A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity

Alexandra Vtyurina
sasha.vtyurina@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Charles L.A. Clarke
claclark@gmail.com
University of Waterloo
Waterloo, Ontario, Canada

Edith Law
edith.law@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Johanne R. Trippas
johanne.trippas@unimelb.edu.au
University of Melbourne
Melbourne, Australia

Horaţiu Bota
horatiubota@gmail.com

ABSTRACT

Voice-based assistants have become a popular tool for conducting web search, particularly for factoid question answering. However, for more complex web searches, their functionality remains limited, as does our understanding of the ways in which users can best interact with audio-based search results. In this paper, we compare and contrast user behaviour through the representation of search results over two mediums: text and audio. We begin by conducting a crowdsourced study exposing the differences in user selection of search results when those are presented in text and audio formats. We further confirm these differences and investigate the reasons behind them through a mixed-methods laboratory study. Through a qualitative analysis of the collected data, we produce a list of guidelines for an audio-based presentation of search results.

ACM Reference Format:
Alexandra Vtyurina, Charles L.A. Clarke, Edith Law, Johanne R. Trippas, and Horaţiu Bota. 2020. A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity. In *Proceedings of the 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20), September 14–17, 2020, Virtual Event, Norway*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3409256.3409822

1 INTRODUCTION

Voice-based interaction systems, and voice-activated personal assistants in particular (e.g., Amazon Alexa), are steadily increasing in popularity [12]. In 2018, Forbes predicted that voice queries would make up to 30–50% of all web searches by 2020 [12].

Current state-of-the-art voice search systems perform well for factoid or simple questions, where an exact answer can be read out loud and easily digested by the listener [26]. For more complex questions, a voice assistant may redirect its user to a companion app (usually phone-based), where search results will be displayed on the screen. In the latter case, the transition interrupts the user's experience by shifting from an audio to a visual interaction modality. If the user is occupied with a primary activity such as driving, where their eyes and hands are engaged, it might be infeasible or even dangerous for the user to attend to their screen-based device. One of the most attractive features of a voice assistant is its ability to support hands-free interaction and multitasking [22].

Conversational search evolved as a recent trend, allowing users to access information by conversing with an automatic system. Though conversational systems may provide a more enjoyable interaction [24], they could also lead to an increased task completion time [19] and raise privacy concerns [2]. In this paper, we consider design challenges for a voice-based web search interface, rather than a conversational search system.

Despite the popularity of voice interfaces, little is known about how people perceive voice-only search captions. There has been extensive research on conventional text-based search interfaces [15] and the visual representation of a search engine result page (SERP) [9, 10, 27]. A SERP is typically represented as a list of captions, where each caption has a title, URL, and a brief summary (or "snippet") describing a particular web page. In this work, we aim to determine experimentally what features make an audio caption "good", and why. In particular, we investigate the following research questions:

- **RQ1:** Does the medium (text/audio) over which search results are delivered affect: (i) the user's search result preference, and (ii) the user's perceived workload?
- **RQ2:** Does the complexity of a search task affect: (i) the user's search result preferences, and (ii) the user's perceived workload over different mediums?
- **RQ3:** What aspects of audio-based search results are important for the accurate assessment of relevance by the user?

To answer these questions, we conducted an experiment in two parts (referred to as *AMT* and *LAB*), and analyzed the data using a mixed-methods approach [28]. In the *AMT* study, we asked 69 crowdworkers to judge the relevance of search results—presented in text or audio format—for six search tasks. In the *LAB* study, we invited 36 people to participate in a controlled laboratory experiment, where they were asked to judge the relevance of search results presented in text and audio formats and participate in a set of semi-structured interviews to discuss how the presentation of audio captions affects their experience. In both studies, we varied

ICTIR '20, September 14–17, 2020, Virtual Event, Norway

ACM ISBN 978-1-4503-8067-6/20/09…$15.00
https://doi.org/10.1145/3409256.3409822

the complexity of the search tasks to account for the variability of experience that it can produce [4].

Results show that there is a significant difference in users' search result preference depending on whether the search results are presented in text or audio format (*RQ1.1*). We also demonstrate that processing audio captions incurs a significantly higher perceived workload compared to text-based captions, addressing *RQ1.2*. Regarding *RQ2*, our analysis does not reveal a significant interaction between the complexity of a search task and the medium, both in terms of search result preference and perceived workload. Finally, we address *RQ3*, by concluding with a set of guidelines for designing audio captions for search results.

2 RELATED WORK

2.1 Visual web search interfaces

Researchers investigated various aspects of SERP appearance in browsers [15]. Clarke et al. [9] and Rose et al. [27] identified caption features that make SERP more attractive to users. Clarke et al. [9] found that users preferred the presence of query terms and a lack of redundancy in captions. Rose et al. [27] found that people preferred full sentences and genre cues (e.g., "official site") in search captions. Others studied the desired snippet length for search captions. Cutrell and Guan [10] varied snippet length for navigational and informational search tasks, finding that longer snippets are detrimental to the former and beneficial to the latter. Kaisser et al. [17] confirmed these findings—different answer lengths are preferred depending on the query type. In this work, we confirm that aspects such as pre-determined snippet length, fully formed sentences, and lack of redundancy across caption parts are highly important for user perception of audio-based search results.

2.2 Voice interfaces

Voice and text are two inherently different interaction mediums. Studies suggest that people consume information differently depending on whether they see or hear it. In a user study with a voice-based application for simple tasks (e.g., checking email, retrieving weather forecast), Yankelovich et al. [36] found that vocabulary, information organization and flow may not translate well between two mediums, and concluded that information in voice interfaces should be organised differently from visual ones. A log analysis study by Guy [13] showed that voice web search queries were longer and used richer natural language compared to text queries. Pointing out the rapid growth of voice user interfaces, Murad et al. [23] reflected on design guidelines for visual and voice interfaces, noting that a high cognitive load poses design challenges for the latter. Demberg et al. [11] showed that preferences for a voice-based interactive system can vary depending on the usage scenario. In their study, a more complex system was preferred when users could fully focus on the interaction, while a simpler system was preferred when users were preoccupied (e.g., while driving).

2.2.1 Conversational search. Several papers studied approaches to *conversational search*, where users communicate with a search system through multi-turn natural language exchanges . Radlinski and Craswell [25] proposed a theoretical framework aimed to formalise interactions with a conversational search system. Thomas et al. [31] and Trippas et al. [33] each created a spoken conversational search dataset from observing two people communicating through an audio channel to complete search tasks. Both papers illustrated approaches to web search through conversation.

The intricacies of presenting search results over audio were further explored by Chuklin et al. [8]. In a crowdsourced study, they varied the prosody features (pauses, speech rate, pitch) of sentences containing answers to factoid questions and found that emphasizing the answer phrase with lower speaking rate and higher pitch increased subjective informativeness of the audio clip. Trippas et al. [34] compared user preferences for longer vs. shorter search result summaries, when they were displayed as text or played as audio. Tombros and Crestani [32] found that users processing document summaries (i.e., top-scoring sentences) achieved the highest performance when the results were presented as text, were more focused when the summaries were read to them over the telephone in human voice compared to in-person, and had difficulties spotting query keywords when listening to text-to-speech generated document surrogates compared to a human voice. Winters et al. [35] explored how non-speech sounds increase user engagement with social media visual content. In this work, we study factors that affect user perception of audio search captions and the considerations one should make when designing a voice-based web search system.

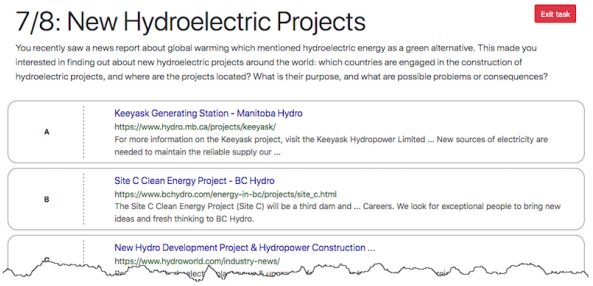
2.2.2 Search interfaces for screen-reader users. Users of screen-readers face a unique set of challenges when accessing visual information [29]. Accessibility research provides us with an insight into important aspects of perceiving information through an audio channel. For example, Abdolrahmani et al. [1] showed that when judging the credibility of a web page, blind users rely more on content as opposed to visual appeal, which is used more heavily by the sighted study participants. When exploring search results, blind participants navigated over result headings by using screen reader shortcuts, and after skimming through the titles, focused on the selected snippets. Both blind and sighted participants considered the source an essential factor for selecting a given search result. Additionally, visually impaired smart speaker users requested the ability to control the audio output settings, such as speech rate [2]. In this work, we find some of the same aspects being of importance for sighted users when processing audio search captions.

2.3 Task complexity

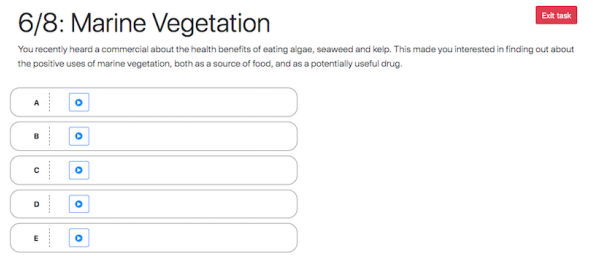
In the research and development of information retrieval systems, search tasks have been designed to investigate the interaction between user behaviour and the difficulty of the search task [18]. One framework for constructing tasks is the Taxonomy of Learning [3], which specifies six levels of increasing cognitive complexity as: *remember, understand, apply, analyse, evaluate*, and *create*. Prior research in visual text search has shown that more complex tasks lead to greater levels of search interactivity, for example, through increased clicks, queries, and time on task [18]. In this work, we chose three task levels to account for variance in user behaviour.

3 STUDY DESIGN

To address our research questions, we conducted a two-part user study. In the first part (*AMT*), we aimed to explore potential differences between user choices when search results were presented



(a) Text condition



(b) Audio condition

Figure 1: For each task five search results were presented in text or audio formats.

in text or audio format in a crowdsourcing setup. We found that there was a significant difference in user preferences depending on whether the search results were presented in as text (called the "Text" condition in the remainder of the paper) or as audio (the "Audio" condition). We further designed and conducted a follow-up laboratory experiment (*LAB*) where we were able to confirm the differences in user choices discovered in the *AMT* study. Additionally, we collected rich qualitative data explaining the challenges in perception of the audio captions. Throughout both parts of this investigation (*ATM* and *LAB*), we used the same search tasks and interfaces as our building blocks, which we go on to describe in the upcoming sections.

3.1 Search tasks

Search task complexity was shown to affect user behaviour [18]. To account for it, we selected six search tasks of varying complexity. Following prior research [4, 33], we adopted three levels of complexity from the taxonomy defined by Anderson and Krathwohl [3]: *remember (R), understand (U)*, and *analyse (A)*. To provide our participants with a detailed description of the supposed information need, we used backstories Specifically, we used tasks 2, 7, 9, 18, 34, 39, and 140, from Bailey et al. [4]. For example, task 2 asked about "the potential health benefits of marine vegetation" with a task complexity level of *"understand"* (U).

3.2 Search results

For each search task, we collected five unique search results. To generate these search results, we submitted the "search topic" as a search query to Google and collected the 1st, 5th, 10th, 50th, and 100th search results, with the assumption that the 1st, 5th, and 10th results will be more relevant than 50th and 100th. We skipped results linked to PDF files and instead collected the next ranked result. For queries that yielded less than 100 results, we used the last one as 100th result. For each result, we collected the displayed title, URL, and snippet.[1]

3.3 Interfaces

To study the differences between text and voice representations of search captions, we created Text and Audio interfaces, as shown in Figure 1. Both interfaces displayed the task topic, followed by its corresponding backstory from Bailey et al. [4]. Below, five search results were displayed in random order to ameliorate participants' bias towards the top-ranked result [16]. For each task, we instructed participants to select three results: one they considered to be the most useful (i.e., the one they would click on first), the second most useful, and the least useful one. We denoted results using letters A-E to avoid the confusion between notations "best" and "first". The bottom portion of the page displayed three sets of radio buttons, with options A-E, where the participants could make their selection.

Text Condition. For the Text condition (Figure 1a), we reproduced the general look of Google's search engine result page with similar fonts and colors, to make the interface more familiar to participants. To restrict the information available to users, we made the captions non-clickable.

Audio Condition. The search captions in the Audio condition (Figure 1b) were displayed through five identical play/stop buttons. The Audio interface provided a possibility to start and stop audio playback. We refrained from providing users with any additional control or information about playback (e.g. current position) to emulate the limited control over audio playback one might experience with a voice-only system (e.g. such as when driving).

We generated audio captions by combining the components of each search caption: the title, top-level domain of the URL, and snippet. We replaced ellipses in the snippets with periods. To produce audio clips, we used Google's TTS engine with en-US-Wavenet-A voice.[2] We recorded 30 audio clips — five for each of the six search tasks — with duration ranging from 11 to 29 seconds (median=16, IQR=6). Figure 2 illustrates a text result, and a corresponding audio result. We expected such representation of audio search results to be less than ideal [36]; nevertheless, it suffices for understanding the users' perception of audio-based search captions and what features are important in designing high-quality audio captions.

3.4 Procedure

To answer the research questions outlined above, we designed a user study in two parts: a crowdsourced study (*AMT*) and a controlled laboratory experiment (*LAB*). Both studies were approved through the ethics approval process at the University of Waterloo for research involving human participants.

The design for both studies crossed two main factors: medium (two levels – Text and Audio) and complexity (three levels – remember, understand and analyse). We counterbalanced the order of

[1]This dataset with task details is available at https://github.com/sashavtyurina/audio-serp-ictir-2020.
[2]https://cloud.google.com/text-to-speech/

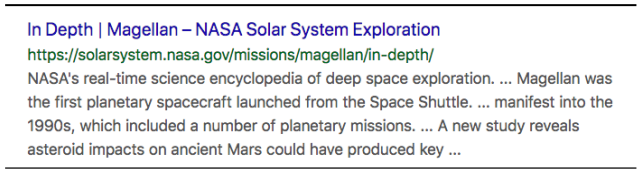


Figure 2: Text snippet and a corresponding audio snippet. The audio result is generated by concatenating the text result's title, the word "From", the text result's domain, and the text result's snippet.

search tasks and interfaces, rotating them in a Latin square design, such that each task occurred with every interface. Each participant was exposed to both Text and Audio conditions. For each task, participants were asked to select: (1) the most useful result (i.e., the one they would click on first); (2) the second most useful result, and (3) the least useful result for the task.

Part 1 - AMT. We began by conducting a within-subject crowdsourced study on Amazon Mechanical Turk.Each participant completed two tasks in Text condition and three in Audio condition. One of the Audio tasks served as a quality check as described below. The study took on average 21 minutes. Data from crowdworkers who failed the quality check was excluded, but all crowdworkers were paid $10 regardless of the quality of their submissions.

To ensure a high quality of submissions, we restricted the participant pool to workers with approval rating 95% or higher, who have completed more than 1,000 HITs, and lived in the US. Additionally, we included an "golden" task, which was presented at the end of the study and always in Audio condition. The search results for this task included 1st and 5th hits from Google, and three non-relevant results. We considered a submission to be of an acceptable quality if the two relevant results were selected as the two most useful ones. Finally, we discarded judgements of the workers who did not listen to all clips in the Audio condition.

Part 2 - LAB. In the second part of the study, we aimed to examine the differences in selected search results in Text vs. Audio conditions to confirm our *AMT* study findings. Additionally, we investigate the reasons for the discovered differences. The *LAB* study followed a similar procedure to the *AMT* study, but with added semi-structured interviews to capture the fine-grained information about users' perception of the audio captions.

After providing their consent, participants completed one training tasks with each interface (Text and Audio). After completing each experimental task, participants completed the NASA-TLX questionnaire — a scale to subjectively assess mental workload [14], measuring mental, physical, and temporal demand, performance, effort, and frustration (we omitted the "physical demand" scale since no physical exertion was involved). After each task, in a short semi-structured interview, participants were asked about the reasons for selecting the most useful result and challenges in perceiving the audio captions. Finally, we asked participants to recall the results they chose as the two most useful ones. Upon completing all six tasks, in a post-study interview, we asked participants about their general impressions of the audio captions and how they could be improved. The study took on average 44 minutes. All participants were reimbursed $10 for their time.

		AMT	LAB
Gender	Male	45	25
	Female	24	11
Age	18-25	8	17
	26-35	31	18
	36-45	19	1
	46-55	6	0
	56+	5	0
Own Smart Speaker	Amazon Echo	23	6
	Google Home	7	12
	None	36	22
	Other	3	0
Use Voice Search	Multiple Times a Day	12	3
	Once a day	5	2
	Multiple Times a Week	19	4
	Once a Week or Less	33	27

Table 1: Participants characteristics.

3.5 Participants

Table 1 illustrates the characteristics of the participants for the *AMT* and *LAB* studies. After removing submissions that did not pass our quality check, we collected data from 69 crowdworkers. For *LAB* study, we recruited 37 participants from the local university, of which the data for one person, who did not fully understand the instructions, were excluded.

4 QUANTITATIVE FINDINGS

4.1 Differences in ranking

To address *RQ1.1*, we study user's search result preference in Text and Audio conditions and answer the following questions:

- Do users make fewer choices that reflect the *true ranking* of results in the Audio condition compared to the Text condition?
- Is the probability of choosing the highest-ranked result as the most useful result lower in the Audio condition compared to the Text condition?

Number of result choices consistent with true ranking. In our experimental setup, participants were asked to select the most, second-most, and least useful results from the five results presented to them. In this setup, we assert that participant choices are consistent with the true ranking of the results (i.e., the ranked result position on Google's SERP) if they have the same relative order. In other words, if their most useful result choice was the top-ranked

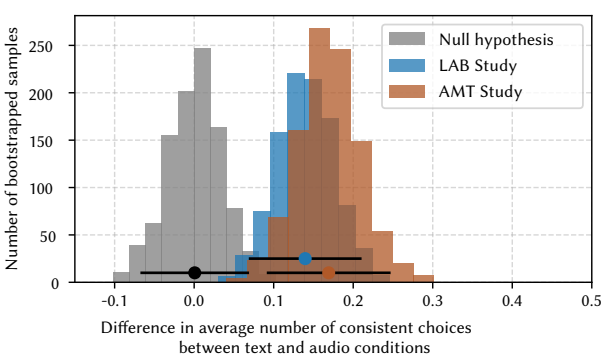


Figure 3: Average difference in number of consistent result choices, between text and audio, under the (a) Null hypothesis, and in the (b) LAB study and AMT study. An average difference higher than 0 means more consistent choices in the text experimental condition compared to the audio condition. Markers indicate mean average difference and 95% confidence intervals for the mean value. Both the Lab study and the AMT study suggest that the text condition leads to more consistent choices compared to the audio condition.

Google result, we assert that the choice is consistent with the result's true ranking. Similarly for their second-most useful result choice if it was the second-highest ranked Google result (in our case, the second-highest ranked Google result is the result at rank five on Google's SERP), and for the least useful result if it was the lowest-ranked Google result presented to them. Therefore, in each of their tasks, our study participants could make between 0 and 3 choices consistent with results' true rankings For example, in our definition, selecting results with true ranks $[1, 5, 100]$ as most, second-most and least-useful is equivalent with making 3 consistent result choices, whereas selecting results with true ranks $[10, 1, 50]$ is equivalent with 0 consistent choices. Consequently, we aim to determine whether participants make fewer consistent choices in the Audio condition compared to the Text condition.

To test whether differences between our experimental conditions (Text or Audio) are meaningful, we bootstrap a test statistic using data collected in our experiments — in this case, we bootstrap the average difference in the number of consistent choices between the two conditions [6]. To achieve this, we compute the number of consistent choices in both Audio and Text conditions, using our experimental data, then repeatedly ($N = 1000$) sample with replacement from the two conditions, subtract the two samples (i.e., Text samples minus Audio samples) and then compute the average difference between the two samples. We then repeat this procedure ($M = 1000$). This method allows us to compute the sampling distribution of the average difference in the number of consistent choices between the two conditions. Similarly, to compute the distribution of the average difference under the null hypothesis (i.e., when there are no differences between experimental conditions), we conduct the same procedure but sample only from the Text condition.

Figure 3 shows the results of our bootstrap test. Both the *LAB* data and the *AMT* data suggest that users make more choices consistent

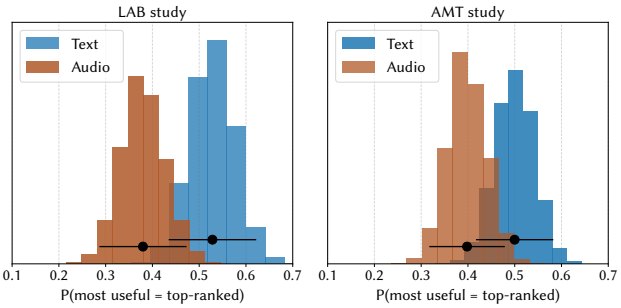


Figure 4: Probability of selecting the top-ranked Google result as *most useful* (Text and Audio conditions). *LAB* study (left) and *AMT* study (right). Markers indicate mean probability and 95% confidence intervals for the mean value.

with result true ranking in the Text condition compared to the Audio condition, on average — for the *LAB* study, the mean average difference is 0.17 (SD = 0.03), whereas for the *AMT* study, the mean average difference is 0.16 (SD = 0.03). This is indicated by the fact that the distribution for the average difference statistic (in both studies) is entirely positive. Furthermore, the mean average difference and its associated 95% confidence interval, in both studies, is entirely positive — under the null hypothesis this difference is expected to be 0 — and therefore we reject the null hypothesis of no differences between Text and Audio conditions with respect to the average number of consistent choices at the 95% confidence level. These findings suggest that, on average, participants make one more result choice consistent with result true ranking in the Text condition compared to the Audio condition every five selections (mean average difference ~ 0.2). The measured difference is unlikely due to chance or noise in our experimental data.

Probability of selecting the top-ranked result as most useful. In addition to differences in the average number of consistent choices, we also analyzed differences in users' ability to identify the most useful (i.e., highest ranked Google result) result in both Text and Audio conditions. To this end, we modelled the probability of choosing the highest ranked result as most useful, in both conditions, using logistic regression. Specifically, we modelled $log(\frac{p}{1-p}) = \alpha + \beta \cdot isAudio$, where p is the probability of the most useful result being the top-ranked Google result, and $isAudio$ is an indicator variable for the audio condition, and fit one separate model for each study. After fitting the models, we use the regression coefficient distributions to compute the probability of selecting the top-ranked result as most useful, in each of the experimental conditions, across the two studies we conducted. Figure 4 shows the distribution of these probabilities, together with their mean values and associated 95% confidence intervals.

In both our studies, differences between Text and Audio conditions related to users' ability to identify the top-ranked search result as most useful are present, but not significant. As shown in Figure 4, confidence intervals for the mean probability of identifying the top-ranked result overlap in the two conditions. This finding leads us to think that even with the rough brute force audio representation, users are able to successfully select the best result

compared to the familiar text interface. We therefore assume that further improvements in audio representation of document surrogates can result in the emergence of successful voice-based search interfaces.

Effects of task complexity. To answer *RQ2.1*, we investigate interaction between task complexity and presentation medium (Text or Audio) in terms of users' ability to identify the most useful result. We extend our regression analysis from the previous section to include additional factors that encode our manipulations of task complexity. Specifically, we modelled the log-odds of selecting the top-ranked result as most useful using: $log(\frac{p}{1-p}) = \alpha + \beta \cdot isAudio + \delta \cdot complexity + \gamma \cdot isAudio \cdot complexity$ (where complexity is encoded using a dummy variable with two levels). We note that, although complexity has a main effect on the probability of selecting the top-ranked result as most useful (with the *Understand* complexity level leading to fewest most useful choices consistent with true result ranking), our analysis did not reveal an interaction effect between task complexity and the medium.

4.2 Perceived workload

In this section, we address *RQ1.2* and *RQ2.2* and investigate whether the task complexity and the medium influence the users' perceived workload. In the *LAB* study, after completing each task, we asked participants to fill out a NASA-TLX questionnaire [14]. We omitted the physical scale since the task did not assume any physical exertion. The mental and temporal demand, effort, and frustration scales in the NASA-TLX range from 0 (low) to 100 (high); and the performance scale ranges from 0 (good) to 100 (poor).

We found that participants estimated that Audio tasks were more demanding than Text across all scales. Since the scores were not normally distributed, we used the Wilcoxon Signed Rank test (W) to check whether there are significant differences in scores between Text and Audio conditions. We found that there were significant differences between all five scales and medium effect size (calculated using Cohen's d) [20], as shown in Table 2. These results support prior findings of audio interfaces being more cognitively demanding [29].

To estimate whether the complexity of the tasks had an effect on the estimated workload, we used a linear mixed-effect model [5] with the medium and the task complexity as main factors, and participant ID and task ID as random factors. We did not find that the task complexity significantly contributed to the difference in NASA-TLX scores, or that there was an interaction between the task complexity and the medium.

5 QUALITATIVE FINDINGS

In this section, we answer *RQ3*: "What aspects of audio-based search results are important for the accurate assessment of relevance by the user?" To address this question, we conducted a set of semi-structured interviews as part of the *LAB* study.

After completing each task, we asked participants what attracted them in the most and second most useful results, and whether they could recall the results they selected as the two most useful ones. After each Audio task, we asked participants about the challenges in comprehending audio captions and how they could be improved.

Table 2: NASA-TLX results for the *LAB* study. Wilcoxon Signed Rank (W) test showed that for all scales the differences in scores between Audio and Text conditions are unlikely due to chance. Cohen's d (d) values correspond to medium effect size [20].

TLX Scale	Text		Audio		W	p	d
	Med	IQR	Med	IQR			
Temporal	22.5	36.25	45.0	45.0	464.5	< 0.001	0.61
Mental	32.5	41.25	55.0	40.0	992.0	< 0.001	0.56
Effort	30.0	40.0	55.0	40.0	1028.0	< 0.001	0.58
Perf.	20.0	25.0	30.0	35.0	1140.5	< 0.001	0.52
Frustration	20.0	30.0	40.0	40.0	791.5	< 0.001	0.60

During the study, with participants' consent, we recorded the interviews. Three researchers then analysed the transcribed interviews and jointly developed a codebook using the method of affinity diagramming [21]. In this section, we report on findings and observations that resulted from this analysis. We outline the participants' perceived challenges regarding the audio results, including some of the behavioural patterns that could be important in designing voice-based search systems.

5.1 Navigation shortcuts

When discussing the selected results with participants, we noticed an interesting trend. Participants tended to refer back to the results using a word, a phrase, or the result's source. Often it was a short "handle" that they associated with the result while listening to/reading it. For example, P17 said, *"The first one is Zimbabwe one, and... I think I clicked the Philadelphia one."* Similarly, P13 said, *"The last one was about Tunisia"*. Interestingly, the "handle" was not always topically relevant to the task at hand, rather it could be a word or a phrase that stood out to the participant, for example, P11: *"The best one was the brief history one"*. Twenty-seven people used a single word to refer to a result they saw/heard at least once during the experiment.

Twenty people used a multi-word phrase for the same purpose. For example, P11 said, *"The second one was the Tallest Buildings in North America"*. The source could also serve as a "handle", and twenty-nine people used the source to talk about the results, such as P13: *"I think the third one was ScienceDirect"*, and P2: *"I chose the NASA one as the best one, and then the one from "the weather network" as the second best one"*. Additionally, twelve people talked about a specific search result referring to a genre of the result, such as P11: *"It's something of a research study"*, and P9: *"The best one was from a travel website"*.

In an end-to-end voice-based web search system, the users will ultimately select a result to hear fully, the voice equivalent of clicking on a result. Additionally, one can envision a scenario in which a user might ask to hear a certain caption again. To facilitate smooth navigation and to understand which result the user is referencing, the system should be aware of the contents of the results it returns, providing a clear and natural way for referencing them.

5.2 Challenges with audio results perception

Each participant in the *LAB* study completed three Audio tasks with five results per task, listening in total to fifteen audio clips. We generated the audio results from text captions by concatenating the title, top-level domain, and the summary as demonstrated in Figure 2. Below we discuss the challenges in the perception of audio results raised by our participants.

Uncertainty about caption structure. The structure and contents of the captions should be made clear. Some users found it challenging to understand how the audio captions were constructed and what information to expect from them. In particular, some participants had difficulty distinguishing between the title, URL, and the snippet when listening to the audio results. P2 said,*"The URLs and the sources they kind of like blended into actual information"*. When P5 was asked whether the URL played had an effect on the choice of the best result, they were surprised replying: *"Was there a URL there?"* Perhaps this problem could be mitigated by amending the captions to clearly indicate the roles of the constituent parts, or by varying the prosody of the generated audio [8].

Monotonicity of the audio. Prosodic features of the audio captions should be varied. Seven of our participants reported that monotonous audio was difficult to comprehend. As P18 says, *"It was very monotone, washing over me"*. Furthermore, different audio features can be used to separate the components of the result. According to P17, *"Sometimes it's hard to know whether it's talking about the source or if it's the summary. So just having that distinction by pausing a little bit... would be really helpful"*. Future work could explore the influence of varying pitch, speaking speed, and pauses on the comprehension level. Similar concerns motivated Chuklin et al. [8] and Winters et al. [35] who used sonification techniques to attract attention to certain text and increase user engagement with content.

Uncertainty about clip duration. Users should be made aware of the duration of the audio captions. Another source of uncertainty was the unknown length of the audio captions. P6 compared the experience to Instagram videos: *"I couldn't tell when it was going to stop... It's why Instagram videos suck — you can't see how far along you are in the video"*. P10 put forward an idea of starting a clip with an audio signal, where the volume would indicate how long the clip will be. Perhaps a length of the signal, rather than volume, can be used to achieve this goal.

Abbreviations. Abbreviations and punctuation should be avoided whenever possible. As noted by eight of our participants, URLs consisting of several subdomains (e.g., "plus.maths.org",), or containing abbreviations (e.g., "AMNH" standing for "American Museum of Natural History") were difficult to parse and were a cause of frustration. For example, P11 says, *"...when somebody's speaking like double-u double-u double-u dot Wikipedia, you're like noooo. Probably not the easiest"*. However, the source of the result was an important consideration, with thirty-two participants mentioning that they paid attention to the source when making their relevance judgements. Using the name of the website can be considered an alternative way to represent the source. P13 provides an example: *"Just give me the name of the website, just say 'Wikipedia', just say 'NASA', whatever it was, I don't need the URL"*.

Truncated sentences. Though truncated sentences are used to save screen real-estate in visual search, they can be a cause of disruption in audio-only environment. Fourteen people noted that sentences cut off abruptly before communicating important information about the result. P13 said, *"It started to talk about the planets and then it went to dot dot dot and... I feel like they were getting there. So the 'dot dot dot' was not in the right place"*. The clipped sentences made it hard to judge the relevance of the result. P18 provides an example: *"This one on the ScienceDirect using algae and marine vegetation looked like it could have been promising, but then it cuts off, so not sure"*. In contrast, truncated sentences were not mentioned as a disruption for the text interface. As part of future work, we suggest experimenting with snippets consisting of full sentences as well as sentences that form a coherent story to ease the comprehension of audio captions.

Repetitions. Audio captions should avoid repetitions. According to our participants, repetitive terms tend to make the experience frustrating. We found that such repetitions may occur due to different reasons. First, a snippet — normally the longest part of the search result caption — might contain repetitive terms, as noted by P1: *"It was pretty annoying because it started off with something like 'action plan'... 'implementation of the action plan', just kept saying those couple of words again and again. So that was frustrating."* Additionally, repetitions may be caused by the overlapping terms between the different parts of the caption. For example, P13 said: *"He said the URL, or something like that, and then he repeated the title which was the exact same thing as the URL"*. Interestingly, no such comments were made for text condition, though the content was identical, which leads us to assume that audio is a more sensitive medium in this respect.

Cognitive load. Finally, as supported by NASA-TLX responses in Table 2, we observed that our participants considered tasks in the Audio condition to be more mentally demanding than the ones in the Text condition. Due to the linear and non-persistent nature of audio, fifteen people noted that they had to pay constant attention to the audio captions to not miss an important part. For example, P19 indicated, *"I had to carefully listen to the audio. And when I'm listening to audio, I feel like this is the only chance I'm listening to it"*. Skimming through results was impossible in the Audio interface, which was noted by sixteen people, who said that reading through results felt faster than listening to them. P12 provided an example: *"I can browse through the results quicker visually. And I'm able to pick out keywords"*. It is not unlikely that the level of mental effort is dependant on the users' working memory: prior work found an effect between the level of working memory and the outcome of a search process [7]. Lack of control over the pace of the speech was pointed out as a downside of the audio captions by eight participants. This aspect was previously discussed by Abdolrahmani et al. [2], who reflected on the need for more advanced features for voice assistants. Such functionality was recently introduced by Amazon, enabling Alexa to speak faster or slower on user's request [30].

6 CONCLUSION

In this paper, we investigated whether the medium (Text or Audio) over which search result captions are presented has an effect on users result preference and their perceived workload. The findings

of a crowdsourced (*AMT*) and a laboratory (*LAB*) studies confirm that there is indeed a significant difference between the choices users make depending on the medium: user relevance judgements in Text condition are significantly more consistent with the results' *true* ranking than those in Audio condition. However, the differences in picking only the most useful result were not significant. Additionally, the results of the *LAB* study showed that the user-perceived workload was significantly higher when working in Audio condition compared to Text condition. We did not find an effect of task complexity on either the users' search result preferences or the perceived workload.

Though our analysis did not reveal an interaction between the medium (Text or Audio) and the task complexity, however, further research is required to check this assumption. Notably, the queries we used came from a standardised dataset and were not reflective of the users' information needs, in addition, the lack of the full search process (query reformulation and access to full documents) could have produced such effect. Notably, our qualitative analysis of rich interview data with the *LAB* study participants revealed a number of aspects that should be considered by designers of future end-to-end voice-based search systems. Such systems should:

- Be aware of the content the system is returning. For navigational purposes, users require a shorthand method for referring to search results, such as the name of the source (e.g., "Play more from NASA.") or the type of the source (e.g., "Let's hear more from the travel website.").
- Clearly indicate the constituent parts of the search caption: a title, a URL, and a snippet. Beyond its use for navigation, a clear statement of the source might help users to assess the quality and authoritativeness of the results, particularly for more cognitively demanding tasks.
- Clearly indicate the duration of the audio clip representing a search result. Users should be aware of caption length to assist them in deciding whether to stop playback or to listen until the end.
- Use prosodic features to avoid monotone voice. Appropriate breaks and changes in pitch can help emphasise the structure and highlight the keywords.
- Avoid abbreviations in the search results.
- Avoid truncation... Results should be reported in full sentences.
- Avoid repetitive terms in the audio results.

REFERENCES

[1] Ali Abdolrahmani and Ravi Kuber. 2016. Should I trust it when I cannot see it? Credibility assessment for blind web users. In *18th SIGACCESS*. 191–199.

[2] Ali Abdolrahmani, Ravi Kuber, and Stacy M. Branham. 2018. Siri talks at you: An empirical investigation of voice-activated personal assistant (VAPA) usage by individuals who are blind. *ACM SIGACCESS* (2018), 249–258.

[3] Lorin W. Anderson and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

[4] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. *SIGIR* (2015), 625–634.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48.

[6] Michael R Chernick, Wenceslao González-Manteiga, Rosa M Crujeiras, and Erniel B Barrios. 2011. Bootstrap methods.

[7] Bogeum Choi, Robert Capra, and Jaime Arguello. 2019. The Effects of Working Memory during Search Tasks of Varying Complexity. In *CHIIR*. ACM, 261–265.

[8] Aleksandr Chuklin, Aliaksei Severyn, Johanne R. Trippas, Enrique Alfonseca, Hanna Silen, and Damiano Spina. 2019. Using Audio Transformations to Improve Comprehension in Voice Question Answering. In *CLEF (CLEF'19)*.

[9] Charles L.A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. 2007. The influence of caption features on clickthrough patterns in web search. *SIGIR* (2007), 135–142.

[10] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: An eye-tracking study of information usage in Web search. *CHI* (2007), 407–416.

[11] Vera Demberg, Andi Winterboer, and Johanna D Moore. 2011. A strategy for information presentation in spoken dialog systems. *Computational Linguistics* 37, 3 (2011), 489–539.

[12] Forbes. 2018. Okay, Google, Will Voice Be The Future Of Search? Retrieved October, 2019 from https://www.forbes.com/sites/nicolemartin1/2018/11/06/ok-google-will-voice-be-the-future-of-search/

[13] Ido Guy. 2018. The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries. *ACM TOIS* 36, 3 (2018), 30:1–30:28.

[14] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[15] Marti Hearst. 2009. *Search User Interfaces*. Cambridge University Press.

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. *28th SIGIR* 51, 1 (2005), 154–161.

[17] Michael Kaisser, Marti A. Hearst, and John B. Lowe. 2008. Improving search results quality by customizing summary lengths. *ACL* June (2008), 701–709.

[18] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-Ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. In *ICTIR*. 101–110.

[19] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *CHIIR*. 121–130.

[20] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[21] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[22] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *CHI*. ACM, 5286–5297.

[23] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. 2018. Design guidelines for hands-free speech interaction. In *20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 269–276.

[24] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. 2017. " Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *CHI Extended Abstracts*. 2853–2859.

[25] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *CHIIR*. ACM, 117–126.

[26] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. 2018. Conversational Query Understanding Using Sequence to Sequence Modeling *(WWW '18)*. 1715–1724.

[27] Daniel E. Rose, David Orr, and R. G P Kantamneni. 2007. Summary attributes and perceived search quality. *16th WWW Conference* (2007), 1201–1202.

[28] Paulette M. Rothbauer. 2008. Triangulation. In *The SAGE Encyclopedia of Qualitative Research Methods*, L. Given (Ed.). SAGE Publications, 893–894.

[29] Nuzhah Gooda Sahib, Anastasios Tombros, and Tony Stockman. 2012. A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *JASIS* 63, 2 (2012), 377–391.

[30] theverge.com. 2019. Now you can choose how fast Alexa talks on your Amazon Echo. Retrieved January, 2020 from https://www.theverge.com/2019/8/7/20757749/amazon-alexa-talk-faster-slower-speed-echo

[31] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and Alignment in Information-Seeking Conversation. In *CHIIR*. ACM, 42–51.

[32] Tassos Tombros and Fabio Crestani. 2000. Users' perception of relevance of spoken documents. *JASIS* 51, 10 (2000), 929–939.

[33] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. *CHIIR* (2017), 325–328.

[34] Johanne R Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. 2015. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *38th SIGIR*. ACM, 991–994.

[35] R Michael Winters, Neel Joshi, Edward Cutrell, and Meredith Ringel Morris. 2019. Strategies for auditory display of Social Media. *Ergonomics in Design* 27, 1 (2019), 11–15.

[36] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *CHI*, Vol. 95. 369–376.