Investigating How Conversational Search Agents Affect User's Behaviour, Performance and Search Experience

Mateusz Dubiel
University of Strathclyde
Glasgow, Scotland, UK
mateusz.dubiel@strath.ac.uk

Martin Halvey
University of Strathclyde
Glasgow, Scotland, UK
martin.halvey@strath.ac.uk

Leif Azzopardi
University of Strathclyde
Glasgow, Scotland, UK
leifos@acm.org

Sylvain Daronnat
University of Strathclyde
Glasgow, Scotland, UK
sylvain.daronnat@strath.ac.uk

ABSTRACT

Voice based search systems currently do not support natural conversational interaction. Consequently, people tend to limit their use of voice search to simple navigational tasks, as more complex search tasks require more sophisticated dialogue modelling. In this paper, we explore how people's search behaviour, performance and perception of usability change when interacting with a conversational search system which supports natural language interaction, as opposed to a voice based search system which does not. Previous research has demonstrated that a voice based search system's inability to preserve contextual information leads to user's dissatisfaction and discourages further usage. We conducted an interactive study comparing a simulated conversational search system against a slot-based voice search system. We hypothesise that the conversational system, with its ability to preserve and maintain conversational state, will lead to greater satisfaction. Our results indicate that participants prefer the conversational system over current voice based system, the conversational search system leads to significantly faster search task completion times, and significantly greater usability.

KEYWORDS

Conversational Search; Dialogue-Systems; Voice Interfaces

ACM Reference Format:
Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2018. Investigating How Conversational Search Agents Affect User's Behaviour, Performance and Search Experience. In *Proceedings of ACM SIGIR CAIR Workshop (CAIR'18).* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/xx.xxx/xxx_x

CAIR'18, July 2018, Ann Arbour, Michigan USA

ACM ISBN xxx-xxxx-xx-xxx/xx/xx…$15.00
https://doi.org/xx.xxx/xxx_x

1 INTRODUCTION

Systems that support voice search (illustrated in Figure 1) are becoming increasingly more popular and widespread [1]. However, due to technical limitations, current voice based systems often lack the capacity to maintain a conversation. Examples of some of the most popular systems that support voice search include: Microsoft Cortana, Google Now and Apple Siri. These state of the art systems provide limited: (*1*) state tracking (i.e. estimating user's goal in a conversation), (*2*) anaphora resolution (i.e. resolving references to earlier items in conversation), and (*3*) have problems with clarifying user's intent. All of these issues currently make interaction through voice both an unnatural and tedious activity in many cases (cf. [10, 26, 31, 33]). However, Cook et al. [8] and Lison and Meena [24] suggest that voice based interaction could be efficiently used for a range of well defined information retrieval tasks, if conversation with a system is sufficiently natural.

In this paper we aim to explore how a "conversational" search agent (**CSA**), an agent that preserves the context of the dialogue and maintains state, affects user's behaviour, performance and perceived system usability when compared to a current state of the art voice based search system (**VSS**) which does not. Based on [8, 24] we hypothesise that **CSA** will:

H1 : lead to faster task completion and better performance
H2 : be more usable and less cognitively taxing
H3 : lead to a more positive search experience

To examine our hypotheses we conducted a Wizard of Oz (WOZ) study [11] in which we simulate **CSA** and (**VSS**). **VSS** uses a slot-filling architecture, which requires users to provide information to fill in slots that are assigned with semantic class labels [30], simulating current state of the art approaches. While, the hypothesised **CSA** has context awareness and memory of past interactions. Contrary to **VSS** which requires intent to be provided upfront in a pre-defined way (i.e. providing slots as prompted by the system), the **CSA** allows for information to be provided in a free-form natural language over several turns. Participants were asked to complete simulated leisure search tasks using both **CSA** and (**VSS**). Performance, and perceived usability were measured using conversation logs as well as standard usability Brooke et al. [5] and task load questionnaires [14].

[1] For Example see https://preview.tinyurl.com/voice-tech-trend-18

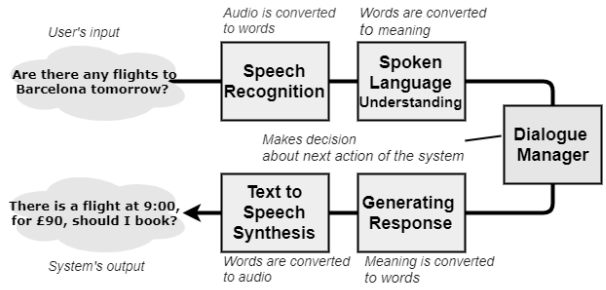

Figure 1: Architecture of a Voice Search System. In current study all of the elements of both simulated systems (i.e. VSS and CSA) with the exception of 'Text to Speech Synthesis' module are simulated by the Wizard (the lead researcher).

Our findings indicate that a more natural, human-like dialogue with users not only leads to the hypothesised benefits, but also results in distinct search behaviour and interaction. We show that the conversational search agent (**CSA**) is significantly more usable than the state of the art voice search system and offers a more positive user experience which is manifested in more positive sentiment and use of simple, and more understandable language.

2 RELATED WORK

Given the recent improvements in Automatic Speech Recognition (ASR) which is considered to have reached a 'human-like' performance [43] - and Synthetic Speech (SS) which is almost indistinguishable from human speech [21] [41] - the focus of research in the field of artificial intelligence has switched to Natural Language Understanding (NLU) and Dialogue Management (DM) components of voice search systems. However, regardless of the technological improvements, voice search systems still struggle with tasks that require multiple conversational turns which leads to users dissatisfaction and infrequent use (cf. [19][20],[18][26], [10]). The challenges of voice interaction have recently captured the attention of the IR community and have led to discussion on the nature of human-like conversational search and issues that need to be addressed before robust conversational search agents can be developed cf. [17].

Theory of Conversational Search: Radlinski and Craswell [32] defined a conversational search agent as a system for retrieving information, where there is a mixed initiative between the user and the agent, and the agent's actions are selected in response to the user's needs within the context of the conversation (considering short and long term knowledge). Radlinski and Craswell further outlined five properties a conversational search agent should possess - of which they highlighted that such a system needs to 'have memory to maintain state'. Given that current voice systems lack memory of past interactions and struggle to maintain conversational state (cf. [33]), in this paper we consider the importance of these properties within a conversational agent.

Trippas et al. [38] proposed a search framework for a spoken conversational search system and highlighted the cognitive challenges that a user faces when dealing with such an agent. In the more recent work, Trippas *et al* expand their framework by analysing behaviour of pairs of human participants engaging in collaborative search [16]. In contrast to that work, our study aims to analyse behaviour interacting with a simulated search system. In our study, via use of interactive search experiment, we explore how the workload associated with using the different systems impacts on task completion rates and participant's perception of system usability.

Interactive Conversational Search Studies: Another line of research on conversational search addresses the problem of designing an efficient and user-friendly search system. Lee et al. [22] focused on system's personality and discoverability of its affordances. In a Wizard of Oz study, Lee *et al* evaluated an early prototype of a conversational search system. Lee's experiment involved two participants (one who simulated the search system and the other that conducted the search). The focus of the study was on understanding how to ascertain the participants' needs, so that they can be applied in the development of such systems. Similarly, Trippas et al. [39] carried out a study aimed at investigating mixed initiative conversational behaviour for information search in an acoustic setting (e.g. voice only). The study involved 13 participants who completed a series of tasks with different levels of complexity. Trippas *et al* found that an increase in task complexity was linked to more queries being issued by participants.

Vtyurina et al. [40] explored what search experience would look like if a truly conversational system existed. In Vtyurina *et al*'s study, 21 participants who used a text-based interface, were asked to complete 3 different search scenarios; one scenario carried out with an existing commercial system, one with a human "expert", and another one with a "wizard" simulating the system. The results indicated that participants did not mind using a voice based search system as long as it was usable.

In the research presented in this paper, in a lab study, 22 participants complete a series of complex search tasks by interacting with two simulated search systems. To make the simulation more realistic the feedback is provided via synthetic voice.

Conversational System Design: Thomas et al. [37] explored the role of style in conversation seeking tasks. Their study, based on the analysis of recordings of interactions of strangers working on assigned tasks, shows that people tend to align their conversational style with intermediaries. In this study, we focus on the impact of different interaction styles i.e. free-form dialogue (offered by conversational agent) and constrained interaction (offered by slot-based voice system) on participant's behaviour.

Schulz et al. [36] proposed a state tracking model which enables a user to compare different results during conversational search. The proposed model assigned dialogue acts of new user utterance to the frames created during the dialogue - with each frame corresponding to an individual goal. The attempt was described as first step to create a memory-enhanced search system that can understand when users refer to older topic in the conversation, and provides user with accurate feedback thanks to understanding the context of their request. Related to the above research is the 'Frames' corpus that was created to study the role of memory in voice search tasks [1]. Our study expands upon this prior work by employing a simulated conversational search agent, where we evaluate how natural dialogue, in which memory of the conversation is used and the state of the need is preserved, affects search behaviour,

performance and perceived usability of the conversational search agent.

In this paper, we evaluate a system that presents participants with results in a free-form natural-language dialogue and measure to which extent its performance varies from the slot-based voice system.

3 METHOD

In order to explore how user's behaviour, performance and search experience change when interacting with a conversational search agent (CSA) when compared to a voice based search system (VSS) ; we conducted a user study where participants undertook four simulated leisure tasks (described later in the current section). A Wizard of Oz (WOZ) [11] setup was used to compare and contrast the agent and the state of the art system. This setup has been previously used in several related papers [39, 42] and is often used when evaluating natural language based interfaces (cf. [27]).

Our motivation for choosing the WOZ framework was based on its ability to simulate a human-like 'conversational system' - a technology which does not currently exist - and test its impact on participant's performance and behaviour. The merit of using WOZ framework is that it enables us to "acquire causal knowledge through controlled variation" cf. [9] - hence why we used WOZ setup to test our main hypothesis i.e.'conversational search is more usable and useful than currently available slot-based voice interfaces' in a controlled lab environment.

Experimental Setup: The experimental setup is diagrammatically presented in Figure 2. A within-subjects experiment was designed involving two systems: (*i*) a voice based search system (**VSS**) and (*ii*) a *simulated* conversational search agent (**CSA**), where participants completed two search tasks per system.

To reduce any priming effects, a Latin Square design[23] was applied to rotate the order in which search tasks and systems were presented to the participants. The experiment was conducted in a controlled lab environment (a quiet office, with a comfortable chair and desk). Each system was simulated using a mock-up setup, where the experimenter controlled the responses of the search systems. Dependent variables used in the study are: participants' perception of system usability (measured with SUS questionnaire [5]), work load (measured with NASA TLX questionnaire [14]), task completion times, number of conversational turns, length of conversational turns, number of repetition requests, and savings per booking made; while our independent variables were the two voice search systems, i.e. **VSS** and **CSA**. We chose our dependent variables to provide data to test our research hypotheses, i.e.(**H1**): Using conversational search agent lead to quicker task completion and better performance than the baseline voice system; (**H2**): Using conversational search agent is less cognitively taxing than the baseline system ; (**H3**): Using conversational search agent leads to more positive search experience then the baseline system.

Before the experiment, participants were briefed about the structure and purpose of the study, and provided informed consent to participate. Ethics approval for the experiment was granted by the Department of Computer and Information Sciences, University of Strathclyde (application no.611). During the experiment, participants were instructed to complete four search tasks (described in more detail in current section). Participants were informed that neither **VSS** nor **CSA** provide any visual feedback and that the interaction will be exclusively voice-based. The wizard (the experimenter in control of the systems) was using a computer to initiate search tasks and switch between the systems once tasks were completed. Participants were informed that their interactions with the system would be recorded.

Participants: Participants were recruited via social media, university mailing lists and flyers posted at a campus of two major UK universities (i.e. University of Strathclyde and University of Glasgow). The study inclusion criteria were that participants were at least years 18 years old, native English speakers and should have no hearing impairment. We decided to focus on native English speakers to limit the potential impact of language competence (native/non-native) on participants' ability to understand the synthetic voice with a local accent [2] that was used in the study. In total, 22 participants took part in the study. There were 9 females and 13 males. The age of participants ranged from 18 to 65 years (M = 28.67, SD = 9.9, Med = 28). The majority of participants (N = 15/22) has used a voice search system at least once before. On average participants who reported to use voice search have been doing so for 2 years (SD= 1.5). All participants were given a 10GBP shopping voucher for taking part in the experiment. The average completion time of the experiment was 35 minutes.

Simulated Leisure Tasks: We used simulated leisure tasks as they provide participants with information need that requires completing several steps in order to be satisfied. This is in contrast with simple factoid queries, frequently carried out by users [12], that can be resolved with a single query (e.g. 'What is the population of Chile?'). Our premise behind using simulated tasks was to provide a background context that our participants could easily relate to (in our study we apply Borlund's approach [4] into a leisure context to create our tasks).

Our leisure tasks required participants to find information about date of departure, destination and cost of a flight. Furthermore, our leisure tasks offer a possibility to measure task completion success (i.e. Was the the cheapest flight within the specified time booked?) and to assess participant's performance by looking at objective factors (i.e. task completion time, number of conversational turns, duration of conversational turn and number of words spoken.). This is important, since the focus of our investigation is on comparing two different modes of interaction, namely 'conversational search' which involves active participation of user and the system, and voice based search in which information needs to be provided upfront. (see Figure 4 for sample conversations).

Participants were asked to complete four casual leisure search scenarios - 2 for the **VSS** and 2 for the **CSA**. The simulated leisure tasks were based on the evaluation model proposed by Borlund [4]. The model relies on using plausible, real-life, search scenarios that participants can easily relate to. In our study, the task for each of the search scenarios was based on leisure activities, e.g. to book a series of one-way flights to travel around Europe. The place of departure was always fixed. In order to provide participants with search intent, participants were given each search scenario which contained four pieces of key information, namely: (*1*) place of departure, (*2*) date

[2]The synthetic voice Heather was provided by Cereproc Ltd. http://cereproc.com

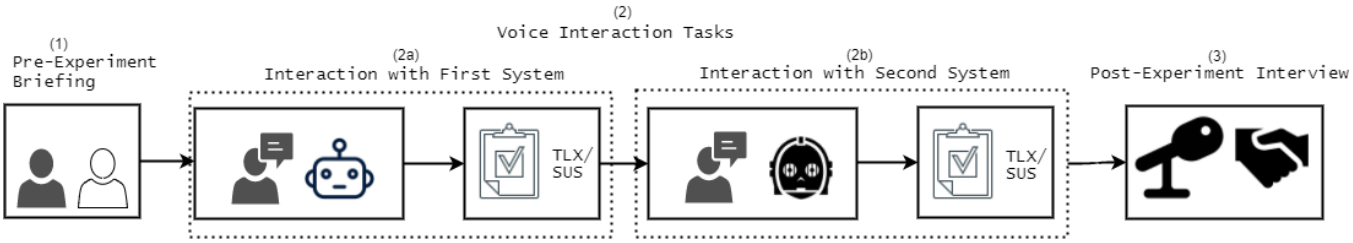

Figure 2: Illustration of Experimental Stages. The experiment consisted of three main stages: (1) Pre-experiment Briefing, (2) Voice Interaction Tasks and (3) Post-experiment interview. At Stage (1) participants were informed about what will happen during the experiment and asked to fill in a demographics questionnaire. Stage (2) consisted of two search sessions (one with system VSS and one with CSA); each session consisted of two search tasks and was followed by completing NASA TLX [14] and SUS [5] questionnaires. In order to reduce priming effects, for each participant the sequence of tasks was altered using a Latin Square Design [23]. Finally, during the last stage (3), participants were invited to take part in a post-experiment interview and were thanked for their participation.

of departure, (*3*) budget and (*4*) preferences. An example search scenario is shown in Figure 3.

Participants were instructed not to interrupt the system, and speak only when the system finished its turn. Every search task began with a pre-recorded prompt that welcomed participants and asked them to submit their search query. The task was considered completed once a participant found and reserved a flight that matched the search criteria provided to them in simulated search scenarios.

Search Systems: Two search systems were used in the study. The state of the art voice based search system (**VSS**), was based on a "slot-filling" architecture which represents the current state of the art. The **VSS** system was designed based on the design recommendations outlined in [28]. At the beginning of each interaction **VSS** system provided participants with a welcome message and presents them with its functionalities. Participants were asked to provide their search criteria, namely "destination airport", "date of travel", and "available budget". (Note: the departure airport was always fixed). The **VSS** system also provided an example to help them formulate their query: ''For example, you can say I'm travelling to London on the 2nd of December and my budget is 100 pounds". The conversational search agent (**CSA**) started with a brief greeting: Hello, how can I help you?' after which participant was supposed to provide their search query. It should be noted that for comparative purposes, interaction time was measured from the moment that participant started to speak after they were greeted by the system. Interaction with **CSA** system was not constrained in any way and participants were free to provide information in any order and the system would ask them follow up questions to clarify their intent. Whereas the **VSS** system could only process a query once all of the requested information was provided.

The prompts used in both of the systems were prepared in accordance with the guidelines outlined in [7]. The prompts were made to resemble a natural spoken discourse by, use of appropriate cohesive devices (pronouns and discourse markers), adhering to the principles of information structure (providing new information at the end of the utterance, and applying Grice's 'Cooperative Principle' [13] (making assumptions about inferences that users will draw from the prompts). During the interaction the Wizard (the lead researcher) played pre-recorded prompts using a GUI. For any unexpected participant responses a live-speech synthesis tool provided by Cereproc Ltd. [2] was used.

Interaction Design: Both **VSS** and **CSA** systems were operated by a Wizard (the lead researcher). The behaviour of **VSS** system was designed to resemble a system that is based on a slot-filling architecture which requires participant to provide the required information in a structured way so that their query can be processed. As opposed to **VSS**, **CSA** was designed to imitate an unscripted, natural human-human conversation. The characteristics of natural human-human conversation used to simulate **VSS** were based on features outlined in [3] i.e. 'ability to parse and non-fully-sentential grammar of spoken language', 'resolution of anaphora and ellipsis','keeping state of dialogue','ability to resolve hesitations, false starts and repairs' and 'ability to infer information from conversational context. We also implemented use of relative questions to clarify participant's intent (e.g. 'Do you have a budget in mind?') This is similar to the approach adopted in [6].

Figure 4 shows an example of interaction of participant and **VSS**. The system begins with introducing itself to the participant and informs them how to use it. Contrary to **VSS** , **CSA** does not follow any strict interaction script, instead, it initiates the conversation with a question.

Logs: Participants' interactions with each of the systems (i.e. **VSS** and **CSA**) were recorded, and then analysed to extract a series of objective measures including: task completion time, task completion success (i.e. whether a flight was booked meeting all the specified criteria), number of question-response pairs (i.e. conversational dyads), the average length of participant's turn (i.e. number of words spoken), number of options explored by participants (to determine if the cheapest flight was chosen), and vocabulary used by participants (to evaluate participant's sentiment towards the system).

Sentiment Analysis: "Vader" [15] sentiment analysis module of the Natural Language Tool Kit (NLTK) python library [25] was used to analyse participant's utterances . The result of the sentiment analysis is a set of polarities composed of "positivity", "negativity",

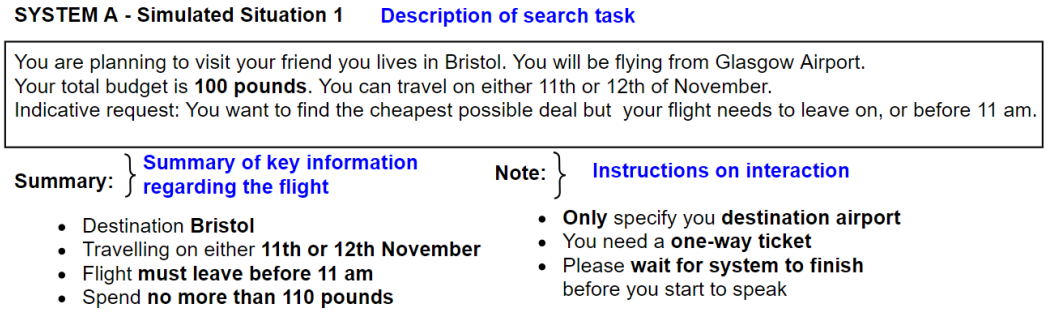

Figure 3: Simulated Search Situation 1. Participants are provided with a printed description of a simulated task situation which contains background information (top), a summary of search criteria (bottom left) and note on how to interact with the system (bottom right). The most important details are provided in bold. Participants can consult the description anytime during search task.

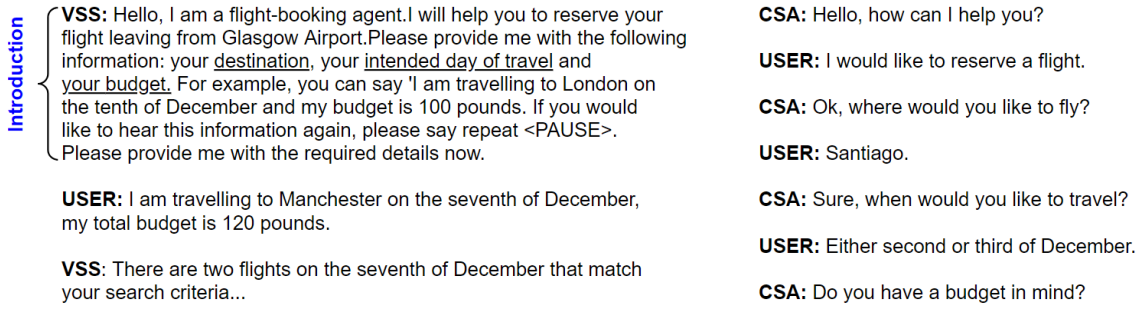

Figure 4: Transcript of recording of search tasks carried out using VSS (left) and CSA (right). VSS first 'introduces itself and explains how to make a reservation; participant is informed how to provide the information to fill the required slots (underlined). CSA allows for a free-form conversation without imposing rules on how participant should ask questions, or provide information. Note: Introduction time is not included in task completion time.

"neutral" and "compound" scores. The compound score: is a normalised, weighted composite score that is useful for giving a single measure of sentiment for a specific utterance.

4 RESULTS

When reporting results, unless stated otherwise, we used the Wilcoxon Signed Ranked test as most of our data was not normally distributed. The object of our comparison are the two systems **VSS** and **CSA**. Where appropriate we also report some differences between tasks within the system.

4.1 Search Behaviour and Performance

Search Behaviour: Table 1 provides high level summary statistics of the interactions with each of the systems for each search session. We report the number of dyads (adjacent conversational turns between user and the system), the total number of words uttered by a user, the number of words per turn and per user, along with the total turn duration in seconds, and turn duration per user in seconds.

We observed that participants on **VSS** interacted fewer times, with 155 dyads (i.e. pairs of conversational exchanges between the system and participant) when compared to the **CSA**, with 180 dyads in total. While there was no significant differences ($p = .0219$) in terms of dyads, the number of words spoken by participants was significantly different ($p < .001$), where on average, 18.87 words per turn were spoken on **VSS** while only 10.48 per turn on **CSA**. This suggests that the **CSA** led to more natural dialogue (i.e. shorter and more frequent conversational turns cf.[34]), while on **VSS** participants had to repeat themselves more and provide more context (as the system did not maintain conversational state).

Performance In terms of success and performance, Table 2 reports how participants spent the budgets allocated to them for each of the scenarios and how successful they were in choosing the best available flight. When it came to exploring best available flight options, the majority of participants made more savings with **CSA** while speaking less. When interacting with **CSA**, participants asked for fewer repetitions. For **VSS** we observed that each time that there was a request for repetition, participants did not proceed to explore different search options, effectively missing out on the cheapest flights.

Task Completion Times: On **VSS** the tasks completion time was 215 seconds on average, while on **CSA** the task completion time was approximately 117 seconds. The average task completion

Table 1: Dialogue statistics for both systems.'' - signifies p < .01**

	Dyads** (Total)	Participants' Words (Total)	Words Per Turn **	Words Per User	Total Users' Turn Duration (s)	Turn Duration Per User (s)
VSS First Attempt	84	1656	19.71	78.85	664	31.62
VSS Second Attempt	71	1269	17.87	60.43	499	23.76
VSS Total	155	2925	18.87	139.29	1163	55.38
CSA First Attempt	95	1067	11.23	50.8	399	19
CSA Second Attempt	85	820	9.64	39	258	12.28
CSA Total	180	1887	10.48	89.86	657	31.28

Table 2: Savings Made. Note: The maximum amount possible to spend is not equal to total budget available to participant. The best outcomes are provided in bold.

	Available Budget	Money Spent	Participants who Chose Cheapest Option	Possible Savings Missed
VSS 1	2310	2210	13/21	295[3]
VSS 2	2520	1800	9/21	120
VSS Total	4830	4010	22/42	315
CSA 1	2520	1710	20/21	30
CSA 2	2100	1695	18/21	15
CSA Total	4620	**3405**	**38/42**	**45**

time was found to be significantly different for both systems (p = .00012). When it comes to change in completion times between the first (T1) and the second task (T2) - for **VSS**, there was a significant drop in time taken to complete T2 from 136 to 117 seconds on average (p = .012), however for **CSA** while the time to complete fell from 64 to 52 seconds, there was no significant difference (p=.046).

As presented in Figure 5, participants took longer to familiarise themselves with the baseline system, VSS, and to explore its functions. Conversational system, CSA, however, was more straightforward to use and did not curtail participants' interaction. **VSS** also outperformed **CSA** in terms of average task accuracy.

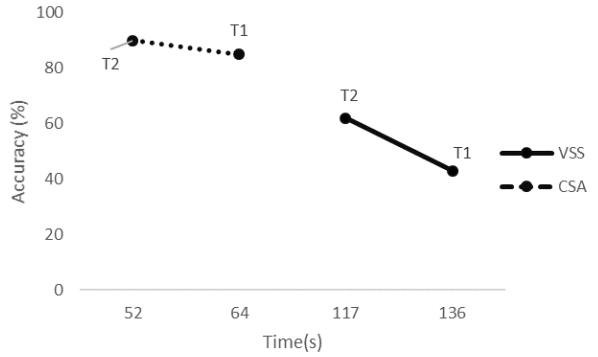

Figure 5: Task completion times and booking accuracy for VSS and CSA for first (T1) and second attempt (T2).

4.2 Workload and System Usability

To examine whether the workload associated with using the different systems we compared the results from the NASA TLX questionnaire in Table 3. The results show that the **CSA** led to lower cognitive workload than **VSS** for all six aspects under investigation (i.e. 'mental demand', 'physical demand', 'temporal demand', 'performance', 'effort' and 'frustration').

Overall, every participant found the **CSA** system less taxing to use than the **VSS** system. The median NASA TLX score for the baseline system, (**VSS**), was 29.5 as compared with 14 for the conversational agent (**CSA**). Note, the lower the score the less demanding in terms of workload the system is. There was a statistically significant difference (p < .001) in overall perception of both systems. On the level of individual dimensions, significant differences were found for "mental demand" (p = .0002), "performance"[4] (p = .025), the "effort" required (p = .00032), and "frustration" (p = .037).

It appears that participants on the **VSS** had to more closely monitor and keep track of their state and the state of the system. This made it increasingly difficult for participants to make direct comparisons between search results i.e. comparing search results by changing a given search aspect (e.g. 'Show me flights next day', 'Show me some cheaper flights' etc.). On the other hand, **CSA** system allowed participants to use search commands such as 'find the cheapest option' or 'I want a flight like this but it needs to leave earlier'. This 'memory' feature of **CSA** system significantly reduced the number of items that participants had to memorise, which we attribute to lower mental demand and lower effort.

Table 4 reports SUS scores for each system, where for **VSS** the median was 81.25 whereas for **CSA** the median was 92.5. This difference was statistically significant (p = .003). Note, the higher the score the more usable the system. The score achieved by CSA corresponds to approximately top 5th percentile of SUS scores [5], whereas VSS falls into the top 30th - 25th percentile bracket.

Taken together, the results from the NASA TLX and SUS suggest that participants found **CSA** more usable than the **VSS** .

4.3 Sentiment Towards the System

Table 5 presents information regarding participants' attitude towards the systems **VSS** and **CSA**. Sentiment analysis carried on transcripts of participants interactions with the systems, indicates that, on average, participants displayed more positive sentiment

[4]Performance score has been inverted for comparison purposes
[5]based on score interpretation guidelines provided by [35]

Table 3: Comparison of workload indexes for VSS and CSA systems. The scores are measured on a 0-100 scale, the lower score the better. '*' - indicates p < 0.05, '' - indicates p < 0.01. Note for the 'performance score' has been inverted for comparison purposes. The best results are provided in bold.**

		Mental**	Physical*	Temporal	Performance*	Effort**	Frustration*	Overall Score**
VSS	M/SD	32.72/19.8	10.45/11.84	22.5/16.46	22.5/22.35	37.27/24.77	29.55/25.21	30.73/17.43
	Med/IQR	27.5/32.5	5/6.25	17.5/11.25	15/25	30/37.5	22.50/26.25	29.5/18
CSA	M/SD	**15.68/13.21**	**6.81/3.29**	**18.18/13.93**	**12.27/10.99**	**15.9/13.42**	**16.59/17.68**	**17.27/10.95**
	Med/IQR	**10/21.25**	**5/5**	**12.5/21.25**	**5/10**	**12.5/15**	**10/13.75**	**14/12.5**

Table 4: SUS Scores, '' - signifies p < .01.**

	Mean	SD	Std Error	Median**	Q1	Q3	IQR
VSS	73.29	21.59	4.6	81.25	61.87	87.5	25.67
CSA	87.5	17.08	3.64	92.5	83.12	100	16.88

towards **CSA**. There is a statistically significant difference in positive sentiment between each of the systems both on the level of individual tasks, i.e. VSS1 vs. CSA1, (p = .001) VSS2 and CSA2 (p = .0001) as well as overall (p < .0001). For the above comparisons, Bonferroni adjusted *alpha* was .0083.

Table 5: Participants' Sentiment Towards the System. Note: Sentiment scores are ratios for proportions of participant utterances that fall into particular sentiment category, i.e. 'positive','negative' or 'neutral'. All of the categories sum up to 1. '' indicates p < .01. The best outcome is provided in bold.**

	Negative Sentiment M(Med)	Neutral Sentiment M(Med)	Positive Sentiment M(Med**)
VSS 1	1.08% (0%)	89.21%(89.73%)	9.7%(10.27%)
VSS 2	1.12% (0%)	90.2%(91.15%)	8.67% (8.85%)
VSS Total	1.1(0)%	89.71%(89.73%)	9.19%(10.27%)
CSA 1	6.34% (0%)	67.16%(76.1%)	**26.5%(23.9%)**
CSA 2	3.28% (0%)	65.53%(67.55%)	**31.19%(32.45%)**
CSA Total	4.77%(0%)	66.33%(76.1%)	**28.9%(23.89%)**

4.4 Post-study Interview

After completing interactive tasks, participants were asked to comment on their experience and indicate which system they preferred. The majority of participants (18/22) indicated **CSA** as their preferred system. Justifying their choice of **CSA** , participants mentioned that: it was natural to use, helped them to accomplish their tasks easily, and help to clarify their intent easily. For example, **P4** said: *'It required less listening and it did not speak that much so there was not that much information for me to remember.',* while **P10** commented: *'[It was] much more natural. As a result of that it just feels less stressful. You can achieve what you need to do in less time compared to the second system [**VSS**].'* , and **P15** pointed out that: *'It [**CSA**] was much more easy to use, I found it much more intuitive. I did not have to remember any details to get the best result'.*

Participants who chose **VSS** said that they preferred command and control task of interaction offered by that system and found it more predicable to to use. For instance, **P12** commented : *'I liked how it [**VSS**] was predictable. When I learned how to ask a question. I knew exactly what to do. Whereas, with the second one [**CSA**] I was worried that not understand what I was asking.'*

5 DISCUSSION

Based on experimental results obtained in our study we can make the following observations. Firstly, with respect to our **(H1)**: 'Using **CSA** leads to lower completion times and better performance than using (**VSS**)' - we observe (as discussed in Section 4.1) that the **CSA** did enable participants to: complete their search tasks quicker, choose the best available options, and effectively, save money. We regards to hypothesis - **(H2)**: **CSA** is more usable and less cognitively taxing then **VSS** - from the NASA TLX and SUS scores (presented in the Section 4.2) we saw that this was also the case. For our last hypothesis **(H3)**: Using **CSA** leads to more positive user experience than using **VSS** - we also observed from sentiment analysis, reported accounts and questionnaire data that participants had a more positive experience with the **CSA**.

In the results we observe that participants could improve their performance with **VSS** with more practice (see Figure 5). However, adaptability comes at the price of accuracy. We observe that participants' performance on **VSS** is inferior (slower and less accurate) than when interacting with conversational agent which allows them to perform search tasks quicker and more accurately. To examine these trade-offs and differences in behaviour longitudinal studies examining prolonged exposure to both the **VSS** and **CSA** styles of interaction are required. However until working **CSA** style interfaces are available, a WOZ study, like the one used in this paper, provides the best possibilities to examine these trade-offs.

Another observation is that while interacting with **CSA** our participants used more positive language and displayed more courteous attitude towards the agent then when speaking to the baseline system. We observed that while interacting with **CSA** participants thanked the agent more frequently and used more polite language (see Section 4.3 for the results of sentiment analysis). This may indicate that more natural conversational style (free-form language) encouraged our participants to approach the agent in a more personalised, human-like way. In our experiment we focused mainly on analysis of participants language at the level of syntax (types of words used). However, it would be interesting to see if positive attitude towards the agent is also reflected in changes in voice quality, i.e. its tone and pace. Such cues could be used as an implicit way to measure to the performance and satisfaction with the system.

During our experiment, we observed that a more dynamic conversational style, i.e. conversation with more frequent and shorter turns led to better accuracy and performance of participants (as discussed in Section 4.1). Drawing on principles of pragmatics in human-human communications where meeting is 'negotiated' as the effect of active interaction between interlocutors [29], we suggest that in order to improve search experience a conversational agent should be more inquisitive and focus on clarifying user's intent rather than merely retrieving the required key words.

6 CONCLUSIONS

Our findings provide empirical evidence that suggests that, overall, conversational search systems are more user friendly and efficient than the current state of the art-systems based on slot-filling. Our proposed conversational agent offers interaction experience that resembles human-human conversations: it is less constrained to use, leads to lower cognitive workload, encourages use of natural language, and incites positive sentiment. All these merits suggest, that, in the future, development of conversational agents should focus on making them more responsive and less reliant on a fixed interaction protocol to ensure the best user's experience when searching for information.

REFERENCES

[1] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems. *arXiv preprint arXiv:1704.00057* (2017).

[2] Matthew P Aylett and Christopher J Pidcock. 2007. The cerevoice characterful speech synthesiser sdk. In *IVA*. 413–414.

[3] Markus M Berg. 2014. *Modelling of Natural Dialogues in the Context of Speech-based Information and Control Systems.* AKA.

[4] Pia Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research. An International Electronic Journal* 8, 3 (2003).

[5] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[6] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems.. In *KDD*. 815–824.

[7] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice user interface design.* Chapter 10.

[8] Malcolm J Cook, Charles Cranmer, Robert Finan, Andy Sapeluk, and Carol-Ann Milton. 2017. 15 Memory load and task interference: hidden usability issues in speech interfaces. *EPCE: Volume 1: Transportation Systems* (2017).

[9] Benjamin R Cowan et al. 2014. Understanding speech and language interactions in HCI: The importance of theory-based human-human dialogue research. In *Designing speech and language interactions workshop, ACM conference on human factors in computing systems, CHI.*

[10] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services.* ACM, 43.

[11] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. *Knowledge-based systems* 6, 4 (1993), 258–266.

[12] Mateusz Dubiel. 2018. Towards Human-Like Conversational Search Systems. *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval* (2018).

[13] Grice et al. 1975. Logic and conversation. *1975* (1975), 41–58.

[14] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index). *Advances in psychology* 52 (1988), 139–183.

[15] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media.*

[16] Lawrence Cavedon Hideo Joho Johanne R. Trippas, Damiano Spina and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search. *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval* (2018).

[17] Hideo Joho, Lawrence Cavedon, Jaime Arguello, Milad Shokouhi, and Filip Radlinski. 2017. First International Workshop on Conversational Approaches to Information Retrieval (CAIR'17). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1423–1424.

[18] Julia Kiseleva and Maarten de Rijke. 2017. Evaluating Personal Assistants on Mobile devices. *arXiv preprint arXiv:1706.04524* (2017).

[19] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* ACM, 45–54.

[20] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval.* ACM, 121–130.

[21] Nicole Kobie. 2018. Google's new voice is as good as your own. (2018).

[22] Sang-su Lee, Jaemyung Lee, and Kun-pyo Lee. 2017. Designing Intelligent Assistant through User Participations. In *Proceedings of the 2017 Conference on Designing Interactive Systems.* ACM, 173–177.

[23] Everet Franklin Lindquist. 1953. Design and analysis of experiments in psychology and education. (1953).

[24] Pierre Lison and Raveesh Meena. 2014. Spoken dialogue systems: the new frontier in human-computer interaction. *XRDS: Crossroads, The ACM Magazine for Students* 21, 1 (2014), 46–51.

[25] Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1.* Association for Computational Linguistics, 63–70.

[26] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, 5286–5297.

[27] Michael McTear, Zoraida Callejas, and David Griol. 2016. Evaluating the conversational interface. In *The Conversational Interface.* Springer, 379–402.

[28] Ditte Mortensen. 2017. How to Design Voice User Interfaces. *Interaction Design Foundation* (2017).

[29] Yuko Nakahama, Andrea Tyler, and LEO LIER. 2001. Negotiation of meaning in conversational and information gap activities: A comparative discourse analysis. *TESOL quarterly* 35, 3 (2001), 377–405.

[30] Roberto Pieraccini, Evelyne Tzoukermann, Zakhar Gorelov, J-L Gauvain, Esther Levin, C-H Lee, and Jay G Wilpon. 1992. A speech understanding system based on statistical representation of semantics. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Vol. 1. IEEE, 193–196.

[31] Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2016. 'Do animals have accents?': talking with agents in multi-party conversation. (2016).

[32] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval.* ACM, 117–126.

[33] Stuart Reeves. 2017. Some conversational challenges of talking with machines. (2017).

[34] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction.* Elsevier, 7–55.

[35] Jeff Sauro. 2011. Measuring usability with the system usability scale (SUS). (2011).

[36] Hannes Schulz, Jeremie Zumer, Layla El Asri, and Shikhar Sharma. 2017. A Frame Tracking Model for Memory-Enhanced Dialogue Systems. *arXiv preprint arXiv:1706.01690* (2017).

[37] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and Alignment in Information-Seeking Conversation. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval.* ACM, 42–51.

[38] Johanne R Trippas. 2015. Spoken conversational search: Information retrieval over a speech-only communication channel. In *Proceedings of the 38th International ACM SIGIR Conference.* ACM, 1067–1067.

[39] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval.* ACM, 325–328.

[40] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* ACM, 2187–2193.

[41] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model. *arXiv preprint arXiv:1703.10135* (2017).

[42] Maria Wolters, Kallirroi Georgila, Johanna D Moore, Robert H Logie, Sarah E MacPherson, and Matthew Watson. 2009. Reducing working memory load in spoken dialogue systems. *Interacting with Computers* 21, 4 (2009), 276–287.

[43] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. The Microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on.* IEEE, 5255–5259.