Spoken Conversational Search: Speech-only Interactive Information Retrieval

Johanne R. Trippas
1st year PhD student
School of Computer Science and Information Technology
RMIT University, Melbourne
Johanne.trippas@rmit.edu.au
Supervisors: L. Cavedon, M. Sanderson, D. Spina

ABSTRACT

This research investigates a new interface paradigm for interactive information retrieval (IIR) which forces us to shift away from the classic "ten blue links" search engine results page. Instead we investigate how to present search results through a conversation over a speech-only communication channel where no screen is available. Accessing information via speech is becoming increasingly pervasive and is already important for people with a visual impairment. However, presenting search results over a speech-only communication channel is challenging due to cognitive limitations and the transient nature of audio. Studies have indicated that the implementation of speech recognizers and screen readers must be carefully designed and cannot simply be added to an existing system. Therefore the aim of this research is to develop a new interaction framework for effective and efficient IIR over a speech-only channel: a *Spoken Conversational Search System* (SCSS) which provides a conversational approach to defining user information needs, presenting results and enabling search reformulations. In order to contribute to a more efficient and effective search experience when using a SCSS, we intend for a tighter integration between document search and conversational processes.

Keywords

Spoken Retrieval; Search Result Summarization; Conversational Search; Voice Search; Interactive Information Retrieval

1. MOTIVATION

Speech-based applications are becoming more prominent and are increasingly accepted among the wider population. Google documented in 2010 that 25% of queries on Android devices were submitted by voice[1]. Even though much research has been conducted into supporting search by voice input, only a few studies have focused on the presentation of information where no display is used [17]. Speech output is easily understood for factoid-style queries (e.g, "Who is the president of the United States?") by systems such as Siri, Google Now or Cortana. However, when users seek answers to non-factoid style queries, the system falls back on displaying a result list on screen such as in a multimodal system [18]. Nevertheless, there are many different scenarios where a speech-only user interface is preferred, such as when operating machinery [6, 7]; when no screen or keyboard is available [24]; when users are on the move [16, 21]; or when using wearable devices [5].

More importantly, some user groups such as visually impaired users [17], people with dyslexia, or people with limited literacy skills are disadvantaged in accessing information on screen. Visually impaired users have been using screen reader software for many years, however this software is still often difficult and frustrating to use because the content is mainly expressed visually and is only accessible via a mouse.

Listening to search results over audio is very taxing for users since audio is a temporal medium and does not leave any traces to which to user can refer [14, 23]. Since speech is a linear medium, it is also challenging to present complex structures. Thus, it is difficult to convey large amounts of information via audio without overloading the user's short-term memory [14, 18, 21]. It has also been shown that word frequency and speech rate have an effect on short-term serial recall [11]. In particular, we seek a better understanding of how to present search results over audio while not overwhelming the users with information [21], nor leaving users uncertain as to whether what they have covered the information space [22]. However, we believe that conveying information through an interactive channel will alleviate some complexities which are associated with speech and will allow users to find information in an efficient and effective manner.

The proposed research will advance the knowledge base by:

- Developing new interaction models for IIR over a voice-only communication channel.
- Determining new methods for providing summary-based result-presentation for unstructured documents.
- Providing an understanding of which strategies and techniques for SCSS are best for users.

Thus, the proposed research will transform search over a voice-only communication channel by using an inherently interactive and conversational search experience.

2. RESEARCH QUESTIONS

The overall aim of the proposed research is to investigate a new framework and interaction model for efficient and effective information retrieval over a speech-only communication channel: a *Spo-*

[1] https://www.youtube.com/watch?v=DtMfdNeGXgM

CHIIR'16, March 13–17, 2016, Chapel Hill, North Carolina, USA.

ACM ISBN 978-1-4503-3751-9/16/03.

DOI: http://dx.doi.org/10.1145/2854946.2854952

ken Conversational Search System (SCSS) which provides a conversational approach to determining user information needs, presenting results and enabling search reformulation.

Researchers have defined the information-seeking process in many different ways [15]. Nevertheless, there are several distinct stages to a SCSS which follow the information-seeking behaviours: establishing the intent of a user's initial query; allowing the user to conversationally interact with search results; and interpreting query reformulations [17].

The proposed research seeks to answer the following research questions:

- Are there any existing **interaction models** which fit the SCSS?
- What are effective techniques to **present query results** using audio so users can efficiently locate items, determine their relevance, provide feedback, and refine their query?
- What are effective techniques to **structure the conversation** interaction to minimize cognitive load in order to support the user in the information seeking processes with search engines?
- Are there **differences** between visually impaired and sighted users in the interaction with a SCSS?

3. RESEARCH METHODOLOGY AND PROPOSED EXPERIMENTS

This section presents a brief overview of the results of experiments to date. It also discusses the research methodology and experiments for the proposed research.

3.1 Results so Far

A short paper was published at SIGIR 2015 with preliminary results of a study which analyzed the result description (information abstract on a Search Engine Result Page) for query results over a speech-only communication channel [20].

The impact of the search result summary length in speech-based web search was investigated and these results were also compared to a text baseline. A crowdsourcing platform was used as a data collection tool. The χ^2 goodness-of-fit test [13] was used to assess whether changing the result summary had an effect on user preference. Table 1 shows that users preferred full text summaries rather than their truncated counterpart. For example, 57% of the users would recommend full text summaries to a friend.

The data showed that while users preferred longer and more informative summaries for text presentation, for single-faceted queries users preferred shortened summaries for audio. For multi-faceted summaries for audio, user preferences were not as clear, suggesting that more sophisticated techniques are required to handle these complex queries.

The above experiment allowed us to gather information about user preferences and query judgements. However, as mentioned in Section 1, audio output can be a very taxing medium for users, so we propose to extend this experiment to capture workload related factors. Capturing the workload would allow us to better understand and design the presentation of search summary results over audio. One of the tools we plan to use is the NASA TASK Load Index (NASA-TLX) [10].

A work in progress paper was accepted and presented at the First International Workshop on Novel Web Search Interfaces and Systems (NWSearch'15) co-located with CIKM 2015. The paper discusses the future directions regarding a novel spoken interface targeted at search result presentation, query intent detection, and interaction patterns for audio search [19].

Table 1: Exit questionnaire results for preferences in the search engine result summaries.

Exit Question	Text		Audio	
	Summary	Truncated	Summary	Truncated
Recommend to a friend	572▲(57%)	434 (43%)	529 (51%)	512 (49%)
Easier to find relevant result	548▲(54%)	458 (46%)	514 (49%)	527 (51%)
Gave better result	576▲(57%)	430 (43%)	539 (52%)	502 (48%)
More efficient to use	529 (53%)	477 (47%)	499 (48%)	542 (52%)

▲ indicates statistical significance with $p < .01$.

3.2 Research Methodology and Experiments

This research project aims to develop new interaction models for IIR over a speech-only communication channel. It has been suggested that investigating usability is necessary for interactive speech development and that one should not translate interactive speech theories literally into practice [4]. For this reason, the development of interaction models during my PhD will be an iterative process using mixed-methods. We will use an observational experiment to form our hypothesis which will then be tested in a Wizard of Oz experiment as explained in Section 3.2.2. In parallel we will analyze interaction logs as explained in Section 3.2.1.

The methodologies and experiments described in this section will be carried out for both visually impaired and sighted users to understand if there are any interaction differences between these two user groups.

3.2.1 Interaction Log Analysis

Interaction logs do not record the user's intention and motivation. Nevertheless, the interaction logs have the advantage that they capture the "real search process" of the user [12].

We have access to interaction logs from a speech-only interaction system which is used by people with visual impairments. The interaction logs are provided by an industry partner[2] who developed a system in which users can search for audio books, podcasts, or news. All the interactions are performed solely over speech.

For the first iteration of the interaction model of a speech-only interaction system, we are analyzing the interaction history with a focus on the linguistic history which records the surface language such as speech acts [4]. Our aim is to decode the user and system utterances to better understand the interaction control of the system. This first iteration of the interaction model will be made for frequent users and non-frequent users. We are also analyzing the logs to investigate data such as query length, query terms, session times, speed of speech, query categories, query reformulation, and GPS coordinates.

It has been suggested that voice system usage behaviour differs in unfamiliar environments. For example, users might change their behaviour because of privacy concerns or social appropriateness [2]. Thus, the GPS coordinates in the interaction logs might provide insight into how people behave when they are not in their home environment.

3.2.2 Observational and Wizard of Oz Experiments

We address a broader and less restrained way of speech than spoken dialogue systems which allows more complex information to

[2] http://www.realthing.com.au

be conveyed. Therefore it is important to understand and predict the interactions between the user and the system [1]. It is suggested that Wizard of Oz (WOZ) methodologies are relevant for iterative development and evaluation of interactive interfaces [4]. However, WOZ methodologies can only be considered if certain pre-conditions are met [9]. One of these pre-conditions will be challenging for our experiment, namely that it must be possible to simulate the future system. Thus, for the first iteration of user studies we will use an observational methodology to discover what kind of language models users expect to use and to hear from a SCSS and how a search might be conducted in a fully audio setting instead of a WOZ experiment.

In the initial observational experiment, one participant will be the SCSS and the other will be the user. This observational setup will have resemblances to collaborative search. The data from the observational experiment allows us to develop a new iteration of the interaction model for spoken conversation search as explained in Section 3.2.1. Hence, the observational experiment will model user interaction to understand their linguistic behaviour [8] and dialogue patterns [4].

WOZ experiments will be conducted once we develop a better understanding of the users' search behaviour and language model and are able to simulate the SCSS. In a WOZ experiment, the human is simulating the system (the wizard). The wizard simulates the spoken interaction with the user who thinks they are interacting with a real system. The WOZ experiments allow us to conduct performance evaluations while the system is being simulated. These simulations may include different restrictions on whether a automatic speech recognizer is used or a text-to-speech module is used. The following parameters can be collected during a WOZ: average utterances used per turn; average number of turns per for wizard and participants together; and vocabulary to inform the next iteration of the interaction model for speech.

When participants are involved in a user study, they will first answer a pre-test questionnaire to obtain user profile information. Then the empirical user test will be conducted to gain knowledge of how participants would interact with the application and to discover problems. The user study will be recorded to allow for data to be analyzed after the test. Once the user study is finished, the participants will complete a post-test questionnaire and a Likert-scale questionnaire [21] and the tester will conduct a semi-structured interview. These measures will evaluate user satisfaction, usefulness of the system and naturalness of the system specifically for the presentation of query results and the structure of the conversation. The analyzed information of the user study will lead to another iteration of an improved interaction model for speech. This allows us to understand how to structure the conversation interaction and how the query results should be presented.

3.2.3 New Interaction Models for Speech

The findings presented in our earlier work emphasize the importance of developing techniques that can both predict when a query needs to be refined and provide suggestions for refinement to a conversational interface [20]. We will develop an interaction model which uncovers the linguistic structure such as speech acts, references and discourse segments [4]. Demands placed on a user could lead to reduced performance and could translate into a slower response and increased errors [3].

To our knowledge, no existing interaction models fit the SCSS and hence we started identifying interactions in an existing log (explained in Section 3.2.1) to either adapt existing models with these findings or develop a new model.

4. REFERENCES

[1] J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent. Toward conversational human-computer interaction. *AI magazine*, 22(4):27, 2001.

[2] S. Azenkot and N. B. Lee. Exploring the use of speech input by blind people on mobile devices. In *Proc. SIGACCESS*, 2013.

[3] C. Baber, B. Mellor, R. Graham, J. M. Noyes, and C. Tunley. Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech Communication*, 20(1):37–53, 1996.

[4] N. O. Bernsen, H. Dybkjær, and L. Dybkjær. *Designing interactive speech systems: From first ideas to user testing*. Springer, 1998.

[5] E. Chang, F. Seide, H. M. Meng, C. Zhuoran, S. Yu, and L. Yuk-Chi. A system for spoken query information retrieval on mobile devices. *Speech and Audio Processing, IEEE Transactions on*, 10(8):531–541, 2002.

[6] V. Demberg and A. Sayeed. Linguistic cognitive load: implications for automotive uis. In *Proc. of AutomotiveUI 2011*, 2011.

[7] V. Demberg, A. Winterboer, and J. D. Moore. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539, 2011.

[8] L. Dybkjaer, N. O. Bernsen, and W. Minker. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1):33–54, 2004.

[9] N. M. Fraser and G. Gilbert. Simulating speech systems. *Computer Speech Language*, 5(1):81 – 99, 1991.

[10] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[11] C. Hulme, S. Roodenrys, R. Schweickert, G. D. Brown, S. Martin, and G. Stuart. Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(5):1217–1232, 1997.

[12] B. J. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Inf. Process. Manag.*, 42(1):248–263, 2006.

[13] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224, 2009.

[14] J. Lai and N. Yankelovich. Speech interface design. In *Encyclopedia of Language & Linguistics (Second Edition)*, pages 764–770. Elsevier, 2006.

[15] G. Marchionini and R. White. Find what you need, understand what you find. *International Journal of Human-Computer Interaction*, 23 (3):205–237, 2007.

[16] L. J. Najjar, J. J. Ockerman, and J. C. Thompson. User interface design guidelines for speech recognition applications. In *Proc. of IEEE VRAIS '98*, 1998.

[17] N. G. Sahib, D. Al Thani, A. Tombros, and T. Stockman. Accessible information seeking. *Proc. of Digital Futures*, 12, 2012.

[18] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. "your word is my command": Google search by voice: A case study. In *Advances in Speech Recognition*, pages 61–90. Springer US, 2010.

[19] J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Results presentation methods for a spoken conversational search system. In *First International Workshop on Novel Web Search Interfaces and Systems (NWSearch'15)*, 2015.

[20] J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Towards Understanding the Impact of Length in Web Search Result Summaries over a Speech-only Communication Channel. In *Proc. of SIGIR'15*, pages 991–994, 2015.

[21] M. Turunen, J. Hakulinen, N. Rajput, and A. A. Nanavati. *Evaluation of Mobile and Pervasive Speech Applications*, pages 219–262. 2012.

[22] S. Varges, F. Weng, and H. Pon-Barry. Interactive question answering and constraint relaxation in spoken dialogue systems, 2006.

[23] N. Yankelovich and J. Lai. *Designing speech user interfaces*. 1998.

[24] N. Yankelovich, G.-A. Levow, and M. Marx. Designing speechacts: Issues in speech user interfaces. In *Proc. of the SIGCHI'95*, pages 369–376, 1995.