Presenting Web Search Results over a Speech-Only Channel with Minimal Cognitive Load

Owen Versteeg, Claudia Hauff (TU Delft)

June 2020

Abstract

While voice assistants have exploded in popularity over the last decade, they still have many issues. Among these is the issue of result presentation: how do you speak results to the user? Prior research has investigated how cognitive load relates to result presentation and other methods for result presentation, but highlighted a lack of research about result presentation over speech and called for further investigation. [12] This research answered that call, and investigated the research question *How can web search results be presented over a speech-only channel such that the users' cognitive load is minimized?* We tested three methods: 1) one-source, one-shot responses, 2) question-and-answer responses, and 3) multi-source, one-shot responses. In all cases, single-source responses were preferred to multi-source responses. Of the two types of single-source responses tested (question-and-answer style responses and one-shot responses), both types had a similar average score. Our recommendation to developers of conversational search technology would be to allow user choice between these two response types. We would only recommend multi-source responses as a fall-back option in the case of a low estimated probability of result relevance.

Introduction

In the last decade, voice assistants have exploded in popularity. Today, they are ubiquitous, having moved from the smartphone to the desktop computer, the speaker, and even the lightbulb [9]. Fierce competition has driven not only features but also improvements in understanding a wide range of accents and languages. But grow as they may, voice assistants are primarily used today for simple tasks, and many are embarrassed to use them in public. [1] Even in 2020, many tasks remain either out of reach or difficult to manage.

Conversational search is the use of search engines in a conversation-based format; often this is "spoken" or read aloud on a phone or laptop speaker. While some conversational problems are fairly simple ("set an alarm for 3pm"), others are extremely complex. In particular, the problem of spoken conversational search: search results can be complex and difficult to represent in a spoken format. Even a query that may seem unambiguous - "how far away is Paris" has many answers: 510km by car, seven hours' drive (but six outside rush hour!) 267 miles as the bird flies, 3 hours by train - just to give a few. Other questions, such as "how much does the train cost to Marseille" depend on a massive number of factors: what day, which stations, what class, what train, who's riding, discounts, and sold out trains. As a result, current systems tend to either ignore many web search queries or present them on a display. As conversational search is often used when displays are not an option, neither of these outcomes is desirable. Furthermore, the "cognitive load" (the amount of working brain memory required by the user) should be minimized to encourage real-world use. So how should search results be presented?

Existing research in the field of Information Retrieval has tested solutions available to consumers [14], tested how cognitive load relates to result presentation [6] and presented models for testing search systems [8], demonstrating the importance of result presentation methods. Recent research has even proposed methods to test speech-based methods for presenting search results [13]. Recently, Trippas and Spina et. al investigated result presentation methods in search and concluded among other things that they could not find clear definitive results and that they believe "there is a need of further research on

how to present results over audio" [12]. This call for further research provides strong motivation for the research question of this paper: *How can web search results be presented over a speech-only channel such that the users' cognitive load is minimized?*

In the existing literature, the two primary factors in the quality of result presentation systems are cognitive load [6] and user preference [14]. Thus, the research question has been divided into two sub-questions:

- What are the preferred methods to present complex speech-only results?
- What is the cognitive load of these methods?

In this research users were selected from the general population using Prolific [2] and tested using a Likert scale (see Methodology), provided with a randomized Bing API+Voice Recognition+Speech Synthesis based [3] voice search assistant (Experimental Setup), providing data indicating that one-source responses were significantly preferred and raising further questions about response structure and format (Results and Discussion) but allowing us to conclude the general superiority of one-source responses to a multi-part Q&A style and present avenues for future work (Conclusion and Future Work.)

Methodology

Technology: In order to construct the research as described above, I constructed a voice assistant system the publicly available Microsoft Bing APIs, using the open-source Microsoft Research project Macaw [15] as inspiration for system design. Within this setup, the Web SpeechRecognition API was used for recognizing speech to send to the voice assistant backend. For frontend speech synthesis, the SpeechSynthesis interface of the Web Speech API will be used. An information retrieval system similar to that described by Chen, Fisch et. al was used [3].

Participants: Study participants were recruited using the online study participant system Prolific. Participants were compensated at a rate of US$10.00 per hour or greater for their time. Native English speakers from the US and UK were selected. Two Prolific system selection criteria were applied: acceptance rate >90% and minimum number of past submissions 50.

Study setup: Participants were presented with a web site interface to the experimental setup elaborated in the section Experimental Setup. Participants were informed of their consent and privacy options for the study as well as given a short testing interface to ensure the proper functioning of their microphone and audio setup. Participants were informed on the rough duration of the study (approximately 20 minutes.)

Experimental Setup

Pre-screening: Any user sent to the study had to first undergo a pre-screening step, where they were instructed to listen for a random number (played over their device speakers) and speak this number aloud to continue. This allowed removal of any participants that could not properly use their microphone or speakers. This pre-screening step also filtered out any Web browsers that were incompatible with the study software.

User interface: A web page, providing only a microphone indicator and speaker indicator to facilitate simple setup. Changing background colors+patterns (for the colorblind) used to assist users in detecting the "different" voice assistants (colors to distinguish different voice assistants is a method used previously in the literature.) No results displayed on-screen.

Experimental design: Each participant was asked to come up with original questions and ask them to three "different" voice assistants, distinguished with unique icons and labels:

1. A voice assistant providing one-source responses

2. A voice assistant providing one-source question-and-answer responses

3. A voice assistant providing multi-source responses

Users rated their satisfaction with the voice assistants, and then asked a new question to all three voice assistants again, this time while playing a red dot game to simulate higher cognitive load. Both surveys used a required Likert scale 1-10 slider.

The game placed a red dot on the screen titled "Click Me!". Users were requested to click the dot as soon as it appeared; once they clicked, it would disappear, and then return after a random amount of time to a random location on the screen. User "scores" for this game were the fastest response time.

Stages: Every user takes all four stages

1. Stage 1, hardcoded responses, no distraction: User was presented with one of two randomly chosen topics and instructed to investigate it using the voice assistants presented to them. The question was displayed at the top of the screen. After having received responses from all three voice assistants, the user was shown a 1-10 Likert scale slider to rate each. Responses were hardcoded and pre-written.

2. Stage 2: hardcoded responses, distraction: identical to Stage 1, but with added distraction: the red dot game described above.

3. Stage 3: free responses, no distraction: identical to Stage 1 with the exception of responses: responses will no longer be hardcoded and now will be provided by the Bing API.

4. Stage 4: free responses, distraction: identical to Stage 3, but with the red dot game described above.

Topics: Two TREC topics (#303 Hubble Telescope Achievements and #363 Transportation Tunnel Disasters) were selected from the broader base of TREC topics. These two topics were selected due to their suitability for the study (length, complexity, and ease of rephrasing.)

Sample size: One-way repeated measures ANOVA analysis ran with test data (sample size 20). Results: Sample size of 20 total is sufficient. As the study is complex (requiring microphone, speaker, and active user participation) the study participant group was expanded to 40 to allow for participants with technical difficulties or who did not complete the study.

Use of methods from literature: This experimental setup uses methods from the existing literature. Specifically, the use of multiple voice assistants distinguished in various ways (e.x. color), rating assistants using a Likert scale, the TREC topics list, and using the "red dot game" to increase cognitive load have all been used before in the literature [6], [7].

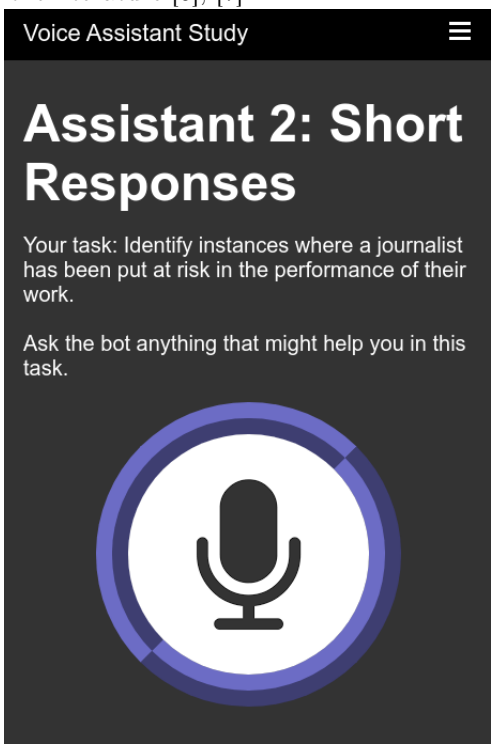

Figure 1: Screenshot of user study interface on mobile device.

Example User Queries

In this section we provide a few example queries:

"What has the Hubble Space Telescope found out about the early universe?"

"What was the worst tunnel disaster in the Southeast UK?"

"What was the worst Swiss tunnel disaster?"

"What has the Hubble Space Telescope done to increase our knowledge of the universe?"

"What was the tunnel disaster with the most fatalities?"

"What is the smallest thing that Hubble has found?"

"How many moons has Hubble found?"

"How many tunnels have collapsed in New York?"

Responsible Research

This experiment was performed with funding provided by the TU Delft, which all authors are affiliated with. The authors declare no conflict of personal interest.
Study participants were informed at the beginning of the study of the extent of data collection, as well as given the option to not participate. Any potential participant that did not consent did not have any data stored.
This research followed the Responsible Research guidelines of the TU Delft.

Results

0.1 Data Filtering

To ensure the integrity of the data, a number of participants were removed from the study at various stages due to filtering steps. Participants were removed for incompatible hardware or software or not following instructions. Incomplete study results were also filtered out.

As a result, the initial population (40 individuals recruited from Prolific) was reduced to 21 individuals. All participants, including those filtered out, were compensated for their time at the same rate.

0.2 Data Processing and Analysis

Data was processed using Wolfram|Alpha and Javascript standard libraries.

Data analysis was performed using a 1-way Repeated Measures ANOVA test where relevant, and measuring significance at the $p < 0.05$ significance level. This is standard for this data type, when observations are performed on the same individual. [5]

0.3 Response Length

The average response length, and standard deviation, in characters for the three bots is given below, rounded to the nearest tenth of an integer.

	Bot A	Bot B	Bot C
Average:	218.0	246.7	1091.0
Stddev:	49.9	87.6	231.8

A 1-way Repeated Measures ANOVA test was performed; as can clearly be seen here, Bot C provides the longest responses in all cases ($p < .00001$). This was as expected.

0.4 Likert Ratings

Each bot was rated after each step by the users. Here, we compile the average Likert scale ratings (1-10) for all four steps for each bot, rounded to the nearest tenth:

Bot A	Bot B	Bot C
53.2	49.1	39.4
15.9	19.2	13.4

This is summary data; the full data is located in the Appendix.

A 1-way Repeated Measures ANOVA test was performed; Bot C is significantly less preferred than the other bots ($p = 0.006151$).

As can be clearly seen, the lowest rated bot is the multi-source bot. This is likely due to its long and complex responses, which take far longer to read than the other bots, regularly taking most of a minute to read. The highest rated bot is the single-source bot; this is likely due to its short responses and minimal user interaction required.

0.5 Cognitive Load

As described in the Experimental section of this paper, we utilized a Red Dot Game to find user reaction times, using this to help calculate cognitive load.

Unfortunately, after testing with a 1-way Repeated Measures ANOVA test, we found the

results of this dot game were inconclusive and not significant (p=0.791725).

Full results for this dot game, including scores, are located in the Appendix.

Discussion

As shown by the data in the results section Likert Ratings, the bots with a single source were clearly the most preferred bot by the users. We believe that this is due to multiple factors: firstly, the use of a single data source is simple to understand, and secondly the response length (as seen in the section Response Length) is clearly the lowest for these bots.

Of interest is that even shorter multi-source responses appear to have been disliked more than single-source responses. Unfortunately there are too few data points for this (n=5) to draw conclusions. Further research is encouraged.

Also of interest is the comparison between the two single-source bots. While the difference between average Likert scores was not significant (p = 0.559501) the data does suggest that users have substantial individual preferences between the two. While the average ratings only differ by four points, multiple users gave the two bots ratings 30 points or more apart, indicating that they significantly preferred one over the other.

Based on this, we would recommend the use of saved user preferences for voice assistants, allowing the user to choose between a one-shot response and a question-and-answer type response. Such a system would be able to increase user Likert ratings by over ten points compared to a plain Bot A approach.

Unfortunately, the data for Cognitive Load, as seen above, was not significant. As a result, we could not conclude anything about the cognitive load. The likely reason for this is due to the limitations of proctoring and running a study over the Internet. In-person timing of reaction times would likely yield more consistency.

In the absence of significant results for the dot game, our only conclusions on cognitive load may come from response length. As shown above, the multi-source responses were significantly longer than the single-source responses. We would encourage further research on cognitive load, with a focus on in-person testing.

For developers in the field of voice assistants, we make the following recommendation: There are times when the accuracy of the short single-source summary method can suffer. As such, we believe that for certain topics, the use of the method of Bot C may be best. A "blended" method, where the Bot A method is primarily used, with the Bot C method as a fallback for "tricky" topics, may be best.

In relation to prior works, there is unfortunately not much available in the field of voice result presentation (thus the call to research by Trippas and Spinas et al), so we cannot say if these results are in or out of line with prior works.

Conclusion and Future Work

In this study, we answered a call in the literature to investigate result presentation methods. We compared three result presentation methods, represented to the users by three different bot profiles. We built a system to compare these bots, represented by different patterns and colors, using the Microsoft Bing Web Search APIs and a browser front-end study interface. After examining response length, accuracy, Likert ratings, and cognitive load factors, we can conclude that the method of single-source result presentation (demonstrated by Bots A and B in this research) was most preferred by the users in all circumstances, as well as providing attractive accuracy, response length, and cognitive load.

As mentioned in the Discussion, two result clusters were found: those preferring question-and-answer style, and those preferring one-shot responses. Due to this, we suggest a "mixed" style of result presentation, where the primary method of result presentation is single-source one-shot, falling back to multi-source responses in the case of a low estimated probability of result relevance.

As noted in the Discussion, even shorter multi-source responses appear to have been disliked more than single-source responses. We call for future research in this area, and an investigation of response length vs complexity.

Due to difficulties conducting response time tests, as well as unclear data, we were unable to draw any statistically significant conclusions about the cognitive load of any methods. We encourage future research in this area, as well as measuring response times in person as opposed to over a web-based study.

Future research is also needed to investigate the possible advantages of a "mixed" style of result presentation such as we propose above. This research would involve the usage of probabilistic models to estimate the odds of the top document being relevant to the query. [4]

Appendix

Study Data: Likert Scores

Headings are formatted as follows:

Bot A of study section 1 is labeled as A/1.

Study sections: 1 = pre-written source, no distraction; 2 = pre-written source, distraction; 3 = open web sources, no distraction; 4 = open web sources, distraction.

Score given as an integer 0-100.

A/1	**B/1**	**C/1**	**A/2**	**B/2**	**C/2**	**A/3**	**B/3**	**C/3**	**A/4**	**B/4**	**C/4**
70	90	55	66	90	55	10	9	10	10	10	10
70	100	65	80	43	70	60	1	54	57	15	19
80	90	0	80	90	0	0	0	0	0	0	0
39	35	43	33	26	38	53	24	46	55	58	55
66	85	49	70	71	46	40	68	26	45	84	20
62	35	7	72	42	31	75	0	39	66	81	67
55	80	90	40	20	80	40	10	20	40	15	40
29	60	70	55	75	40	60	80	61	55	80	40
70	80	85	70	80	85	35	45	55	35	45	55
57	29	31	50	41	36	0	0	74	11	3	0
100	78	44	100	84	44	0	33	0	0	0	0
65	53	71	60	80	64	89	58	26	90	40	20
98	70	34	70	58	56	12	0	4	4	18	13
60	50	90	80	65	65	86	52	69	88	41	39
29	37	34	33	27	15	10	0	2	8	3	6
70	87	62	72	65	58	36	39	27	38	40	34
20	44	78	31	61	70	44	66	46	66	73	33
89	86	70	85	80	80	11	12	9	16	32	23
100	74	30	100	41	29	0	0	0	100	40	38
87	97	58	74	64	38	73	62	17	59	51	10
80	91	30	90	84	40	41	43	60	40	57	10

Study Data: Dot Game Scores

Scores (given in each cell) are the best time in integer milliseconds to complete the dot game at each stage.

Time A/2	Time B/2	Time C/2	Time A/4	Time B/4	Time C/4
1385	1829	1156	1044	1048	900
1483	832	912	875	881	734
1287	1838	1108	1566	1362	2255
647	1047	1098	968	2114	837
910	852	776	1402	3201	903
886	771	1439	941	883	982
1278	3362	1525	7795	1645	1551
942	2568	823	920	776	861
927	1151	1523	1718	1300	790
944	1458	1036	1186	1336	1374
1892	2590	2485	1368	1349	7774
1602	2238	809	2507	1002	1017
1490	793	998	872	984	1245
916	893	673	803	784	706
926	1797	871	696	864	2769
1633	1182	1102	1269	831	1430
1109	1219	986	838	1507	925
866	2090	1455	2002	1975	10961
1142	1993	875	1090	1076	1130
862	973	1046	994	953	1100
1300	1056	1180	1008	931	1112

References

[1] Voice assistant anyone? yes please, but not in public!, 6 2016.

[2] Prolific | quickly find research participants you can trust, 6 2020.

[3] Chen, D., Fisch, A., Weston, J., and Bordes, A. Reading wikipedia to answer open-domain questions, 2017.

[4] Fuhr, N. Probabilistic models in information retrieval. *Comput. J. 35*, 3 (June 1992), 243–255.

[5] Hageman, J. What is the difference between simple anova and repeated measure anova?, 11 2017.

[6] Harvey, M., and Pointon, M. Searching on the go: The effects of fragmented attention on mobile web search tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2017), SIGIR '17, Association for Computing Machinery, p. 155–164.

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

[8] Radlinski, F., and Craswell, N. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (New York, NY, USA, 2017), CHIIR '17, Association for Computing Machinery, p. 117–126.

[9] Reinvent. Alexa light bulb time lapse, 2019.

[10] Trippas, J. R., Spina, D., Cavedon, L., Joho, H., and Sanderson, M. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New York, NY, USA, 2018), CHIIR '18, Association for Computing Machinery, p. 32–41.

[11] Trippas, J. R., Spina, D., Cavedon, L., and Sanderson, M. 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* (New York, NY, USA, 2017), CHIIR '17, ACM, pp. 325–328.

[12] Trippas, J. R., Spina, D., Sanderson, M., and Cavedon, L. Results presentation methods for a spoken conversational search system. In *Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems* (New York, NY, USA, 2015), NWSearch '15, Association for Computing Machinery, p. 13–15.

[13] Vtyurina, A. Towards non-visual web search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (New York, NY, USA, 2019), CHIIR '19, Association for Computing Machinery, p. 429–432.

[14] Vtyurina, A., Savenkov, D., Agichtein, E., and Clarke, C. L. A. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI EA '17, Association for Computing Machinery, p. 2187–2193.

[15] Zamani, H., and Craswell, N. Macaw: An extensible conversational information seeking platform, 2019.

Related Works

A short selection of related and relevant works that may be of interest:

Informing the Design of Spoken Conversational Search: Perspective Paper by Trippas, Spina et. al [10]

How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis by Trippas, Spina et. al [11]

A Theoretical Framework for Conversational Search, Radlinski et. al: [8]

Exploring Conversational Search With Humans, Assistants, and Wizards, Vtyurina et. al [14]