Towards Understanding the Impact of Length in Web Search Result Summaries over a Speech-only Communication Channel

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

Damiano Spina
damiano.spina@rmit.edu.au

Mark Sanderson
mark.sanderson@rmit.edu.au

Lawrence Cavedon
lawrence.cavedon@rmit.edu.au

School of Computer Science and Information Technology
RMIT University, Melbourne, Australia

ABSTRACT

Presenting search results over a speech-only communication channel involves a number of challenges for users due to cognitive limitations and the serial nature of speech. We investigated the impact of search result summary length in speech-based web search, and compared our results to a text baseline. Based on crowdsourced workers, we found that users preferred longer, more informative summaries for text presentation. For audio, user preferences depended on the style of query. For single-facet queries, shortened audio summaries were preferred, additionally users were found to judge relevance with a similar accuracy compared to text-based summaries. For multi-facet queries, user preferences were not as clear, suggesting that more sophisticated techniques are required to handle such queries.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]; H.3.3 [**Information Search and Retrieval**]

Keywords

Spoken retrieval, search result summarisation, crowdsourcing

1. INTRODUCTION

Speech-based web search (i.e., posing search queries using voice rather than a keyboard/touchscreen interface) is increasingly ubiquitous, particularly through the use of mobile devices. Several systems (e.g., Siri, Google Now, Cortana) can speak a reply to "factoid"-style search queries (e.g., "How high is Everest?"). If no factoid answer exists, the systems revert to displaying a ranked list of results on screen. However, there are situations where a full speech-only interface is preferable, such as while driving a car [6, 7], when there is no screen or keyboard available [20], when users are mobile [12, 16, 18], or when using wearable devices [4]. Moreover, visually presented results may be inaccessible for certain populations, such as visually impaired users [14, 17] or people with limited literacy skills.

Presenting lists of search results over a speech-only communication channel presents a number of challenges; in particular, simply speaking the textual component of a standard search results list has been shown to be ineffectual [17]. The serial nature of the speech/audio channel makes it difficult for users to "skim" back and forth over a list of results (a standard process in browsing a visual list).

Few studies have investigated techniques for effective presentation of web search results via a speech-only communication channel (hereafter referred to as audio) [7]. The present study seeks to address this. In particular, we seek a better understanding of how to present search results over audio while not overwhelming the users with information [18], nor leaving users uncertain as to whether what they heard covered the information space [19].

The length of a spoken search result summary plays a crucial role in the success or failure of presenting search results over audio. A short summary might not yield enough information to judge whether the retrieved document is relevant or not; in contrast, a more descriptive summary might take too long to be played and thereby diminish user experience. Thus a trade-off is necessary between a short summary and a longer, more descriptive summary.

The present study investigated these trade-offs via a crowdsource-based interactive experimental design. The aim of the study was to develop a baseline of understanding about the audio result summary length users prefer. This study aimed to answer the following research questions:

- What is the impact of search result summary length in a spoken retrieval scenario?
- Do users prefer a longer or shorter summary?

2. METHODOLOGY

Our experiments used a crowdsourcing platform to present queries and search results to users. Result summaries of various length were presented in text or audio form. Summary length was either a *full* Google-length summary or a *truncated* version extracted from the original summary. Users were asked to select a result that best addressed the query, and were also presented with post-task and exit questionnaires. The present study used the CrowdFlower crowdsourcing tool [9, 10] to provide user data for the tasks and questionnaires.

SIGIR'15, August 09–13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767826.

Table 1: Examples of full and truncated summaries for faceted and single-faceted queries.

Query	Type	Full Summary	Truncated Summary
`old town scottsdale` Find restaurants in Old Town Scottsdale, AZ.	faceted	Downtown Scottsdale / Old Town Scottsdale. Downtown Scottsdale pulses with a vibrant energy all its own. By day, shopping, gallery-hopping and dining...	and residents alike to enjoy art galleries, specialty retail...
		Where the Old West Meets the New West. Downtown Scottsdale is the ultimate destination for visitors and residents alike to enjoy art galleries, specialty retail...	all its own. By day, shopping, gallery-hopping and dining...
`what was the name of elvis presley's home` What was the name of Elvis Presley's home?	single-facet	The unique association between Elvis Presley and Graceland, his home in Memphis, ... Early on, Elvis became known around the world by his first name alone...	Elvis Presley and Graceland, his home in Memphis, ... Early
		Graceland? Learn fun Graceland facts, and get the official answers to common questions about the home of Elvis Presley! ... How did Graceland get its name?	Graceland facts, and get the official answers to common

2.1 Experimental Design

We first describe the task users undertook, followed by the queries, search engine results summaries, post-task and exit questionnaires, and use of text as a baseline for audio. The crowdsourcing setup for presenting results lists to users and collecting judgements is also described.[1]

Users were presented with a task which consisted of three queries, corresponding lists of result summaries, a post-task questionnaire, and an exit questionnaire. Users were asked to read three query descriptions and read/listen to the summaries, before stating their preferred result description in the questionnaires. One set of summaries was the full length, one set was truncated.

2.1.1 Queries

The task was designed to reflect common search tasks on the Web: query topics from the Text REtrieval Conference (TREC) 2013 Web Track were used [5]. The queries in the track were manually selected from logs of a commercial search engine [5]. Since this was a preliminary study, a subset of twenty queries from the TREC 2013 Web Track dataset were used. An assessment of the queries indicated two categories, *single-facet queries* (queries with a clear intent) and *faceted queries* (typically broader in intent and represented with subtopics). It was decided to investigate whether these categories impacted on result summary preference.

The present study included seven single-facet and thirteen faceted queries. Table 1 shows an example for each type of query. Only informational subtopics were selected for the present study, since they have a primary interpretation which is reflected in the description field.

2.1.2 Search Engine Results Summaries

Each query was sent to the Google search engine and Google-generated text summaries were extracted for the top five search results.[2] The summaries were converted into a spoken synthetic voice (audio). Instructions were added to each summary set: *These are summaries of the top results. Select the summary that leads to the information you are looking for.* A list-rank number was added to the front of each summary to allow easy identification of the users' selection. Table 1 shows a sample summary before and after the conversion for the task.

Truncated versions of the original Google-generated summaries were created manually. Here, a contiguous subset of nine words was selected from each full summary. Nine was found to be a little less than half the length of a standard Google-generated summary. For this initial work, manual summaries were created to avoid bias introduced by poor automatic truncation, which may negatively impact user perception. Human judgement was assumed to be the best way to preserve the meaning of the summary.

The presentation of the twenty search queries was randomized with the use of a Latin square design. Each user saw three queries per task. The order in which the users were presented with the result description (original summary vs. truncated summary) was rotated. These steps were implemented to avoid learning effects triggered by usage order and by users becoming accustomed to a synthetic voice [1, 11].

A problem reported with crowdsourcing is that users try to receive payment without completing the task properly [3]. Therefore, every task was populated with a *Gold Question* to help with data integrity and to detect if the participant was paying attention to the task [2]. The Gold Questions in this study used queries with clear pre-determined answers. Users were presented with three query descriptions and corresponding summaries. However, one of these summaries was populated with unrelated summary results. A user that was not able to identify that the summaries were not related to the query had their judgements discarded.

Though crowdsourcing provides a significant and easily accessible sample, we discovered that despite the precaution of releasing the tasks at different times and on different days, the results showed that many users completed every task. As a result, a modification was made and instead of running a new task for every set of queries, the queries were grouped into batches of ten and restrictions placed on how many batches users could judge.

2.1.3 Post-task and Exit Questionnaires

Post-task questionnaires are frequently implemented to assess the system–task interaction and gather user feedback on their experiences with using a particular system to complete a particular task [11]. Since no validated questionnaire has been published for studying user reaction to audio summaries, we used questionnaires adapted from previous studies.

Users completed the post-task questionnaire three times for each of the queries given. The post-task questionnaire (see Table 2) consisted of five questions on a five-point Likert scale (1–5); one question on query judgement with multiple choice answers (6); one question on how the participant listened to the audio with tick boxes (7); and a text box for further comments (8).

In addition, [11] suggests conducting a questionnaire at the end of the completed task to capture comparisons for within-subjects studies. Thus users were also presented with an exit questionnaire. Using a dynamic panel, the exit questionnaire was available only to users who were successful in answering a Gold Question. The exit questionnaire was used to measure users' preferences for information exploration using different result description configurations.

[1] All experiments were performed under Ethics Application BSEH 10-14 at RMIT University.

[2] Only the top five were presented to keep the audio task manageable.

Table 2: Post-task questionnaire questions.

Post-task Questionnaire
1. The search results I heard are informative.
2. The search results give me a good overview of the available options.
3. The search results give me enough information to select the most relevant result.
4. The search results are presented in a way that is easy to understand.
5. I am confident I can recall the search results that I heard.
6. Which search result would you select to hear further information for?
7. Which statement describes how you listened to the audio?
8. Further comments.

The exit questionnaire was analysed with the help of responses from the post-task questionnaire.

2.1.4 *Using Text as Baseline for Audio*

Tasks were paired, whereby one task's summaries were audio and the other's were text. The text output was used to create a baseline measure of the system, facilitating analysis of the difference in preference between audio and text [11], enabling us to compare audio against the text baseline.

2.2 Users

CrowdFlower allows *contributors* (users who submit tasks to CrowdFlower) to place constraints on the users assigned to a task. The following constraints were put in place for the present study:

- Only users with an IP address from Australia, Ireland, New Zealand, the UK, and the USA were allowed to participate in order to maximise the likelihood that users were native English speakers or or had a high level of English.
- Users were able to participate only once in a particular task to maximise the worker pool.
- Users who took less than sixty seconds to complete the whole task were discarded on the basis that it would take users more than sixty seconds to listen to/read the summaries.

Although users were not permitted to participate more than once in a task which had the same set of queries, they were allowed to participate in tasks with different queries. A minimum of 36 users were recruited for each given task [13].

When a participant did not answer the Gold Question successfully, that participant's submission was discarded. These users were also not allowed to participate in later tasks. It was found that 11.8% of all users did not answer the Gold Question successfully, their submissions were discarded.

3. RESULTS

The exit questionnaire was analyzed using the χ^2 goodness-of-fit test to compare the distribution of scores across two levels. Results are shown in Table 3.

The χ^2 goodness-of-fit test [11] was used to assess whether changing the result summary had an effect on user preference.

Users were asked (in the post-task questionnaire) which summary made the users want to know more about the underlying document.[3] These judgements were analysed with the two-sample Kolmogorov-Smirnov test (*KS test*) to determine whether two given samples follow the same distribution [15]. The tasks compared the result 'click' distributions where the length of the summary was manipulated. The tasks were conducted in pairs: audio and text [11].

[3]This is equivalent to asking which result they would *click* in a traditional search engine result page (SERP).

Table 3: Exit questionnaire results for preferences in the search engine result summaries. $^{\blacktriangle}p < .01$.

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

The KS test showed that for truncated summaries, only two out of twenty queries were there different distributions for audio-based versus text-based summaries. For full-length summaries and three out of twenty summaries resulted in different distributions for audio versus text. In general, the same distribution was found for query judgements between the text baseline and audio indicating that users made similar query judgements regardless of the presentation style being audio or text.

3.1 Preferred Length of Text Summaries

Table 3 shows that users tend to prefer full text summaries rather than their truncated counterpart. For instance, 57% of users would recommend full text summaries to a friend and 57% indicated that full summaries gave better results. The χ^2 goodness-of-fit tests were statistically significant ($p < .01$) for three exit questions in relation to the use of the original summary for presenting text results, indicating that this information exploration style was preferred.

3.2 Preferred Length of Audio Summaries

Results in audio summaries do not indicate a clear preference between full and truncated (preferences differ at most by only 2%). The χ^2 goodness-of-fit tests were not statistically significant ($p > .05$) for any of the exit questions about presenting audio results. No statistically significant difference were found for faceted queries. However, for single-facet queries using audio, all exit questions where statistically significant ($p < .05$) with a user preference for truncated summaries.

The KS test revealed that only one out of seven single-facet queries judgements was statistically significant ($p < .05$) when comparing truncated audio to the truncated text baseline. The KS test was not statistically significant for any faceted queries in audio, with one exception.

Users reported that overall it was easier to recall truncated audio summaries (54.4%) than full audio summaries (49.9%). Moreover, fewer users reported that they had to listen to the audio more than once (16.8%) for truncated summaries than for full-length audio summaries (23.7%). Only three users reported that they stopped the audio for truncated summaries, possibly indicating that both the information presented and the length of the information were short enough to avoid cognitive overload.

4. DISCUSSION

The present study investigated whether summaries of shorter length would be preferred for audio presentation as they could avoid overloading users' memory [20]. The exit questionnaire responses demonstrated that, for text summaries, full-length were preferred. However, for audio, no significant length preference was found.

Users reported that truncated summaries for single-facet queries were preferred. Thus for simpler, less ambiguous queries, shorter audio summaries were both effective and preferred. However, for faceted queries, users may have benefited from a more informative audio response even at the cost of listening time.

The single-facet query judgement distribution for both audio and text followed the distribution reported in past work [8] where query results ranked first and second receive most user attention. However, this expected distribution was not reflected in the faceted query judgements; rather, summaries ranked first and last obtained the most attention. This is also of interest: the serial nature of audio seems to lead to a bias towards most-recently-heard results, a behavior not found in visual presentation.

Users left comments in questionnaires. For summaries of faceted queries, they indicated that the summaries were missing key information. This suggests that the way of presenting summaries may differ depending on query intent: short audio summaries may be appropriate for clear intent queries (single-facet), whereas broader intent queries (faceted) may need more complex techniques (e.g., interactive/conversational approaches).

5. CONCLUSIONS

The present paper describes an initial investigation into result summaries for audio-based search. This study aimed to answer the following research questions:

- What is the impact of search result summary length in a spoken retrieval scenario?
- Do users prefer a full or truncated summary?

Differences were observed when result summary lengths were presented in the spoken retrieval scenario. In general, there was no preference for fuller descriptive summaries or for truncated summaries. However, results revealed that different kinds of queries (single-facet vs. faceted) benefited from an optimised summary depending on the type of query.

Extensions of the research include testing based on a larger number of queries and using automated techniques for truncating summaries for audio presentation. (The current method used manual truncation to avoid poor truncation from confounding the results.) More significantly, the results suggest a need for developing more sophisticated approaches to handling result-presentation over audio for faceted queries.

Acknowledgments

This research was partially supported by Australian Research Council Project LP130100563 and Real Thing Entertainment Pty Ltd. The authors wish to thank Bruce Croft who provided valuable feedback.

6. REFERENCES

[1] W. Albert, T. Tullis, and D. Tedesco. *Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies*. Morgan Kaufmann, 2009.

[2] S. Buchholz, J. Latorre, and K. Yanagisawa. Crowdsourced assessment of speech synthesis. *Crowdsourcing for Speech Processing*, pages 173–216, 2013.

[3] C. Callison-Burch and M. Dredze. Creating speech and language data with amazon's mechanical turk. In *Proc. of NAACL HLT CSLDAMT'10 Workshop*, 2010.

[4] 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. *Transactions on Speech and Audio Processing*, 10(8):531–541, 2002.

[5] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. Clarke, and E. M. Voorhees. TREC 2013 Web Track Overview. In *TREC'13*, 2014.

[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] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. of SIGIR '05*, 2005.

[9] G. Jones. An introduction to crowdsourcing for language and multimedia technology research. *Information Retrieval Meets Information Visualization*, 7757:132–154, 2013.

[10] 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 *Proc. of INTERSPEECH'11*, 2011.

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

[12] S. R. Klemmer, A. K. Sinha, J. Chen, J. A. Landay, N. Aboobaker, and A. Wang. Suede: a wizard of oz prototyping tool for speech user interfaces. In *Proc. of UIST'00*, 2000.

[13] B. P. Knijnenburg, M. C. Willemsen, and A. Kobsa. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proc. of RecSys'11*, 2011.

[14] J. Lai and N. Yankelovich. *Speech Interface Design*. Elsevier, 2006.

[15] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Assoc.*, 46(253):68–78, 1951.

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

[17] N. G. Sahib, D. Al Thani, A. Tombros, and T. Stockman. Accessible information seeking. In *Proc. of Digital Futures'12*, 2012.

[18] M. Turunen, J. Hakulinen, N. Rajput, and A. A. Nanavati. Evaluation of mobile and pervasive speech applications. In *Speech in Mobile and Pervasive Environments*, pages 219–262. 2012.

[19] S. Varges, F. Weng, and H. Pon-Barry. Interactive question answering and constraint relaxation in spoken dialogue systems. In *Proc. of ACL'06*, 2006.

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