Crowdsourcing User Preferences and Query Judgments for Speech-Only Search

Johanne R. Trippas
RMIT University
johanne.trippas@rmit.edu.au

Damiano Spina
RMIT University
damiano.spina@rmit.edu.au

Lawrence Cavedon
RMIT University
lawrence.cavedon@rmit.edu.au

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

ABSTRACT

Presenting search results over a speech-only channel (e.g., searching in devices such as Amazon Echo or Google Home) involves a number of challenges due to the transient nature of audio. We present two case studies which used a crowdsourcing methodology to gather user preferences and query judgments. We believe the design used in our previous experiments can inform the design of similar crowdsourcing tasks involving audio.

CCS CONCEPTS

•**Information systems** →**Information retrieval;** *Speech/audio search;*

KEYWORDS

Speech-Only Search; Crowdsourcing; Interactive Information Retrieval

ACM Reference format:
Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. Crowdsourcing User Preferences and Query Judgments for Speech-Only Search. In *Proceedings of ACM SIGIR Workshop on Conversational Approaches for Information Retrieval, Tokyo, Japan, August 2017 (CAIR'17),* 3 pages.

1 INTRODUCTION

Speech-based applications are becoming more prominent and are increasingly accepted among the wider population. Even though much research has been conducted into understanding supporting search by voice input [3], only a few studies have focused on the presentation of information where no display is used [8]. Voice output is easily understood for factoid-style queries (e.g, "Which city is the capital of Japan"). When users seek answers to non-factoid style questions or queries, the information seeking tasks are often complex and cognitively taxing for the users. Sahib et al. [9] have demonstrated that the naïve approach of using a screen reader to synthesize standard search results (designed to be rendered visually) is highly unsatisfactory to users.

CAIR'17, Tokyo, Japan

This leaves the problem, then, of generating appropriate search results for presentation over speech.

Listening to search results over audio is very taxing for users since audio is a temporal medium and does not leave any traces for the user to refer to [7]. Since speech is a linear medium, it is also challenging to present complex structures. Simultaneously, speech can blend in with the surrounding sounds which poses difficulties recognizing and distinguishing the right sounds from the background noise [12]. Thus, it is challenging to convey large amounts of information via audio without overloading the user's short-term memory [7, 12].

Crowdsourcing has become a popular research tool allowing researchers to access a diverse and on-demand population making low-cost human computation resources available [5]. It is also seen as an alternative way to recruit large groups of users who evaluate systems in a remote setting requiring no supervision [6]. Much of the existing research on using crowdsourcing with audio content focuses on collecting utterances from the crowdsource workers (hereafter referred to as workers) [1], using crowdsourcing for speech transcriptions or for Spoken Dialog System evaluation [2]. However, only limited research has been conducted in using crowdsourcing as an evaluation tool for search result presentation over audio in Interactive Information Retrieval (IIR).

This paper compiles two case studies [10, 11] for measuring the impact in user preference of different factors involved in presenting search results over audio via crowdsourcing. Presenting the workers with search results over audio allows us to collect large-scale input on affective reactions and the usability of the IIR system. This paper contributes to experimental design approaches for audio evaluation in IIR over crowdsourcing and proposes different techniques to collect task workload and user behavior through crowdsourcing platforms.

2 COLLECTING USER PREFERENCE AND QUERY JUDGMENTS

We now describe two case studies illustrating the design of crowdsourcing tasks gathering user preferences and query judgments.[1] These tasks were designed to investigate the presentation of search results via an audio-only communication channel [10, 11]. Here, where "displaying" a summary means *playing* a segment of audio.

Spina et al. [10] focused on preference and effectiveness in relation to the *content* of summaries; Trippas et al. [11] investigated

[1] All experiments were performed under Ethics Applications at RMIT University

aspects of preferred and effective *forms* of spoken/audio summaries, specifically, their length. Both works involved a crowdsourcing methodology[2] for (i) collecting preference judgments for multiple versions of audio summaries presented to the workers by comparing two versions at one time, and (ii) measuring how these different audio summaries impacted the workers' relevance judgments.

Figure 1 shows the crowdsourcing interface used to collect worker preferences between two versions of audio summaries. Workers were asked to read a text and then proceed to Question 1. Question 1 displayed two lists of audio summaries, of which only one item on each lists corresponded to a summary generated from the actual document –and the other summary on each list was generated from a completely unrelated document. This quality control mechanism aimed to ensure that workers read the document before preference judging between the two versions of summaries. Next, workers were asked to choose one summary of each list as the most representative of the given document. The option chosen by workers in Question 1 was then populated in Question 2 where workers had to indicate which summary version they preferred. An option of *no preference* was included.

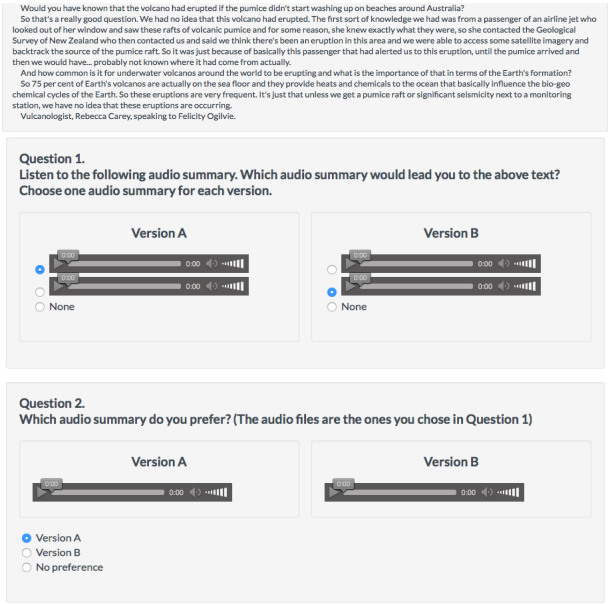

Figure 1: Example of a crowdsourcing task for collecting preference judgments for audio summaries.

Figure 2 shows a fragment of the crowdsourcing task to collect relevance judgments from audio summaries.

Workers were asked to read the query, listen to each search summary result, and assess the relevance of the underlying document for that summary. The audio files were stored on a server and embedded in the web interface using the <audio> HTML tag. A link to the audio URL was included to mitigate problems workers might experience with in-browser audio players.

In the second case study, Trippas et al. [11] used exit questionnaires to collect user preferences between two audio interfaces by

[2]The tasks were performed on CrowdFlower: http://www.crowdflower.com

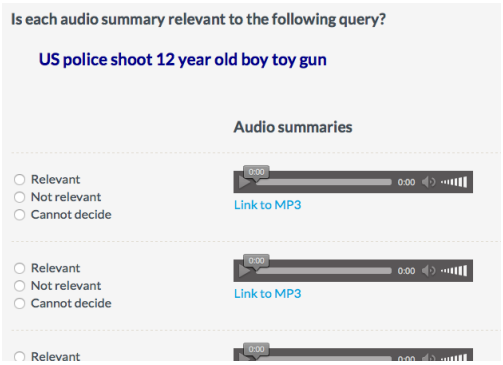

Figure 2: Example of a crowdsourcing task for collecting relevance query judgments using audio summaries.

capturing comparisons for the within-subject study (Figure 3). This exit questionnaire was only available to workers who successfully answered a previous validation phase based on choosing the correct audio summary from a list that included unrelated audio to the query.

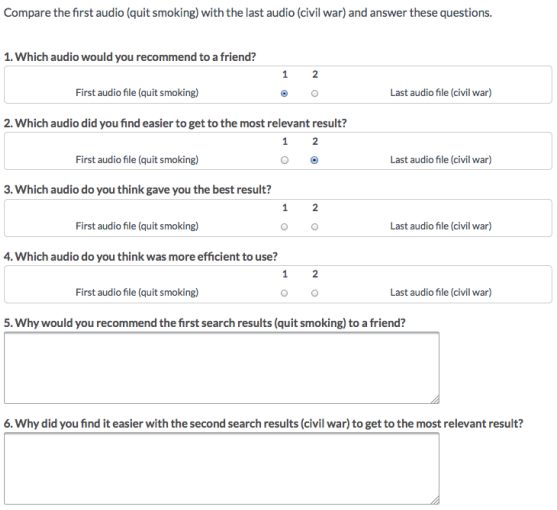

Figure 3: Example of exit questionnaire.

Example screenshots were included at the beginning of each crowdsourcing task so that workers know what to expect. Workers reported that those screenshots were informative.

The experiments were designed with a user-centered design approach. From the first iteration of the task users were involved to validate our design decisions. Several usability studies were performed before launching a pilot study on CrowdFlower. Whenever we launched a pilot CrowdFlower task we explicitly stated that it was a pilot and that feedback was highly appreciated. We received positive feedback and comments which helped the CrowdFlower task design.

The next section analyzes the possible extensions that can be incorporated to the above described methodology to measure task

workload and collect user behavior in speech-only search interactive scenarios.

3 MEASURING TASK WORKLOAD AND COLLECTING USER BEHAVIOR

We firstly describe how to collect task workload information. Secondly, we describe how to collect behavioral data from workers while they are interacting with audio.

Task Workload. The experiments explained in Section 2 allows us to capture user preference and query judgments. However, since audio is a serial multimedia format it would be beneficial to capture workload-related factors. These factors could help us to understand and better design the presentation of search result summaries over audio. In order to capture these factors we propose to use the NASA-Task Load Index (NASA-TLX) which is a multidimensional assessment tool for assessing a system's task workload [4]. The NASA-TLX is performed by filling out rating scales which makes it a practical tool to apply in a crowdsourcing environment where users are performing micro-tasks for remuneration.

Interaction Data. The enhanced capabilities to record and analyze interaction data have driven the research and development of information systems. For instance, current web search engines rely on logged search behaviors to improve search experience [13]. Previous work has shown that search interactions –such as the temporal length of a session or the number of documents clicked– can be effectively collected via crowdsourcing [14].

The linearity of the audio channel is a property that can be exploited to collect further detailed behavioral data. Gathering this behavioral data can be done by keeping track of the events generated during the interaction –e.g., using JavaScript event listeners, which are supported by the most popular crowdsourcing platforms– and thus can record *how* workers listened to the audio.

Event listeners allow us to gather information about whether:

- the audio was completely played
- the audio was paused
- the user has moved or skipped the audio playback to a new position.
- ranges of audio have been played or skipped

These events can also be used for *validation*, e.g., the audio was played during a certain amount of time. These metrics can also be used in conjunction with explicit questions asked to workers for feedback on how they listened to the audio. The quality measure can be that if the event-listener and the question about how they listened to the audio are the same, we expect the workers to be completing the task properly.

Finally, the results of task overload tests such as NASA-TLX can be compared to the actual log interaction data, although how to perform this comparison still remains as an open research question.

4 CHALLENGES

Besides the aforementioned challenge of estimated workload from audio interactions, a major research question to address is whether crowdsourcing can be used in fully interactive audio scenarios, e.g., to evaluate speech-only search systems online. Initiatives such as the ParlAl framework[3] recently announced by Facebook facilitate the integration of chatbot agents in crowdsourcing tasks. However, it is still unclear how this methodology can be effectively extended to the speech-only scenario, e.g., allowing workers to interact with intelligent assistants used in devices such as Amazon Echo or Google Home.

Finally, challenges that are intrinsically related to crowdsourcing still remain present to our scenario and that includes: quality assurance, maximize the worker pool, maintain tasks according to changes in the platform, etc.

5 CONCLUSION

This paper presents two case studies illustrating how a crowdsourcing methodology can be used to effectively collect relevance and preference judgments for a IIR tasks where the search results are presented over audio in a speech format. We believe the design used in our previous experiments can inform the design of similar crowdsourcing tasks for speech-only search. We also propose the use of existing assessment tools such as NASA-TLX and the gathering of low-level interaction signals using event-listeners to measure work load. Finally, we identify a number of open research challenges that needs to be addressed in order to use crowdsourcing in more complex speech-only search interaction scenarios.

ACKNOWLEDGMENTS

This research was partially supported by Australian Research Council Project LP130100563 and Real Thing Entertainment Pty Ltd.

REFERENCES

[1] J. Arguello, S. Avula, and F. Diaz. Using query performance predictors to improve spoken queries. In *Proceedings of ECIR'16*, pages 309–321. Springer, 2016.
[2] M. Eskenazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann. *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment.* John Wiley & Sons, 2013.
[3] I. Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of SIGIR'16*, pages 35–44, 2016.
[4] 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.
[5] G. Jones. *An Introduction to Crowdsourcing for Language and Multimedia Technology Research*, pages 132–154. Springer, 2013.
[6] F. Jurčíček, S. Keizer, M. Gašic, F. Mairesse, B. Thomson, K. Yu, and S. Young. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In *Proceedings of INTERSPEECH'11*, 2011.
[7] J. Lai and N. Yankelovich. *Speech Interface Design*, pages 764–770. Elsevier, 2006.
[8] N. G. Sahib, D. Al Thani, A. Tombros, and T. Stockman. Accessible information seeking. *Proceedings of Digital Futures*, 12, 2012.
[9] N. G. Sahib, A. Tombros, and T. Stockman. A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *Journal of the Association for Information Science and Technology*, 63(2):377–391, 2012.
[10] D. Spina, J. R. Trippas, L. Cavedon, and M. Sanderson. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology*, 2017.
[11] J. 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 *Proceedings of SIGIR'15*, 2015.
[12] M. Turunen, J. Hakulinen, N. Rajput, and A. A. Nanavati. *Evaluation of Mobile and Pervasive Speech Applications*, pages 219–262. John Wiley & Sons, Ltd, 2012.
[13] R. W. White. *Interactions with search systems.* Cambridge University Press, 2016.
[14] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. M. Jose, and L. Azzopardi. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval*, 16(2):267–305, 2013.

[3] https://github.com/facebookresearch/ParlAI