the confederate") the participant only got written information about his/her partners' available dates. In condition C_T ("with a confederate") the project partner entered the room and could give this information himself. Thus, appointments could now be made by the help of both, ALEXA and the project partner. The confederate was part of the research team and was instructed to interact only with the participant, not with ALEXA.

Module 2 ("Quiz Module"): In this more informal interaction the participant had to answer questions of a quiz (e.g., "How old was Albert Einstein?"). He/she was instructed that ALEXA was not able to give the full answer, but could offer support by solving partial steps to get it. In condition Q_A the participant had to answer the questions on his/her own. In condition Q_T the participant and the project partner built up a team supported by ALEXA. The confederate (here again only interacting with the participant, not with ALEXA) was instructed to make command proposals to the participant if frustration due to failures was imminent. The quiz in Q_T was more sophisticated than in Q_A to force cooperation between the two speakers and ALEXA.

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, see Fig. 2. 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). The participant sat on the sofa (right side of the photo in Fig. 2) and interacted with the voice assistant system, placed on the table in the middle. The informed second speaker – Jannik – present only in the two-person variants of each scenario, sat on the armchair (left side of the photo in Fig. 2). The positions were identical for each experiment to ensure comparability.

As voice assistant system, we used the Amazon ALEXA Echo Dot (2nd generation). We opted for a commercial system to allow a fully free interaction with a currently available system. We decided against developing a voice assistant system or using a WOZ(-technique), because we wanted to meet the abilities of current commercial voice assistant systems and did not want to pretend having further capabilities. For this dataset, we declined to do video recording as we wanted to use commercial systems as they are – they do not support video or gaze analyses. Furthermore, the awareness of video recording has the danger that participants behave differently and thus distorting our primary and only analysis modality, the speaking style.

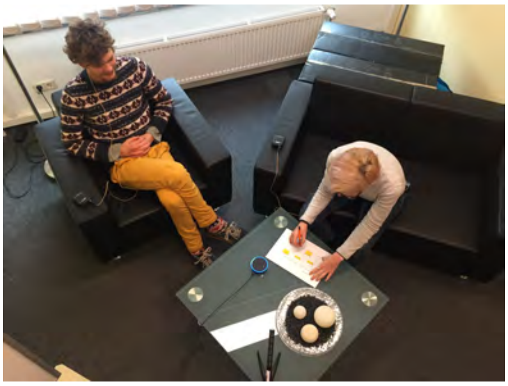

Figure 2: A snapshot of the data collection setup. The informed second speaker – Jannik – (left side) and the participant (right side) are sitting around a table, where the voice assistant (Amazon ALEXA Echo Dot) is located.

To conduct the recordings, we used two high-quality neckband microphones (Sennheiser HSP 2-EW-3) to capture the voices of the participant and the informed second speaker as well as one high-quality shotgun microphone (Sennheiser ME 66) to capture the overall acoustic and especially the output of the voice assistant. The recordings were stored in WAV-format with 44.1 kHz sample rate and 16 bit resolution.

4. Questionnaires

Several psychological questionnaires accompanied the experiment: *Before the experiment*, a short form of a self-defined questionnaire used in (Rösner et al., 2012) was utilized to obtain socio-demographic information as well as the participants' experience with technical systems (Q_1).

After the experiment, some self-defined computer-aided questionnaires were applied (Q_2). The first two of them focused on participants' experiences regarding a) the interaction with the voice assistant and the second speaker in general, b) possible changes in voice and speaking style while interacting with the voice assistant and the second speaker. According to the so-called principle of openness in examining subjective experiences (Hoffmann-Riem, 1980), the formulation of questions developed from higher openness and a free, non-restricted answering format (e.g., "If you compare your speaking style when interacting with ALEXA or with Jannik – did you recognise differences? If yes, please describe the differences when speaking with ALEXA!") to lower openness and highly structured answering formats (e.g., "Did your speed of speech differ when interacting with ALEXA or with Jannik? Yes or No? If yes, please describe the differences!"). This structure allows to examine the degree of participants' awareness of changes in the their voice and speaking style: If they already describe changes in some features (e.g. melody or speed) according to the open, initial questions, a higher degree of awareness is indicated than if they report about differences regarding these features only when they are explicitly asked for in the closed questions.

A third questionnaire focused on previous experiences with voice assistants. Furthermore, AttrakDiff (Hassenzahl et al., 2003) was used to supplement the open questions on self-evaluation of the interaction by a quantifying measurement of the quality of the interaction with the voice assistant (hedonic and pragmatic quality).

The answering of all questionnaires takes about 20 minutes.

5. Dataset Characteristics

VACC contains recordings of 27 German speaking participants, all students at the Otto von Guericke University Magdeburg. The sex is nearly balanced with 13 males and 14 females, the age ranges from 20 to 32 years (24.11 ± 3.32 y). The data collection took about 60 minutes (40 minutes recording and 20 minutes questionnaires) per participant. Table 1 summarises the dataset characteristics. The participants came from different study courses including computer science, engineering science and humanities. Thus, this dataset is not biased towards technophilic students. Regarding the experience with voice assistants, all participants had known Amazon ALEXA before. When asked about experience with ALEXA, only six participants specified that they had used ALEXA prior to this experiment. Five of them used ALEXA rarely for testing, only one participant specified that he uses ALEXA regularly – for playing music. Regarding experience with other voice assistants, additional ten participants indicated prior use. As voice assistants, they indicated Apple SIRI, GOOGLE NOW, or Microsoft CORTANA. Seven of them used these voice assistants seldom, just to try. Only three used them regularly, e.g. for programming a timer. In total, 18 out of 27 participants have prior experience with voice assistants. The nine participants not using any voice assistant before mistrusted the necessity of voice control and expressed data protection concerns when asked for reasons.

Subjects/Experiments	27
Sex	Male 13 / Female 14
Total Recorded Data	17 h 07 min
Experiment Duration	Mean: 31 min
Age	Mean 24 (Std: 3.32) Min: 20; Max: 32
Language	German
Annotation	Transcription, Speaker Events
Supplementary self-reports	evaluation of interaction, AttrakDiff, speaking style, experiences in interacting with voice assistants

Table 1: Dataset Characteristics

Furthermore, we analyzed the participants' technique affinity by asking how often the participant installed new software. We identified a clear distintion of 15 users who familiarise with new software at least once a quarter and 12 users familiarising with new software only 1-2x per year or less often. Interestingly, there is no significant difference between these groups in terms of the joy of computer work (JOY), the easement of work by the help of computers (EASE), weekly computer work (HOURS), or the age of the first use of computers (AGE), see Fig. 3. Comparing technique affinity and prior experience with voice assistants, seven out of nine participants having no prior experience with voice assistants also have less affinity to technology.

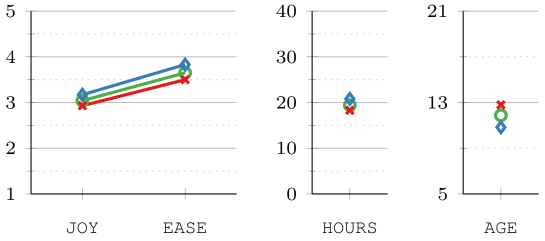

Figure 3: Evaluation of technique affinity regarding joy of computer work (JOY), the easement of work by the help of computers (EASE), weekly computer work (HOURS), or the age of the first use of computers (AGE) for all participants (—○—) as well as technology experienced (—◆—) and technology unexperienced (—✖—).

AttrakDiff is employed to understand how participants evaluate the usability and design of interactive products (Hassenzahl et al., 2003). It distinguishes four aspects (pragmatic quality (PQ), hedonic Quality (HQ) – including the sub-qualities identity (HQ-I) and stimulation (HQ-S), as well as attractiveness (ATT)). Regarding the experience with Amazon ALEXA, PQ, HQ-I, and ATT are perceived as neutral. Thus, it can be assumed that ALEXA provides useful features, it allows participants to identify themselves with ALEXA, and it has a kind of attractiveness. But all of these aspects leave room for improvements. Regarding HQ-S, a slightly negative assessment can be observed, showing that the support of the own needs was inappropriate. This can be justified by the difficulties of the calendar task where ALEXA has deficits. For all four aspects, no significant difference between technology experienced and technology unexperienced participants could be observed.

Furthermore, the participants filled out questionnaires dealing with their experiences of the interaction with ALEXA and the second speaker in general, regarding possible changes in their voice and speaking-style during the interaction with both as well as regarding their previous experiences with voice assistants (see Sec. 4., Q2). Besides the structured part of these questionnaires (e.g., “Have you ever worked with voice assistants aside from ALEXA? Yes or No?”), there were more open and unstructured ones, which had to be answered by using free text fields. For this, the participants used headwords and sentences to describe their experiences and evaluations. These texts made up a total number of 43307 characters. Regarding their speaking style in interacting with ALEXA compared to interacting with the second speaker, a first unsystematic analysis suggest, that participants are aware of differences e.g., in the length of sentences or the accentuation.

As stated in Sec. 2., the two scenarios (Calendar Module, Quiz Module) are either conducted alone or together with an informed speaker. Regarding the duration of the different sequences, it can be stated that for the calendar task, the duration of the first round is remarkably longer together with the informed speaker (submodule C_T). This can be attributed to the effect that in this case, the second speaker is frequently asked about the operation of ALEXA. Regarding the Quiz Module, the submodule condition Q_T (together with the informed speaker) took longer no matter of the order. This was expectable due to harder questions. Surprisingly, if Q_T was conducted after Q_A, it took remarkably longer in comparison to the case when Q_T was conducted before Q_A.

Although aimed at analysing the speaking styles for different scenarios in single and multi-user HCI among the same participants, this dataset can be used for a variety of applications Besides the mentioned characteristics, VACC is a fruitful resource for realistic and natural HCI. It contains different communication phenomena, for instance off-talk, overlaps, laughter, engagement, and emotional reactions. This additional information is currently being annotated using listening evaluations and will be distributed as EXMARaLDA transcripts (Schmidt, 2004).

6. Conclusion

In this paper, a new dataset on natural single- and multi-user HCI is proposed. The focus if this dataset is on the interaction with a commercial voice assistant system and the speaking style while addressing the system. Within the course of the recorded interactions, participants face two different situations with and without a second supportive speaker. Furthermore, the participants' socio-demographic characteristics, their self-assessment of the interaction, their speaking style, as well as a quantifying measurement of the quality of the interaction was gathered via questionnaires. Therefore, VACC captures both, the objectively measurable voice characteristics as well as their subjective assessment. Thus, it allows to correlate voice characteristics and subjective assessments in different situations. In total, 27 subjects took part in the experiment. The mean recording time per person is about 31 m, resulting in 17 hours of recorded material. The dataset will be enriched with additional information gained from post-processing (off-talk, overlaps, laughter).

As VACC aims to represent the same participants in two different scenarios with and without an accompanying speaker and furthermore represents a naturalistic HCI, it allows to analyse the problem of addressing the technical system in these different scenarios. Furthermore, this dataset enables comparisons of user behaviour in general in different scenarios for human-computer interaction and human-human interaction.

Availability

The Voice Assistant Conversation Corpus is available for research purposes upon written request from the authors.

Literature

Dowding, J., Clancey, W. J., and Graham, J. (2006). Are you talking to me? dialogue systems supporting mixed teams of humans and robots. In *AIAA Fall Symposium Annually Informed Performance: Integrating Machine Listing and Auditory Presentation in Robotic Systems*, Washington, DC; United States, October.

Frommer, J., Rösner, D., Andrich, R., Friesen, R., Günther, S., Haase, M., and Krüger, J., (2017). *LAST MINUTE: An Empirical Experiment in User-Companion Interaction and Its Evaluation*, pages 253–275. Springer International Publishing, Cham.

Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003*, volume 57 of *Berichte des German Chapter of the ACM*, pages 187–196. Vieweg+Teubner, Wiesbaden, Germany.

Hoffmann-Riem, C. (1980). Die Sozialforschung einer interpretativen Soziologie – Der Datengewinn. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 32:339–372.

Katzenmaier, M., Stiefelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. of the 6th ACM ICMI*, pages 144–151.

Lee, H., Stolcke, A., and Shriberg, E. (2013). Using out-of-domain data for lexical addressee detection in human-human-computer dialog. In *Proc. NAACL*, pages 221–229, Atlanta, USA, June.

Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., and Otto, M. (2012). LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In *Proc. of the 8th LREC*, pages 96–103, Istanbul, Turkey.

Schmidt, T. (2004). Transcribing and annotating spoken language with exmaralda. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*. ELRA. EN.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Heck, L. (2012). Learning when to listen: Detecting system-addressed speech in human-human-computer dialog. In *Proc. of the INTERSPEECH'12*, pages 334–337, Portland, USA, September.

Shriberg, E., Stolcke, A., and Ravuri, S. (2013). Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. In *Proc. of the INTERSPEECH'13*, pages 2559–2563, Lyon, France, August.

van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proc. of the 7th ACM ICMI*, pages 175–182.

Vinyals, O., Bohus, D., and Caruana, R. (2012). Learning speaker, addressee and overlap detection models from multimodal streams. In *Proc. of the 14th ACM ICMI*, pages 417–424.