Towards Human-Like Conversational Search Systems

Mateusz Dubiel (2nd Year PhD Student)
Supervised by Dr. Martin Halvey and Dr. Leif Azzopardi
Department of Computer and Information Sciences, The University of Strathclyde
18 Richmond Street, Glasgow
mateusz.dubiel@strath.ac.uk

ABSTRACT

Voice search is currently widely available on the majority of mobile devices via use of Virtual Personal Assistants. However, despite its general availability, the use of voice interaction remains sporadic and is limited to basic search tasks such as checking weather updates and looking up answers to factual queries. Present-day voice search systems struggle to use relevant contextual information to maintain conversational state, and lack conversational initiative needed to clarify user's intent, which hampers their usability and prevents users from engaging in more complex interaction activities. This research investigates the potential of a hypothesised interactive information retrieval system with human-like conversational abilities. To this end, we propose a series of usability studies that involve a working prototype of a conversational system that uses real time speech synthesis. The proposed experiments seek to provide empirical evidence that enabling a voice search system with human-like conversational abilities can lead to increased likelihood of its adoption.

KEYWORDS

Conversational Search; Interactive Voice Interfaces; Usability Testing

ACM Reference format:

Mateusz Dubiel. 2018. Towards Human-Like Conversational Search Systems. In *CHIIR'18: 2018 Conference on Human Information Interaction & Retrieval, New Brunswick, NJ, USA, March 11-15, 2018*, 3 pages.
DOI: 10.1145/3176349.3176360

1 MOTIVATION

The recent technological advances in speech technology have contributed to the proliferation of devices that support voice search. Currently, the performance of automatic speech recognition (ASR) is reported to be on a par with human performance [25; 26], while high quality synthetic voices generated with deep neural networks (WaveNet Model) can sound almost indistinguishable from natural speech [2; 18]. Another argument in favour of using speech for information retrieval is its speed (reportedly voice interaction is 3 times faster than texting [21]) and overlearned character [19].

However, regardless of technological improvements and potential to facilitate information retrieval, voice-based interaction with search systems remains sporadic [3; 4] and limited to simple functionalities such as looking for factual information or checking weather updates [7]. Recent evaluation studies of voice search systems [5; 11-13; 16; 24] highlight a number of problems that lead to users' dissatisfaction with voice interaction. Firstly, present day conversational systems struggle with preserving contextual meaning [11; 12; 24], which makes tasks that require several conversational turns either very cumbersome, or impossible to complete. Secondly, voice technology is perceived as unreliable as device does not understand user's intent and irrelevant returns [5; 17]. Finally, users tend to have unrealistic expectations regarding capabilities of voice search systems and lack awareness on how to communicate with them in order to obtain required results, which discourages frequent use of the system and limits its scope[13; 16]

Moore et al. suggested that by making voice search to resemble human-human dialogue it can become a viable alternative of text-based information retrieval [14; 15]. In a similar vein, Radlinski and Craswell [20] suggested a set of conditions that a search system needs to meet in order to be considered conversational. The two main features suggested are (1) 'Conversational Memory', which is required to maintain conversational state and (2) 'Mixed Initiative' that can be used to clarify user's intent and make necessary repairs during the conversation.. In recent years we have seen several attempts to create a conversational system with human-like capabilities [9; 23]. While implementation of deep learning methods resulted in improvement of voice search systems, their performance is still far from human conversational abilities.

The goal of my PhD is to investigate whether enabling voice search systems with human-like conversational abilities can improve their usability. The project is empirical in nature and seeks to provide data obtained via evaluation experiments with real users. My research is expected to advance the knowledge on voice search by:

- Helping to understand users requirements regarding conversational system
- Validating proposed theoretical framework for conversational search
- Providing evidence that systems enabled with conversational memory and initiative can lead to more frequent usage and more functions being explored.

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

ACM ISBN 978-1-4503-4925-3/18/03.
https://doi.org/10.1145/3176349.3176360

2 RESEARCH QUESTIONS

In my research, I seek to gain a better understanding of implications of enabling voice search systems with human-like conversational systems on their perceived usability. In particular, the questions that I seek to answer are:

- ***RQ1***: Do users expect their interaction with voice search system to reflect 'human-human dialogue'?
- ***RQ2***: Is voice search system with conversational memory perceived as more usable as compared with current state of the art system?
- ***RQ3:*** Can we improve user's satisfaction with the voice search system by enabling it with conversational initiative, (i.e. making it more inquisitive)?
- ***RQ4:*** Can real-life implementation of conversational system with human-like capabilities (memory and initiative) lead to improved usability and extend the scope of system's applications to tasks that go beyond checking weather and answering factual queries?

The anticipated contributions of my research are: firstly to elicit users' expectations towards conversational search system, and secondly, based on the obtained results, to propose a set of design guidelines for future conversational interfaces to make them more usable, and, in turn, to improve the prospect of their adoption in the future. Although the focus of the project is currently anticipated to be on contextual awareness and frequency of turn-taking in conversation, features of speech such as, speed and prosody may be included in the analysis (if time permits).

3 RESEARCH METHODOLOGY AND PROPOSED EXPERIMENTS

The methodology applied in my PhD project comprises of both qualitative and quantitative methods gathered from a survey, semi-structured interviews and usability studies. The project consists of 3 main parts: (1) 'Gathering Users' Requirements', (2) 'Voice Interaction Studies', and (3) 'Creating a Prototype of a Closed Domain Conversational Search System'. The goal of Part 1 (already completed) is to elicit users' requirements of conversational systems. Part 2 (currently in progress) is based on usability studies in which a hypothesised conversational system is tested by using a Wizard of Oz (WOZ) framework [6]. Finally, in Stage 3, a prototype of a closed domain conversational search system will be developed based on feedback obtained from Stages 1 and 2, and tested in a usability study.

3.1 Results so Far

3.1.1 Part 1 – Online Opinion Survey. The results of our opinion survey (N = 178) [7] have provided the answer to **RQ1**. The feedback provided by respondents, presented in Figure 1, indicates that the majority of people want their interaction with voice search system to be more human like. However, the opinions are divided when it comes to system's conversational initiative - with less than 50% of respondents who agreed that voice search system should ask more questions. In the answers provided to open questions, many respondents expressed need to for conversational system to have memory of their past interactions, and to ask follow up questions in order to clarify their intent. The insights obtained from the survey informed the design and the scope of 'Voice Interaction Studies' used in Part2 of the project.

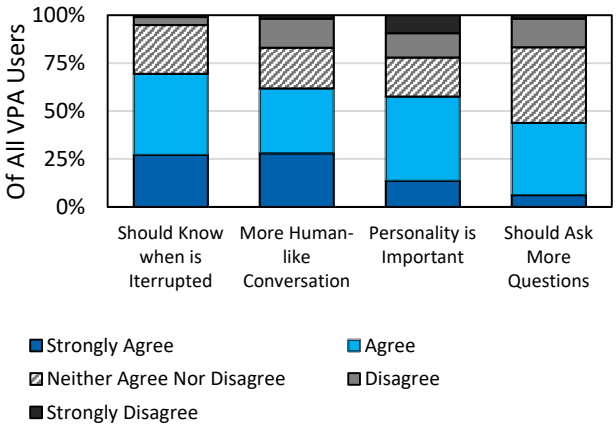

Figure 1: Users' expectations regarding performance of voice search systems. Note: blue highlights correspond to respondents who 'agree' or 'strongly agree' with presented statements. NOTE: 'VPA'= 'Virtual Personal Assistant' (Used as a synonym of voice search system)

3.1.2 Part 2 - Voice Interaction Study Conversational System with Memory Component. We carried out a lab based experiment (N = 12) [8] in which participants were asked to complete four search tasks on two voice search systems (two tasks per system). One of the systems was designed to reflect the performance of current state of the art voice search systems that are based on slot-filling architecture, while the other one was a hypothesised conversational system with memory component. Participants were given two questionnaires, i.e. NASA TLX [10] to assess their cognitive load for each of the system, and System Usability Questionnaire (SUS) [1] to evaluate systems' usability. The findings obtained from the experiment provided us with the answer to **RQ2,** indicating that our proposed conversational system was both more usable and less taxing to use than current state of the art system. The experimental results are provided in Table 1 and Table 2.

Table 1: Comparison of Cognitive Impact of Baseline System and Proposed Conversational System. The Scores are measures on a 0-100 scale, the lower score the better. * – indicates p <0.05, ** - indicates p < 0.01.

	Baseline (M/SD)	Conv. (M/SD)
TLX Score*	**23.26/11.53**	**13.19/10.38**
Mental Demand**	**29.11/6**	**14.21/3.68**
Effort*	30.8/5.9	14.6/3.5
Frustration	30.4/6.5	17.5/5.85
Temporal Demand	17/2.9	16.25/3.69
Performance[1]	16.9/5.8	9.1/2.67

[1] Values for performance have been inverted for comparability reasons

Table 2: Comparison of Usability of Baseline System and Proposed Conversational System. The Scores are measures on a 0-100 scale, the higher the score the better the performance. * – indicates p <0.05,

	Baseline (M/SD)	Conv. (M/SD)
SUS Score*	**77.91/21.31**	**89.37/16.17**

Note: The score of Baseline system falls between the 30th and 25th percentile of top SUS scores, while the score of Conv. system corresponds to the 5th percentile.

3.2 Planned Experiments

During the remaining part of my PhD (Years 2 and 3), I plan to carry out another lab-based experiment that will involve comparing usability of a baseline voice search system with a system enabled with conversational initiative. The experiment will conclude Part 2 of my PhD project. Once the data gathered in Part 2 has been analysed and conclusions drawn, I will proceed to the final stage of my project in which I will create a prototype of a conversational search system and evaluate it in a usability study.

The remaining research activities with brief descriptions and approximate timelines are provided below.

3.2.1 Part 2: Conversational System with Conversational Initiative. The experiment will follow the pattern explained in *Section 3.1.2.* (study designed in WOZ framework). The main focus of the study will be on creating a system that will use incremental dialogue approach, i.e. the system that will actively interact with participants without waiting for their conversational turn to be over, and likewise, the participants will be able to barge in at any point of the conversation. Real time reactive speech synthesis will be used to increase the naturalness of interaction [22]. The goal of the experiment will be to test if increased conversational initiative of the system can improve error recovery and ability to recover from misunderstandings during search task. The results obtained from the experiment are expected to provide the answer to ***RQ3***. The experimental part of the study is planned to run between November 2017 to May 2018 with the aim to write a journal paper by June 2018.

3.2.2 Part 3: Prototype of Human-Like Conversational System. Finally, having investigated both memory component and turn taking aspects of conversational system, I will move on to develop a prototype of a conversational system. The system will be designed based on feedback obtained from both experiments carried out in Part2 and then evaluated in a usability study. Prototyping is expected to be the most time consuming part of my project that is expected to run from summer 2018 to autumn 2019. During that time I will use machine learning techniques to analyse the data gathered in Part 2 of the project and use state of art spoken language understanding, and dialogue management modules (using neural network models e.g. Google Speech API) to create the prototype. The results of prototype evaluation are expected to provide answer to **RQ4,** and conclude my PhD project.

4 REFERENCES

[1] Brook, J., 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry 189*, 194, 4-7.

[2] CABRAL, J.P., COWAN, B.R., ZIBREK, K., and MCDONNELL, R., 2017. The Influence of Synthetic Voice on the Evaluation of a Virtual Character. *Proc. Interspeech 2017*, 229-233.

[3] CAROLINA, M., 2017. Voice Assistant Anyone? Yes please, but not in public! In *Creative Strategies*.

[4] COWAN, B.R., 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*.

[5] COWAN, B.R., PANTIDI, N., COYLE, D., MORRISSEY, K., CLARKE, P., AL-SHEHRI, S., EARLEY, D., and BANDEIRA, N., 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.

[6] DAHLBÄCK, N., JÖNSSON, A., and AHRENBERG, L., 1993. Wizard of Oz studies—why and how. *Knowledge-based systems 6*, 4, 258-266.

[7] DUBIEL, M., HALVEY, M., and AZZOPARDI, L., 2018. What Stops People from Speaking to Machines? A Survey Investigating Barriers to Adoption of Virtual Personal Assistants *(Under Review)* (2018).

[8] DUBIEL, M., HALVEY, M., and AZZOPARDI, L., and DARONNAT, S., 2018. Towards Conversational Search Agents: Investigating how natural language dialogue affects search behaviour, performance and satisfaction. *(Under Review)* (2018).

[9] FUJITA, T., BAI, W., and QUAN, C., 2017. Long short-term memory networks for automatic generation of conversations. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017 18th IEEE/ACIS Conference on* IEEE, 483-487.

[10] HART, S.G. and STAVELAND, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology 52*, 139-183.

[11] KISELEVA, J., WILLIAMS, K., HASSAN AWADALLAH, A., CROOK, A.C., ZITOUNI, I., and ANASTASAKOS, T., 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference* ACM, 45-54.

[12] KISELEVA, J., WILLIAMS, K., JIANG, J., HASSAN AWADALLAH, A., CROOK, A.C., ZITOUNI, I., and ANASTASAKOS, T., 2016. Understanding User Satisfaction with Intelligent Assistants, 121-130. DOI= http://dx.doi.org/10.1145/2854946.2854961.

[13] LUGER, E. and SELLEN, A., 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.

[14] MOORE, R.K., 2015. From talking and listening robots to intelligent communicative machines. *Robots that Talk and Listen–Technology and Social Impact*, 317-336.

[15] MOORE, R.K., 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots* Springer, 281-291.

[16] MOORE, R.K., LI, H., and LIAO, S.-H., 2016. Progress and Prospects for Spoken Language Technology: What Ordinary People Think. In *INTERSPEECH*, 3007-3011.

[17] MOORE, R.K. and MARXER, R., 2016. Progress and Prospects for Spoken Language Technology: Results from Four Sexennial Surveys. In *INTERSPEECH*, 3012-3016.

[18] OORD, A.V.D., DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., KALCHBRENNER, N., SENIOR, A., and KAVUKCUOGLU, K., 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[19] PIERACCINI, R. and RABINER, L., 2012. *The voice in the machine: building computers that understand speech*. MIT Press.

[20] RADLINSKI, F. and CRASWELL, N., 2017. A theoretical framework for conversational search CHIIR.

[21] RUAN, S., WOBBROCK, J.O., LIOU, K., NG, A., and LANDAY, J., 2016. Speech Is 3x Faster than Typing for English and Mandarin Text Entry on Mobile Devices. *arXiv preprint arXiv:1608.07323*.

[22] WESTER, M., BRAUDE, D.A., POTARD, B., AYLETT, M.P., and SHAW, F., 2017. Real-time reactive speech synthesis: incorporating interruptions. *Proc. Interspeech 2017*, 3996-4000

[23] WESTON, J.E., SZLAM, A.D., FERGUS, R.D., and SUKHBAATAR, S., 2017. End-to-end memory networks Google Patents.

[24] WILLIAMS, K., KISELEVA, J., CROOK, A.C., ZITOUNI, I., AWADALLAH, A.H., and KHABSA, M., 2016. Is This Your Final Answer?, 889-892. DOI= http://dx.doi.org/10.1145/2911451.2914736.

[25] XIONG, W., DROPPO, J., HUANG, X., SEIDE, F., SELTZER, M., STOLCKE, A., YU, D., and ZWEIG, G., 2016. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

[26] XIONG, W., DROPPO, J., HUANG, X., SEIDE, F., SELTZER, M., STOLCKE, A., YU, D., and ZWEIG, G., 2017. The Microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* IEEE, 5255-5259.