Supporting Voice-Based Natural Language Interactions for Information Seeking Tasks of Various Complexity

by

Alexandra Vtyurina

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Max L. Wilson, Associate Professor, Mixed Reality Lab, University of Nottingham
Supervisor(s):	Charles L.A. Clarke, Professor, Cheriton School of Computer Science, University of Waterloo Edith Law, Professor, Cheriton School of Computer Science, University of Waterloo
Internal Members:	Ed Lank, Professor Cheriton School of Computer Science, University of Waterloo Mark Smucker, Associate Professor, Department of Management Sciences, University of Waterloo
Internal-External Member:	Oliver Schneider, Assistant Professor, Department of Management Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This dissertation includes first-authored peer-reviewed material that has appeared in conference and journal proceedings published by the Association for Computing Machinery (ACM). The ACM's policy on reuse of published materials in a dissertation is as follows:[1]

> *"Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included."*

The following list serves as a declaration of the Versions of Record for works included in this dissertation.

Portions of Chapter 4:
Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17). Association for Computing Machinery, New York, NY, USA, 2187–2193.
DOI: https://doi.org/10.1145/3027063.3053175
Research presented in Chapter 4 was conducted at Emory University during the first author's stay there as a Visiting Scholar. Prof. Eugene Agichtein and Prof. Charles L.A. Clarke provided research guidance. Denis Savenkov provided assistance conducting the user study. All co-authors contributed in writing the manuscript.

Portions of Chapter 5:
Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 208, 1–7.
DOI: https://doi.org/10.1145/3173574.3173782
Research presented in Chapter 5 was conducted during the first author's internship at Microsoft Research, Redmond. Adam Fourney provided research guidance and participated in qualitative coding of participant utterances. All co-authors contributed in writing the manuscript.

[1] https://authors.acm.org/author-services/author-rights. Accessed in August 2020.

Portions of Chapter 6:
Alexandra Vtyurina, Charles L.A. Clarke, Edith Law, Johanne Trippas, and Horaţiu Bota. 2020. A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '20).
In press.
Research presented in Chapter 6 was conducted under research guidance of Prof. Charles L.A. Clarke and Prof. Edith Law in collaboration with Johanne Trippas and Horaţiu Bota. Horaţiu Bota provided his expertise on statistical analysis. All co-authors contributed in writing the manuscript.

Portions of Chapter 7:
Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19). Association for Computing Machinery, New York, NY, USA, 414–426.
DOI: https://doi.org/10.1145/3308561.3353773
Research presented in Chapter 7 was conducted during the first author's internship at Microsoft Research, Redmond. Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen White provided research guidance. Meredith Ringel Morris provided her assistance and expertise with hiring participants for the survey and the user study. Adam Fourney participated in coding survey responses. All co-authors contributed in writing the manuscript.

Abstract

Natural language interfaces have seen a steady increase in their popularity over the past decade leading to the ubiquity of digital assistants. Such digital assistants include voice-activated assistants, such as Amazon's Alexa, as well as text-based chat bots that can substitute for a human assistant in business settings (e.g., call centers, retail / banking websites) and at home. The main advantages of such systems are their ease of use and – in the case of voice-activated systems – hands-free interaction.

The majority of tasks undertaken by users of these commercially available voice-based digital assistants are simple in nature, where the responses of the agent are often determined using a rules-based approach. However, such systems have the potential to support users in completing more complex and involved tasks. In this dissertation, I describe experiments investigating user behaviours when interacting with natural language systems and how improvements in design of such systems can benefit the user experience.

Currently available commercial systems tend to be designed in a way to mimic superficial characteristics of a human-to-human conversation. However, the interaction with a digital assistant differs significantly from the interaction between two people, partly due to limitations of the underlying technology such as automatic speech recognition and natural language understanding. As computing technology evolves, it may make interactions with digital assistants resemble those between humans. The first part of this thesis explores how users will perceive the systems that are capable of human-level interaction, how users will behave while communicating with such systems, and new opportunities that may be opened by that behaviour.

Even in the absence of the technology that allows digital assistants to perform on a human level, the digital assistants that are widely adopted by people around the world are found to be beneficial for a number of use-cases. The second part of this thesis describes user studies aiming at enhancing the functionality of digital assistants using the existing level of technology. In particular, chapter 6 focuses on expanding the amount of information a digital assistant is able to deliver using a voice-only channel, and chapter 7 explores how expanded capabilities of voice-based digital assistants would benefit people with visual impairments.

The experiments presented throughout this dissertation produce a set of design guidelines for existing as well as potential future digital assistants. Experiments described in chapters 4, 6, and 7 focus on supporting the task of finding information online, while chapter 5 considers a case of guiding a user through a culinary recipe. The design recommendations provided by this thesis can be generalised in four categories: how naturally

a user can communicate their thoughts to the system, how understandable the system's responses are to the user, how flexible the system's parameters are, and how diverse the information delivered by the system is.

Acknowledgements

I would like to thank all people who helped bring this dissertation into existence. First and foremost, I would like to thank my PhD advisors Charlie Clarke and Edith Law for their support (work-related as well as emotional) and help navigating the waters of grad school. I would not be able to do this without you.

I would like to give my gratitude to the members of my examining committee: Max Wilson, Ed Lank, Mark Smucker, and Oliver Schneider. I look back on the day of my defence fondly – the deep discussions we had left me with a feeling of meaningfulness of my work, and the feedback you provided helped me significantly improve the dissertation. I am additionally thankful to Ed and Mark, whose advice and support helped me answer a find balance in life.

I am immensely greatful to my internship advisors who helped me view the world from a different perspective: to Eugene Agichtein – working together with you was an honour, and I will always think of my time in Atlanta with a smile; to Adam Fourney – developing VERSE together made me feel the rare sense of purpose and fulfillment, and working with you made my summers in Redmond a blast. I am greatful to all my co-authors who helped shape this work – Meredith Morris, Horaţiu Bota, Johanne Trippas, Denis Savenkov, Leah Findlater, and Ryen White. I feel very lucky to have had an opportunity to work side by side with you.

I would like to say thank you to all of the wonderful people I have had a pleasure to be around during these years. I trust that we helped keep each other's mental health in check and the caffeine levels consistently high. Bahareh Sarrafzadeh, my first co-author and teacher, thank you for leading me by example; Alex Williams, thank you for being an unwavering optimist and providing your guidance through the Starbucks drinks menu; Mike Schaekermann, thank you for all the laughter and beers we had together; Greg d'Eon and Blaine Lewis, thank you for the roads cycled, boulders climbed, and Rubik's cubes solved; Damien Masson, Nils Lukas, and Glaucia Melo, thank you for the most hilarious of discussions over coffee at C&D; Nathan Harms, thank you for maintaining my morale; David Maxwell and Johanne Trippas – your advice and encouragement were invaluable; Adam Roegiest and Kira Systems crew – thank you for the support and accommodations.

Thank you to my dear friend and co-author Horaţiu Bota for always being there for me. Thank you to my friends Marianna Rapoport and Abel Nieto for being with me through one of the darkest periods of my life and helping me emerge from it ever stronger. Thank you to Uzma Rehman for teaching me the art of self-acceptance. Thank you to Valentina Vaneeva and Alexey Kon for continuously inspiring me to climb higher, run longer, and achieve more. Thank you to my wonderful friends Anya Sorokina, Katya Davydenko, Pasha Skornyakov, Kolya Shtanko, Dasha Mayorova, Anya TX, Evgeny Pasternaque for being the source of laughter and funny dog pictures.

Thank you to my family for believing in me throughout all these years. Thank you to my parents for teaching me perseverance and patience. Thank you to my sisters Masha and Ira for being my role models. Thank you to my grandmother Natalia who spent countless hours helping me with my school homework – I am happy to say it paid off.

Finally, to my loving partner Carmelo Mastrandrea – thank you for believing in me when I did not believe in myself.

Dedication

To Persei, the woof of my life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Popular culture has thoroughly explored the idea of communication with digital assistants using natural language, with HAL 9000 of "2001: A Space Odyssey" and TARS of "Interstellar" providing just a few of many examples. Continued technological improvements have now made this idea a reality – we can finally communicate to computers much like we do with fellow humans. Increasingly, natural language interfaces, and voice interfaces in particular, are gaining popularity. For example, a survey conducted by Google in 2014 revealed that out of 1400 people, 55% of teenagers and 41% of adults used voice search at least once a day [78]. A 2016 study by Ido Guy showed that search queries submitted over voice resemble natural language questions more so than search queries submitted over text [84].

Advances in technology also promoted the popularity of digital assistants, such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google Assistant. Digital assistants first emerged with the release of Siri by Apple in 2011 as an iPhone-based application. At that point a mixed-modality paradigm was presented: Siri could be activated by voice by uttering a wake word, like "Hey Siri". It would attempt to answer user's request through voice, and in case of falling short of completing the request, Siri would revert to displaying visual information on the phone screen. Figure 2.4 illustrates an example of an exchange where a user's request is addressed by displaying a list of search results on the screen. The number of such digital assistants worldwide will grow from 2.5 billion today to 8 billion by 2023, according to estimates by Juniper Research [138].

Over time, digital assistants were programmed to fulfil more and more user commands using voice only, until eventually the concept of a smart speaker was introduced. Amazon's Echo and Google Home are examples of these and are shown in Figure 2.2. A smart-speaker based digital assistant is still paired with a screen-based device, such as a smart phone or a tablet, where visual information can be displayed, but the majority of the interactions are meant to be conducted using voice only. The popularity of smart speakers has been and is continuing to grow consistently. The Nielsen agency reported that in the second quarter of 2018, 24% of US households owned a smart speaker – a 2% increase compared to the first quarter of 2018 [132]. In January 2020, Voicebot.AI presented the results of a survey confirming that over a third of US adult population – nearly 90 million people – owned a smart speaker [40].

In this thesis, I focus on digital assistants. Strictly speaking, programs like Siri and Alexa provide their users with an opportunity to formulate commands in natural language and issue them through voice, by speaking out loud, and receiving an audio, or sometimes text, response. This process often mimics normal human conversation, therefore a term "conversational assistants" has been coined to describe this technology. A trending example of application of digital assistants is "conversational search" – a process which presumes multiple exchanges between a system and a user in order to retrieve relevant information from the underlying document corpus (or in a more general scenario, the internet). However, it has been argued that the term "conversation" assumes a deeper interaction level than that presented by the current technology and that an interaction between a user and a digital assistant is "conversation-like". I agree with this argument, and consider the term "conversation" not applicable to currently available systems. Instead, I prefer to use the broader term "dialogue". The downside of this term is its breadth: many different processes may be considered a dialogue – issuing text commands in a computer console, using graphical user interfaces, as well as issuing voice commands to a digital assistant.

There are a number of ways to communicate with a digital assistant, with text (typing commands) and speech (speaking commands out loud) being the major ones. Throughout this thesis, I will use terms "text-based assistants" and "voice-based assistants" or "voice assistants" to indicate the primary modality in which an assistant operate, whereas the term "digital assistant" will be used to denote a modality-agnostic assistant. I will focus on studying interactions with voice-based assistants, except for the experiments reported in chapter 4 where a text-based exchange with an assistant was used to simplify the experimental setup.

Though currently available voice assistants are a long way away from the functionality (and malice) of HAL 9000, they are highly integrated with smart home devices, enabling their users to control lights, thermostats, door locks, etc. using voice-only commands. Multiple studies found that voice assistants are used for short simple tasks such as setting up alarms clocks, timers, looking up answers to factoid questions [123, 116] with a noticeable fraction of voice assistant users using them for the sake of entertainment and amusement.

Moreover, the integration of voice assistant with smart home devices and the voice-based interaction they provide, serves as an additional layer of accessibility and provide added independence to people with disabilities [5, 142].

Voice assistants are excellent tools made possible by mixing a number of recent technological advances. However, they are not without their limitations. Many marketing campaigns position voice assistants as a friend in a box that you can "just talk to". This creates inflated user expectations which lead to underwhelming experience [116]. After the first period of curiosity and experimentation, the voice assistant usage tends to taper off. Users grow disillusioned by the lack of the assistants' ability to understand their intent, users identify language that works and stick with it to get reliable responses. Continued technological advancements are likely to improve problems like speech recognition and intent identification. **My first focus in this thesis is to investigate how user perception of digital assistants could change if they were not limited by technological capabilities.**

However, even with inflated expectations, voice assistants have proven to be highly popular among various user groups. The leading reason for people using voice-based assistants is hands-free interaction and therefore the opportunity for multitasking they provide [116]. Some use cases include assistance during cooking process, lounging on the couch while choosing a TV channel, and others. A scenario mentioned above, where a user is redirected from voice-based interaction to a screen-based interaction terminates the voice-based exchange and breaks down the intended flow. Presenting information through an audio-only channel is essential for facilitating seamless and complete voice-based interaction. **Therefore, the second goal of this thesis is to investigate ways in which voice assistants can expand their voice-only interaction using currently existing underlying technology.**

1.1 Research Questions

This thesis poses two main research questions:

- **RQ1**: How would users perceive digital agents that could understand them as well as their fellow humans?
- **RQ2**: How can we improve interaction using currently available tools?

In chapters 4 and 5, I tackle RQ1 and explore "the systems that could be" by imitating them and studying how users react and communicate back to such agents. More specifically, I ask:

- **RQ1-a**: Given equal performance, would people choose to communicate with another person or a digital assistant and why?
- **RQ1-b**: What opportunities are opened by the language people use?
- **RQ1-c**: What aspects of digital assistant design are important to consider?

I tackle RQ2 in chapters 7 and 6, where I experiment with systems built using available tools and address the question of how they can be improved. In particular, I examine how the process of web search can be conducted using primarily voice for general population and for people with visual impairments. Therefore I pose the following sub-questions:

- **RQ2-a**: Does the medium (text/audio) over which search results are delivered affect the user's search result preference?
- **RQ2-b**: What aspects of audio-based search results are important for the accurate assessment of relevance by the user?
- **RQ2-c**: How might voice assistants and screen readers be merged to confer the unique advantages of each technology?

At each step of the investigation, I discovered aspects of system design that emerged repeatedly and played a role in how the interaction with a digital assistant is assessed by users. I consider these design elements essential in developing future digital assistants. By consolidating the results from different user studies, I developed the following thesis statement:

> *A productive interaction with a dialogue system critically depends on how naturally a user can communicate their intent to the system, the understandability of the system's responses, the flexibility of the system's parameters, and the diversity of information accessible through the system.*

I describe each of the aspects in details below while table 1.1 indicates the thesis chapters in which each of the aspects is addressed.

	Chapter 4	Chapter 5	Chapter 6	Chapter 7
User's ability to communicate their intent to the system	x	x	x	x
Understandability of system's responses	x		x	
Flexibility of parameters	x		x	x
Diversity of information	x			x

Table 1.1: Diagram of the thesis structure. Each chapter covers at least one interaction aspect.

User's ability to communicate their intent to the system. Naturalness of interaction does not necessarily imply natural language interaction, rather, how an agent's functionality matches users' mental model of it. Here, I note that certain features used in a human-to-human conversation can be used to enhance the naturalness of the interaction between an agent and a user. One such feature is contextual awareness. In chapters 4, 6, and 7, I illustrate that the ability of an agent to keep the memory of the information mentioned previously (by both the user and the agent itself) is a positive aspect. For example, in chapter 4 persistent memory of past interactions allowed participants to forego formulating full self-sufficient questions at every step and instead allowed them to use partial questions relying on previously appearing information. Furthermore, in chapter 7, participants expressed the desire to be able to use the content of an article for navigation, and in chapter 6, many participants implicitly echoed this sentiment by using content words to talk about search results they heard. I also argue that if presented with an opportunity, users will likely take advantage of the functionality borrowing from the human-to-human conversation. To this point, in chapter 5, I demonstrate that most participants did not exhibit reluctance to employ language characteristic of human-to-human communication when interacting with a digital agent.

Understandability of system's responses. Another part of designing a smooth interaction is providing the information back to the user in the form that is easily understood and interpreted by the user. The requirements for aspect may differ depending on the use case. For example, in chapter 4, information source was required to interpret the credibility of the information. However, in chapter 6, I present a different outlook, where the understandability of agent's responses implies incorporating pauses and tones into an audio response.

Flexibility of parameters. To accommodate a variety of user groups as well as individual users, a system ought to offer a number of settings which can be changed as desired. Throughout the studies described in this thesis, I found that aspects such as the response length returned by the agent (chapter 4), the pitch and speed of voice used to generate auditory responses (chapters 6, 7) are based on preferences. Interestingly, while certain parameters such as answer length can be individual, others, such as speech rate, can be specific to a user group – people with visual impairments will likely prefer audio played at a faster rate compared to sighted people.

Diversity of information. Finally, I point out that access to a variety of information is beneficial to the agent's design. In chapter 7, one of the agent's positive aspects was its ability to return information from multiple search parallels (e.g. news and Wikipedia). In a similar vein, participants from chapter 4 expressed a desire to have the agent provide results from opinion aggregating websites in addition to its regular search capabilities.

The remainder of this thesis is structured as follows. In chapter 2, I discuss prior work that impacted, inspired or otherwise influenced my research. I also outline and provide background on the research methods used throughout this dissertation. In chapters 4 and 5, I describe the work focusing on future dialogue systems design. User studies in both chapters are based on simulated agents and are aimed at exploring the behaviour of the users interacting with the agents. In chapters 6 and 7, I explore the design of dialogue systems that can be built with currently available tools. In this exploration, I take two different angles. First, in 6, I investigate the caveats around designing a fully audio-based dialogue system for web search. I outline the key principles that should be followed when presenting search results in the absence of a screen and compare them with existing recommendations for visually displaying web search results. Afterwards, in chapter 7, I explore a similar problem – designing an audio-only dialogue system for web search, however, this time, I focus on designing specifically for people with visual impairments. Finally, I discuss the findings in chapter 8 and outline possible directions for future work in chapter 9.

Chapter 2

Background and Related Work

In this chapter, I discuss key prior and related work that informed, inspired, and impacted the research presented in this dissertation. I begin by discussing the terminology often applied to describe digital assistants, highlighing the differences between human-to-human conversation and human-machine dialogue. I continue by describing standard architectures for digital assistants, followed by common use cases and attitude of users toward digital assistants. I continue by describing the role voice assistants play for people with disabilities in general, and visual impairments in particular. Finally, I outline the work done in the area of designing of speech and text interfaces for web exploration and search, as well as search tasks typically used in user studies.

2.1 Conversation or Dialogue?

Major commercially available systems operate based on prescribed scenarios and are able to function as long as the user's input is within certain predefined boundaries but are otherwise quite fragile. However, in an attempt to appease the buyers, these commercial systems are designed to mimic certain superficial features of a human conversation. For example, a command for checking the weather can be phrased in a variety of different ways: from "What is the weather forecast?" to "Do I need an umbrella today?". Furthermore, the voices of digital assistants become increasingly similar to those of humans. The 2018 Google Duplex project demonstrated an assistant making a reservation at a restaurant by talking to a person [3]. Such imitation has downsides as well as merits. On one hand, this imitation produces high user expectations who may be quickly underwhelmed after the beginning of the interaction [116]. On the other hand, it leads to an anthropomorphisation of such systems – a trait that has been found beneficial in certain medial applications, in particular for mental health [180, 144].

While digital assistants do mimic certain aspects of a human-to-human conversation, an interaction between a person and a computer system can be considered "conversation-like" but it is not a "conversation" in and of itself [140]. Language is something most of us learn from the young age and conversation is the go-to way of interaction using language. A conversation is considered to be one of the most seamless and intuitive ways to conduct an efficient interaction. Herbert Clark, one of the prominent researchers in the area of conversational analysis, suggests that a human conversation is an innately *collaborative* act without a predefined outcome where each party contributes independently towards a mutual goal [54]. A human conversation encompasses a multitude of parameters making it difficult if not impossible to model with available tools [14].

In this thesis, I avoid the term "conversation" to describe an interaction between a human and a computer system. Instead, I use the term "dialogue" throughout this thesis to describe such interaction. The term "dialogue system" is well suited to describe the type of interaction we consider. However, it is also quite broad, encompassing a wide range of interactive system, such as most graphical user interfaces, a computer terminal, a telephone flight booking systems, and even an elevator. The terminology chosen in this thesis, does differ from that of a large body of research. However, even in prominent HCI and IR venues, it is acknowledged that the systems called "conversational" are currently operating in a limited capacity, for example, "conversational search" systems are mostly functioning as speech-based question-answering systems [61]. Perhaps the usage of the term conversation in this context comes from the desire to one day achieve a truly conversational experience with a computer, rather than reflecting the current capabilities of these systems. In the next sections, I outline the main traits of a human-to-human conversation and discuss architecture of many commercial digital assistants.

2.2 Human Conversation

Before we can teach a computer how to converse like a human, we should obtain a thorough understanding and formalization of how humans converse with each other. A field of studies called conversational analysis is devoted to exploring the first step of this process. The process of conversational analysis begins with collecting a large dataset of conversations (audio or video) and transcribing them in great detail, known as Jefferson transcription [95] and shown in Figure 2.1. Such transcription includes such features as precise timestamps up to one tenth of a second, indications of pauses, intonation, volume, body language, etc.

Ken: I started workin etta buck thirty en hour
(0.4)
Ken: en'e sid that if I work fer a month: yih getta buck,h ·h thi[rty ↓fi:ve=
(Dan): [((sniff))
Ken: ='n hour en (·) ev'ry month he uh ( ) he rai[ses you]°( )°]
Dan: [How'dju]g e t th]e jo:b,
(1.0)
Ken: ↑I js wen' down there'n ↓a:st eem for it
(1.8)
Dan: °Cz° la:st week you were mentioning something about th' fa:ct °thet you ↓u[h°
Ken: [I got ul ụ -got (·) lost in one jub=↓Yea:h.
(0.5)

Figure 2.1: An example of transcription made by Gail Jefferson [95] outlining minute characteristics of speech including intonations, pauses, speech overlaps, etc.

After completing the transcription, researchers iteratively analyse the conversations noting the recurring patters and anomalies. Finally, scientists attempt to generalise and formalise the discovered patterns, leading way to understanding the structure of a conversation.

One of the pioneers of conversational analysis, Harvey Sacks, posited that in its basis, a conversation is an sequence of turns [155]. Sacks and colleagues proposed a model of a conversation having 14 characteristics, including:

- speaker change recurs, or at least occurs,
- turn order is not fixed but varies,
- overwhelmingly, one party talks at a time,
- what parties say is not specified in advance,
- turn size is not fixed but varies,
- length of conversation is not specified in advance,
- number of parties is not specified in advance,
- relative distribution of turns is not specified in advance.

A conversation contains a variety of other moving parts. As mentioned before, Herbert Clark postulated that *grounding* is necessary for a successful conversation, i.e. the speakers must agree on the common ground in order to continue the conversation [53]. Clark posited that a conversation is a collaborative act in which people must share information and mutual beliefs in order to continue the conversation. Furthermore, Marilyn Walker, investigated the shift in the initiative during a conversation – depending on the type of the conversation it may belong to a single speaker (e.g. an interview) or be mixed (e.g. a regular conversation) [182]. One of the additional complexities of spoken conversation analysis lays in the choice of units. Written text analysis can be conducted by splitting the text into sentences, a frequent unit of analysis is an *utterance* – in a dialogue, speakers usually exchange utterances. The complexity, however, is in the fact, that turns, and therefore utterances may overlap, one utterance can span multiple turns. In this thesis, the term "utterance" will be used to denote an uninterrupted block of speech from one partner, or interlocutor.

In the spirit of studying human-to-human conversations in order to model human-computer dialogue, Thomas et al. [175] and Trippas et al. [176] created spoken conversational search datasets, in which they observed how two individuals communicated over an audio-only channel to complete web search tasks. Both papers illustrate how people would approach web search through a conversation if the current technical limitations could be ignored. Prior to this work, Radlinsky and Craswell [147] presented a theoretical model for a conversational search system where they presented a detailed interaction flow and desired system's functionality to deliver a satisfactory user experience.

2.3 Architecture of Dialogue Systems

Digital assistants, and more broadly dialogue systems, can be implemented using two main approaches: corpus-based and rules-based. In a corpus-based approach, a system takes advantage of large datasets of prior conversations or exchanges, often between two people. Such datasets may include human-human phone conversations, movie scripts, exchanges between people on online board and forums, chains of tweets. Some such datasets may be created with a particular goal in mind, and crowdsourcing technique may be used to create these datasets.

Corpus-based systems often produce their responses based on the user's prior input either by finding the most similar message, or using encoder-decoder machine learning model to generate a response based on the underlying database. One may guess that such methods would perform poorly in completing a task, such as booking a table at a restaurant for a particular day and time. Corpus-based systems may be best suited to chatbots – often text-based systems with the main goal of maintaining the interaction for as long as possible.Some chatbots are created for a specific task, for example locating a specific item in an e-commerce store, and in this case the corpus-based approach will not be be suitable. Another caveat for corpus-based systems is that they will reflect the data they have seen before. For example, a Twitter bot called Tay, released by Microsoft, learned from the responses of other Twitter users and had to be taken down after a 16 hours online. This was a result of Twitter users tweeting offensive messages at Tay for the sake of entertainment. The bot soon picked up the unfortunate patterns. This phenomenon echoes a broader issue with training data used to produce machine learning models: Biased data produces biased models which in turn produce biased outcomes [43].

Another approach to modelling an interaction is rules-based. ELIZA [183] and PARRY [57] were the pioneer rules-based text-based systems. Both of them were used in the field of clinical psychology and both were based on regular expression rules. Interestingly, PARRY was the first chatbot to pass a Turing test – study subjects could not confidently distinguish transcripts of conversations between two people from transcripts of exchanges produced from an interaction with PARRY.

Modern digital assistants aim at assisting users with a variety of tasks, making it exceedingly difficult to model interactions with regular expressions. To make these interactions as stable as possible and to avoid breakdowns, people designing these digital assistants must make sure that an assistant is capable of handling various scenarios and user input within the boundaries of a given task. Upon receiving user input, the system first attempts to classify the domain of user request, for example setting up an alarm or booking a flight. If a system is designed to operate within the boundaries of a single domain, this step is not completed. After identifying the general area of interest, the system attempts to identify a specific user intent, for example looking up flight schedule or cancelling a previously booked flight. Each intent is usually paired with a set of slots – variables that are required to be filled in to complete the task. For example, if the system identifies that the user is looking to book a flight, the slots may include flight origin and destination, departure and return dates, preferred departure and return time and so on. Upon identifying the intent, the system attempts to extract some of the values from user's original input and then proceeds asking questions to fill the rest of the slots. For examples, if the user's original request was "I would like to book a flight from Toronto to Vancouver", the system may identify the intent to book a new flight and extract the "Toronto" as an origin and "Vancouver" as a destination. It may then proceed to ask the user about the preferred dates of the flight and whether the flight is meant to be one-way or return. The identification of user intents as well as slot value extraction is a product of machine learning models which are trained to be able to process various phrasings.

2.4 Adoption of Digital Assistants

Digital assistants are known by a number of names in the literature: voice-activated personal assistants (VAPAs) [5], intelligent personal assistants (IPAs) [116], personal digital assistants (PDAs), and voice-activated digital assistants (VADAs). The wide adoption of digital assistants started with the release of Siri as a standalone iOS app in 2010. The application was acquired by Apple soon after, and in just over a year, Siri was integrated in the operating system of iPhone 4S. Since then, voice-activated digital assistants have been gaining popularity, and soon were given their own homes in specialized devices.

The year 2014 marked the beginning of the smart speaker era with the release of Amazon Echo – a smart speaker powered by Alexa – a digital assistant developed by Amazon. The trend was quickly followed by many major tech corporations: Google announced a Google Home [63], Microsoft partnered with Harman Kardon to release Invoke powered by Cortana [164], Apple relocated Siri from an iPhone to a Home Pod [136], Samsung has been expected to unveil their smart speaker powered by Bixby [49]. Yandex released Yandex station with assistant Alice [110, 194].

Figure 2.2 shows Amazon Echo and Google Home – two examples of screenless smart speaker powered by digital assistants. Though the majority of smart speakers do not have a screen, some, such as Amazon's Echo Show come with a display, presumably to overcome the challenge of presenting rich information in an audio-only format as well as to enable additional functionality such as video calls [163, 165]. Nielsen reports that in the second quarter of 2018, 24% of US households owned a smart speaker – a 2% increase compared to the first quarter of 2018 [132].

According to an interview study of voice assistant users [116], one of its most appealing aspects is providing the ability to engage in hands-free interaction and multitasking, i.e., engage in another primary activity while keeping a conversation with a voice assistant as a secondary task. This use case comes to shine when a user's eyes and/or hands are engaged elsewhere (for example, walking or driving) [116]. A number of studies investigated what people use voice assistants for. It was found that voice assistants are often used to complete simple atomic tasks, such as controlling smart home appliances, music playback, setting timers and alarms, checking the weather are among the top categories [15, 26, 123, 179, 116, 106, 26].

Figure 2.2: Screenless smart speakers: Amazon Echo and Google Home

In designing voice assistants,companies use the state-of-the-art text-to-speech technology, and more generally, pursue the strategy of mimicking a human-to-human conversation. In an effort to do so, a number of design tools offer a way to develop interactions based on slot-filling approach, whereby machine learning models are employed to categorise user inputs and synthesise appropriate system output.[1] This tactic may be a double-edged sword, as the increasingly natural speech bolsters users' expectations on the functionality of voice assistants, which at the moment is not nearing the level of human-to-human conversation [116].

Another major concern regarding voice assistants is related to Automatic Speech Recognition, or ASR [101, 124, 116]. In particular, problems may arise when recognizing names [170], accents [116], and children's speech [113]. Troubleshooting problematic interaction has been proven to be be challenging, if not impossible.

To illustrate the difficulties in interaction, one can consider an information finding task as an example. Information requests are reported to constitute a major portion of user interactions with voice assistants [26, 106, 15, 123]. Requests for information can span a number of user intents: asking single-faceted factoid questions, such as "When was Albert Einstein born?". Such questions are typically answered by the assistants with high accuracy, producing a single phrase as an answer, e.g. "Albert Einstein was born on March 14 in 1879". However, multi-faceted queries, queries that are ambiguous, queries that can not be answered with a single sentence or phrase, or queries require deeper research are usually where voice assistants may all fall short. A common strategy is to read back a top-ranked paragraph, or an excerpt from Wikipedia, and when prompted for more information

[1]https://developers.google.com/assistant

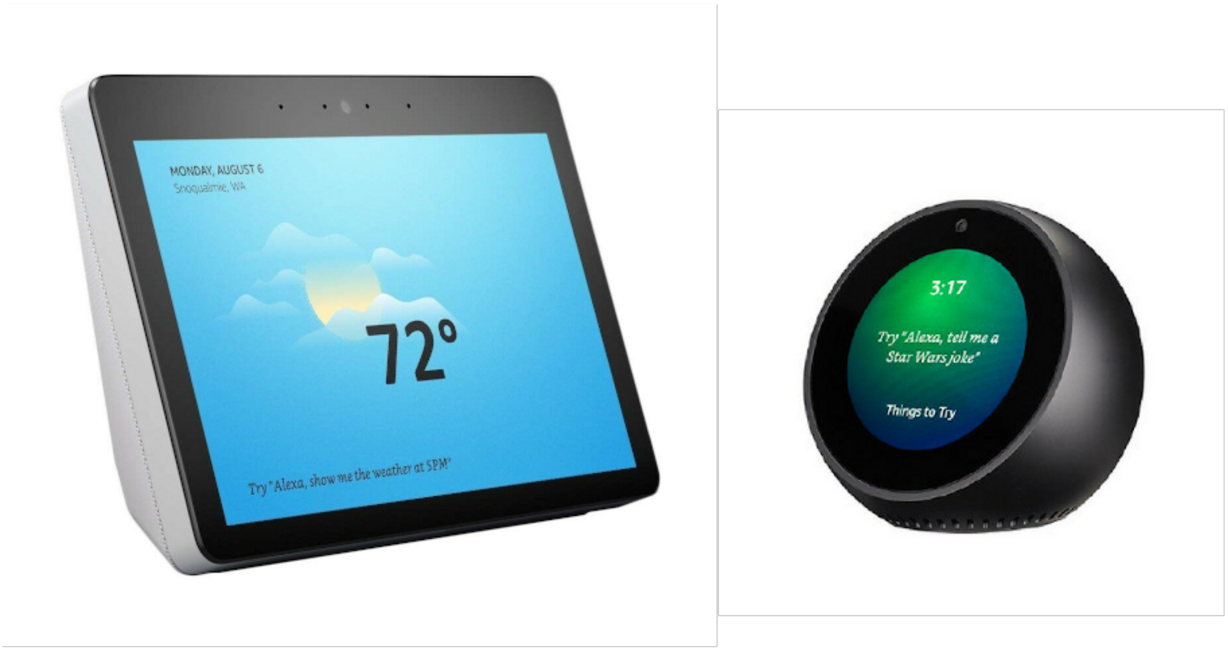


Figure 2.3: Smart speakers with a display for convenient information visualisaion.

there is not much wiggle room. At the time of this writing, Alexa and Google Assistant offer to read more information from the source. When given a more complicated or ambiguous query many assistants will respond with "I cannot answer this right now" or "I searched the web for you, look for the results in the phone app". Both outcomes abruptly end the user interaction leaving a user with no recourse in the first instance, or forcing them to engage with a visual interface in the second. Such experiences disrupt the user's voice experience, forces the user to switch modalities, and can understandably can cause user frustration.

2.4.1 Perception of Digital Assistants

Though increasingly popular, digital assistants have not yet become a social norm yet. Much work has been done in the area of comparing user interactions with a human and a computer. There are varying opinions on the subject. Edwards et al. [68] found no significant differences in how Twitter users treated a social bot, whether it was perceived as a human or not. In turn, Clément and Guitton [56] report that the way bots are perceived varies with the role they play. They found that "invasive" Wikipedia bots received more “polarizing” feedback – both positive and negative – compared to the bots that carried out “silent helper” functions. The similar result is reported by Murgia et al. [130] – Stackoverflow bot receives more negative feedback for false answers when its identity as an automatic program is revealed. Another work by Aharoni and Fridlund [9] reports mixed results from participants who underwent a mock interview with a human and an automatic system. The authors report that there were no explicit differences in the interviewer perception described by the participants, although the authors noticed significant differences in people’s behaviour – when talking to a human interviewer they made greater effort to speak, smiled more, and were more affected by a rejection.

2.5 Auditory Comprehension

Human auditory system is an extremely complex mechanism comprising many moving parts: echoic memory is used for a very short-term sound storage, ascending pathways from the ear to the brain [21]. Sound processing occurs over time and different processes occur at different timestamps. At 20ms a phoneme is recognised, syllabic stress requires 200ms (“melody” vs “melodic”), while 1-2 seconds are needed to recognise the intonation of the sentence, and determine whether a question is being asked.

Every sounds possesses a set of physical characteristics which can be accurately measured. These characteristics include frequency (or pitch) and intensity (or loudness). However, there is more to sound than these physical characteristics. For example, one qualitative feature of a sound is *timbre*, or a quality of sound. Timbre lets us distinguish between sounds produced by different means even if they have the same frequency and intensity (for example, the same note generated by a saxophone or a piano). Timbre is a quality that lets us distinguish between different voices. However, even the “physical” characteristics turn out to carry a qualitative flavour to them, since the comprehension by two individuals of the same sound may vary depending on their training (an opera singer can distinguish slightest pitch variations, while some people have a higher tolerance to louder noises). As such, while sounds have objective measurable physical characteristics, it is important to remember that sound perception varies depending on the individual hearing it.

One important difference between the auditory and visual perception is spatial locality. Vision is inherently spatial. While sound may convey spatial information about it’s source — for example, the intensity difference between the ears can signal direction of the sound – is not primarily a spatial signal, but rather a temporal signal. While visual signals inherently carries spatial information about objects around us, auditory signals is inherently temporal, "with auditory input changing over time" [190], such that a person cannot take a "second look" as one can in reading [90].

While studies of speech perception started in the early 1940s, it is still a relatively new area of investigation. One thing is clear – speech perception and processing involves mechanisms different from regular sound processing. The scientific community cannot come to an agreement on a basic "building block" of speech perception. Simple sound-to-phoneme mapping theory fails to describe this process, because sounds with different physical characteristics can be easily mapped to the same phoneme. People have no problem processing speech produced using different loudness (whisper), pitches (a man or a baby).

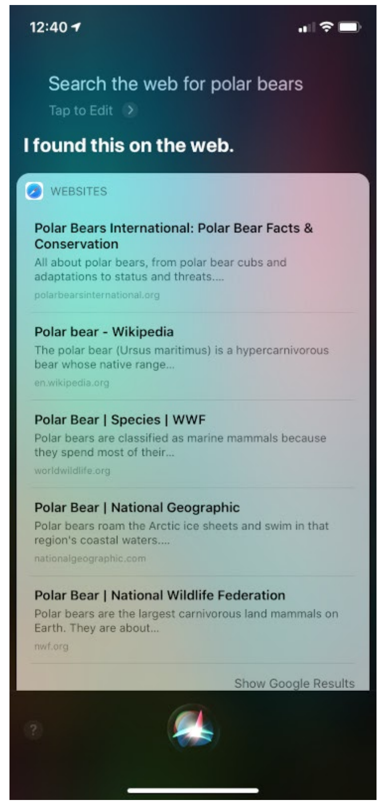


Figure 2.4: Voice assistant Siri reverts to showing information on the screen instead of producing auditory output.

Shannon et al. [168] conducted a "temporal envelope" experiment, in which he and colleagues showed that deteriorated speech, stripped of the frequency values could be easily recognised by the participants. Following this discovery, Binder et al. [32], showed that similar brain activity was incurred by word and non-word sounds, making it likely that the processing of speech is a hierarchical process.

Additionally, the plasticity of our brain allows to reuse visual cortex to process audio signals, thus increasing the characteristics of audio comprehension in people who consume large volumes of information through an auditory channel. Bragg et al. [37] conducted a large-scale study of human listening rates, finding that the mean of intelligible speech was 309 WPM (words per minute) – notably faster compared to the average speaking rate 200WPM [198] (the estimates vary quite a lot from 120 to over 200wpm). The study additionally noted that factors such as age, native language, and sightedness significantly impacted the listening rate. The study exposes a wide variability in the listening abilities throughout the population, suggesting that different experiences need to be tailored to suit each category.

Studies suggest that people consume information differently depending on whether they see it or hear it. After conducting a user study with a voice-based application for simple tasks such as checking email and retrieving a weather forecast, yankelovich1995designing concluded that information in voice interfaces should be organised differently from visual ones. They found that vocabulary, information organization and flow may not translate well between two mediums.

2.6 Voice Interfaces for Accessibility

Though not originally designed for this purpose [129], voice assistants found a wide acceptance in the community of people with disabilities. They were found to provide an additional, or as Pradhan and colleagues put it, "accidental", layer of accessibility [142]. Voice assistants have been noted to support people with visual [7], motor [73], and cognitive impairments [23]. Additionally, prior research showed how they can be used to aid entertainment of young children [75, 114] and support older adults [141, 193, 167]. However, there are still unaddressed challenges in the design of commercial assistants. In this section, I will focus on the obstacles faced by people with visual impairments.

2.6.1 Accessing Web using a Screen Reader

To access visual information many people with visual impairments use screen readers – software that converts visual information to auditory using synthetic speech. Braille display is another tool that serves a similar purpose. It uses tactile feedback to duplicate the text on screen and is essential for deaf-blind population. Each operating system has a built-in screen reader: Narrator on Windows[2], VoiceOver on iOS[3] and OSX[4], TalkBack on Android[5]. Additionally, NVDA[6] and JAWS[7] are stand-alone screen readers that can be freely accessed or purchased. A number of studies exposed a variety of challenges faced by people with visual impairments on the Internet [27, 131, 107, 156, 128, 30].

While guidelines are in place for accessible web design (WCAG[8]), similar accessibility problems persist on the Internet as websites become increasingly more complex and adopt

[2]https://support.microsoft.com/en-ca/help/22798/windows-10-complete-guide-to-narrator
[3]https://www.apple.com/ca/accessibility/iphone/vision/
[4]https://help.apple.com/voiceover/mac/10.14/
[5]https://support.google.com/accessibility/android/answer/6283677
[6]https://www.nvaccess.org/
[7]https://www.freedomscientific.com/Products/Blindness/JAWS
[8]https://www.w3.org/WAI/standards-guidelines/wcag/

new technologies [85]. To illustrate, in 1999 Jonathan Berry outlined found that a lack of text descriptions for images "excluded" screen reader users from accessing the Web and poor page design led to frustration and getting lost. Nearly two decades later, problems remain. Lack of text descriptions for pictures is a widespread issue [77, 31, 160] and one that causes the most user frustration [107].

The "exclusion" aspect is further emphasised in comparison studies between people with visual impairments using screen readers and sighted participants. People with visual impairments were found to spend more time and effort on web-based tasks compared to sighted participants [156, 29]. To gain access to inaccessible content, expert screen reader users also employed advanced techniques [31, 29], including using multiple screen readers, accessing HTML source code, probing (clicking on the link to quickly return back), using search within a page to reach otherwise inaccessible content [156, 29, 31, 34].

Web search engines pose additional unique challenges to screen reader users. Sahib et al. [156] found that blind users may encounter problems at every step of information seeking, and showed lower levels of awareness of some search engine features such as query suggestions, spelling suggestions, and related searches, compared to sighted users. Although these features were accessible according to a technical definition, using them was time consuming and cumbersome [139]. Likewise, Bigham et al. [29] found that blind participants spent significantly longer on search tasks compared to sighted participants, and exhibited more probing behaviour (i.e., "a user leaves and then quickly returns to a page" [29]) showing greater difficulty in triaging search results. Assessing trustworthiness and credibility of search sources can also pose a problem. Abdolrahmani et al. [6, 4] found that blind users use significantly different web page features from sighted users to assess page credibility.

2.6.2 Novel Screen Reader Designs

Another problem with screen reader web access are pages overloaded with banners, ads, menus, as well as unclear structure of the content that makes the task of navigating the web time consuming and strenuous [74, 35, 117]. One approach to overcome this problem is to segment pages into semantically sound sections, and provide access to those sections as opposed to individual HTML elements. Such segmentation can be accomplished based on linguistic [117], visual [83], and hybrid features [92]. Such high-level segmentation reduces the effort and time necessary to navigate web pages using screen readers, and is generally found appealing by users.

Traditional screen readers provide sequential access to web content. Stockman et al. [172] explored how this linear representation can mismatch the document's spatial outline, contributing to high cognitive load for the user. To mitigate this issue, prior research has explored a variety of alternative screen reader designs [148], which is briefly outlined below.

One approach is to use concurrent speech, where several speech channels simultaneously vocalize information [82, 201]. For example, Zhu et al.'s [201] Sasayaki screen reader augments primary output by concurrently whispering meta information to the user.

A method for non-visual skimming presented by Ahmed et al. [10] attempts to emulate visual "glances" that sighted people use to roughly understand the contents of a page. Their results suggest that such non-visual skimming and summarization techniques can be useful for providing screen reader users with an overview of a page.

Khurana et al. [99] created SPRITEs – a system that uses a keyboard to map a spatial outline of the web page in an attempt to overcome the linear nature of screen reader output. All participants in a user evaluation completed tasks as fast as, or faster than, with their regular screen reader.

Another approach, employed by Gadde et al. [74], uses crowdsourcing methods to identify key semantic parts of a page. They developed DASX – a system that transported the users to the desired section using a single shortcut based on these semantic labels; as a result, they saw performance of screen reader users rise significantly. Islam et al. [92] used linguistic and visual features to segment web content into semantic parts. A pilot study showed such segmentation helped the user navigate quickly and skip irrelevant content. Semantic segmentation of web content allows clutter-free access, at the same time reducing the user's cognitive load.

2.6.3 Voice-controlled Screen Readers

Prior work has also explored the use of voice commands to control screen reader actions. Zhong et al. [200] created JustSpeak – a solution for voice control of an Android OS. JustSpeak accepts user voice input, interprets it in the context of metadata available on the screen, tries to identify the requested action, and finally executes this action. The authors outline potential benefits of JustSpeak for blind and sighted users. Ahok et al. [18] implemented CaptiSpeak – a voice-enabled screen reader that is able to recognize commands like "click ⟨name⟩ link," "find ⟨name⟩ button," etc. Twenty participants with visual impairments used CaptiSpeak for the task of online shopping, filling out a university admissions form, finding an ad on Craigslist, and sending an email. CaptiSpeak was found to be more efficient than a regular screen reader. Both JustSpeak and CaptiSpeak reduce the number of user actions needed to accomplish a task by building voice interaction into a screen reader. Chapter 7 investigates a complementary approach, which adds screen-reader-inspired capabilities to VAs, rather than adding voice control to screen readers.

Screen readers usually accept keyboard-based (or gesture-based on touch screens) input and provide audio feedback to the users. Screen readers supply the users with a number of shortcuts or gestures ranging from simple ones for basic features to more complicated ones for advanced features. As such, learning screen reader functionality has a steep learning curve. Adopting new complex technology such as a screen reader can be difficult for people who lost their sight later in their lives. To alleviate this issue, researchers have explored screen readers accepting voice-based input from users. For example, Capti-Speak [18] uses a dialogue model to convert natural language commands to keyboard shortcuts. It was found more efficient compared to a conventional screen reader. JustSpeak [200] and VoiceNavigator [58] enable universal access to Android devices through voice commands. Both were found superior in performance and preferable to conventional screen readers.

2.6.4 Issues with Design of Voice Assistants

Commercial voice assistants are modeled after a human-to-human conversation, striving to provide a "frictionless" user experience [38]. However, in the process of mimicking a human conversation, the needs of people with disabilities are not considered. The analyses of Mukkath et al. [129] and Branham et al. [38] discuss how striving for a natural conversational experience may harm the user experience of people with disabilities and in particular, people with visual impairments. For example, the commercial guidelines for voice assistant designers encourage short conversational turns in order to keep the cognitive complexity low. This conflicts with the way people with visual impairments prefer to give commands – by providing complex and detailed commands [5]. Abdolrahmani et al. [5] additionally report that verbose feedback produced by voice assistants can be redundant and time-consuming. Cowan et al. [59] also discuss the push-back that may occur around the conversational nature of voice assistants, noting that while some people would prefer to personify Siri and chat with it, others would prefer to give it commands. The "natural" speed of a conversation is often considered frustratingly slow by people who are blind [5, 142] due to the fact that people who are blind can often comprehend speech at much higher rates compared to sighted people [37] as well as process multiple audio streams simultaneously [201].

Another concern is speech recognition errors, which is especially acute when dealing with children's speech, deaf or hard of hearing, and people with disabilities because of variability in their speech patterns. Correcting speech recognition errors requires tedious work [142] and is often done using a visual interface. Such modality switch between audio interaction and touch-based interaction represents an significant interruption for voice assistant users for whom hands-free interaction can be essential [59, 116].

The issue of data privacy in interactions with a voice assistant can manifest itself in several ways. First, users might be concerned about the handling and misuse of their personal information by the company-developer of a voice assistant at hand [59]. Additionally, the design of voice assistants may make it difficult to use voice assistants in public without disclosing personal information or passers by overhearing the details the user's personal affairs [5].

2.7 Visual Interfaces for Web Search

As the number of indexed documents on the web is estimated to be in billions [181], web search becomes a ubiquitous and essential tool for navigating the web [157, 89, 189]. Over the years, researchers proposed several frameworks of information seeking process. The model proposed by Marchionini and White [119] consisted of the following stages:

- recognise, accept, and formulate the problem,
- express the problem to a search system in a form of a query,
- examine results,
- reformulate the problem,
- use results.

Current search system strive to support their users at each step of search process. In this section, I will focus on the process examining search results and, in particular, different solutions that have been proposed to aid this process.

A significant body of research was produced in an effort to optimise web search. In their books, Marti Hearst [89] and Max Wilson [189] provide an detailed overview of the past efforts. Over the years of research, the design and the information displayed on the search engine results page, or SERP for short, have undergone a number of changes. However, its main component is a ranked list of results [89, 36]. Traditionally, a SERP consists of a list of documents, or hits. Each hit, is represented by a document surrogate – a combination of the document's metadata, such as page title, url, and a snippet – a brief extract of the relevant part of the document.

Several researchers examined how the design of SERP and its parts can impact searcher preference. For example, Clarke et al. [55] used clickthrough inversions to identify caption features that make search engine result pages more attractive to searchers. They analysed pairs of adjacent search results, where the lower-ranked result received more clicks, to produce a set of guidelines for displaying search results. According to these guidelines, all query terms should appear in the title when possible, but if they appear in the title, they need not be present in the snippet. Additionally, URLs should be displayed in a manner that emphasizes their relevance to the query. Rose et al. [152] pursued a similar goal of identifying positive features of search results. They conducted an online survey, in which participants were asked preference questions about a set of displayed editor-generated captions. Their findings include that users preferred full sentences in the snippets, rather than incomplete sentences, and that user trust was increased by the presence of genre cues (e.g., "official site").

Other researchers focused on manipulating parts of document surrogates. For example, Aula [19] found that presenting a document summary in a form of a bulletted list increased user performance, and boldedness decreased it. Special attention has been given to studying the desired snippet length for search results. Cutrell and Guan [62] varied snippet length for navigational and informational search tasks, finding that longer snippets are detrimental to the former and beneficial to the latter. They used an eye-tracking methodology to determine that the longer snippet tends to draw user attention to itself, whereas the URL, which plays an influential role in navigational tasks, does not receive the same attention for informational tasks.

Following this work, Kaisser et al. [97] found that different answer lengths are preferred depending on the query type. They also found that crowdworkers could successfully predict the desired answer length given a query. Maxwell et al. [122] later investigated how the varying snippet length impacted user experience. They suggested that longer snippets were considered more informative and clear, as well as led participants to engage with the results more, though there was little change in objective accuracy measures. Paek et al. [134] experimented with interfaces presenting varying amount of information.

A number of researchers also explored alternative organisation of SERP. Dumais et al. [66] tried displaying the results grouped by theme instead of a list. White et al. [184] found that displaying a list of highly ranked document sentences rather than summaries can be beneficial. Among other techniques for information organization is Sarrafzadeh et al.'s work [158] that compared visualization of data through networks and hierarchically organized data, concluding that networks led to user reading the underlying document significantly less. Many decisions in these areas are made based on the spatial locations on the page, and how the user's gaze is distributed on the page [44].

As the variety of types of information available on the internet grew to include web pages, maps, images, etc., so did the need to intelligently organise them. In his work, Horatiu Bota [36] investigated the aggregation of different data types and presenting them on a single page.

2.8 Search Task Complexity

Search tasks are a key component in the research and development of information retrieval systems. These tasks provide the goal that users need to achieve with their search and are often used by the researchers to investigate different interaction behaviours depending on the difficulty of the search task [192, 98, 45].

Many researchers in IR have constructed tasks based on the Taxonomy of Learning [16], which allows for investigating tasks from the perspective of cognitive complexity. This taxonomy is traditionally used in educational settings but has more recently been adopted by information retrieval researchers [93]. The Taxonomy of Learning specifies six levels of cognitive complexity as: *remember, understand, apply, analyse, evaluate*, and *create* as seen in Table 2.1.

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 [98, 192, 17]. Furthermore, research by Trippas et al. [177] showed that searchers may engage more in different parts of the search process depending on the task complexity.

Alternative taxonomies split search tasks into different categories. Broder [41] slice the types of search task along different axis according to their intent, presenting three classes: navigational (immediate intent to reach a particular site), informational (intent to acquire some information assumed to be present on one of more web pages), and transactional (the intent is to perform some web mediated activity). According to this taxonomy too user preferences may differ. To this end, Cutrell [62] showed that preferences for design of displayed search results may differ depending on the task at hand. They showed that longer snippets increased searcher efficiency for informational tasks, but had an opposite effect on navigational tasks.

Table 2.1: Task complexity from the Taxonomy of Learning Objectives [16]. In the experiments described in this thesis, three levels were used: Remember, Understand, Analyse (highlighted in bold).

Dimension	**Definition**
Remember	Retrieving, recognising, and recalling relevant knowledge from long-term memory.
Understand	Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarising, inferring, comparing, and explaining.
Apply	Carrying out or using a procedure through executing or implementing.
Analyse	Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organising, and attributing.
Evaluate	Making judgments based on criteria and standards through checking and critiquing.
Create	Putting elements together to form a coherent or functional whole; reorganising elements into a new pattern or structure through generating, planning, or producing.

2.9 Chapter Summary

In this chapter, I reviewed the research that impacted various aspects of this dissertation. The chapter began with an outline of the history of voice-activated digital assistants, their acceptance and usage. It continued with the examining specifics of auditory comprehension and its differences from visual comprehension. It then explored how voice assistants, and more generally, voice interfaces benefit people with visual impairments. Finally, the chapter provided a brief review of search task complexity taxonomies and visual search engine interfaces.

Chapter 3

Methodology

This thesis uses a variety of research methods to answer its research questions. This chapter explains and reflects upon the methods used in the experiments presented in the remaining chapters. I discuss reasons for choosing these specific methods as well as merits and shortcomings of each.

3.1 Wizard-of-Oz

As with much of human behaviour, we tend to purposefully or unintentionally adapt the language we use to match that of our dialogue partner – we speak differently with a toddler, a peer, and an elderly person [76, 51]. This phenomenon also occurs during human-computer dialogues. For example, Eva Luger and colleagues [116] discovered that users of Siri "learn" to phrase their commands in a way that is understood by the assistant and use this language repeatedly afterwards to arrive at the satisfactory results. Because of the limited functionality current digital agents offer, such dialogues are noticeably simpler than human-human ones. While the latter employ complex turn-taking techniques and use secondary communication channel to demonstrate engagement, the former are mostly limited to exchanging explicit commands and requested information. This renders extensive human-human dialogue datasets collected throughout decades of linguistics research unsuitable for the task of analysing human-computer dialogues.

As the research community pushes forward the frontiers of digital assistant research, a multitude of questions arises – would people be polite to computers when speaking to them? Would people use complex language if a dialogue system could correctly react to it? How would children interact with a embodied robot? Many of these questions can be generalised into "How would people behave if a computer could do X". One way to study a question like this is by asking experiment participant to imagine the hypothetical situation and provide their feedback. However, what people *say* and what people *do* famously differ from each other [143]. Another way to approach a question like this is by implementing the system with capabilities in question and observing participants' interaction with it. This choice can prove to be time- and resource-consuming and in certain cases the hypothetical system can be beyond the technological state-of-the-art and simply cannot be implemented at the time.

To remedy situations described above and enable the research of hypothetical or implementation costly systems, a technique known as "Wizard-of-Oz", or WoZ for short, is used. During WoZ studies, subjects are told that they are interacting with an automated system, whereas in fact the system is partially or fully powered by a human operator, or a wizard. Such setup provides researchers with a full control over the information is delivered to the study participants as well the manner in which it is delivered. In a situation like this, participants are likely to exhibit a similar behaviour, for example, use a similar language, that they would when interacting with a fully automated system. While WoZ is popular in studies on natural language systems, for example digital assistants, it can be applied to conduct research of any intelligent interface where the "intelligence" is mimicked by the wizard.

Though the Wizard-of-Oz approach is a popular technique for studying future intelligent systems, a number of concerns should be addressed before choosing to use this methodology. One of the main limitations of WoZ, pointed out by Fraser and Gilbert [72], is participant deception that is frequently involved in the studies. In their WoZ implementation guidelines, Fraser and Gilbert bring up the issue of a potential embarrassment of the participants when the deception is uncovered. From an engineering perspective, Fraser and Gilbert consider an example of a voice dialogue system and discuss the possibility of controlling multiple aspects of the wizard behaviour. Such aspects include restricting the freedom of wizard's output to a set of predefined choices, distorting the wizard's voice during communication with the user. In addition, the design of a WoZ study can be manipulated so that the wizard is in the participant's sight. An example of this setup can be found in the study by Yarosh et al. [196] where children participants interacted with a voice assistant that was manipulated by a person sitting across the table from them behind an opaque separator such that only their upper shoulders and face remained visible. In their 2012 review study, Riek found that WoZ studies of social robots were often used to emulate such aspects as verbal and non-verbal behaviour as well as navigation and mobility skills [151]. The same method is sometimes used in rapid prototyping to simulate parts of a system that have not yet been implemented.

In this thesis, the WoZ approach is used in studies described in chapters 4 and 5 both of which study functionality that was not a part of digital assistants at the time of the studies. In chapter 4, an advanced search system is mimicked by the means of the wizard answering the study participants' requests and questions with the same accuracy that another human would. This technique allowed to isolate and study a single variable – how does the perception of a user change when a person is communicating with a computer compared to another person. In this case, the limitations included a learning effect of the wizard, i.e. the wizard's knowledge of the topics may improved over the course of the experiment resulting in later subjects receiving higher quality of information or receiving it faster. In chapter 5, the WoZ approach was used to simulate a high accuracy digital cooking assistant and study the language study participants used when interacting with it. Whereas a human-level accuracy could not be guaranteed when employing a truly autonomous system. One of the limitations in this experiment was also a potential variability of the system's responses, however the preset answer options were used to reduce the potential variability and alleviate this limitation.

3.2 Workload Assessment

To quantify the amount of effort required to complete a task using an interface, one may use the notion of a workload. Quantifying workload helps estimate how taxing a certain task or interacting with a system is for the user. One of the most popular ways to assess workload is by administering a questionnaire called NASA Task Load Index, or NASA-TLX for short [88]. It was originally developed in 1988 by Sandra G.Hart and colleagues and aimed at assessing the workload imposed by a variety of tasks in aircraft industry. Over years NASA-TLX became widely popular as a method of workload assessment during interaction with a variety of interfaces [87].

The goal of NASA-TLX is not only to provide an assessment of a perceived workload but also identify the factors that most contribute to it. To this end, NASA-TLX uses the scores of the following six scales to evaluate the overall workload:

- mental demand (MD),
- physical demand (PD),
- temporal demand (TD),

- performance (OP),
- effort (EF),
- frustration (FR).

Because each of the six factors above can contribute differently to the final workload score, in her research, Hart proposes to weigh the scales in order of their perceived importance by the study subjects. As such, the NSA-TLX questionnaire is administered in two steps: (1) a subject chooses a score from 0 to 20 for each of the scales above, and (2) the subject conducts fifteen pairwise comparisons of the scales above rank the scales in order of their importance. User scores are converted into a 0 - 100 point scale and weighted accordingly, yielding the final workload score ranging between 0 and 100 points.

In the decades since its development, NASA-TLX has been widely used in a variety of areas and has sustained a number of modifications. The most popular modification is Raw NASA-TLX, or RTLX, where the second step of pairwise scale comparison is omitted and scales are considered to be contributing to the workload equally. In the meta analysis conducted by Hart in 2006, it is noted that different studies have shown that RTLX is more, less, and equally as sensitive as the original NASA-TLX [87]. Other, less frequent modifications, include reformulating the description of the scales according to the context of the task at hand. In her meta-analysis, Hart points out that such modifications are undesirable without additional validation of the scales used.

One of the potential drawbacks of NASA-TLX is the ambiguity in interpretation of the workload scores. In other words, there are no anchor scores pointing to "acceptable" workload, or a so-called "red line" above which the workload is considered too high. To alleviate this issue, a Grier and colleagues [81] conducted a meta-analysis of research studies using any variant of NASA-TLX. The meta-analysis showed that the majority of reported workload scores ranged between 26.08 and 68 points. However, Grier [81] notes that subjects' boredom and frustration also contribute to the workload score. For example, in two studies where subjects had to do nothing but wait, the reported workload scores were 12.0 and 14.8. Whereas the lowest score in the analysis was 6.21 for an air-traffic control task.

In this thesis, RTLX is employed in chapters 6 and 5. The studies described in these chapter do not make an attempt to describe the workload in absolute terms but rather to compare the workload changes between different experimental conditions.

3.3 Crowdsourcing

Crowdsourcing is a practice where freelance workers, called crowdworkers, complete tasks posted by requesters on the crowdsourcing platform. It became especially popular when the rise of machine learning algorithms necessitated the need in large amounts of human-labelled data. Crowdsourcing is also a popular research tools for studies and experiments that require participation from a large number of people. For example, d'Eon in 2019 used crowdsourcing as a tool for identifying people's perception of fair pay in a collaboration task [64].

Crowdsourcing tasks can be voluntary or paid. Voluntary crowdsoucring tasks, often called citizen science, contribute to a bigger project and offer an opportunity to the crowdworkers to contribute their skill and knowledge in order to progress on the project. In this case, people are incentivesed by the project's success, or engaged by its gamified nature. For example, Foldit project [], founded by David Baker is puzzle game aimed at analysing the possible protein folding algorithms submitted by the players. Zooniverse [169] is another example – a collection of citizen science projects covering areas including space, biology, medicine, and humanities. Both Foldit and Zooniverse have proven to be immensely successful, yielding results that would other require decades of affiliated researchers' work to achieve.

On the other hand, paid crowdsourcing tasks differ in offering monetary incentive to crowdwourkers. Platforms like Amazon Mechanical Turk (AMT) and Appen (former Figure Eight and Crowdflower) are the more popular ones. Such platform offer an easy way for requesters to connect with crowdoworkers. Requesters post tasks, called Human Intelligence Tasks, or HITs for short. Each HIT has a set of associated requirements for crowdworker eligibility that could include geographic region, native language, performance rating, and others. Once a crowdworker completes a HIT, the requester has an option to accept the work and pay the fee, or reject the work because of a low submission quality. A crowdworker's performance rating then is based on the percentage of accepted HITs they have completed. On the other hand, crowdworkers can choose HITs they would like to complete, for example based on the minimum payment [188].

Because of its low entrance threshold, crowdsourcing has become a full-time work for many. Social communities and forums contain a multitude of tips and automation scripts to streamline the work, list untrustworthy requesters, and share overall experiences. However, the payment workers receive per hour is often below minimum wage due to factors like requesters under-estimating their HIT's duration, time spent searching for HITs, working on rejected HITs [187, 86]. In this work, my colleagues and I aimed to provide fair payment based on a $15 per hour wage.

Due to the nature of crowdsoucring, there may be a significant chance of low quality submissions. To avoid this, requesters may set minimum conditions for people who are able to accept and complete their tasks, such as performance rating. In addition to the minimum requirements, a popular practice is to include a “golden task”, or an attention check task, where there exists a single unambiguosly correct answer. An example of such task can be asking to enter a specific word in a text box. Submissions that fail attention check tasks help requesters leverage the final quality of the collected data.

3.4 Other Methods

In addition to the methods described above, other methods were used to facilitate experiments which contributed to this thesis: online survey, usability study, controlled experiments, System Usability Scale (SUS).

3.4.1 Controlled Experiment

Often research questions are based around a hypothesis – a statement that proved or disproved by running an experiment. A hypothesis is usually centered around a specific aspect, or an independent variable. In order to test the hypothesis, an experimenter should change that variable while keeping all other conditions fixed. For example, in chapter 4 one of the hypotheses is as follows: people will rate the same system differently depending on whether they think it is automatic or human-powered. In order to test whether this hypothesis is supported, only one variable should change – whether the subjects think they are interacting with an automatic or a human-powered system while keeping all other factors constant. By changing the value of the independent variable, experimental conditions are created. In order to avoid what is called a “carryover effect” – a situation where the order of conditions influences the outcome of the experiment, Latin square design is often used to counterbalance the order of experimental conditions. Latin square design provides a way to rotate experimental conditions to avoid order effects, yet does not require to test all possible combinations of independent variables. the controlled experiments utilising Latin square design are used in chapters 4 and 6.

3.4.2 Usability Study

At times, when there is not a specific hypothesis to be tested, a controlled experiment approach is not applicable. This is the case in chapters 5 and 7, where the goal of the

Table 3.1: Diagram of methods used in this thesis.

	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Wizard-of-Oz	x	x		x
NASA-TLX		x	x	
Crowdsourcing			x	
System Usability Scale		x		x
Controlled experiment	x		x	
Usability study	x			x
Online survey				x

experiment is to evaluate the system, uncover its design flaws, and study the behaviour of its users. In the case of a usability study, a study subject is asked to complete a proposed task using a system at hand, while the experimenter provides the instructions and observes the behaviour of the subject. During a usability study, the experimenter may collect quantitative as well as qualitative information about the interaction. While qualitative information may include insights for systems improvements, quantitative information may focus on measurable metrics, such as completion time, number of interactions, etc.

3.4.3 Online Survey

Controlled experiments and usability studies are often conducted in-person and require resources such as the experimenter's time, allocated space and tools. While on one hand, these factors contribute to collecting rich and in-depth information, they also limit the number of potential participants. A survey administered online provides a solution in cases when the sample of participants needs to be relatively large. Even though a survey does not allow for a deep dive into the answers provided by the respondents, a large number of respondents leads to capturing a wider spectrum of experiences and opinions. An online survey is used in chapter 7 to collect information about pros and cons of screen readers to inform the design of Verse.

3.4.4 System Usability Scale (SUS)

A System Usability Scale, or SUS for short, is a questionnaire combining ten statements. A participant is instructed to choose a response on a five point scale ranging from 0 to 4, or from *"strongly agree"* to *"strongly disagree"*. After all responses have been selected, the scores are added and multiplied by 2.5 to produce the final SUS score ranging from 0 to 100. In contrast with NASA-TLX scores, the absolute values of SUS have been interpreted to denote the quality of the system at hand. As such, any system scored 68 or higher is considered to have above average usability, conversely a system scoring below 68 is considered to be below average. In this thesis, SUS is used in chapters 5 and 7 to ensure that the design of the proposed systems was positively perceived by the study subjects.

Chapter 4

Exploring Conversational Search With Humans, Assistants, and Wizards

Digital assistants are often used for everyday tasks such as smart home controls and music playback. Finding information online and looking up answers to questions is another prominent use case [15, 26]. However, commercially available systems have limited capabilities and often respond with the highest ranking answer provided by the search engine. In cases when an answer is not available, the search results are shown on the screen of a paired device as shown in figure 2.4 and discussed in chapter 2. Researchers working in the area of conversational search are putting efforts into mitigating scenarios like this by designing systems that would be able to narrow down the area of user's interest by engaging them in a dialogue, much like a librarian would with a person with a broad interest in mind. Radlinsky and Craswell in [147] proposed a theoretical model of a search system, describing all scenarios and functions it should be able to process in order for users to achieve their goals. However, no such system has been developed yet, therefore it is users will see dialogue as a beneficial way to finding information online. This chapter describes an experiment with a simulated search system, providing information to the user via a conversation-like exchange.

4.1 Motivation

As discussed in chapters 1 and 2, voice-controlled and well as text-based assistants recently became a solid part of the market. Both technologies are becoming increasingly integrated into people's everyday lives. Voice-activated digital assistants are predominantly used for simple tasks such as controlling music, manipulating smart home appliances, and setting up a timer. On the other hand, text-based chat-bots are often designed to perform a single type of task [56, 68]. Nevertheless, both technologies are also being used for information seeking tasks. Web search is consistently found in the top few categories of use cases for voice assistants, while text-based chats are integrated into commercial websites to provide immediate support to the users [15, 26, 179].

Voice assistants are designed to mimic superficial aspects of human-to-human conversation, leading some people to perceive them as digital friends [145]. On the other hand, the process of designing a dialogue system is usually quite scripted and is based on intent recognition and slot-filling. Though the voice assistants' rules-based design does not impact their ability to successfully answer most factoid questions, the challenges begin when a users' intent becomes more complex and requires deeper engagement. The need to expand the abilities of a conversational system is reflected in the work of Braslavsky et al. [39] who underlined the need for asking clarification questions in order to make an information-seeking conversation more efficient. The work by Radlinsky and Craswell [147] describing a theoretical framework for a dialogue-based search system makes a step towards designing dialogue systems capable of maintaining a conversation-like exchange.

While the development of text- and voice-based dialogue systems is underway, the users are making sure to motivate the progress. A growing proportion of web search queries are formulated as natural language questions [135, 112, 20] with an average length of a search query growing from 2.35 terms in 1998 [94] to 3.2 terms in 2016 [84]. This phenomenon can be partially explained by the increased usage of voice interfaces [185] and better question-answering technology embedded in search engines. As such, the demand for natural language interfaces for search is emerging. However, before jumping into implementing additional features for conversational-like search systems, it is important to gain a better understanding what the users' expectations are when interacting with a truly intelligent conversational search agent. It is equally important to anticipate how users might behave when faced with a conversational search system since user feedback is critical for system evaluation and improvements. In this chapter, I focus on text-based interaction and begin to explore the first question of this thesis: **RQ1**: *"How would users perceive digital agents that could understand them as well as their fellow humans?"* In particular, I address the following more specific questions:

- **RQ1-a**: Given equal performance, would people choose to communicate with another person or a digital assistant and why?
- **RQ1-c**: What aspects of digital assistant design are important to consider?

As a system capable of supporting dialogue-based interactions for search does not exist yet, user preferences and behaviours when interacting with such system can be studies by the means of a Wizard-of-Oz protocol. In this chapter, I describe a study comparing three conversational search systems: an existing commercial intelligent assistant, a human expert and a human disguised as an automatic system. A total of 21 participants were recruited for the study each of whom were faced with 3 complex information search tasks. Participants interacted with each system using a text-based messaging application. The results of the study suggest: (1) people do not have biases against automatic conversational systems, as long as their performance is acceptable; (2) existing digital assistants cannot be effectively used for complex information search tasks; (3) by addressing requests from users, even current search systems might be able to improve their effectiveness and usability, with feasible modifications.

In the remainder of this chapter, I discuss the rationale behind choosing to use a Wizard-of-Oz protocol, as well as the search task selections for the study. I go on to describe each of the three conversational search systems and the evaluation metrics used in the study. Finally, I conclude with a discussion of the study's findings.

4.2 Study Design

Three conversational agents were devised to address the research questions above: Wizard, Human, and Automatic. Each agent was assigned a photo to reflect their nature, as seen in Figure 4.1. A total of 21 participants were recruited – all graduate and undergraduate students at Emory University in Atlanta, GA (8 female, 13 male) – to complete three different search tasks selected from the TREC Session track 2014 [46] as seen in Table 4.1. Each participant completed three search tasks and interacted with all three agents, completing one task with one agent. The order of tasks as well as agents was rotated according to a Latin square design resulting in 9 groups (3 agents x 3 tasks). The participants were given a brief description of each agent (discussed below) but were not given any specific instructions on how to communicate with any agent and therefore were free to interact with them in any way they chose. By omitting the instructions, the author and colleagues hoped to observe the way users would choose to interact with an "ideal" agent. Additionally, the absence of specific instructions served to elicit interaction patterns in agreement with users' expectations for each agent type. The participants were allowed to spend up to 10 minutes working on each task, after which they were asked to move on a topical quiz. Each quiz included three questions designed for a specific topic. After seeing the topical quiz questions, the participants were not allowed to talk to the agent again. Doing so ensured that the participants did not have a set of predefined points to cover and questions to ask the agent. After the study was completed, the answers to the topical quizzes submitted by the participants were analysed and evaluated on a scale from 0 to 2.

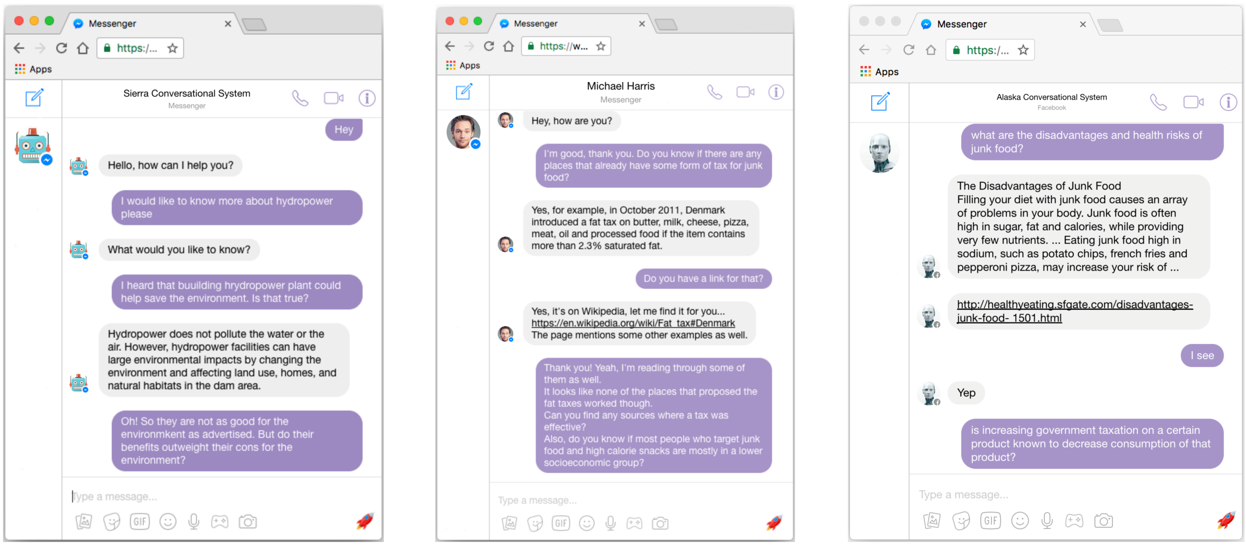
After completing the topical quiz, the participants filled out a questionnaire, where they were asked to rate their overall experience with the agent (on a scale 1 to 5). Participants were also asked what features in particular they liked and disliked about the agent and whether they were able to find all the information they were looking for. Upon completing all the tasks, the participants were explained that one of the agents was powered by a human through a Wizard-of-Oz setup. After being debriefed, participants were asked to choose which system they liked best and why.

Although throughout much of this thesis, I focus on voice-based interaction with dialogue systems, these interactions can introduce unwanted error into an experiment. For example, automatic speech recognition (ASR) errors may contribute to a participant giving the system a lower score. The voice itself can be seen as a confounding variable, since prior research showed that preference for voices varies depending on the individual [149, 102]. To avoid introducing potential mistakes, the interaction in this study was implemented through a text-based exchanges on Facebook Messenger.[1] Participants used a Facebook account created specifically for the purpose of the study. Message history was cleared prior to every experiment. Considering that the purpose of this experiment was to evaluate users' perception of the agents, the study's findings are not specific to the text-based interaction. However, care should be taken in expanding these findings to voice-based dialogue systems, as factors such as response delay and automatic voice characteristics may have an effect on the perception of an agent by the user.

4.2.1 Wizard Agent

This thesis explores human behaviour in human-computer conversational communication. However, there are currently no general purpose intelligent conversational search systems, that could be used for the purposes of this experiment. Therefore one was "faked" by substituting the backend with a person (the author of this thesis and a colleague played

[1] www.messenger.com



(a) Wizard agent (b) Human agent (c) Automatic agent

Figure 4.1: Three agents for conversational search: (4.1a) Wizard, (4.1b) Human, and (4.1c) Automatic.

the role of the wizard behind the curtain). However, the participants were told that it was an experimental *automatic* system, thus following a general Wizard-of-Oz setup. I will be further referring to this system as the Wizard agent, and the person in the backend as the Wizard. The Wizard had previously done the research about the topics of the three search tasks prior to the experiment and compiled a broad set of passages covering most of the aspects of each topic. At the time of the experiment, the Wizard tried to find the best passage to reply to the participant's question or comment. However, in cases where such passage could not be found, the Wizard would reply with a passage retrieved from web search, or write a new passage. In case the participant's question or comment was ambiguous, the Wizard was allowed to ask a clarification question to better identify the information need of the participant.

The Wizard agent was allowed to maintain the context of the conversation, respond to vague questions, understand implied concepts, and provide active feedback in form of clarification questions when needed (all of these capabilities do not yet exist in commercial systems). At the same time, by partially restricting the Wizard to a pre-compiled set of passages, it was possible to maintain the consistency of answers between participants, i.e., for the same question any participant would receive a similar answer. By analyzing the ways the participants communicated with the Wizard agent, the insights were gained about strategies people use in a human-computer dialogue for solving complex tasks and look for design implications for automatic conversational systems thus addressing RQ1-b.

4.2.2 Human Agent

To explore the differences between human-to-human and human-to-computer communication, a second conversational agent was devised – the Human agent. In this case, the Wizard from the previous setup was still serving as a backend, but the participants were explicitly informed that they were talking to a live person. Another difference was that the Human agent was not restricted to the pre-compiled set of passages but was free to slightly reformulate or revise the passages to better respond to the question. By including both the Human and Wizard agents in the study, it was possible to maintain a constant level of intelligence for both agents, thus comparing not the accuracy of each agent, but rather the participants' attitude and expectations towards a perceived automatic agent compared to a known human. The results of this comparison served as evidence to address RQ1-a.

4.2.3 Automatic Agent

As a means of comparison to an existing conversational agent, a Google Assistant was used as a backend for the third agent. Every message sent by a participant was forwarded to the Google Assistant app, and the response was forwarded back to the participant. Most of the time, the response consisted of an URL and a text snippet. The participants were told that they were interacting with another experimental conversational search system, but were not given any specific information about it. Using a system representative of the state-of-the-art technology made it possible to evaluate its drawbacks, and situations where it failed to respond properly. Collecting participants' feedback about each system resulted in a set of design changes that would improve commercial digital assistants for the task of information finding.

4.3 Search Task Selection

The goal in selecting search tasks was to find ones that were likely to require interaction beyond a single question and answer. For this reason TREC Sessions track was selected. TREC is a yearly conference co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. It consists of several tracks, each

Table 4.1: Description of the tasks used in the study. All the tasks were obtained from TREC Session track 2014 [46].

Topic ID	Topic description
10	Suppose you are writing an essay about a tax on "junk food". In your essay, you need to argue whether it's a good idea for a government to tax junk food and high-calorie snacks.
20	You have decided that you want to reduce the use of air conditioning in your house. You've thought that if you could protect the roof being overly hot due to sun exposure, you could keep the house temperature low without the excessive use of air conditioning.
21	Hydropower is considered one of the renewable sources of energy that could replace fossil fuels. Find information about the efficiency of hydropower, the technology behind it and any consequences building hydroelectric dams could have on the environment.

targeted to study a different application or aspect of information retrieval. Each year TREC organisers supply participants with a test set of documents and tasks. Participants in turn, submit the results of their retrieval systems which are then evaluated by NIST judges.

Sessions track has been a part of TREC since 2010 and is targeted towards evaluating retrieval systems over a search session rather than a single query. More concretely, the tasks included a cross between two facets of search tasks defined by Li and Belkin [109] "product" and "goal quality". The "product" facet represented the end goal of the search task and could have be either "intellectual" – aimed at producing new ideas based on the information learned, or "factual" – locating existing information items. In turn, the "goal quality" item could be presented as "specific goal" or "amorphous goal", or as Ingwersen and Järvelin [91] put it – "well defined" and "ill defined" information needs. By crossing the two facets, four task categories emerged: factual task with specific goals – "known item" search, factual task with amorphous goals – "known subject search", intellectual task with specific goals – "interpretive search", and intellectual task with amorphous goals – "exploratory search". It was not the aim of the study to compare the effects of different search task categories, and therefore all topics from Sessions track were deemed suitable for the purpose. Topics with IDs 10, 20, and 21 were selected as search tasks. Table 4.1 demonstrates a detailed description of each topic, which was provided to the participants

Agent	Human	Wizard	Automatic
Overall satisfaction (max 5)	4.1	3.8	2.9
Able to find information (max 2)	1.5	1.3	1.0
Topical quiz success (max 2)	1.6	1.6	1.3

Table 4.2: Row 1: average satisfaction for each agent; row 2: average rate of success for finding desired information; row 3: average rate of success for answering topical quiz questions.

as a task prompt.

4.4 Results

After running the study, participants' preference and quiz scores were analysed. All of participants' verbal comments were qualitatively analysed to extract commonly occurring sentiment. This section describes the findings.

4.4.1 Overall Satisfaction

After completing each task, participants rated their overall experience of working with each agent on a 1 to 5 Likert scale. Average ratings for each agent are shown in the first row of Table 4.2. The scores were normally distributed for all three systems, making it possible to conduct a paired t-test. The differences in ratings between Human ($M = 4.1$, $SD = 0.8$) and Automatic ($M = 2.9$, $SD = 0.8$) systems as well as Wizard ($M = 3.8$, $SD = 0.7$) vs. Automatic systems were statistically significant with respectful t-statistic and p-values being $t(20) = 5.06$, $p < 0.01$ and $t(20) = 3.5$, $p < 0.001$. While the difference between the Human vs. Wizard systems was not significant with $t(20) = 1.6, p = 0.1$. The lack of significant difference in the scores between Wizard and Human systems does not imply that the systems are not different, rather it signifies that for the given context and population, the systems performed similarly. A number of variables could influence these results including the information need at hand, the relatively young age of participants (younger people may have different preferences than older generations), and the environment (laboratory experiments are rarely able to reconstruct real life scenarios).

In the final questionnaire, after completing all the tasks, participants were asked which system they liked the most. Out of 21 people, 8 people preferred the Human agent, 6 –

the Wizard agent, 4 – the Automatic agent, 2 people said they would use the Wizard or the Human depending on their goals, and 1 person said they would choose between the Human and the Automatic agent depending on the goal.

After completing each task participants were asked whether they were able to find all the information they were looking for. Each answer was coded on a 0-2 scale (0 - no, I couldn't; 1 - partially; 2 - yes, I found everything I needed). Average results for each agent are shown in the middle row of Table 4.2.

4.4.2 Topical Quiz Success

After completing each task participants were asked three questions about the topic. Each of the answers was evaluated on a scale 0-2, where 0 meant no answer, 1 - poor answer, 2 - good answer. On average, participants showed a similar level of success with each agent. The average user ratings for each agent are shown in the bottom row of Table 4.2.

4.5 Qualitative findings

I now turn to qualitative results, reporting the themes uncovered during the analysis of verbal comments participants provided after the study was completed. The participants' comments broke down into three areas: the ability of an agent to maintain the context between conversational turns, trustworthiness of the information provided by an agent, and the social acceptance of using a human agent for search tasks.

4.5.1 Conversational Context

Within a conversation, people expect that the main topic of the discussion is maintained, and they tend to ask short questions, omitting the subject, or referring to the subject using pronouns. Formulating a full question takes effort and is unnatural. For the Automatic agent, anaphora resolution did not always work, which annoyed the participants (Figure 4.2 provides an example). For example, P19 had this feedback about the Automatic agent: *"It didn't use contextual information so there was no way to expand on the previous answer it gave me."* Conversely, when interacting with Human and Wizard agents, participants pointed out the ease of use, because their partially stated questions were understood and relevant answers were returned.

4.5.2 Trustworthiness of Information

Even though the Automatic agent did not always return a relevant result, it received approval from our participants for providing sources of its answers. P7 said: *"I [...] like to be able to verify the credibility of the sources used."* Out of 21 participants, 13 people said that being able to access the URL allowed them to assess the trustworthiness of the source and therefore to accept or reject the answer. On the other hand, in spite the Human and Wizard agents returning more relevant results, they were both criticized for not providing the sources of their answers.

4.5.3 Social Acceptance

When dealing with the Human agent, four out of 21 participants reported feeling uncomfortable talking to a person, thought more about the social norms, were afraid to ask too many questions, were not sure how to start and end a conversation. P15 reflected on this aspect saying: *"you have to think about social norms, asking too much, being too stupid, not giving them enough time to respond, troubling them."* This additional burden of interacting with another person further motivates research in the area of automated digital agents.

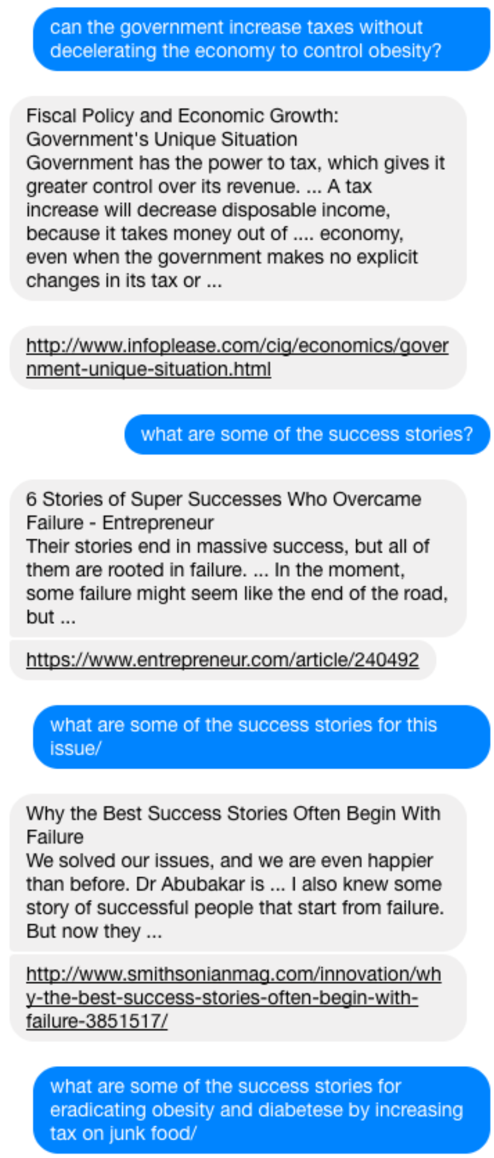
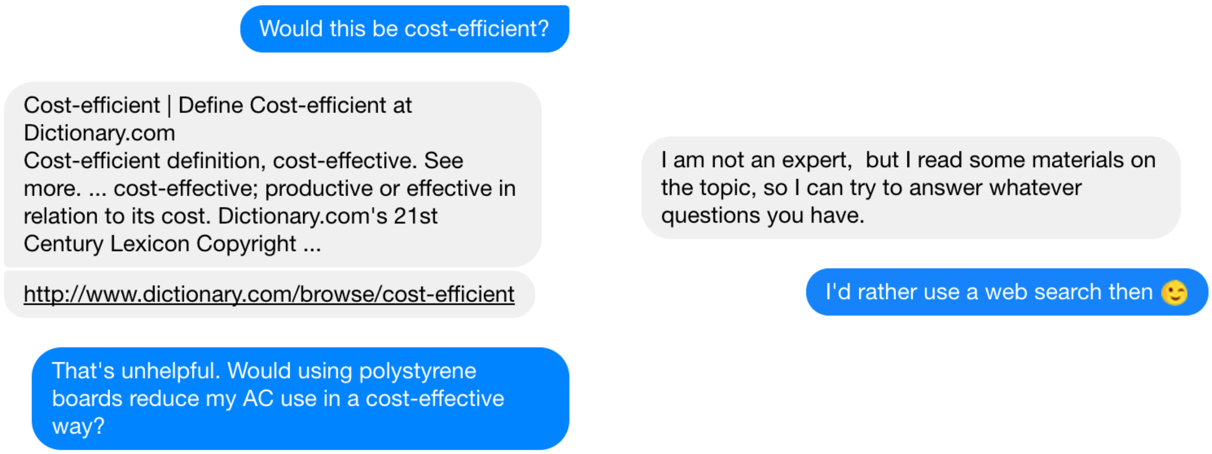


Figure 4.2: Automatic system (gray background) fails to maintain context, which causes the participant 15 (blue background) to reformulate his question twice.

4.6 Discussion and Design Implications

Below I describe the list of design recommendations for a conversational search agent, that



(a) Explicit user feedback could be used to recover from failure. Part of a conversation between participant 12 (blue background) and Automatic system (gray background).

(b) A participant prefers web search to talking to a person. Part of a conversation between participant 7 (blue background) and Human agent (gray background).

Figure 4.3: Examples of user behaviour when interacting with Automatic and Human agents.

according to the study results will improve the
user experience. Additionally, I draw a connection between each of the design recommendations and the four interaction aspects constituting the statement of this thesis: how naturally a user can communicate their intent to the system, the understandability of the system's responses, the flexibility of the system's parameters, and the diversity of information accessible through the system.

Context. Maintaining a context of the conversation to enable short questions and comments is crucial to user experience since formulating long sentences each time feels unnatural and takes longer.

Answer sources. Finding relevant and precise answers is important. But trustworthy information sources are equally important, and their absence may diminish the credibility of the system. While the Automatic agent supported each answer with an URL, Human and Wizard did not, unless specifically asked. Answer source were of could be a deal breaker for some participants, as shown in Figure 4.3b.

Feedback. One crucial difference between conversational search and conventional web search is an opportunity for the user to provide the system with explicit feedback. It is likely to contain essential information that may help the system to get back up from failure and improve upon the previous result. For example, in Figure 4.3a a user says *That's unhelpful.* and rephrases her query. Feedback processing may also be of help in case a user decides to switch the focus of the search. It may also produce rich data for user satisfaction evaluation and can make a rival for the implicit relevance feedback used in web search engines.

Opinion aggregation. According to the participants, sometimes what is needed is the *experience* of other people in similar situations. A good conversational search system should be able to aggregate opinions and present them to the user in a short summary, perhaps explaining each one. P17 said: *"It would be nice if I could see a summarisation of different opinions that there exist – from different sources."*

Direct answers vs. expanded information Regarding this aspect, our participants split into two camps: those who preferred getting direct answers to the question provided, and those who preferred also getting a broader context. Those expecting concise answers, were unhappy that the answers returned by the systems were too long (even for Wizard and Human agents), and preferred to have their questions answered directly with minimum extra information. On the other hand, those who favoured longer answers, said they preferred talking to a person, who would recognize their true information need (beyond the immediate question) and provide the relevant information.

The findings discussed above illustrate examples of the four interaction aspects outlined in the statement of this thesis. The desire to keep context and ability to manipulate the flow of the dialogue by providing feedback to the system follow directly from the fact that a conversation between two people is seen as a collaborative act where each participant contributes towards a mutual goal. The inability of automatic systems to engage in this behaviour causes the lack of *naturalness* in phrasing one's queries. Further, the desire to know the source of information helps users *understand* the provenance of the response and allows for an accurate evaluation of received information. This information may further impact whether or not the user finishes their search or continues to explore other sources. Participants' feedback also provided an example of *personal preferences* in the answer format – long and expanded with related information vs short and to the point. Importantly, these preferences may change depending on the information need or the environment. Finally, the study participants also expressed desire to have access to a *variety* of sources, perhaps even to a compilation of opinions about the topic at hand.

The above study was conducted using text-based messaging and future research may find it useful to confirm the findings during interactions with voice-based systems. Additionally, the participants' preferences with regards to the context should be explored. Prior research showed that people may not be comfortable with using voice-based digital assistants in public [125]. It is also possible that voice or text may be preferred for personal search tasks.

4.7 Chapter Summary

This chapter described a user study with three text-based dialogue systems for web search. It compared participant behaviour when talking to a human expert, a commercial automatic system, and a perceived automatic system secretly controlled by a person (implemented through a Wizard-of-Oz protocol). The observations showed that people do not have biases against automatic systems and are glad to use them as long as their expectations about answer accuracy are met. Furthermore, automatic systems may be preferred in certain social contexts.

Chapter 5

Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants

The previous chapter illustrated that people do not have prominent biases against interacting with text-based digital agents for the purposes of information finding. This chapter continues to investigate the idea of people's comfort when communicating with a digital agent but in a different context: the agent is voice-based instead of text-based, and the participants are asked to complete a culinary recipe instead of a web search task. The study goes beyond analysing the general sentiment about the agent and delivers a fine-grained analysis of the language used by the participants during their interactions with the voice-based digital agent.

5.1 Motivation

As discussed in chapter 2, a conversation between people is rich and full of subtle verbal and non-verbal cues. We use body language and tone to indicate that it's our partner's turn to talk and use facial expressions to show that we are not satisfied with the information we received. When we talk to computers, however, the protocol is much stricter and provides little space for behaviours that are commonplace in human-to-human conversation. In the majority of cases, commercially available agents support interactions that follow the ⟨trigger word, question, answer⟩-protocol. For example, a weather forecast request can go as follows: *"Hey Siri"* followed by *"What is the weather like today?"* with a response: *"It's currently cloudy and 4 degrees"*. Over time people adapt to this way of interaction and phrase their queries and questions in a way that an agent can understand best [116]. Such communication protocol differs significantly from a human-to-human conversation and may cause trigger word fatigue – a phenomenon where the necessity to pronounce a trigger word before each request causes user frustration.

In the previous chapter, I described a study that focused on analysing user attitudes towards a digital agent for the task of information finding – one of the top use cases for commercial voice-based digital assistants. According to recent surveys, kitchens are in the top of the list of smart speaker placement. In their study, Graus et al. showed that users of smart speaker frequently employ them at mealtime to set timers or to manage short processes related to cooking [79]. Commercial voice-based digital agents, such as Google Home and Amazon's Echo, implement the functionality of walking a user through a recipe step by step [63, 171]. In this chapter, I describe a study which investigates how users would interact with a voice-based digital agent able to understand them as well as a human.

To tackle this problem, the author and colleagues ran a high-fidelity Wizard-of-Oz study in which people were asked to interact with a voice-based digital agent as they prepared a simple culinary recipe. This chapter explores potential interactions that occur in the moments following, or in lieu of, users' explicit *"trigger word, question, answer"* triples. The goal of the study was to analyse the language that was used by the participants and answer the question: *RQ1-b: What opportunities are opened by the language people use?*

In order to conduct such an investigation, a unit of analysis needs to be defined. A challenge with analysing transcripts of verbal exchange between two or more parties is that, unlike in written text where words and sentences follow one another, in situations when two or more parties are involved, participants may interrupt each other and speak at the same time [154]. During the analysis conducted in this study, a unit of analysis was considered to be a continuous uninterrupted speech from one of the parties, called an *utterance*. Although the discussed study employs a high-fidelity prototype of a voice-based digital assistant, a dialogue between the users and the system had very well defined turns and no interruptions occurred at any time.

As participants engaged with the agent, it was observed that their utterances divided into two groups: (1) explicit – requests that are clearly phrased as questions (e.g., *"What do I do next?"*, *"What else?"*), or as imperative commands (*"Read me the next step."*, *"Next step."*), and (2) implicit – utterances that could not be definitively interpreted in isolation, i.e. without the information about previous exchanges, intonation, or timing.

1. U: Alright. Uh... What ingredients do I need?
2. A: One-half teaspoon of chili powder.
3. *U: Okay.*
4. A: One-eighth teaspoon of dried oregano.
5. *U: Oregano. Okay.*
6. A: One pinch onion powder.
7. *U: Okie-doke.*
8. A: Cayenne pepper to taste.
9. *U: Sounds good.*
10. U: Alright, can you repeat that?
11. A: Cayenne pepper to taste.
12. U: Uh, I meant all the ingredients.

Figure 5.1: An example exchange between a user (U) and the agent (A) during the Wizard-of-Oz study. Italicised user utterances are implicit conversational cues – utterances that advance the conversation and move the user closer to their goal, without the user asking an explicit question nor giving an imperative command. An implicit cue cannot be definitively interpreted without seeing the history of preceding exchanges.

For example *"Okay"* to signal one's readiness to proceed to the next step, or asking to confirm the correctness of an instruction by repeating it. For the remainder of the chapter, I will refer to these two groups as explicit and implicit cues. For example, in Figure 1, I consider highlighted utterances to be the implicit cues. In other words, the user intents for utterances 3, 5, 7, 9 could not be clearly identified if they were stripped of the surrounding context, and are therefore considered to be implicit. On the other hand, utterances like 10 leave no doubt as to what the user intent is, even if not provided with any additional context.

The remainder of the chapter is structured as follows. I describe the Wizard-of-Oz experiment and introduce the taxonomy of intents of verbal conversational cues for a task-oriented dialogue. I continue to describe the different purposes of short affirmative utterances (e.g., *"Okay."*), as well as conversational cues that repeat the system's previous response. I conclude the chapter by presenting design implications for future voice-based dialogue systems.

5.2 Study design

In this work, a high-fidelity Wizard-of-Oz simulation was developed. A simulated voice-based agent was used to study the role of conversational cues in guided task scenarios. I describe the protocol and apparatus below.

5.2.1 Apparatus

To study user interactions with a culinary assistant, a simulated assistant was developed using a Wizard-of-Oz protocol. (The success of this experiment depended on the fidelity of the Wizard-of-Oz simulation.) The main goal of the study was to simulate the limited capabilities of commercial systems in terms of the answers they are able to provide. The purpose of the experiments was *not* to study a well-rounded conversationalist that could also guide users through cooking a recipe. Conversely, the simulation was designed so that the assistant could support one narrowly defined task.

One of the main considerations in developing of the assistant was to minimize latency and ensure consistency of responses across participants. To this end, a preset list of computer-synthesized audio responses was developed from which an experimenter – the Wizard – could select the appropriate one. The response list included each of the recipe's ingredients, each sentence from the list of recipe instructions, as distinct candidate answers. Additionally, relevant culinary definitions were included (e.g., *"a pinch"*, *"to taste"*, etc.), meta-information about the recipe (*"What is the cooking time?"*, *"number of servings"*), as well as a *"no answer"* response to handle questions that fell out of scope of the current task. The Wizard was allowed to type a free-form response, in case an unexpected but related question was asked. However, a post-experiment analysis showed that this option was used mainly to produce *"yes"* and *"no"* responses. Overall the design of the simulated assistant was successful – no participants reported suspecting that they were interacting with a simulation.

5.2.2 Procedure

A total of 10 participants (6 male, 4 female, average age 30) were invited to engage with a simulated conversational assistant with the goal of preparing a simple culinary recipe. Out of the 10 participants, 2 reported having used an intelligent assistant earlier that day, 5 – earlier that week, and 1 each – earlier that month, more than a month ago and never. 8 people said they usually enjoyed cooking, and 6 said they cooked often.

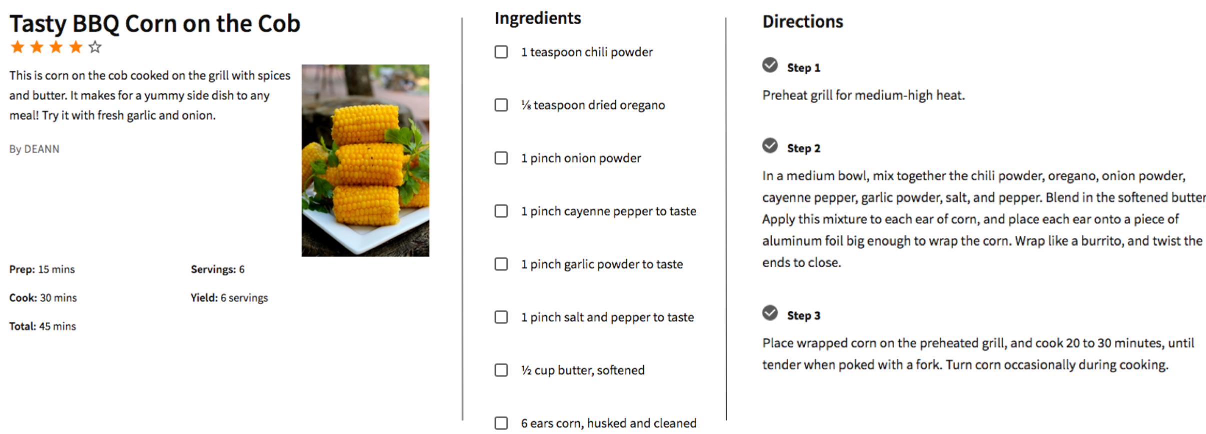


Figure 5.2: Recipe for corn on the cob given to the participants

The experiment took place in an office at Microsoft Research, Redmond, USA. Participants were briefed upon arrival, but were not instructed on what commands to use when communicating with the conversational agent, nor were the participants informed that the agent was a simulation. Instead, the participants were simply instructed to communicate with the agent in a way they felt was natural in order to prepare a spice rub recipe.[1] This recipe was chosen because it includes numerous preparation steps and ingredients, but makes limited use of cooking surfaces or appliances, i.e., it is ideal for a laboratory environment. The printed recipe was shown to the participants in the beginning of the experiment, so that they could familiarise themselves with the process and the goal. By showing the recipe to the participants, the experiment flow imitated a scenario where a person cooks a previously chosen recipe with the help of a smart assistant. The printout of the recipe was also available to the participants to consult in case they got stuck and could not proceed with using the voice assistant, but none of the participants used this option.

To conduct a high-fidelity simulation of the cooking process, participants were provided with all the required recipe spices as well as additional ones to simulate a kitchen pantry with a variety of items present, and avoid making the ingredient selection obvious. The participants were also provided with a corn cob, as well as necessary items such as a bowl, spoon, and foil. The experimental environment did not allow us to use real butter, or grill. Instead of butter, participants were provided a child's wooden cube sized as a butter stick saying "butter" on it. All participants were instructed to pretend to use this butter cube

[1] http://allrecipes.com/recipe/17338/tasty-bbq-corn-on-the-cob/

while cooking. Additionally, participants were asked to omit the last step of cooking the recipe – the grilling of the corn – due to the absence of the grill at the experiment location.

The experiment began with a participant saying the phrase "start cooking", and concluded when the participant completed the penultimate step of the recipe (the final step involved grilling the corn on a barbecue). During the experiment, all interactions between the agent and the participant were mediated via a speakerphone which relayed user utterances to an operator seated in another room. The operator then selected responses from a preset list, which were then played back to the participant in a computer-synthesized voice.[2] All participants' actions were audio and video recorded.

Upon concluding the recipe, participants were asked to complete the NASA Task Load Index (TLX) [88] and System Usability Scale (SUS) [42] questionnaires. Given the study's research focus, and the simulation aspect of this experiment, these questionnaires served primarily as a check to ensure that the simulation was of sufficient quality and completeness to warrant the further investigation of the subtler aspects of the human-agent interaction. Finally, the authors conducted semi-structured interviews and debriefed participants about the simulation.

5.3 Results

In this section, I review the results of the TLX and SUS evaluations, then describe the most common explicit requests and implicit cues that were observed in the study.

5.3.1 General Impressions

	MD	PD	TD	OP	EF	FR	SUS
Median	22.5	7.5	42.5	12.5	20	25	84.25
IQR	13.75	5	43.75	20	21.25	16.25	13.125

Table 5.1: The scores given by the study participants to the culinary assistant. TLX scale scores range from 0 to 100, while SUS scale scores from 1 to 5.

All 10 participants successfully completed the recipe, taking an average of 6.56 minutes (min = 3.22, max = 8.57) and 19 conversational turns (min = 9, max = 27) to reach the

[2] https://responsivevoice.org/

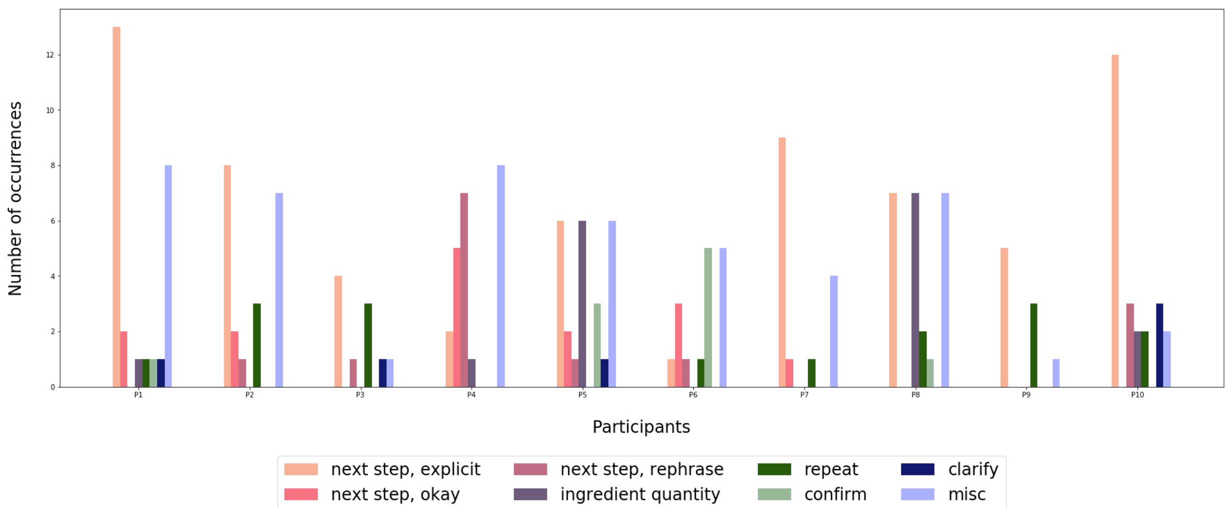


Figure 5.3: Distribution of utterance categories across participants

final step. The simulation received favorable scores as shown in Table 5.1 on both the TLX and SUS scales. Notably, participants reported low frustration and low effort on the TLX. Likewise, participants reported high levels of confidence, and low levels of inconsistency via the SUS. Taken together these findings suggest that the simulation was of sufficient quality and completeness to effectively ground the analysis that follows.

5.3.2 Types of User Utterances

In an initial briefing, participants were instructed to speak with the agent naturally, as if they were conversing with another person. In fact, only a single explicit command was mentioned to participants: *"start cooking"*, that activated the system. Given these limited constraints, participants very quickly adopted a highly conversational style of dialogue, rich with implicit cues. As an example, after the agent read the very first ingredient to P6, she simply responded with "*Okay*", then waited for the agent to continue listing the second ingredient (as seen in Figure 5.1).

The frequency and richness of these cues presented a phenomenon for a further detailed investigation. Throughout the rest of this chapter, the term *"implicit cue"* is used to denote an utterance made by the user, that, taken out of its context (i.e. the previous exchanges, intonation, and conversation timing), could not be definitively interpreted. Utterances for which the intent can be clearly identified, for example "next step", are called explicit.

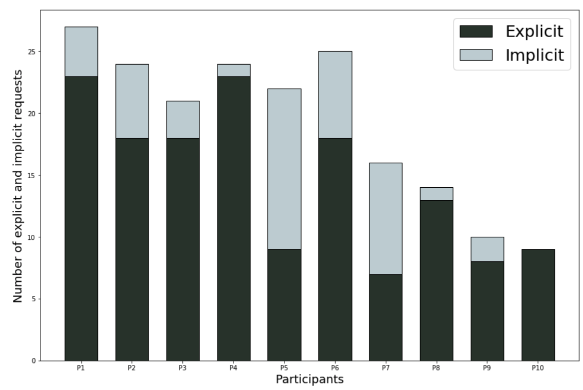


Figure 5.4: Distribution of implicit and explicit utterances across participants

Participants were not asked about the intents that corresponded to each of their utterances. However, the analysis relied on the fact that the simulation was powered by a person, who was able to infer the intent from the conversational context. The high SUS scores and low effort evaluated using TLX (Table 5.1) serve as evidence of the high quality of the correctness of the identified intent.

To understand the prevalence of implicit cues, the author and a colleague counted the number of cases in which a system's response was triggered by implicit cue and an explicit request, as defined earlier. To do that the system's responses were manually labelled as either resulting from an explicit question or statement (e.g., *"what is the next step?"*), or from an implicit cue (e.g., "*Okie dokie*"). To this end, two researchers independently labelled 50 agent's responses as implicit or explicit, achieving high inter-rater reliability (Cohen's $\kappa = 0.92$). One researcher then continued to label the remaining 142 responses. In the end, 46 of agent's responses (24%), were deemed to have been initiated by an implicit cue. Figure 5.4 shows a distribution of absolute numbers of implicit cues (light grey) and explicit requests(dark grey) across all ten participants. It can be seen that although the fraction of implicit cues and explicit requests differ from person to person, the implicit cues are present in the interactions of 9 out of 10 participants.

While labelling the aforementioned interactions, it was observed that implicit cues serve numerous intents. I continue this section with examining the intents of implicit cues and explicit requests that occurred during the course of the study.

5.3.3 Explicit Requests

Throughout the course of the study, the system produced 192 answers, in response to users' utterances. A total of 46 these responses were initiated from an explicit query, question, or an imperative command. I briefly describe five most popular intents, representing (58.3%) of all explicit user requests.

Explicit Next
The most common request simply asked for the next ingredient or step of the recipe (Figure 5.3, blue). These questions constituted 32.8% of all user requests, and varied greatly in their phrasing: E.g., "*What is step number two?*", "*Next*", "*What else?*", etc.

Ingredient Quantity
Comprising 8.9% of all requests, were those asking about ingredient quantities (Figure 5.3, orange). Examples include: "*How much chili powder do I need?*", "*How much of each do I need?*". The lexical variety was quite low for this category.

Repeat
7.8% of all requests, were for the agent to repeat a prior response (Figure 5.3, green; i.e., incorporated in "misc"). Examples include: "*Can you please repeat what you just said?*", or "*Can you repeat that?*".

List Ingredients
A smaller fraction of all requests – 6.2% – were to list all, or some subset of the ingredients (Figure 5.3, green; "misc"). Examples include: "*Okay, thank you. Do I need any other ingredients?*", "*Okay. Is there anything else?*", and "*Alright. So... I've got a few spices in front of me. Can you tell me the spices I need to make the rub for the corn?*". As these examples illustrate, these requests varied considerably in diversity and complexity.

Definitions
Finally, 2.8% of all requests asked for the definitions of ingredients, or cooking procedures (Figure 5.3, green; "misc"). Examples include "*What is an ear of corn?*", "*What do you mean to taste? How much should I put?*", "*How high is medium-high heat is?*".

5.3.4 Implicit Intents

When interacting with the simulated agent, nine of ten participants initiated responses using some form of implicit cues at least once. These cues served a number of distinct intents, as described below.

1. A: Garlic powder to taste.
2. U: Yup. Done that.
3. A: Salt and pepper to taste.
4. U: I don't have salt. Can I omit the salt?
5. A: Yes.
6. *U: Okay. Pepper's in already.*
7. A: Step number 3: blend in the softened butter.
8. U: So I pretend to do that now?
9. *U: Okay, everything's blended.*
10. A: Step number four: apply this mixture to each ear of corn, and place each ear onto a piece of aluminum foil big enough to wrap the corn.
11. *U: Okay. So I applied it to the corn.*
12. A: Step number five: Wrap like a burrito, and twist the ends to close.
13. *U: Ends are twisted.*

Figure 5.5: Example of a user (U) showing readiness to proceed by repeating previous system response (utterances 6, 9, 11 and 13).

Implicit Next

Given the nature of the task, requests for the next step, or next ingredient were the most common. However, 15.6% of the requests for the next step or ingredient were not phrased explicitly. Rather, in 7.8% of cases, participants used short positive utterances, such as *"yup"*, *"alright"* to signal that the current step was completed and they were ready to proceed. Figure 5.1 shows an example of such interactions, as the user's utterances *"3: Okay."* and *"7: Okie-doke."* signal that they are prepared to continue.

Purpose	All	Paraphrase / Repeat	Okay	Explicit
Next	93	15	15	63
Acknowledgement	48	16	32	n/a
Memory	32	19	13	n/a

Table 5.2: Distribution of user utterances requesting next item on the list (Next), showing acceptance of previous system response (Acknowledgement), and utterances spoken to keep short term memory updated (Memory).

In another 7.8% of cases, participants would paraphrase the step they have just completed, to signal that they were ready to go on to the next step, expecting the system to read the next instruction or ingredient in response. Interaction of this type are outlined in Figure 5.5, where in utterances 6, 9, 11, and 13 the participant is describing the last completed instruction in his own words, showing that he is done with this step and is ready to move on.

The first row of Table 5.2 illustrates the counts of next step requests using short positive utterances, paraphrase and explicit questions.

Grounding Behavior

During the experiment, the experimenters noticed, that although the participants did not know what parts of their speech the system could and could not understand, they would still respond to the system's statements. The purpose of these responses in a human-to-human conversation is to let the other speaker – in our case the system – know that the information has been processed and accepted, and that the dialogue may continue. In the literature, this has been referred to as grounding behavior [54].

Grounding behavior can also be exhibited using short positive utterances, as well as partial, or verbatim repetitions of the previous content. These behaviors have been called "acknowledgements", "demonstration" and "display" [54]. Grounding cues closely resemble those of the *implicit next* category and are chiefly differentiated by how they are manifested in a conversational turn. For example, utterance 3 in figure 5.6 shows a participant paraphrasing the agent's prior response, then using a short affirmative phrase ("*Okay*"), and finally, without pause, proceeding to explicitly ask about the next ingredient. This timing pattern precludes these cues from having an *implicit next* intent.

The second row of Table 5.2 gives counts of different types of grounding behavior that has been observed in the study (an utterance is considered to be a repetition if it either partially or fully, repeats a system response verbatim or paraphrased). During the study, 8 of 10 participants repeated a system response out loud, at least once, while preparing the recipe.

Rehearsing Behavior

Another curious phenomenon presented itself when participants talked to themselves while they were in the process of completing a step. With this sort of "memory rehearsal" behavior people refresh and maintain items in their short-term memory [60], which is believed to rely on the same pathways as language and speech. Consequently, people often narrate recipes, as Figure 5.7 demonstrates. In that case, the participant was repeating the name of the ingredient he was looking for, while he was looking for it.

1. U: How much onion powder?
2. A: One pinch onion powder.
3. *U: One pinch. Okay,* and how much oregano?
4. A: One-eighth teaspoon of dried oregano.
5. *U: Okay.*

Figure 5.6: Example of acknowledgement by the user (U) with okay's and repetitions (utterances 3, 5).

1. A: Garlic powder to taste.
2. *U: Garlic powder...*
3. *U: Garlic powder to taste...*
4. *U: Garlic powder to taste... Okay, one second.*
5. *U: Garlic powder... Garlic powder...*

Figure 5.7: Example of a user (U) repeating the response to himself while completing the step (utterances 2, 3, 4, 5)

Clarifications and Confirmations

Additionally, response repetitions came as clarifying questions (Figure 5.9). Whenever people didn't understand the system's response, had doubts about its correctness, or needed more detailed information, they would often repeat a part of the system's response that was not clear, expecting it to provide more thorough explanation.

Closely related to clarifying questions are those that seek confirmations. Such user utterances occurred 10 times (5.2%) throughout the experiment. Their purpose was to confirm user s belief about a step in the recipe. An example is listed in Figure 5.8. Here the first part of utterance 2 is reiterating previous content and user s actions, while the second part serves as a cue for confirmation.

5.4 Implications, Limitations and Future Work

As it has been shown above, implicit cues constitute a large portion of interactions between a user and a digital agent. A dialogue system that is able to recognize and act upon these requests will enable its users to interact using a more human-like style of language, yielding

1. *U: So I applied all the ingredients on the corn, and then applied the softened butter and wrapped it with the aluminum. Right?*
2. A: Correct.
3. U: Perfect. What's next?

Figure 5.8: Example of a user (U) confirming an existing belief about a recipe step (utterance 1).

1. A: One quarter cup butter, softened.
2. *U: One quarter of the butter?*
3. A: One quarter cup butter, softened.
4. U: One quarter cup butter. Okay.

Figure 5.9: Example of a user (U) asking for clarification on the previous response (utterance 2).

high satisfaction scores even when constrained to a simple response model (e.g., limited to sentence selection for question answering [197]).

The results of the study lead us to answer the research question posed in the beginning *RQ1-b: What opportunities are opened by the language people use?* According to the outcomes of the analysis of user utterances, it may be possible to understand when the user is prepared for the next step in the recipe, when the user has heard the agent's response and understood it vs when the user is not certain about what they heard, and finally, by monitoring the rehearsing behaviour it may be possible to infer the step of the recipe the user is working on at the moment even if they did not use the agent the entire time. The results build on top of the findings described in chapter 4 and indicate that as long as an agent it responding correctly, people are ready to speak to a computer in the same way they speak to another person. The digital agent employed in the study supported participants' way to communicate with it, the high scores of SUS and NASA TLX illustrate that this was an important quality, further supporting the statement of this thesis – users' ability to communicate their intent to the system. One of the participants did not use any implicit cues, and instead spoke to the agent using only explicit requests. This shows that people may choose a different mode of interaction with a digital assistant, in this case the choice could be between an agent that reacts to implicit cues, or one that can only process explicit requests.

In the Wizard-of-Oz study, participants did not need to issue a trigger word to initiate interactions, and I believe that this property is one reason for observing such a high frequency of short conversation cues such as "*Okay*" and "*Yup*". One of the advantages of the simulation was that it was listening to the user at all times, which could be challenging to implement in practice (there are both technical limitations and privacy concerns).

However, in the current study, most of these implicit cues followed shortly after an agent's prior responses. Leaving the microphone on for a few moments after each response may be an acceptable compromise and could allow for a more seamless dialogue flow.

Despite having different intents, many implicit cues and utterances transcribe into the same lexical representation. However, contemporary virtual assistant frameworks follow a pipeline architecture, transcribing user utterances prior to doing intent classification [1, 2]. In such an architecture, correct classification of these conversational cues will be challenging if not impossible. To extend their functionality, frameworks and SDKs should include information about prosody, and other acoustic features. These features have already proven to be valuable in improving the detection of dialogue acts [173, 100], which can be seen as a similar – but coarser-grained – taxonomy of spoken intents. Likewise, it will be important for situated agents to model a user's attention, either through acoustic features alone [69], or through gaze, so as to facilitate addressee detection. This will allow more conversational cues to be captured in the first place, by allowing the mic to stay on between utterances, and perhaps by eliminating wake words altogether.

A limitation of this study is its singular focus on recipes might raise questions about the generalisability of the findings. Prior work has studied the importance of implicit cues in human-to-human task-oriented dialogues over a range of tasks [79, 80]. The current Wizard-of-Oz study shows that the importance of these cues extends to at least one class of human-agent task-oriented dialogue: cooking while interacting with a voice-based digital assistant. A culinary recipe is effectively a set of instructions leading to a result. While the study's focus is limited to the task of following a recipe, its conclusions may be expanded to cover other instructions-like scenarios [11, 33, 121]. Though the study protocol employed a simulated agent, a number of steps were taken to ensure that the simulation was convincing, and was close in fidelity to existing voice-based digital assistants. To this end, it is reasonable to expect to see categories like implicit next, grounding, clarification, and confirmation, in other human-agent task-oriented dialogues.

5.5 Chapter Summary

Current voice-based conversational assistants mostly abide by the ⟨*trigger word, question, answer*⟩ paradigm, which constrains user interactions, and a number of implicit conversational cues are missed as a result. This work studied a set of common implicit cues exhibited by users of a simulated voice-based dialogue assistant for the task of cooking a culinary recipe. The study analysis described these cues and their intents in detail and provided a set of design implications for designing task-oriented dialogue systems.

Chapter 6

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

In the last two chapters, I discussed the question of how users would interact with digital assistants that could understand and/or respond to them as well as humans. In the next chapters, I will consider the question of how the design of current dialogue systems can be improved with existing tools. This and the following chapter are focusing on *RQ2: how can we improve interaction with voice-based digital assistants using currently available tools?*

In this chapter, reports on the results of a study aimed to investigate how search results can be presented to the user over an audio-only channel. The study compares and contrasts the representation of search results over two mediums: text and audio. The study consists of two parts. The first part is a crowdsourcing-based experiment exposing the differences between searcher's perceptions and understanding of text and audio search results. Motivated by these results, the second part of the study further investigates the reasons behind these differences through a mixed-methods laboratory study. After discussing the results of both parts of the study, I reflect on how a search system's awareness of the content of the search results can facilitate an extended user interaction and propose a set of guidelines for the design of audio search results. Finally, I propose a set of guidelines for the design of audio search results.

6.1 Motivation

Voice-based dialogue systems have seen a steady increase in recent years. A 2014 Google survey indicated that 55% of teenagers and 41% of adults used voice search at least once a day [78]. The popularity of voice-based digital assistants (e.g., Google Assistant, Amazon Alexa, or Siri) is continuing to grow substantially [40]. In 2018 Forbes predicted that voice queries would make up to 30–50% of all web searches by the year 2020 [71]. While this forecast may not have come true, one thing is clear — voice search is on the rise.

Current state-of-the-art voice search systems perform well for factoid or simple questions, where an exact answer or a top-ranked paragraph can be read out loud and digested by the listener with little difficulty [150]. As discussed in chapter 1 and illustrated in Figure 2.4, 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 effectively interrupts the user's experience by shifting it from an audio modality to a visual one. Depending on the situation (e.g., if the user is occupied with a primary activity, such as driving, where their eyes and/or hands are engaged), it might be infeasible or even dangerous for the user to attend to their screen-based device. Indeed, according to prior research, one of the most attractive features of voice assistants is their ability to support hands-free interaction and multitasking [116].

More recently, there has been a revival in studying voice-only search interactions, with an increase in attention from the natural-language processing community [175, 178]. These research efforts aim to identify the high-level patterns that the searchers follow. However, little is known about how people perceive voice-only search results. The studies described in this chapter aim to address this gap.

The predecessor of a voice-based web search interface is a text-based web search interface displayed on a screen. Hearst [89] provides an overview of the historical changes to visual web search interfaces over time. Researchers in this field examine the representation of a search engine results page (SERP) — historically represented as a list of captions, each corresponding to a web page. In turn, each caption includes a title, the page URL, and a brief summary (or "snippet") of the content of the page. Much research has been conducted in relation to the visual representation of these captions [62, 118, 19, 55, 152]. In this work, I juxtapose the findings discovered in their to the findings discovered during my investigation of audio-based search interfaces. Though prior work demonstrated that translating a text interface to an audio one can be problematic [195], the current study makes a first step towards a *fine-grained* understanding of the features that makes an audio caption "good".

In this chapter, I outline a set of design guidelines aiming at presenting a set of search results (in the absence of a single correct answer) using a voice-only channel. The study addresses the following research questions *RQ2-a: Does the medium (text/audio) over which search results are delivered affect the user's search result preference?* and *RQ2-b: What aspects of audio-based search results are important for the accurate assessment of relevance by the user?*

To answer these questions, an experiment in two parts (further referred to as *AMT* and *LAB*) was conducted, employing quantitative and qualitative methods and analysed the collected data through a mixed-methods approach [153]. In the *AMT* study, 69 crowd-workers were asked to judge the relevance of a set of pre-selected results presented in text, audio, or image format, for six search tasks. The *LAB* study focused on text and audio interfaces. A total of 36 people were invited to participate in a controlled laboratory experiment, during which they were asked to judge the relevance of the same set of search results presented in text or audio format. The *LAB* study also included semi-structured interviews with the participants, which provided rich insight into the aspects of audio-based web search results presentation that need to be considered in the design of future voice-based search systems. In both studies, the complexity of the search tasks varied to account for the variability of experience that it can produce [22].

The results of this study answer *RQ2-a*, showing that there is indeed a significant difference in users' search result preference depending on whether the search results are presented in the Text or Audio condition. The study findings also demonstrate evidence that the Audio condition leads to a significant increase in the searcher's perceived workload compared to the Text condition. The analysis does not reveal a significant interaction between the complexity of a search task and the medium in terms of users' search result preference or their perceived workload. The qualitative results of the study address *RQ2-b* and lead to a set of design guidelines for audio captions for web search results.

6.2 Study Design

To address *RQ2-a* and *RQ2-b*, a two-part user study was conducted. The first part (further referred to as *AMT*, short for "Amazon Mechanical Turk"), aimed to explore potential differences between searcher choices when search results were presented in text or audio (i.e., medium) in a crowdsourcing setup. The concept of crowdsourcing studies is described in detail in chapter 3. The results of the experiment suggested that there was a significant difference in searcher preferences depending on whether the search results were presented in a Text or Audio condition. To explore this phenomenon in more detail, a follow-up laboratory experiment (further referred to as *LAB*) was conducted where the differences in searcher choices discovered in the *AMT* study were confirmed. Additionally rich qualitative data was collected explaining the challenges in perception of the audio results.

Throughout both parts of this investigation (*ATM* and *LAB*), the same search tasks and interfaces were used. In the following sections, I describe these parts as well as the procedure for the *AMT* and *LAB* parts of the study.

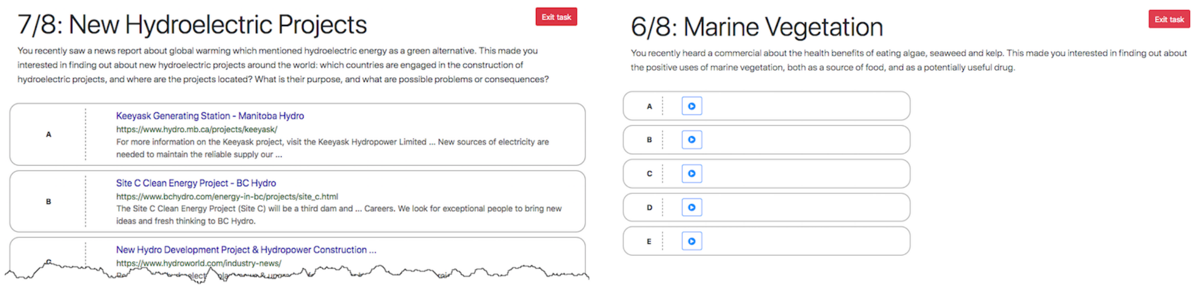
6.2.1 Search Tasks

Considering that the complexity of search tasks can have a substantial effect on user behaviour, six search tasks of varying complexity levels were selected. Following Bailey et al. [22] and Trippas et al. [176], the current study adopted the following three levels of complexity from the taxonomy defined by Wu et al. [192]: *remember (R), understand (U)*, and *analyse (A)*. To provide the study participants with a detailed description of the supposed information need, backstories proposed by Bailey et al. [22] were used. Table 6.1 presents the tasks that were used as well as their corresponding backstories and complexity levels.

ID	Task Topic	Backstory	Complexity
140	Which planet was researched extensively by spacecraft Magellan?	Gazing up into the night sky you see some of the planets come out. It would be great to look at some close up pictures of the planets. You remember hearing about the voyage of the spacecraft Magellan, and wonder which planet it allowed scientists to explore.	Remember
034	How tall is CN tower in Toronto?	Every city seems to have at least one really big building in it these days, but depending on how long ago the building was built, it may no longer seem quite so big. Growing up in Toronto, Canada, the biggest building by far was the CN Tower. How tall was it?	Remember
009*	What year was the phonograph invented?	You've been talking with your niece about changes in music technology. You realize you don't know much about anything from before you were born, and you'd like to know when the phonograph was invented.	Remember

002	Marine Vegetation	You recently heard a commercial about the health benefits of eating algae, seaweed and kelp. This made you interested in finding out about the positive uses of marine vegetation, both as a source of food, and as a potentially useful drug.	Understand
018	Antarctica Exploration	On the T.V. news last night, you saw footage of scientists in Antarctica. There seemed a surprising number of people there. This got you wondering what scientific expeditions or projects are under way in Antarctica, and what is planned.	Understand
007	Hubble Telescope Achievements	You recently heard that a new space telescope is to be launched in 2021 to replace the Hubble Space Telescope. This made you curious about the accomplishments of the Hubble Telescope since it was launched in 1991. How did it produce new better quality data than was previously available? How did it increase human knowledge of the universe?	Analyse
039	New Hydroelectric projects	You recently saw a news report about global warming which mentioned hydroelectric energy as a green alternative. This made you interested in finding out about new hydroelectric projects around the world: which countries are engaged in the construction of hydroelectric projects, and where are the projects located? What is their purpose, and what are possible problems or consequences?	Analyse

Table 6.1: Search tasks used in the study



(a) Text condition (b) Audio condition

Figure 6.1: For each task five search results were presented in text or audio formats. The Text interface reproduced the general look of Google's search engine result page. The links were not clickable to restrict the participants' access to information. The Audio interface supports only the basic functionality to emulate the voice-only environment of a voice assistant where a searcher has limited control over audio playback, such as one might have in a system controlled only by voice.

6.2.2 Search Results

For each search task, as seen in Table 6.1, the system displayed five search results to the participants. To generate these search results, the "search topic" was submitted as a search query to Google and the 1st, 5th, 10th, 50th, and 100th search engine results were collected, with the assumption that the 1st, 5th, and 10th result will be more relevant than 50th and 100th. Results linked to PDF files were skipped and replaced by the next ranked non-PDF document. For queries that yielded less than 100 results, the last one was used to represent the 100th result. For each result, the following information was collected: the displayed title, URL, snippet, and screenshot of the underlying web page.[1]

6.2.3 Interfaces

This study explores the differences between searcher behaviour and perception of results when those are presented in either Text or Audio format.[2] Both conditions displayed the task topic, followed by its corresponding backstory from [22]. Below the backstory, five

[1]This dataset is available at https://github.com/sashavtyurina/audio-serp-ictir-2020.

[2]The *AMT* part of the study also included an Image interface where the results were presented in the form of the snapshot of the underlying documents. However, this chapter focuses on comparing text and audio interfaces and does not report on the results of the image interface.

search results were displayed, as shown in Figure 6.1. The order of results was randomized to ameliorate participants' bias towards the top-ranked result [96]. For each task, participants were instructed 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. Each result was denoted using letters A-E to avoid the confusion between notations "best" and "first". For each task, 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 The Text condition (Figure 6.1a) reproduced the general look of Google's search engine result page with similar fonts and colors, to make the interface more familiar to participants. In contrast to the standard Google search engine results, the captions were made non-clickable to restrict the information available to the participants.

Audio Condition The search results in the Audio condition (Figure 6.1b) were displayed through five identical play/stop buttons. The audio interface was kept minimalistic. It did not provide participants with an option to pause or jump to a different position of the audio, nor did it display the information about the time elapsed, or percentage of the clip played. We chose to support only the basic functionality to emulate the voice-only environment of a voice assistant where a searcher has limited control over audio playback, such as one might have in a system controlled only by voice.

Prior research suggests that direct translations of text interfaces to audio ones are suboptimal for human comprehension of the system [195], however to the best of our knowledge, no prior work explored the translation of search result captions to audio format. As a starting point, we chose to heavily base the audio representation of search captions on the text content. As such, we generated audio results by combining the search results' caption components: the title, top-level domain of the URL, and the snippet. We replaced ellipses in the snippets with periods. To produce audio clips, we used Google's text-to-speech engine using voice en-US-Wavenet-A.[3] We recorded 30 audio clips — five for each of the six presented tasks — with duration ranging from 11 to 29 seconds (median=16 seconds, IQR=6 seconds). Figure 6.2 illustrates a text result, and a transcript of the corresponding audio result. We instrumented both interfaces by logging user interactions with them. Specifically, we recorded interactions with the audio results — which results were played and for how long.

[3] https://cloud.google.com/text-to-speech/

Table 6.2: 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.

In Depth | Magellan – NASA Solar System Exploration https://solarsystem.nasa.gov/missions/magellan/in-depth/ NASA's real-time science encyclopedia of deep space exploration. ... Magellan was the first planetary spacecraft launched from the Space Shuttle. ... manifest into the 1990s, which included a number of planetary missions. ... A new study reveals asteroid impacts on ancient Mars could have produced key ...
In depth. Magellan - Nasa solar system exploration. From solarsystem dot nasa dot gov. Nasa's real-time science encyclopedia of deep space exploration. Magellan was the first planetary spacecraft launched from the Space shuttle. Manifest into the 1990s, which included a number of planetary missions. A new study reveals asteroid impacts on ancient Mars could have produced key.

6.2.4 Procedure

The design for both studies crossed two main factors: medium (two levels) and complexity (three levels). The order of individual tasks (six tasks, two for each complexity level) and mediums was counterbalanced, rotating them in a Latin square design, such that each task occurred with every medium. Each participant was exposed to both Text and Audio conditions. Participants in the *AMT* study completed two tasks in Text condition and two tasks in Audio condition. Participants in the *LAB* study completed three tasks in Text and three tasks in Audio condition. For each task, participants were asked to select three search results: (1) the most useful result, the one they would click on first for the task; (2) the second most useful result, and (3) the least useful result for completing the task.

Both *AMT* and *LAB* studies were approved through the ethics approval process for research involving human participants at the University of Waterloo.

Part 1 - *AMT*.

Each participant completed two tasks in the Text condition and three in the Audio condition. One of the audio tasks served as a quality check as described in Section 6.2.4.

The study took on average 21 minutes.

Quality control

As described in section 3.3 of chapter 3, data collected using crowdsourcing may contain noise. To ensure the quality of the data collected in this study, the participant pool was restricted to workers who had approval rating 95% or higher, have completed more than 1,000 tasks, and lived in the US to ensure a high level of English proficiency. To ensure high quality of submissions, a "golden task" was included as an attention check (task 009 in Table 6.1). This task was presented as the final task for every participant in the Audio condition. The search results for this task included 1st and 5th hits from Google, and three results that belonged to other topics, not presented on the previous pages. A submission was considered to be of acceptable quality if the two relevant results were selected as the most useful and the second most useful. As a measure of additional quality assurance, the judgements of the workers who did not click on all audio clips in the Audio condition were discarded. All crowdworkers were paid 3.50 regardless of the quality of their submissions.

Part 2 - LAB.

The second part of the study aimed to carefully examine specific differences that exist between searcher perception of search results in Text vs. Audio conditions, as well as the reasons for these differences. This part of the study followed a similar procedure as the AMT study.

After providing their consent and filling out a short demographics survey, participants were given two training tasks — one in Text and one in Audio condition — to become familiar with both interfaces and the flow of the experiment. After completing each task, participants competed the NASA-TLX questionnaire — a scale to subjectively assess mental workload [88]. The scale measures mental, physical, and temporal demand, performance, effort, and frustration. The "physical demand" measure was omitted since there was no physical exertion involved. After each task, the experimenter conducted a short semi-structured interview. During each interview, participants were asked to explain what attracted them in the result they selected as the most useful one, whether their decision was affected by any particular part of the result or keywords. Finally, the experimenter asked participants to recall the results they chose as most useful and as the second most useful. Upon completing all six tasks, the experimenter conducted a post-study interview, in which participants were asked about their general impressions of the audio results they heard and how they thought audio results could be improved. The study took on average

44 minutes. All participants were reimbursed $10 for their time. The experiment was audio recorded to facilitate transcription and analysis of participants' interview responses.

6.2.5 Participants

Table 6.3 illustrates the characteristics of the study participants for the *AMT* and *LAB* parts. After removing submissions that did not pass the quality check, data from 69 crowd-workers remained. 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.

Table 6.3: Participants characteristics in AMT and LAB studies.

		AMT	**LAB**
Number of people	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

6.3 Quantitative Findings

This section examined the effect of the interaction medium (text or audio) on the participants' search results preferences as well as on their perceived workload from both *AMT* and *LAB* studies.

6.3.1 Differences in Ranking

In this section, I describe the results of the study findings regarding differences in users' ability to identify useful results between the Text and Audio conditions. Specifically, whether searchers make fewer choices that reflect the *true ranking* of results in the Audio condition compared to the Text condition.

Number of result choices consistent with true ranking

In the experimental setup, participants were asked to select the most, second-most, and least useful results from the five results shown to them. In this setup, participant choices are considered to be 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 Google result, 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 (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, the study participants could make between 0 and 3 choices consistent with results' true rankings (e.g., 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, the analysis aims to determine whether participants make fewer consistent choices in the Audio condition compared to the Text condition.

To test whether differences between the experimental conditions (Text or Audio) are meaningful, a test statistic was bootstrapped using data collected in the experiments — in this case, the average difference in the number of consistent choices between the two conditions were bootstrapped [48]. To achieve this, the number of consistent choices in both Audio and Text conditions is computed, using the 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. Then this procedure is repeated ($M = 1000$). This method allows 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), the same procedure is conducted using sampling from Text condition.

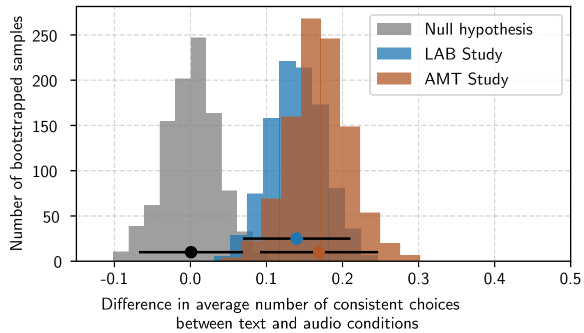


Figure 6.2: Average difference in number of consistent result choices, between text and audio, under the **(a)** Null hypothesis, and in the **(b)** LAB study and AMT study. An average difference higher than 0 means more consistent choices in the text experimental condition compared to the audio condition. 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.

Figure 6.2 shows the results of our bootstrap test. Both the *LAB* data and the *AMT* data suggest that searchers make more choices consistent 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.

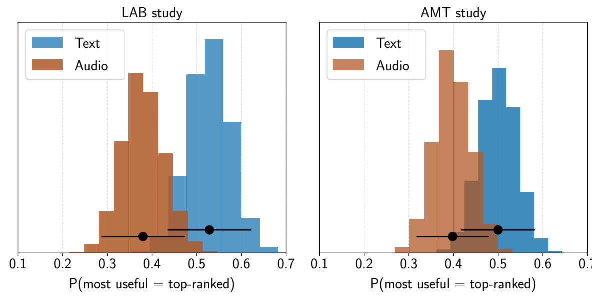


Figure 6.3: Probability of selecting the top-ranked Google result as *most useful* (Text and Audio conditions). *LAB* study (left) and *AMT* study (right). Markers indicate mean probability and 95% confidence intervals for the mean value.

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 searchers' 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,[4] 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 6.3 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 searchers' ability to identify the top-ranked search result as most useful are present, but not significant. As shown in Figure 6.3, confidence intervals for the mean probability of identifying the top-ranked result overlap in the two conditions.

[4]We use `PyMC3` with the No-U-Turn Sampler to fit our models, for each iteration generating 10000 samples, burning the initial 2000 samples. We note that all our regression coefficient estimates converge using this sampling approach.

Effects of task complexity

To study the interaction between task complexity and result medium (text or audio) on searchers' 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.

6.3.2 Perceived Workload

In this section, I report on the results of investigation of whether the task complexity and the medium (text or audio) influenced the searchers' perceived workload.

In the *LAB* study, after completing each task, we asked participants to fill out a NASA-TLX questionnaire [88]. 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 1 (low) to 20 (high); and the performance scale ranges from 1 (good) to 20 (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 Mann-Whitney-Wilcoxon test (MWW) 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, as shown in Table 6.4.

To estimate whether the complexity of the tasks had an effect on the estimated workload, we used a linear mixed-effect model [25] 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.

Table 6.4: 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 [103].

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

6.4 Qualitative Findings

This section tackles *RQ2-b*: "What aspects of audio-based search results are important for the accurate assessment of relevance by the user?" As part of the *LAB* study, the experimenter conducted a set of semi-structured interviews, where participants were asked about their impressions of the results they heard/saw, what attracted them in the results they selected as the best and second-best ones, and whether they could remember the two results they selected as the most useful ones. After each task presented in the Audio condition, the experimenter asked whether the audio results were challenging to comprehend, and if so, what made it challenging. The participants were also asked how they envisioned changing the representation of the audio results.

During the study, with participants' consent, the interviews were audio recorded for the ease of transcription later. Three researchers analysed the transcribed interviews and jointly developed a codebook using the method of affinity diagramming [108]. In this section, I report on the findings and observations that resulted from this analysis. I outline the participants' perceived challenges of the audio results, including some of the behavioural patterns that could be important for consideration and design of future voice-based search systems.

6.4.1 Navigation Shortcuts

When discussing the selected results with participants, an interesting trend emerged. 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 describing the type of the underlying page, 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 searchers will ultimately select a result to hear more from, the voice equivalent of clicking on a result. Additionally one can envision a scenario in which a searcher might ask to hear a certain caption again. To facilitate smooth navigation and to understand which result the searcher is referring to, the system should be aware of the contents of the results it returns, providing a clear and natural method for referencing them.

6.4.2 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 6.2. Below we discuss the challenges in perception of audio results that were raised by our participants.

Uncertainty about result structure. The structure and contents of the results should be made clear. Some searchers found it challenging to understand how the audio results were constructed and what information to expect from them. In particular, some participants pointed out that it was difficult to distinguish 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 in to 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 results to clearly indicate the roles of the constituent parts, or by varying the prosody of the generated audio as discussed in above.

Uncertainty about clip duration. Searchers should be made aware of the duration of the audio results. Another source of uncertainty was the unknown length of the audio clips. P6 expressed their unhappiness with it, comparing the experience with 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 the 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.

Monotonicity of the audio. Prosodic features of the audio 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, difference 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 the work of Chuklin et al. [52], who used prosody modifications to verbally "highlight" the answer inside a longer paragraph. Later, Winters et al. [191] used different audio generation features to signal the sentiment of the underlying text.

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. Truncated snippet sentences can be a cause of disruption for the participants. Fourteen people mentioned that sentences would cut off abruptly before communicating important information about the result. As 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, this snippet truncation disruption was not mentioned for the text interface. As part of future work, we suggest experimenting with snippets consisting of full sentences to ease the comprehension of audio.

Repetitions. Audio results 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. A recommendation against repetitions in text search results was also outlined by Clarke et al. [55], who list as one of their guidelines: "When query terms are present in the title, they need not be repeated in the snippet".

6.4.3 Cognitive Load

Finally, as supported by NASA-TLX responses in Table 6.4, 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 results 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 searcher's working memory: prior work found an effect between the level of working memory and the outcome of a search process [50].

Lack of control over the pace of the speech was pointed out as a downside of the audio results by eight participants. This aspect was previously discussed by Abdolrahmani et al. [5], who reflected on the need of more advanced features for voice assistants. Such functionality was recently introduced by Amazon, enabling Alexa to speak faster or slower on user's request [174].

6.5 Chapter Summary

The study described in this chapter addresses *RQ2: How can we improve interaction using currently available tools?*. Commercial voice-based digital assistants are often used for web search, however, they are limited in the amount of information they are able to return via an audio-only channel without displaying any additional information on the screen. The results of the study described in this chapter suggest that although there are differences in the preference of search results chosen in Text and Audio interfaces, the differences in selecting the most useful result are not significant. This finding leads one to conclude that even the imperfect representation of web search results through an audio channel allows for a fairly accurate selection of the web search results, although at a price of a much higher cognitive load. 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 it is returning. For navigational purposes, searchers 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 result: a title, a URL, and a snippet. Beyond its use for navigation, a clear statement of the source might help searchers 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. Searchers 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 truncated sentences. Results should be reported in full sentences.
- Avoid repetitive terms in the audio result.

Chapter 7

VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search

In the previous chapter, I began addressing *RQ2* by discussing how existing digital agents can expand the amount of information being returned by representing search results through a voice-only channel. In this chapter, I turn to discussing how by providing access to a wider range of information through an audio-only channel digital assistants can become not only a convenience but also an accessibility tool for people with visual impairments. The study described in this chapter addresses *RQ2-c: How might voice assistants and screen readers be merged to confer the unique advantages of each technology?* Below, I describe the design and report on the results of two studies: an online survey with respondents who are legally blind and a design-probe study of a voice-based web search interface working in a conjunction with a screen reader.

7.1 Motivation

People with visual impairments are often early adopters of audio-based interfaces, with screen readers being a prime example. Screen readers work by transforming the visual content in a graphical user interface into audio by vocalizing on-screen text. They are an important accessibility tool for blind computer users – so much so that every major operating system includes screen reader functionality (e.g., VoiceOver[1], TalkBack[2], Narrator[3]), and there is a strong market for third-party offerings (e.g., JAWS[4], NVDA[5]). Despite their importance, screen readers have many limitations. For example, they are complex to master, and depend on the cooperation of content creators to provide accessible markup (e.g., *alt text* for images).

Voice-activated digital assistants (VAs), such as Apple's Siri, Amazon's Alexa, and Microsoft's Cortana, offer another audio-based interaction paradigm, and are mostly used for everyday tasks such as controlling a music player, checking the weather, and setting up reminders [179]. In addition to these household tasks, however, voice assistants are also used for general-purpose web search and information access [123]. In contrast to screen readers, VAs are marketed to a general audience and are limited to shallow investigations of web content. Being proficient users of audio-based interfaces, people who are blind often use VAs, and would benefit from broader VA capabilities [142, 5].

The work presented in this chapter explores opportunities at the intersection of screen readers and VAs. I describe the results of an online survey with 53 blind screen reader and VA users, aimed to investigate the pros and cons of searching the web using a screen reader-equipped web browser, and when getting information from a voice assistant. Based on these findings, a prototype of a tool called VERSE (Voice Exploration, Retrieval, and SEarch). The prototype augments the VA interaction model with functionality inspired by screen readers to better support free-form, voice-based web search. I continue with describing a design probe study of VERSE, and discuss future directions for improving eyes-free information-seeking tools.

In the following sections I cover the online survey, the functionality of VERSE, and the VERSE design probe study. I conclude by discussing the implications of the presented findings for designing next-generation technologies that improve eyes-free web search for blind and sighted users by bridging voice assistants and screen readers paradigms.

7.2 Online Survey

To better understand the problem space of non-visual web search, an online survey addressing three topics was designed:

[1]https://www.apple.com/accessibility/mac/vision/
[2]https://support.google.com/accessibility/android/answer/6283677
[3]https://www.microsoft.com/en-us/accessibility/windows
[4]https://www.freedomscientific.com/Products/Blindness/JAWS
[5]https://www.nvaccess.org/

- What challenges do people who are blind face when conducting information searches using screen readers?
- What challenges do people who are blind face when conducting information searches using VAs?
- Do people who are blind envision the integration of screen readers and VAs to provide stronger task support, and, if so, in what ways?

7.2.1 Survey Design and Methodology

The survey consisted of 40 questions spanning five categories: general demographics, use of screen readers for accessing information in a web browser, use of digital assistants for retrieving online information, comparisons of screen readers to digital assistants for information seeking tasks, and possible future integration scenarios (e.g., voice-enabled screen readers). When asking about the use of screen readers and digital assistants, the survey employed a recent critical incident approach [70], in which participants were asked to think of recent occasions they had engaged in web search using each of these technologies. The survey respondents were then asked to describe these search episodes, and to use them as anchor points to concretely frame reflections on strengths and challenges of each technology.

The survey respondents were adults living in the U.S. who were legally blind and who used both screen readers and voice assistants. The recruitment was conducted using the services of an organization that specializes in recruiting people with various disabilities for online surveys, interviews, and remote studies. While the online questionnaire was designed to be accessible with most popular web browser/screen reader combinations, the partner organization worked with participants directly to ensure that content was accessible to each individual. In some cases, this included enabling respondents to complete the questionnaire by telephone. The survey took an average of 49 minutes to complete, and participants were compensated $50 for their time.

Two researchers iteratively analyzed the open-ended responses using techniques for open coding and affinity diagramming [108] to identify themes.

7.2.2 Participants

A total of 53 respondents completed the survey. Participants were diverse in age, education level, and employment status as seen in Table 7.1. All participants reported being legally

Table 7.1: Demographic characteristics of survey respondents.

Gender	Male	25
	Female	28
Age	18-24	5
	25-34	17
	35-44	12
	45-54	9
	55-64	6
	65-74	4
Education level	Some high school, no diploma	1
	High school or GED	4
	Some college, no diploma	17
	Associate degree	7
	Bachelor's degree	12
	Some graduate school, no diploma	1
	Graduate degree	11
Occupation	Employed full time	21
	Employed part-time	6
	Part-time student	4
	Full-time student	6
	Not currently employed	10
	Retired	3
	Unable to work due to disability	3
Experience with VAs	Over 3 years	35
	Between 1 and 3 years	17
	Under 1 year	1
Devices used	Touchscreen smartphone	53
	Laptop	46
	Tablet	29
	Desktop computer	27
	Smart TV	21
	Smart watch	11

blind, and most had experienced visual disability for a prolonged period of time ($\mu = 31.6$ years, $\sigma = 17$ years). As such, all but three respondents reported having more than three years of experience with screen reader technology. Likewise, most of the participants were early adopters of voice assistant technology. Out of 53 survey respondents, 35 respondents reported having more than three years of experience with such systems and 17 had between one and three years of experience.

More generally, the survey respondents were active users of technology. 40 participants reported using three or more devices on an average day including: touchscreen smartphones, laptops, tablets, desktop computers, smart TVs and smart watches.

7.2.3 Findings

According to the survey results, respondents made frequent and extensive use of both virtual assistants and screen reader-equipped web browsers to search for information online, but saw certain shortcomings in both methods. Moreover, the survey participants reported that transitioning between screen readers and voice assistants introduced its own set of challenges and opportunities for future integration. This section first details broad patterns of use, then presents specific themes around the technologies' advantages and challenges.

General Patterns of Use

Respondents said they used their voice assistants regularly on a variety of devices: smartphones, smart speakers, tablets, laptops, smart TVs, smart watches, and desktop computers. The most popular assistant used on a smartphone was Siri (used by 51 people), followed by Google Assistant and Alexa. Fewer people used assistants on a tablet, but a similar pattern emerged, with Siri being the most popular, followed by Alexa and Google Assistant. Amazon Echo was the most popular smart speaker among our respondents, followed by Google Home and Apple Home Pod. The most popular assistant on laptops and desktops was Cortana, followed by Siri. Siri and Alexa were the most popular assistants on smart TVs (the Apple TV and Amazon Fire TV, respectively).

Most of the respondents were active searchers: when asked how often they searched for answers or information online, most said they performed online searches multiple times a day. They also actively used voice assistants as their search tools with over half using a voice assistant to conduct an online search multiple times a day. The most popular devices for searching the internet were touchscreen smartphones, laptops, touchscreen tablets and desktop computers. The data described above attests to the fact that our survey participants use VAs frequently overall and specifically for information search. Table 7.2 displays a detailed breakdown of each of the aspects discussed above, however.

Theme 1: Brevity vs. Detail

The amount of information provided by voice assistants can differ substantially from that returned by a search engine. Voice assistants provide a single answer (suitable for simple question answering but not for exploratory search tasks [186]), that may be short and provide limited insight (P1078: *"a virtual assistant will only give you one or two choices, and if one of the choices isn't the answer you are seeking, it's hard to find any other information."*, P959: *"you just get one answer and sometimes it's not even the one you were looking for. "*, P1148: *"a lot of times, a virtual assistant typically uses one or two sources in order to find the information requested, rather than the entire web"*, P1027: *"It can be difficult to have things elaborated on"*). Whereas using a search engine a user is provided with a number of different sources, is able to triage the search results, and can access more details if needed (P1027: *"information can be gathered and compared across multiple sources"*, P960: *"you can study detailed information more thoroughly"*).

But those details come at a price – using a screen reader a user has to cut through the clutter on web pages before getting to the main content (P1035: *"you don't get the information directly but instead have to some times hunt through lots of clutter on a web page to find what you are looking for"*, P1140: *"the information I am seeking gets obfuscated within the overall web design of the Google search experience. Yelp, Google, or other information sites can be over designed or poorly designed while not taking any of the WCAG standards into consideration"*), while VAs provide a direct answer with minimal effort (P1058: *"The assistant will read out information to me and all I have had to do is ask"*).

The upside of using a search engine is *"the ability to see more information that you think might be essential for you to know"* (P960), whereas when using a VA *"the responses are to the point and a bit scripted"*(P659), and *"information ⟨...⟩ tends to be truncated."* (P931).

In cases when an assistant performs a web search, the user is forced to interact with the pair screen-based device (phone or tablet) to read through results. Such interaction breaks the voice-based experience and forces the user to switch modalities (P1083: *"virtual assistants normally pick the nearest place to you and require you to look at an app for more information"*, P945: *"If the answer is complicated, it requires using a screen reader anyway."*)

Table 7.2: General patterns of voice assistants use.

Devices for VAs	Phone	53
	Smart speaker	34
	Tablet	18
	Laptop	15
	Smart TV	13
	Smart watch	7
	Desktop	5
VAs used on smartphone	Siri	51
	Google Assistant	23
	Alexa	18
	Cortana	3
VAs used on tablet	Siri	18
	Google Assistant	8
	Alexa	8
Smart speakers used	Amazon Echo	29
	Google Home	14
	Apple Home Pod	1
VAs used on laptops and desktops	Cortana	17
	Siri	8
Searching online	Multiple times a day	41
	Multiple times a week	9
	Once a day	2
	Multiple times a month	1
VAs as search tool	Multiple times a day	29
	Once a day	7
	Multiple times a week	11
	Once a week or less	6
Device for online search	Smartphone	45
	Laptop	41
	Tablet	23
	Desktop	23

Theme 2: Granularity of Control vs Ease of Use

The survey participants widely recognized that voice assistants were a convenient tool for performing simple tasks, but greater control was needed for in-depth exploration (P56: *"They are good for specific, very tailored tasks."*). The notion of control came up for all stages of performing a search: query formulation, results navigation, and information management. When using voice assistants, *"you have to be more exact and precise as to the type of information you are seeking."* (P1148), *"say what you're looking for in just the right way so that you will get the desired results"*(P1078), participants noted. Screen readers also provide the freedom of exploring the search results using various navigation modes (P1035: *"you can navigate by heading landmark or words"*, P1078: *"It's easier to scan the headings with a screen reader when searching the web"*, P459: *"one is able to navigate through available results much faster than is possible with virtual assistants."*). Additionally, using a screen reader, users can customize multiple settings (speech rate, pitch) to fit their preferences – a functionality not yet available in voice assistants (P950: *"sometimes you can get what you need quicker by going down a web page, rather then waiting for the assistant to finish speaking"*). Such dexterity of screen readers comes at a price of having to memorize many keyboard commands or touch gestures, whereas VAs require minimal to no training (P56: *"you don't have to remember to use multiple screen reader keyboard commands"*).

Theme 3: Text vs Voice

Speaking a query may be faster (P1027: *"typing questions can take more time"*), less effortful (P945: *"It is easier to dictate a question rather than type it."*), and can help avoid spelling mistakes (P682: *"You do not know how to spell everything"*). Albeit, speech recognition errors can cancel out these benefits (P944: *"I can type exactly what I want to search for and don't have to edit if I'm heard incorrectly by the virtual assistant."*) and even lead to inaccurate results (P1066: *Virtual assistant often 'mishears' what I am trying to say. The results usually make no sense.*) Especially prone to misrecognition are queries containing *"non-English words, odd spellings, or homophones"* (P1140). Environmental conditions can create additional limitations on the modality of input and output (P926: *"it [voice interaction] is nearly impossible in a noisy environment, such as a crowded restaurant. Even when out in public in a quiet environment, the interaction may be distracting to others."*). Environmental limitations of voice assistant interaction have also surfaced as a user concern in prior work [67].

Theme 4: Portability vs Agility

Assistants are either portable – such as Siri on an iPhone (P960: *"Its in your pocket practically all the time, and you can literally talk to it and it will give you an answer quickly."*), or are always ready to use – like smart speakers (P1025: *"I can be on my computer doing an assignment and ask Alexa"*). They are hands-free and allow multitasking (P920: *" especially helpful if I have my hands dirty or messy while cooking"*). On the other hand, to use a screen reader one needs to spend time setting up the environment before performing the search (P959: *"It takes more time to go to the computer and find the browser and type it in and surf there with the results"*). But once set up, screen readers provide an agile environment, allowing a different type of multitasking – virtual multitasking (P659: *"You are able to multitask on the computer whereas a virtual assistant is sequential"*).

Theme 5: Information management and reuse

Another common theme in the survey responses was the lack of information management functionality in voice assistants. The survey participants pointed out that information found using a voice assistant does not persist – it vanishes as soon as it is spoken (P1036: *"[with a screen reader] I am able to easily go back through what I just read. With something like Siri and Cortana, you can't. You have to listen very carefully because they won't go back and repeat"*). Additionally, sharing information with third party apps is impossible to achieve using a VA (P927: *"[with the screen reader] I can copy and paste the info into a Word document and save it for future use."*).

Theme 6: Incidental vs. Intentional Accessibility

One of the valuable features of voice assistants is their voice-first design. Voice assistants are contrasted with screen reader technology, in that they are not translating visual content to audio, but are accessible by virtue of their audio-based design. There is no problem of inaccessible[6] content in voice assistants, and no assistive technology that may crash (P56: *"You don't have to worry about dealing with inaccessible websites."*). Such an approach *"levels the playing field, as it were (everyone searches the same way)."*(P942).

[6] Importantly, for people with other than visual impairments, the accessibility considerations will differ.

Theme 7: Transitioning between modalities

Another theme worth noting is transitioning from a voice assistant to a screen reader. To study this part of respondents' experience, a recent critical incident approach was used and the questions asked participants to describe a case when they started by asking a VA a question, but then switched to using a search engine with a screen reader. Reasons for switching mentioned in participants' incident descriptions included failure of speech recognition, especially when non-trivial words were involved, voice assistants not returning enough details in the answer, returning a non-relevant answer, or no answer at all. When asked about the ideal scenario for a transition between a voice assistant and a screen reader, respondents suggested persisting the assistant's responses by sending an email, or continuing in-depth search with a screen reader (P1078: *"A virtual assistant could give you basic information and then provide a link to view more in depth results using a screen reader."*), and performing in-depth search upon user's request (P1037: *"[A voice assistant] would ask you if you wanted more details. If you replied yes, it would open a web page such as google and perform a search"*).

7.3 VERSE

Inspired by the survey findings and the aforementioned related work, I created VERSE (Voice Exploration, Retrieval and SEarch), a prototype situated at the intersection of voice-based virtual assistants and screen readers. Importantly, VERSE serves as a design probe, allowing to better understand how these technologies may be merged, and how such systems may impact voice assistant-based information retrieval. In this section I describe VERSE in detail. Later, I present the results of a design probe study.

7.3.1 Overview

When using VERSE, people interact with the system primarily through speech, in a manner similar to existing voice-based devices such as the Amazon Alexa or Google Home Assistant. For example, when asked a direct question, VERSE will often respond directly with a concise answer (Figure 7.4a). However, VERSE differs from existing agents in that it enables an additional set of voice commands that allow users to more deeply engage with content. The commands are patterned on those found in contemporary screen readers, for example, allowing navigation over a document's headings.

As with screen readers, VERSE addresses the need to provide shortcuts and accelerators for common actions. To this end, VERSE optionally allows users to perform gestures on a companion device such as a phone or smart watch (see Table 7.3). For most actions, these companion devices are not strictly necessary. However, to simplify rapid prototyping, the microphone activation was limited to gestures, rather than allowing activation via keyword spotting (e.g., "Hey Google"). Specifically, microphone activation is implemented as a double-tap gesture performed on a companion device (e.g., smartphone or smartwatch). Although hands-free interaction can be a key functionality for VA users [116], a physical activation is a welcomed ancillary, and at times, a preferred option [5]. There are no technological blockers for implementing voice-only activation in future versions of VERSE.

The following scenario illustrates VERSE's capabilities and user experience.

7.3.2 Example Usage Scenario

Alice recently overheard a conversation about the Challenger Deep and is interested to learn more. She is sitting on a couch, her computer is in another room, and a VERSE-enabled speaker is on the coffee table. Alice activates VERSE and asks "What is the Challenger Deep?". The VERSE speaker responds with a quick answer – similar to Alice's other smart speakers – but also notes that it found a number of other web pages, Wikipedia articles, and related searches (Table 7.4a). Alice decides to explore the Wikipedia articles ("Go to Wikipedia"), and begins navigating the list of related Wikipedia entries ("next") before backtracking to the first article, this time rotating the crown on her smartwatch as a shortcut to quickly issue the *previous* command (Table 7.4b).

Alice decides that the first Wikipedia article sounded good after all, and asks for more details ("Tell me more"). VERSE loads the Wikipedia article and begins reading from the introduction section (Table 7.4c), but Alice interrupts and asks for a list of section titles ("Read section titles"). Upon hearing that there is a section about the Challenger Deep's history, Alice asks for it by section name ("Read history").

Finally, Alice wonders if there may be other useful resources beyond Wikipedia, and decides to return to the search results ("Go to web results"). As before, Alice rotates the crown on her smart watch to quickly scroll through the results. Alice identifies an interesting webpage from the list VERSE reads out to her, and decides to explore it more deeply on her phone ("Send this to my phone"); the chosen web page opens on her iPhone (Table 7.4d), where Alice can navigate it using the phone's screen reader.

Table 7.3: Mapping of voice commands and corresponding gestures in VERSE. Commands marked Wiki are only valid after a user asks VERSE to read a certain Wikipedia article. To return to all results, the user should switch verticals by saying *"go to ⟨source⟩"*. Sources and levels of granularity are covered in more detail in section 7.3.3.

Voice commands	**Phone gestures**	**Watch gesture**	**Action**
(Activation gesture)	Double tap with two fingers	Double tap with one finger	VERSE opens mic
"Search for *⟨query⟩*"	-	-	VERSE issues the query to Bing API, speaks an answer
"Cancel"	One tap with two fingers	One tap with one finger	Stop voice output
"Go to *⟨source⟩*"	Up/down swipe	Up/down swipe	Previous/next search source
"Next"/"Previous"	Right/left swipe	Right/left swipe or rotate digital crown	Next/previous element
"Tell me more"	Double tap with one finger	-	Provide details if available or open link in the phone app
"Navigate by *⟨level⟩*" (Wiki)	Up/down swipe	Up/down swipe	Previous/next granularity level
"Next"/"Previous" (Wiki)	Right/left swipe	Right/left swipe or rotate digital crown	Next/previous element of the article
"Read section *⟨section name⟩*" (Wiki)	-	-	Read the first section with a partial heading match
"Read section titles" (Wiki)	-	-	Read all titles of the current Wikipedia article

7.3.3 VERSE Design Elements

The design of VERSE was informed by the themes that emerged in the survey. Below I discuss how VERSE directly addresses four of the six themes. The remaining two themes – Text vs. Voice, and Portability vs. Agility – are not directly relevant to VERSE's current focus on voice-based only interaction.

VERSE: Brevity vs. Detail

The design of VERSE addresses *Theme 1* by providing users with an instant concise answer (in a manner similar to voice assistants) as well as an opportunity to explore multiple sources returned by a web search engine (breadth), and then to engage with individual content items and web sources (depth).

With respect to breadth, VERSE leverages the Bing.com search API[7] to collect results across numerous search verticals: web pages, facts, entities, dictionary definitions, Wikipedia articles, news stories, images, videos, and related search queries. Facts are instant answers provided by the Bing Search API, while entities are nodes extracted from the Bing Knowledge Graph.[8] If facts, entities, or dictionary definitions are present, VERSE reads them out similar to existing voice assistants, then follows by summarizing the results available in other verticals (Table 7.4a).

With respect to depth, VERSE allows voice- and gesture-based navigation of Wikipedia articles. Wikipedia was chosen as it has rich data, is often included among the top results, and has a consistent page structure that facilitates semantic navigation. When a desired Wikipedia article is selected, the user can say "tell me more", or perform an alternative gesture (Table 7.3) to get a quick summary of the article (e.g., the number of sections and words), then hear narration of the first section. At any time, the user can ask for an overview of the article's sections (*"read section titles"*), and can ask for a given section by name (*"read ⟨section name⟩"*). When asking to read a named section, VERSE will begin reading the first section which heading contains the words in the user's request. For example, in Table 7.4c, Alice asks *"read History"*. In response VERSE finds the first section with the word *"history"* in it and starts reading it.

[7] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/
[8] https://www.microsoft.com/en-us/bing/apis/bing-entity-search-api

Alice: *(activation)* What is the Challenger Deep?

VERSE: The Challenger Deep is the deepest known point in the Earth's seabed hydrosphere. With the depth 10898 to 10916m by direct measurements from submersibles and slightly more by sonar bathymetry. I also found 2 entities, 9 web pages, 8 related search queries, 10 videos, 10 Wikipedia articles.

(a) Initial search

Alice: *(activation)* Go to Wikipedia.

VERSE: I found 10 Wikipedia articles. The first one is Challenger Deep, Wikipedia. The Challenger Deep is the...

Alice: *(activation)* next article

VERSE: Article 2. Deepsea Challenger, Wikipedia. Deep Sea Challenger, DCV1, is 7.3 metre, 24 feet...

Alice: (*rotate smartwatch crown to return to previous element*)

VERSE: Article 1. Challenger Deep, Wikipedia. The Challenger Deep is the deepest ...

(b) Results navigation

Alice: *(activation)* Tell me more.

VERSE: I parsed the Wikipedia article called Challenger Deep. It has 16 sections and 3127 words. Here's the first section. Introduction. The Challenger Deep is the deepest...

Alice: *(activation)* Read section titles.

Verse: Introduction. Section 1: History of depth mapping from the surface. Section 2: Descents. Section 2.1...

Alice: *(activation)* Read History.

VERSE: Okay, here's the section number 1, titled History of depth mapping from the surface. Over many years, the search for the point of maximum depth...

(c) Wikipedia exploration

Alice: *(activation)* Go to web results.

VERSE: I found 9 web pages. The first one is Challenger Deep, Wikipedia.

Alice: *(rotates smartwatch crown to advance to next result)*

VERSE: Page 2. Deepest ocean Challenger Deep, Extreme Science. Challenger Deep got its name from...

Alice:*(activation)* Send this to my phone.

VERSE: Opening this link in your phone app. *(Web page opens within VERSE app)*

Alice: *(uses VoiceOver gestures to navigate on the page)*

(d) Transition between modalities

Table 7.4: Example usage scenario of VERSE

VERSE: Granularity vs. Ease of Use

To address *Theme 2* from the survey findings, VERSE allows users a quick and easy way to navigate between search results using either voice commands or touch gestures. By saying *"next"* or *"previous,"* the user is able to move on to the next element in the selected search vertical (Table 7.4b). A similar effect is achieved by swiping right and left on a companion device (Table 7.3). These gestures mirror those used by screen readers on popular smart phones.

To switch between different search verticals, a user can say *"go to ⟨search source⟩"* (e.g., *"go to Wikipedia"* or *"go to news"*). VERSE will respond with the number of elements found in the new vertical and start reading the first element (Table 7.4b). Alternatively, the user can swipe up or down to move along the available search verticals. After each swipe VERSE will announce the name of the search vertical by saying it out loud. Upon arriving at each vertical, the user can swipe left or right to navigate between the items in this vertical. If the user tries to move past the last or the first item in the vertical, VERSE plays a short sound indicating that there are no more items.

Finally, when exploring Wikipedia articles, VERSE also supports screen-reader-inspired navigation modes (by headings, sentences, paragraphs, and words). The navigation mode then impacts the granularity of navigation commands and gestures, such as *"next"* and *"previous"*. Without loss of generality, one can switch modes by saying *"navigate by headings"*, or can swipe up or down on a companion device to iterate between modes

– again, these gestures are familiar to people who use screen readers on mobile devices. Similar to switching between the search verticals, VERSE will announce the granularity level by saying “sections”, “paragraphs”, “sentences”, or “words”.

VERSE: Incidental vs Intentional Accessibility

VERSE addresses *Theme 5* by submitting user queries, and retrieving results via the Bing.com search API. This allowed the design of VERSE to be a truly audio-first experience consistent with existing voice assistants, rather than attempting to convert visual web content to auditory format. Likewise, the connection to Wikipedia allows VERSE to focus on the article’s main content rather than on visual elements. This behaviour is consistent with the concept of semantic segmentation [92]. It also mirrors the style of the brief summaries narrated by existing virtual assistants, but allows convenient and efficient access to the entire article content.

VERSE: Transitioning between Modalities

Finally, VERSE addresses *Theme 6* by giving users an opportunity to seamlessly transition between voice-based interaction and a more traditional screen-reader-equipped web browser. If the user requests an in-depth exploration of a web resource that is not Wikipedia, VERSE will open its url within the VERSE phone application. The user can then explore the web page using the device’s screen reader. From this point onward, all gestures are routed to the default screen-reader until a “scrub” gesture is performed[9] or a new voice query is issued. Gesture parity between VERSE and popular screen readers ensures a smooth transition. This interaction is illustrated in Table 7.4d.

7.4 Design Probe

After developing the initial prototype and receiving an approval from the ethics board, 12 blind screen reader users were invited to use VERSE, and to provide feedback their feedback. In the following sections I detail the procedure, describe the participants, and present participant feedback.

[9] A standard VoiceOver gesture for “go back”.

7.4.1 Procedure

Participants completed consent forms, provided demographic information, then listened to a scripted tutorial of VERSE's voice commands and gestures. Each participant was asked to use VERSE to complete two search tasks, and to think aloud as they engaged with the system. One of the tasks was pre-specified and the same for all participants. Specifically, participants were asked to find two or three uses for recycled car tires. This task has previously been used in investigations of speech-only dialogue search systems [177], is characterized as being of intermediate cognitive complexity, and occupies the "Understanding" tier of Krathwohl's Taxonomy of Learning Objectives [16]. Completing the task requires consulting multiple sources or documents, [22], and is thus difficult to perform with contemporary voice assistant. In a second task, participants were asked to express their own information need by searching for a topic of personal interest. Half the participants began with the fixed task, and half began with their own task. Each task had a time limit of 10 minutes.

This design was not meant to formally compare search outcomes on tasks of different difficulties – indeed, we had no control over the difficulty of self-generated tasks. Rather, the fixed task ensured that the participants used a variety of strategies for a moderately complex information need, whereas the self-generated task ensured that a variety of information needs were presented for which we had no advance knowledge. Together, this provided a varied set of experiences with the system that would provoke interesting opportunities for observation and comment.

Regardless of task order, the first search session required participants to use a smart phone for gesture input, while the second session used a smart watch. This order of introduction reflects anticipated real-world use where phones would be the primary controller, with watches an optional alternative.

Throughout the tasks, participants were encouraged to think aloud. Following the completion of both tasks, participants completed the System Usability Scale (SUS) questionnaire [42]. Finally, the experimenter conducted an exit interview, prompting participants to provide open-ended feedback and suggestions. Participants' comments during the study, and their responses to the interview questions, were transcribed and analyzed by two researchers using a variation of open coding and affinity diagramming [108].

7.4.2 Participants

A total of 12 blind screen reader users (4 female, 8 male) were recruited through a mailing list in the local community. Participants were reimbursed $50 for their time. The participants' transportation costs to the study location were offset by up to $50. The study lasted about an hour.

Participants' average age was 36.6 years old (σ = 13.8 years). Seven reported being totally blind and five were legally blind but had some residual vision. Ten participants had their vision level since birth, and two reported having reduced vision for 15 or more years. Participants had an average of 18.5 years of experience with screen readers (σ = 7.6 years), and 5.7 years of experience with voice assistants (σ = 2.5 years). For comparison, at the time that the study was conducted, Apple's Siri had been available on the market for 6.9 years, suggesting that the participants were indeed early adopters of this technology.

7.5 System Usability

All participants successfully completed the fixed search task, which required that they identify at least two uses of used car tires. Though it was difficult to apply a common measure of completeness or correctness for user-chosen queries, participants indicated satisfaction with VERSE's performance, as is reflected in open-ended feedback, and in responses to items on the System Usability Scale.

VERSE received a mean score of 71.0 (σ = 15.5) on the System Usability Scale. To aid in interpretation, note that this score falls slightly above the average score of 68, reported in [159], and just below the score of 71.4, which serves as the boundary separating systems with "Ok" usability from those with "Good" usability, according to the adjective rating scale developed by Bangor et al. in [24]. Breaking out individual items, it was found that most participants found VERSE to be "*easy to use*" (median: 4, on a 5-point Likert scale), and its features were "*well integrated*" (median: 3.5). Likewise, participants "*felt very confident using the system*" (median: 4), and reported that they would "*use the product frequently*" (median: 4). These results suggest that the VERSE prototype reached a sufficient quality to serve as a design probe, and to ground meaningful discussions of VERSE's capabilities.

7.6 Participant Feedback

Participants commented on VERSE throughout use, and answered questions about the prototype in an exit interview. Here, participants' feedback was generally positive, and largely aligned with their responses to SUS items, described above. For instance, participants reported that the system was easy to learn, given prior experience with screen readers ("*if we're talking about screen reader users, they kind of know what they are doing, I think it would be fairly easy,*" P4). In this capacity, VERSE's gesture accelerators were especially familiar ("*the touch experience doesn't feel that different from VoiceOver (...) I think I would have probably figured them out on my own,*" P3; "*[Y]ou're just using the same gestures as VoiceOver, and that, in itself, is comprehensive.,*" P5).

Participants also found that VERSE extended the capabilities of voice assistants in meaningful ways, increasing both the depth and breadth of exploration. For instance, P4 reported:

> *"The information it gives is quite a bit more in-depth. [...] There was one time I asked Siri something about Easter eggs. Siri said 'I found this Wikipedia article, do you want me to read it to you?' [...] It only read the introduction and then stopped, and I think [VERSE] could come in so that you can read whole sections."*

Likewise, P7 reported:

> *"[VERSE] gives you a lot more search options like web pages, or Wikipedia. Even though the smart speaker I use [Echo] has some ability to read [Wikipedia], I can't get back and forth by section and skip around. In that way, it's an improvement. I like it."'*

However, participants were more mixed about how VERSE compared to traditional screen readers. For instance, P7 noted "*screen readers are a lot more powerful*", whereas P6 noted "*I like it better than desktop screen readers, but I would probably prefer phone screen readers.*" VERSE was never intended to replace screen readers, and was instead focused on extending the web search and retrieval capabilities of voice assistants with screen-reader-inspired functionality. This point was immediately recognized by P5, who noted:

> *"I think [VERSE and Screen Readers] are fundamentally different. There's just no way to compare them. Screen readers aren't for searching for stuff, they are about giving you control."*

Restricted to the domain of web search and retrieval, VERSE was found to confer numerous advantages. P10 commented that, compared to accessing web search with a screen reader, VERSE was "*Much better. This gives you much more structure.*" P3 elaborated further:

> *"Most screen readers and search engines do use headings, [...] but it's hard to switch [search verticals]. This is different and kind of interesting. It seems to put you at a higher level."*

This sentiment was echoed by P5, who explained:

> *"One thing that immediately caught my eye was that different forms of data were being pulled together. When you go to Google and you type in a search you just get a stream of responses. [VERSE] gathers the relevant stuff and groups it in different ways. I really did like that."*

Additionally, participants expressed a strong interest in voice, often preferring it to gestural interaction. For instance, P8 stated "*Just using voice would be fine with me.*", while P7 noted:

> *"I preferred voice integration. There were times where it's just going to be faster to use my finger to find it, but mostly [I preferred] voice."*

Other participants offered more nuanced perspectives, noting that gestures were advantageous for high-frequency navigation commands. ("*I liked being able to use the gestures. [With voice] it would have been 'next section', 'next section.'* ", P6; "*I liked the gestures. I will spend more time with gesture, but getting this thing started with voice is beautiful.*", P9).

Nevertheless, participants reported concerns that voice commands were difficult to remember (e.g., "*I didn't find the system complicated. I'd say the most complicated part is the memorization of [...] the voice commands.*", P3). To this end, participants expressed a strong desire for improvements to conversation and document understanding. For instance, P3 expressed "*I should just have the ability to use [a] more natural voice like I'm having a conversation with you.*" Likewise, P5 explained:

> *"I'm most passionate about the whole language understanding part, where I [would like] to say 'read the paragraph that talks about this person's work' and it should understand."*

Recent results in machine reading comprehension and question answering [133] may provide a means of delivering on this promise; this remains an important area for future work.

Finally, all 12 participants preferred using the phone over the watch. Several factors contributed to this preference including: the limited input space of the watch ("*I've got fat fingers [...] and on that device feels very cumbersome*", P9), a power-saving feature that caused the screen to occasionally lose focus ("*It was a little annoying [when] I lost focus on the touch part of the screen*", P3), and latency incurred by the watch's aggressive powering-down of wireless radios ("*The watch wasn't bad, but it lagged a little. That was my chief complaint." P7*).

In sum, participants were generally positive about the VERSE prototype, and expressed interest in its continued development or public release. The design probe further revealed that participants were especially positive about voice interaction, and the expanded access to web content afforded by VERSE. While one could hypothesise that watch-based interaction would be an asset (given that watches are always on hand), their appeal is diminished by the limitations of current form factors and hardware. Conversely, extending the conversation and document understanding capabilities of VERSE is a desirable avenue for future work.

7.7 Discussion

This chapter aimed to answer *RQ2-c: How might voice assistants and screenreaders be merged to confer the unique advantages of each technology?* The described investigation described above consisted of an online survey with 53 blind web searchers, and collecting user feedback about a system prototype informed by the survey findings.

The survey results revealed that screen readers and voice-based digital assistants present a series of trade offs spanning dimensions of brevity, control, input modality, agility, incidental accessibility, and paradigm transitions. The respondents reported that transitions between the technologies can be especially costly. The prototype aimed to eliminate these trade offs and costs, by adding screen reader-inspired capabilities to a voice-assistant. An alternative approach would have been to augment a screen reader with voice and natural language controls, which, as noted earlier, has been explored in prior literature [18, 200]. The decision was made to opt for the former since voice assistants are an emerging technology that open a new point in the design space, while also avoiding challenges with legacy bias [126]. For example, VERSE redefines search results pages by adding summaries, and by mapping screen reader navigation modes to search verticals. These features were received positively by design probe participants. Future work could compare VERSE to screen readers (or voice assistants) in a controlled laboratory study to determine if participants' stated preferences are reflected in measurable reductions in task performance time or other performance metrics. Additionally, coexistence and complementary nature of voice assistants and screen readers brings up new research questions raised by the survey findings such as whether these two technologies should remain separate, be merged into a single technology, or be more carefully co-designed for compatibility.

The findings support the statement of this thesis. Participants were able to easily communicate their intent to the system using gesture controls implemented by VERSE which were equivalent to those used by VoiceOver (a screen reader all of the participants had much experience using). Additionally, people wanted to see more natural language commands being recognised and the ability to navigate in the articles based on their content. The request for the flexibility of parameters is supported by the survey responses as well as prior literature related to information accessibility for people with visual impairments: the comprehension speed by people who are blind can be much higher compared to sighted people. Furthermore, the choice of voice could differ from that preferred by sighted population. Finally, the diversity of information provided by VERSE was pointed out by the study participants as beneficial, and survey respondents pointed out the inability to access a variety of resources as one of the main shortcomings of voice-based assistants. The same aspect was seen as an advantage of screen readers.

The online survey served as a data collection tool to inform the design of VERSE. Alternative ways for collecting high-quality qualitative feedback could be used, including interviews and contextual inquiries. The conducted survey used a "recent critical incident" approach [70], paired with open-ended survey questions, which provided rich data and served as a way to reach a large and geographically diverse audience.

Voice-based digital assistants are frequently used to complete tasks beyond web search and retrieval. In these settings, a similar set of VA limitations are likely to arise. For example, a VA might read recent messages, or help compose an email, but is unlikely to provide granular navigation of one's inbox folders. Generalizing VERSE to scenarios beyond web search is an exciting area of future research.

Furthermore, other user communities may also benefit from VERSE. For instance, sighted users may wish to have expanded voice access to web content when they are driving, cooking, or otherwise engaged in a task where visual attention is required – especially if VERSE were enriched with the document and conversation understanding capabilities discussed earlier. VERSE may also benefit other populations with print disabilities, such as people with dyslexia, who also have challenges using mainstream search tools [127]. In addition, all the survey participants were based in the U.S. Understanding the voice search needs of people from other regions [28, 146] is a valuable area of future work.

Finally, rather than accessing raw HTML, VERSE leverages APIs for Bing and Wikipedia to provide an audio-first experience. This is similar to other smart speaker software applications known as "skills." For general web pages, VERSE encounters the same challenges with inaccessible content as traditional screen readers. Given the broad appeal of smart speakers, it is possible that experiences such as VERSE could motivate web developers to consider how their content would be accessed through audio channels. For example, a recent proposal[10] demonstrates how web developers can tag content with Schema.org's `speakable` HTML attribute to help direct the Google Assistant to the parts of an article that can be read aloud.

7.8 Chapter Summary

In this chapter, I described a study aimed at understanding what challenges people with visual impairments face when searching for information online and how voice assistants can be improved by borrowing functionality from screenreaders. The investigation consisted of two parts: an online survey with 53 legally blind adults, and a design probe study of the system prototype, called VERSE, with twelve blind participants. This chapter continued to answer *RQ2* for the specific use case by people with visual impairments. It concludes the series of investigations of current and potential future voice-based digital assistants. In the next chapter, I summarise and discuss the findings of all the investigations that constituted this thesis, provide clear design guidelines, and outline potential directions for future work.

[10] https://developers.google.com/search/docs/data-types/speakable

Chapter 8

Discussion

In this thesis, I described research studies aimed to investigate how voice-based dialogue systems should be designed for different use cases and different user groups. In writing of this thesis, I was guided by two main research questions:

- RQ1: How would users perceive digital agents that could understand them as well as their fellow humans?
- RQ2: How can we improve interaction using currently available tools?

Commercially available dialogue systems are designed to mimic superficial aspects of human-to-human conversation and are able to make jokes[1], recognise variety of natural language requests, and have a human-like voice all of which lead to high user expectations. However, because of their rules-based nature, assistants may generally be quite brittle and unable to maintain an interaction that goes beyond a set of pre-defined topics or intents. Experiments described in chapters 4 and 5 of this thesis entertain and explore the idea of a digital agent that is capable of interacting with a person fluently, correctly recognising and reacting to explicit requests as well as implicit behaviours. The results of these experiments show that most people are embracing the human-like interaction and respond in a way they would to another human. Moreover, the results suggest that in some cases users prefer speaking to a computer over speaking to another person to avoid the embarrassment or social burden. This sentiment is also supported by prior work of Lucas et al. who demonstrated that people may be more likely to open up and show their emotions in an interaction with a computer than with a person [115].

[1] https://www.digitaltrends.com/home/funny-things-to-ask-alexa/, accessed April 2021

On the other hand, a more rigid command-based approach can be preferred over a human-like interaction by certain users. A small fraction of study participants in chapters 4 and 5 preferred an automated agent or did not exhibit behaviour characteristic to a human interaction. This was a signal that not all users may enjoy and prefer dialogue systems mimicking human behaviour. Prior work by Branham et al.[38] further supports this argument by demonstrating that people with visual impairments find command-based interfaces easier compared to conversation-like interaction.

While the level of existing technology prevents developers from creating an agent capable of a human level interaction, existing dialogue systems can be advanced using available tools. Chapter 7 shows that a dialogue system can benefit greatly by expanding the amount of information is provides the users access to. Current voice-based digital assistants have a big potential to serve as a universal interface, as they are already used as accessibility tool by people with visual impairments [5], as well as by sighted people when during situational impairments – such as hands or eyes occupied by a different task [116].

While designing a voice interface requires a different set of considerations than designing a graphical interface [195], the experiment described in chapter 6 shows that for the specific use-case of information search, the amount and type of content required to judge the relevance of a search result is similar for both audio- and text-based interfaces. The study found no significant difference between audio- and text-based interfaces for selecting a single most relevant search result. Prior work shows that users usually do not examine a large number of search results before making their selection [199, 8, 96], therefore identifying the best search result out of the top few would be the most frequent task in a real-life scenario.

8.1 Design Recommendations for Voice-based Dialogue Systems

The design recommendations that follow from the results of this thesis, are centered around four interaction aspects appearing in the thesis statement: user's ability to **c**ommunicate their **i**ntent to the system (CI), **u**nderstandability of system's **r**esponses (UR), **f**lexibility of system's **p**arameters (FP), and **d**iversity of **i**nformation provided by the system (DI).

To enable users to easily communicate their intent to a voice-based dialogue system, based on the findings of the experiments comprising this thesis, I provide the following recommendations:

- the system should be able to recognise a wide variety of natural language commands for every action that it is able to take (CI);

- the system should be aware of the content it is returning. It should have a reading comprehension module to help users navigate the content of an article, it should be able to discern the follow up questions asking for expanded information or clarification. If returning a list of items, it should be able to navigate along that list based on the content and order of the items. A system will benefit from understanding the users' intents phrased in explicit as well as implicit ways. However, this ability should be weighed against user's priorities and privacy preferences (CI);
- the system should provide evidence and sources of the information it is returning (UR);
- if a system's response consists of several parts, these parts should be separated from each other using sounds (earcons) or intonations (UR);
- the length of a system's responses should be predictable. Either by setting the length to be constant, or by indicating the end of the answer using sound (UR);
- the content returned over audio channel should not break and should represent a coherent piece of text (UR);
- the system should enable the users to adjust the majority of its parameters including: audio settings (how robotic or human it is, the pitch and speed, gender in case voice is chosen to be human-like). The length of the answer should be adjustable. Any privacy settings should be transparent and adjustable. If a system is operating in a variety of settings (mic on/off), these settings should be also adjustable. A system should be able to hand off controls to another application where the user can explore more freely (FP);
- the system should strive to provide users with access to a variety of information: search verticals, full wikipedia articles, aggregate and summarise opinions, access various sources (DI).

The system should be aware of the content it is returning to enable its users to refer back to that content for follow up questions, elaboration requests, and to not have to phrase their question as one including many details. This finding is supported by results of chapters 4, 6 and 7.

8.2 Future work

Considering the findings discussed throughout this thesis, in this section I reflect on potential directions future work could pursue.

8.2.1 Continuous Interaction

The strict protocol of ⟨*wake word, question, answer*⟩ causes the interaction with a voice assistant to be disjoint and causes a phenomenon called "wake word fatigue" – a repetition of the wake word multiple times in a row causes frustration. The alternative way to build an interaction with a voice-based digital assistant is to allow the microphone to be left on. This setting can be troublesome in several ways. Separating commands directed to the assistant from other sounds happening in the household is uniquely challenging. Especially so if multiple people are present next to the assistant. Arguably an even bigger concern may be related to privacy – keeping a "spying" device is not an appealing option. Future work could explore the middle ground between the two options: having the microphone on for a certain amount of time when several interactions in a row are likely to happen. Considering the variability in users' preferences discussed earlier in this thesis, users of voice assistants are likely to favour different settings. Additionally, individual use cases and scenarios may require adjustments. Understanding the taxonomy of users' preferences and use cases will inform potential future system design. An important part of this investigation will include an exploration of how interactive systems can make their settings known to and understood by the users. A mismatch between user expectations and system activity can lead users to mistrust the system.

An interaction between a person and a voice-based digital assistant is generally one-sided and initiated by the user. In other words, all system's actions are completed as a response to the user. The area of mixed-initiative interaction is focusing on organising an interaction between a human and a computer where each side contributes what is best suited at the most appropriate time [13]. Recent work began exploring the settings in which voice assistants' proactive behaviour would be welcomed by the user [47]. The results indicated that additional sensors will be required to determine the user's interruptibility (e.g. position in the room, movement). Such sensors would carry an additional privacy burden on the user, especially in cases when the user is not fully familiar with how data is being collected and stored. An investigation of how to effectively reveal data collection processes to the user will increase people's understanding of how their device work, and will open way to more relaxed interaction protocols with voice assistants.

8.2.2 Customisation of Parameters

Considering the multitude of parameters affecting the behaviour of voice assistants, an effective way to customise these parameters is required. Prior work explored how user behaviour, and gaze distribution in particular, connect to user's preferences of a graphical user interface design [12]. A similar approach of monitoring user behaviour to determine user preferences may be used to adjust the settings of a digital assistant. An alternative route to explore is for a digital assistant to proactively ask about the preferred settings in opportune moments, simultaneously achieving the goal of revealing settings which a user may not otherwise have known about. Such investigation should also consider whether user preferences may change over time or depending on the user's activity, mood, or surrounding context.

8.2.3 Comparison of Command-based with Conversation-like Approaches

The majority of experiments described in this thesis focused on information search as a use case for digital assistants. Conversational search is a popular area of research proposing that a user engages in a back-and-forth dialogue with a system to iteratively narrow down the information need and finally retrieve the necessary materials. This process is akin to a conversation with a librarian who helps the user to discover the desired resources. As mentioned earlier, the setting of a conversation-like exchange with an automated system may not be the best fit for all populations, especially the needs of people with disabilities should be considered, or use-cases. In such cases, a command-based voice digital assistant could be considered as an alternative. Future work could compare two types of interactions: conversation-like and command-based. It should outline scenarios where each one is preferred and investigate whether the benefits one has over the other. Such comparison is applicable not only for the task of information search, but for broader use-cases as well.

8.2.4 Universal Interface

Graphical and visual content often suffers from lack of accessibility for people with disabilities, and people with visual impairments in particular. Voice-first design overcomes these problems since all of its content is audio-based and therefore accessible for hearing individuals. In chapter 7, I described experiments with VERSE – a prototype of an information seeking system for people with visual impairments. During the process, a number design implications were discovered that catered to people with visual impairments. Future work could explore how a system like VERSE could be used by sighted people. Such investigation could focus on various aspects such as situations in which sighted people benefit from interacting with a voice-only search system, what types of information is being sought, whether and how the system design would change to satisfy sighted users. The phenomenon in which general population benefits from solutions designed for people with disabilities is commonly known in HCI as "curb cutting". Some examples of curb-cutting are audio announcements of stations in buses and closed captions. The results of experiments that include a wide variety of user types would begin making steps towards shaping a universal interface for information search.

8.2.5 Parallels with Visual Iinterfaces for Search

As the design of voice interfaces for search moves forward, the researchers may find it beneficial to draw parallels with investigations of visual search interfaces and explore whether similar findings are applicable to voice-based interfaces. For example, a number of researchers explored how user behaviour, such as cursor activity, can signal which search results people found relevant to their information need. Future work should explore which factors signal that a user is or is not satisfied with the search results they received. Such factors could include user's responses to the system, however this opportunity would only become available under a relaxed interaction protocol where the microphone stays turned on after the interaction. Similarly questions regarding the length of a summary of a search result, the number of results returned, the ways to indicate the relevance of a search result to one's information need have all been thoroughly explored for visual search interfaces. As voice-based interfaces for search develop, analogous questions will arise and will need to be answered in order to improve the interaction.

As I discussed prior, the main appeal of voice interfaces is the fact that they enable people to multitask. For example, voice-based interfaces can be used while driving, cycling, or walking – settings in which the users are not able to look at or interact with a screen-based device. However, when designing for multitasking, an important factor to consider is users' cognitive load, or in other words, how effectively they can complete multiple tasks at the same time. Information delivered over voice requires incurs more mental load on the listener than information delivered visually. This factor makes developing voice-based interfaces for multi-tasking all the more challenging. Recent work studied usability of using completing popular tasks using Siri while driving [105]. The results showed that Siri did not provide a truly eyes-free experience and was not safe to use for most drivers. In the next section, I describe an experimental design that may be followed by researchers pursuing this direction in the future.

8.2.6 Voice Interface for Driving

Drivers around the world have long been discouraged from using their phones behind the wheel and postpone phone interactions until it is safe to do so. However, with people spending more and more time on their daily commute, there is a desire to utilise the time spent in the car productively. This prompted an emergence of a number of built-in car voice assistants [111]. Researchers examined a wide variety of tasks that could be performed using voice-only interaction mode. For example, Martelaro et al. [120] conducted a study where participants worked on a presentation on their way to work, Large et al. [104] focused on comparing distractions imposed by a phone conversation and interactions with a simulated voice assistant, and a number of studies attempted to predict opportune moments to talk to the driver [166, 162, 161].

As discussed in previous chapters, at times voice assistants revert to displaying information on a screen of a companion app. A study by Larsen et al. [105] showed that providing information to users in a visual way was distracting to their driving and caused crashes in a car simulator. Therefore a need for in-car voice-only assistant becomes apparent. However, audio distraction alone can be detrimental to driver's attention. The experimental design outlined below aims at quantifying the distractedness caused by conducting voice-based search while driving. The results of a study following this setup should provide insight into how well drivers can complete search tasks of varying complexity while behind a wheel.

Measuring the vehicle

Studies investigating various aspects of distracted driving are divided into ones involving real driving and ones using driving simulators. When running a study in real driving conditions, measures should be taken to not put at risk the participants as well as other drivers and passengers on the road. On the other hand, when using a driving simulator, participants may pay less attention to the road than they would in real life. In both cases, prior work employed a number of measures to quantify driver's control of the vehicle as well as driver's physical state.

To estimate the driver's control of the vehicle while driving a car, data from CAN bus (Controller Area Network) can be used. CAN bus system incorporates a various ECUs (Electronic Control Units) of the car and facilitates necessary exchange of information between ECUs. CAN is standard in automotive machinery and is included in cars, buses, ships, planes, etc. CAN loggers capture information about the vehicle including steering wheel position, pedal pressure, speed, and possibly others. The signals available from a driving simulator cover all of the main ones but may vary depending on the exact software used.

As measuring the vehicle parameters can help estimate the quality of driving, having access to the driver's physiological indicators can help estimate their levels of stress and distractedness. To this end, various measures have been used: heart rate and heart rate variability [104], galvanic skin response (GSR), eye movements [65]. Driver's workload can also be evaluated using NASA TLX and it's derivative – DALI (Driving Activity Load Index) [137].

In the presented study setup, I propose to use measures collected from the driving simulator: vehicle speed, steering wheel position, accelerator pressure, and vehicle position in the lane. Additionally, I propose to use NASA TLX to evaluate driver's subjective workload.

Reference tasks

In aiming to measure how cognitively demanding a certain task is, the natural way is to compare it with other tasks. To this end, various secondary task are usually compared to standard ones. One of the popular tasks is called N-back digit recall and consists of a participant listening to a sequence of digits and repeating a digit that occurred N digits ago. For example, during 0-back digit recall a participant would repeat a digit they had just heard, while during a 1-back digit recall task a participant would repeat the previous digit. Typically, 2-back digit tasks are used in studies investigating driver workload. In the current study, I propose to use 2-back digit recall to serve as a comparison with a task of voice-only web search.

Alternative comparison tasks may also include other activities drivers normally engage in. For example having a conversation with a passenger, talking on the phone, setting up GPS navigation, controlling the music, or listening to radio or a podcast.

Voice assistant for search

In designing the voice assistant, I did not pursue the goal of creating a conversational user experience. In contrast, I set out to explore an voice-based command-based interface. The design of the voice-based interface was informed by the findings of chapters 6 and 7.

The proposed interfaces possesses the following functionality:

- submit a new query,
- navigate to the next and previous search result,
- read the document,
- skip ahead or back 3 seconds,
- increase or decrease speech rate,
- repeat previous phrase,
- cancel current action.

After a user submits a new query, the system returns a list of five search results and starts reading out the first result, or an "instant answer" if one is available in the same manner as VERSE did in chapter 7. Then the user can continue exploring the results by pronouncing any of the above commands.

When synthesising audio clips for a search result, I used insight based on our study described in chapter 6. In particular, a single audio result was generated by combining the name of a website and a summary with sentences amended from the website's content. Additionally, to emphasise the website name and separation between different composite parts of a result, I used a lower pitch setting to generate the website name, and added a 0.5 second pause before the beginning of the summary. To emphasise the end of a search result, I added a "beep" sound in the end of each summary. To manipulate these prosodic features, I used Google's Text-to-Speech (TTS).

To enable alternative to voice controls, as was seen beneficial in chapter 7, I used the buttons of the steering wheel as controls for the assistant. Additionally, I implemented a push-to-talk protocol: the driver needed to push and hold a button while they talk and release it once they were done pronouncing the command. All buttons were labelled in case the participants forgot the initial briefing.

To facilitate search results retrieval, I used Google Cloud services: Text-to-Speech and Google Search. When using Google Search, I somewhat unexpectedly, found out that specifying domains ".com" and ".gov" enabled me to search a significant portion of the web. To map user utterances to intents I used Facebook's free service called Wit.ai service[2].

[2] https://wit.ai

Figure 8.1: The author uses our laboratory setup to test the study setup. The microphone on the right is used to communicate with the voice assistant. The steering wheel is a part of Logitech Momo set. The view on the screen is showing Beam.NG “East USA” route.

Simulated driving setup

To avoid the possibility of physical injury to participants and the experimenter, I simulated a driving environment in the laboratory. For this purpose, I used a BeamNG.drive [3] game installed on a Windows 10 PC via Steam [4]. I used a free-roaming mode which allowed participants to drive at their own speed and no other traffic present on the roads. I used a loop road on East Coast USA map which was not extremely challenging, though still had a three turns and began and finished with a stop sign.

I used Logitech Momo racing force feedback wheel and pedals to simulate the process of driving. I also used an external condenser microphone placed right next to the steering wheel to ensure the high quality of speech recognition but a far field microphone could be used instead.

Study design

The selected driving route forms a closed loop. Therefore a single condition per loop can be assigned. The first loop is used for practice – getting used to the controls and the route. The next three loops are used for each condition: (1) driving with no distraction,

[3] https://www.beamng.com/

[4] https://store.steampowered.com/app/284160/BeamNGdrive/

(2) driving with 2-back digit recall, and (3) driving with voice-based search. To keep track of user load, participants can be asked to fill out NASA TLX after each loop. Throughout the experiment data should be collected with respect to the steering wheel position, gas and brake pedal positions, and car position within lane as described above. To understand whether search task complexity plays a role in driver's cognitive load, the complexity of the search tasks can be varied from *Remember* to *Analyse* categories according to Bloom's taxonomy used in chapter 6.

8.3 Chapter summary

In this chapter, I discussed the findings of each research chapter of this thesis, how they help answer the high level research questions posed in the Introduction. Furthermore, I outlined how the findings of each chapter were reflected in the statement of this thesis. Finally, I reported on potential directions for future work.

Chapter 9

Conclusion

This thesis aimed at exploring how people may interact with digital assistants capable of human-like interaction and how existing digital assistants can be improved using currently available levels of technology. The experiments described throughout this thesis resulted in a set of design guidelines for digital assistants that can be divided into four categories reflected in the statement of this thesis:

> *A productive interaction with a dialogue system critically depends on how naturally a user can communicate their intent to the system, the understandability of the system's responses, the flexibility of the system's parameters, and the diversity of information accessible through the system.*

9.1 Summary of Contributions

The research presented in this dissertation makes the following contributions.

Comparative analysis of user perception of human-powered and automated text-based dialogue systems. Chapter 4 outlines the details of a user study in which participants conducted web search tasks using automatic and human-powered assistants. Unbeknownst to the participants, one of the automated assistants was human-powered. The results of the study suggest that many users embrace the human-like interactions with an automated assistant. Furthermore, in certain cases an interaction with a digital assistant can be preferred to that with another person due the absence of social norms that should be followed. A list of design recommendations for digital assistants is proposed as a result of this study.

Analysis of user language during interactions with a voice-based digital assistant. Chapter 5, describes a detailed analysis of the language used by study participants while guided through a culinary recipe by a voice-based digital assistant implemented using a Wizard-of-Oz protocol. The results of the analysis indicate that although the digital assistant possessed basic functionality, participants nevertheless used highly conversational language typical to that of a human-to-human interaction. Additionally, the chapter lays out a list of opportunities for how this language can be taken advantage of to improve the human-assistant interaction.

Design guidelines for synthesising audio-only search results. Chapter 6 reports on results of two user studies comparing user search result preferences when the search results are presented in a visual or an auditory interface. The findings demonstrate that the search results selected using the text interface are more consistent with the ground truth ranking than results selected using the audio interface. However, in selecting a single most relevant result, both interfaces result in similar choice consistency suggesting that the content of the web search results used in traditional visual interfaces is enough to make an accurate selection. The chapter concludes with a list of design guidelines for presenting web search results using an audio-only interface.

Incorporating features of screen readers into a voice-based digital assistant. Chapter 7 discusses the idea of extending the capabilities of a voice-based digital assistants with the functionality of a screen reader – the most popular accessibility tool for people who are visually impaired. The chapter describes the implementation and usability testing of a prototype based on the insight collected from an online survey of 56 people who are blind. The results promote the extended functionality of voice-based digital assistants borrowed from existing accessibility solutions.

The results of the experiments described in this dissertation demonstrate that digital assistants can be of help whether or not they support human-like way of interaction or are command-based. Further, digital assistants are a valuable technology that has many ways to benefit general population during situational impairments and multitasking, users with visual impairments. In certain cases digital assistants can even be preferred to a human operator. As technology develops, users will adapt and adjust their behaviours as they do today. However, when developing future digital assistants, system designers should take a variety of interaction aspects into consideration and allow for users to customise the behaviour of digital assistants in their preferred way.

References

[1] Alexa skills kit. https://developer.amazon.com/alexa-skills-kit, 2017. [Online; retrieved 5-Jan-2018].

[2] The cortana skills kit. https://docs.microsoft.com/en-us/cortana/getstarted, 2017. [Online; retrieved 5-Jan-2018].

[3] Google duplex: An ai system for accomplishing real-world tasks over the phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html, 2018. [Online; last accessed 3-Apr-2021].

[4] Ali Abdolrahmani and Ravi Kuber. Should I trust it when I cannot see it? Credibility assessment for blind web users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 191–199, 2016.

[5] Ali Abdolrahmani, Ravi Kuber, and Stacy M Branham. "Siri talks at you": An empirical investigation of voice-activated personal assistant (VAPA) usage by individuals who are blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 249–258. ACM, 2018.

[6] Ali Abdolrahmani, Ravi Kuber, and William Easley. Web search credibility assessment for individuals who are blind. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 369–370. ACM, 2015.

[7] Ali Abdolrahmani, Kevin M Storer, Antony Rishin Mukkath Roy, Ravi Kuber, and Stacy M Branham. Blind leading the sighted: Drawing design insights from blind users towards more productivity-oriented voice interfaces. *ACM Transactions on Accessible Computing (TACCESS)*, 12(4):1–35, 2020.

[8] Mustafa Abualsaud and Mark D Smucker. Patterns of search result examination: query to first action. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1833–1842, 2019.

[9] Eyal Aharoni and Alan J Fridlund. Social reactions toward people vs. computers: How mere lables shape interactions. *Computers in human behavior*, 23(5):2175–2189, 2007.

[10] Faisal Ahmed, Yevgen Borodin, Andrii Soviak, Muhammad Islam, IV Ramakrishnan, and Terri Hedgpeth. Accessible skimming: faster screen reading of web pages. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 367–378, 2012.

[11] G. Aist, J. Dowding, B. A. Hockey, M. Rayner, J. Hieronymus, D. Bohus, B. Boven, N. Blaylock, E. Campana, S. Early, G. Gorrell, and S. Phan. Talking through procedures: An intelligent space station procedure assistant. *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, 2:187–190, 2003.

[12] Hosam Al-Samarraie, Samer Muthana Sarsam, and Hans Guesgen. Predicting user preferences of environment design: a perceptual mechanism of user interface customisation. *Behaviour & Information Technology*, 35(8):644–653, 2016.

[13] James E Allen, Curry I Guinn, and Eric Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999.

[14] James F Allen, Donna K Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. Toward conversational human-computer interaction. *AI magazine*, 22(4):27–27, 2001.

[15] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. Music, search, and iot: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(3):1–28, 2019.

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

[17] Jaime Arguello, Wan-Ching Wu, Diane Kelly, and Ashlee Edwards. Task complexity, vertical display and user interaction in aggregated search. *Proceedings of Conference*

on Research and Development in Information Retrieval (SIGIR), pages 435–444, 2012.

[18] Vikas Ashok, Yevgen Borodin, Yury Puzis, and IV Ramakrishnan. Capti-speak: a speech-enabled web screen reader. In *Proceedings of the 12th Web for All Conference*, page 22. ACM, 2015.

[19] Anne Aula. Enhancing the readability of search result summaries. 2004.

[20] Anne Aula, Rehan M Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 35–44. ACM, 2010.

[21] Bernard J Baars and Nicole M Gage. *Cognition, brain, and consciousness: Introduction to cognitive neuroscience.* Academic Press, 2010.

[22] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. User variability and ir system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 625–634. ACM, 2015.

[23] Matthias Baldauf, Raffael Bösch, Christian Frei, Fabian Hautle, and Marc Jenny. Exploring requirements and opportunities of conversational user interfaces for the cognitively impaired. In *Proceedings of the 20th International Conference on human-computer interaction with mobile devices and services adjunct*, pages 119–126, 2018.

[24] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[25] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[26] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–24, 2018.

[27] Jonathan Berry. Apart or a part? access to the internet by visually impaired and blind people, with particular emphasis on assistive enabling technology and user perceptions. *Information technology and disabilities*, 6(3):1–16, 1999.

[28] Apoorva Bhalla. An exploratory study understanding the appropriated use of voice-based search and assistants. In *Proceedings of the 9th Indian Conference on Human Computer Interaction*, pages 90–94. ACM, 2018.

[29] Jeffrey P Bigham, Anna C Cavender, Jeremy T Brudvik, Jacob O Wobbrock, and Richard E Ladner. Webinsitu: a comparative analysis of blind and sighted browsing behavior. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 51–58. ACM, 2007.

[30] Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. Webinsight:: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 181–188. ACM, 2006.

[31] Jeffrey P Bigham, Irene Lin, and Saiph Savage. The effects of not knowing what you don't know on web accessibility for blind web users. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 101–109. ACM, 2017.

[32] Jeffrey R Binder, Julie A Frost, Thomas A Hammeke, Patrick SF Bellgowan, Jane A Springer, Jackie N Kaufman, and Edward T Possing. Human temporal lobe activation by speech and nonspeech sounds. *Cerebral cortex*, 10(5):512–528, 2000.

[33] Dan Bohus and Alexander Rudnicky. Larri: A language-based maintenance and repair assistant. *Spoken multimodal human-computer dialogue in mobile environments*, pages 203–218, 2005.

[34] Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV Ramakrishnan. More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, page 13. ACM, 2010.

[35] Yevgen Borodin, Jalal Mahmud, IV Ramakrishnan, and Amanda Stent. The hearsay non-visual web browser. In *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, pages 128–129, 2007.

[36] Horatiu S Bota. *Composite web search*. PhD thesis, University of Glasgow, 2018.

[37] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. A large inclusive study of human listening rates. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[38] Stacy M. Branham and Antony Rishin Mukkath Roy. Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 446–458, New York, NY, USA, 2019. Association for Computing Machinery.

[39] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. What do you mean exactly? Analyzing clarification questions in cqa. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 345–348, New York, NY, USA, 2017. Association for Computing Machinery.

[40] Bret Kinsella. Nearly 90 million u.s. adults have smart speakers, adoption now exceeds one-third of consumers, 2020. URL: https://voicebot.ai/2020/04/28/nearly-90-million-u-s-adults-have-smart-speakers-adoption-now-exceeds-one-third-

[41] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM New York, NY, USA, 2002.

[42] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[43] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[44] Georg Buscher, Susan T Dumais, and Edward Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 2010.

[45] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. Differences in the use of search assistance for tasks of varying complexity. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 23–32, 2015.

[46] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. Overview of the trec 2014 session track. Technical report, DTIC Document, 2014.

[47] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. Hello there! is now a good time to talk? Opportune

moments for proactive interactions with smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–28, 2020.

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

[49] Monica Chin. We still haven't seen the galaxy home — and that's a good thing., 2020. URL: https://www.theverge.com/2020/2/14/21136224/samsung-galaxy-home-bixby-not-released.

[50] Bogeum Choi, Robert Capra, and Jaime Arguello. The effects of working memory during search tasks of varying complexity. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 261–265, 2019.

[51] Laura Christopherson. Can u help me plz?? Cyberlanguage accommodation in virtual reference conversations. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–9, 2011.

[52] Aleksandr Chuklin, Aliaksei Severyn, Johanne R. Trippas, Enrique Alfonseca, Hanna Silen, and Damiano Spina. Using audio transformations to improve comprehension in voice question answering. In *Proc. of the Conf. and Labs of the Evaluation Forum*, CLEF'19, 2019.

[53] Herbert H Clark and Thomas B Carlson. Hearers and speech acts. *Language*, pages 332–373, 1982.

[54] Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989.

[55] Charles LA Clarke, Eugene Agichtein, Susan Dumais, and Ryen W White. The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 135–142. ACM, 2007.

[56] Maxime Clément and Matthieu J Guitton. Interacting with bots online: Users' reactions to actions of automated programs in wikipedia. *Computers in Human Behavior*, 50:66–75, 2015.

[57] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25, 1971.

[58] Eric Corbett and Astrid Weber. What can I say? addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 72–82, 2016.

[59] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. What can I help you with?: Infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, page 43. ACM, 2017.

[60] Fergus I.M. Craik and Robert S. Lockhart. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6):671–684, 1972.

[61] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM New York, NY, USA, 2018.

[62] Edward Cutrell and Zhiwei Guan. What are you looking for?: An eye-tracking study of information usage in Web search. *Conference on Human Factors in Computing Systems - Proceedings*, pages 407–416, 2007. doi:10.1145/1240624.1240690.

[63] Dan Seifert. Google home review: Home is where the smart is., 2016. URL: https://www.theverge.com/2016/11/3/13504658/google-home-review-speaker-assistant-amazon-echo-competitor.

[64] Gregory d'Eon. Applying fair reward divisions to collaborative work. Master's thesis, University of Waterloo, 2019.

[65] Nicole Dillen, Marko Ilievski, Edith Law, Lennart E Nacke, Krzysztof Czarnecki, and Oliver Schneider. Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[66] Susan Dumais, Edward Cutrell, and Hao Chen. Optimizing search by showing results in context. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 277–284. ACM, 2001.

[67] Aarthi Easwara Moorthy and Kim-Phuong L Vu. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction*, 31(4):307–335, 2015.

[68] Chad Edwards, Autumn Edwards, Patric R Spence, and Ashleigh K Shelton. Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 33:372–376, 2014.

[69] Florian Eyben, Felix Weninger, Lucas Paletta, and Björn W Schuller. The acoustics of eye contact: detecting visual attention from conversational audio cues. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*, pages 7–12, 2013.

[70] John C Flanagan. The critical incident technique. *Psychological bulletin*, 51(4):327, 1954.

[71] Forbes. Okay, google, will voice be the future of search?, 2018. URL: https://www.forbes.com/sites/nicolemartin1/2018/11/06/ok-google-will-voice-be-the-future-of-search/.

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

[73] Natalie Friedman, Andrea Cuadra, Ruchi Patel, Shiri Azenkot, Joel Stein, and Wendy Ju. Voice assistant strategies and opportunities for people with tetraplegia. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 575–577, 2019.

[74] Prathik Gadde and Davide Bolchini. From screen reading to aural glancing: towards instant access to key page sections. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 67–74. ACM, 2014.

[75] Radhika Garg and Subhasree Sengupta. He is just like me: A study of the long-term use of smart speakers by parents and children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–24, 2020.

[76] Howard Giles and Susan C Baker. Communication accommodation theory. *The international encyclopedia of communication*, 2008.

[77] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. "It's almost like they're trying to hide it": How user-provided image descriptions have failed to make twitter accessible. In *The World Wide Web Conference*, pages 549–559, 2019.

[78] Google. Omg! mobile voice survey reveals teens love to talk, 2014. URL: https://googleblog.blogspot.com/2014/10/omg-mobile-voice-survey-reveals-teens.html.

[79] David Graus, Paul N. Bennett, Ryen W. White, and Eric Horvitz. Analyzing and predicting task reminders. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 7–15. ACM, 2016.

[80] Agustín Gravano, Julia Hirschberg, and Štefan Beňuš. Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1):1–39, 2012.

[81] Rebecca A Grier. How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 1727–1731. SAGE Publications Sage CA: Los Angeles, CA, 2015.

[82] João Guerreiro and Daniel Gonçalves. Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing (TACCESS)*, 8(1):2, 2016.

[83] Hui Guo, Jalal Mahmud, Yevgen Borodin, Amanda Stent, and I Ramakrishnan. A general approach for partitioning web page content based on geometric and style information. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 929–933. IEEE, 2007.

[84] Ido Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 35–44, 2016.

[85] Stephanie Hackett, Bambang Parmanto, and Xiaoming Zeng. Accessibility of internet websites through time. In *Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility*, pages 32–39, 2003.

[86] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14, 2018.

[87] Sandra G Hart. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

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

[89] Marti Hearst. *Search user interfaces*. Cambridge University Press, 2009.

[90] AWF Huggins. Temporally segmented speech. *Perception & Psychophysics*, 18(2):149–157, 1975.

[91] Peter Ingwersen and Kalervo Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer Science & Business Media, 2006.

[92] Muhammad Asiful Islam, Faisal Ahmed, Yevgen Borodin, and IV Ramakrishnan. Tightly coupling visual and linguistic features for enriching audio-based web browsing experience. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2085–2088, 2011.

[93] B. J. Jansen, D. Booth, and B. Smith. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, 45(6):643–663, 2009.

[94] Bernard J Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. Real life information retrieval: A study of user queries on the web. In *ACM Sigir Forum*, volume 32, pages 5–17. ACM New York, NY, USA, 1998.

[95] Gail Jefferson. Glossary of transcript symbols. *Conversation analysis: Studies from the first generation*, pages 13–31, 2004.

[96] Thorsten Joachims, Laura A Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Sigir*, volume 5, pages 154–161, 2005.

[97] Michael Kaisser, Marti A. Hearst, and John B. Lowe. Improving search results quality by customizing summary lengths. *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, (June):701–709, 2008.

[98] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-Ching Wu. Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of International Conference on the Theory of Information Retrieval (ICTIR)*, pages 101–110, 2015.

[99] Rushil Khurana, Duncan McIsaac, Elliot Lockerman, and Jennifer Mankoff. Nonvisual interaction techniques at the keyboard surface. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 11. ACM, 2018.

[100] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 45–54. ACM, 2016.

[101] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 121–130. ACM, 2016.

[102] Casey A Klofstad, Rindy C Anderson, and Stephen Nowicki. Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PloS one*, 10(8):e0133779, 2015.

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

[104] David R Large, Gary Burnett, Ben Anyasodo, and Lee Skrypchuk. Assessing cognitive demand during natural language interactions with a digital driving assistant. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 67–74, 2016.

[105] Helene Høgh Larsen, Alexander Nuka Scheel, Toine Bogers, and Birger Larsen. Hands-free but not eyes-free: A usability evaluation of siri while driving. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 63–72, 2020.

[106] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–31, 2018.

[107] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of human-computer interaction*, 22(3):247–269, 2007.

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

[109] Yuelin Li and Nicholas J Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, 2008.

[110] Shannon Liao. Russian search giant yandex made its own smart speaker., 2019. URL: https://www.theverge.com/circuitbreaker/2018/5/29/17405790/yandex-station-smart-speaker-alice-russia.

[111] Shih-Chieh Lin, Chang-Hong Hsu, Walter Talamonti, Yunqi Zhang, Steve Oney, Jason Mars, and Lingjia Tang. Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 531–542, 2018.

[112] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. When web search fails, searchers become askers: understanding the transition. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 801–810. ACM, 2012.

[113] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. Talk to me: Exploring user interactions with the amazon alexa. *Journal of Librarianship and Information Science*, page 0961000618759414, 2018.

[114] Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. Hey google, do unicorns exist? Conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, pages 301–313, 2019.

[115] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

[116] Ewa Luger and Abigail Sellen. Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5286–5297. ACM, 2016.

[117] Jalal U Mahmud, Yevgen Borodin, and IV Ramakrishnan. Csurf: a context-driven non-visual web-browser. In *Proceedings of the 16th international conference on World Wide Web*, pages 31–40, 2007.

[118] Gary Marchionini. *Information seeking in electronic environments*. Number 9. Cambridge university press, 1997.

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

[120] Nikolas Martelaro, Jaime Teevan, and Shamsi T Iqbal. An exploration of speech-based productivity support in the car. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[121] Filipe Martins, Joana Paulo Pardal, Luís Franqueira, Pedro Arez, and Nuno J. Mamede. Starting to cook a tutoring dialogue system. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 145–148. IEEE, 2008.

[122] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. A Study of Snippet Length and Informativeness. pages 135–144, 2017.

[123] Rishabh Mehrotra, A Hassan Awadallah, AE Kholy, and Imed Zitouni. Hey cortana! exploring the use cases of a desktop based digital assistant. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, volume 4, 2017.

[124] Roger K Moore. Spoken language processing: where do we go from here? In *Your Virtual Butler*, pages 119–133. Springer, 2013.

[125] Aarthi Easwara Moorthy and Kim-Phuong L Vu. Voice activated personal assistant: Acceptability of use in the public space. In *International Conference on Human Interface and the Management of Information*, pages 324–334. Springer, 2014.

[126] Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, MC Schraefel, and Jacob O Wobbrock. Reducing legacy bias in gesture elicitation studies. *interactions*, 21(3):40–45, 2014.

[127] Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. Understanding the needs of searchers with dyslexia. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[128] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. With most of it being pictures now, i rarely use it: Understanding twitter's evolving accessibility to blind users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5506–5516. ACM, 2016.

[129] Antony Rishin Mukkath Roy, Ali Abdolrahmani, Ravi Kuber, and Stacy M Branham. Beyond being human: The (in) accessibility consequences of modeling vapas after human-human conversation. *iConference 2019 Proceedings*, 2019.

[130] Alessandro Murgia, Daan Janssens, Serge Demeyer, and Bogdan Vasilescu. Among the machines: Human-bot interaction on social q&a websites. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1272–1279, 2016.

[131] Emma Murphy, Ravi Kuber, Graham McAllister, Philip Strain, and Wai Yu. An empirical investigation into the difficulties experienced by visually impaired internet users. *Universal Access in the Information Society*, 7(1-2):79–91, 2008.

[132] Nielsen. (Smart) speaking my language: Despite their vast capabilities, smart speakers are all about the music, 2018. URL: https://www.nielsen.com/us/en/insights/news/2018/smart-speaking-my-language-despite-their-vast-capabilities-smart-speakers-all-abo html.

[133] Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 647–656, 2018.

[134] Tim Paek, Susan Dumais, and Ron Logan. WaveLens: A new view onto Internet search results. *Conference on Human Factors in Computing Systems - Proceedings*, pages 727–734, 2004.

[135] Bo Pang and Ravi Kumar. Search in the lost sense of query: Question formulation in web search queries and its temporal changes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 135–140. Association for Computational Linguistics, 2011.

[136] Nilay Patel. Apple homepod review: locked in., 2018. URL: https://www.theverge.com/2018/2/6/16976906/apple-homepod-review-smart-speaker.

[137] Annie Pauzié. A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems*, 2(4):315–322, 2008.

[138] Sarah Perez. Report: Voice assistants in use to triple to 8 billion by 2023, 2019. URL: https://techcrunch.com/2019/02/12/report-voice-assistants-in-use-to-triple-to-8-billion-by-2023/.

[139] Helen Petrie and Omar Kheir. The relationship between accessibility and usability of websites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 397–406, 2007.

[140] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[141] Alisha Pradhan, Leah Findlater, and Amanda Lazar. "phantom friend" or" just a box with information" personification and ontological categorization of smart speaker-based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21, 2019.

[142] Alisha Pradhan, Kanika Mehta, and Leah Findlater. Accessibility came by accident: Use of voice-controlled intelligent personal assistants by people with disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 459. ACM, 2018.

[143] Stéphanie A Prince, Kristi B Adamo, Meghan E Hamel, Jill Hardt, Sarah Connor Gorber, and Mark Tremblay. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International journal of behavioral nutrition and physical activity*, 5(1):1–24, 2008.

[144] Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. Embodied conversational agents in clinical psychology: a scoping review. *Journal of medical Internet research*, 19(5):e151, 2017.

[145] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. "Alexa is my new BFF" social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2853–2859, 2017.

[146] Aung Pyae and Paul Scifleet. Investigating differences between native english and non-native english speakers in interacting with a voice user interface: A case of google home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, pages 548–553. ACM, 2018.

[147] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 117–126. ACM, 2017.

[148] Iv Ramakrishnan, Vikas Ashok, and Syed Masum Billah. Non-visual web browsing: Beyond web accessibility. In *International Conference on Universal Access in Human-Computer Interaction*, pages 322–334. Springer, 2017.

[149] Daniel E Re, Jillian JM O'Connor, Patrick J Bennett, and David R Feinberg. Preferences for very low and very high voice pitch in humans. *PloS one*, 7(3):e32719, 2012.

[150] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. Conversational query understanding using sequence to sequence modeling. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1715–1724. International World Wide Web Conferences Steering Committee, 2018.

[151] Laurel D Riek. Wizard of oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1):119–136, 2012.

[152] Daniel E. Rose, David Orr, and R. G P Kantamneni. Summary attributes and perceived search quality. *16th International World Wide Web Conference, WWW2007*, pages 1201–1202, 2007.

[153] Paulette M. Rothbauer. Triangulation. In L. Given, editor, *The SAGE Encyclopedia of Qualitative Research Methods*, pages 893–894. SAGE Publications, 2008.

[154] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735, 1974.

[155] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

[156] Nuzhah Gooda Sahib, Anastasios Tombros, and Tony Stockman. A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers.

Journal of the American Society for Information Science and Technology, 63(2):377–391, 2012.

[157] Mark Sanderson and W Bruce Croft. The history of information retrieval research. *Proceedings of the IEEE*, 100(Special Centennial Issue):1444–1451, 2012.

[158] Bahareh Sarrafzadeh, Alexandra Vtyurina, Edward Lank, and Olga Vechtomova. Knowledge graphs versus hierarchies: An analysis of user behaviours and perspectives in information seeking. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 91–100. ACM, 2016.

[159] Jeff Sauro. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC, 2011.

[160] Andreas Savva. *Understanding accessibility problems of blind users on the web*. PhD thesis, University of York, 2017.

[161] Maria Schmidt, Wolfgang Minker, and Steffen Werner. How users react to proactive voice assistant behavior while driving. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 485–490, 2020.

[162] Maria Schmidt, Wolfgang Minker, and Steffen Werner. User acceptance of proactive voice assistant behavior. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, 2020.

[163] Dan Seifert. Amazon echo spot review: an almost-perfect smart alarm clock, 2017. URL: https://www.theverge.com/2017/12/18/16787600/amazon-echo-spot-alexa-clock-review.

[164] Dan Seifert. Harman kardon invoke review: Cortana gets a speaker of its own., 2017. URL: https://www.theverge.com/2017/10/20/16505468/harman-kardon-invoke-cortana-microsoft-smart-speaker-review.

[165] Dan Seifert. Amazon's echo show 5 is the smart alarm clock to get, 2019. URL: https://www.theverge.com/2019/6/24/18714432/amazon-echo-show-5-alexa-smart-alarm-display-review-specs-price-features.

[166] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. Is now a good time? An empirical study of vehicle-driver communication timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[167] Shradha Shalini, Trevor Levins, Erin L Robinson, Kari Lane, Geunhye Park, and Marjorie Skubic. Development and comparison of customized voice-assistant systems for independent living older adults. In *International Conference on Human-Computer Interaction*, pages 464–479. Springer, 2019.

[168] Robert V Shannon, Fan-Gang Zeng, and John Wygonski. Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of America*, 104(4):2467–2476, 1998.

[169] Robert Simpson, Kevin R Page, and David De Roure. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*, pages 1049–1054, 2014.

[170] Aaron Springer and Henriette Cramer. Play prblms: Identifying and correcting less accessible content in voice interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 296. ACM, 2018.

[171] AllRecipes Staff. Introducing a cool new way to cook: Allrecipes on amazon alexa. http://dish.allrecipes.com/introducing-allrecipes-on-amazon-alexa/, 2016. [Online; retrieved 5-Jan-2018].

[172] Tony Stockman and Oussama Metatla. The influence of screen-readers on web cognition. In *Proceeding of Accessible design in the digital world conference (ADDW 2008), York, UK*, 2008.

[173] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[174] theverge.com. Now you can choose how fast alexa talks on your amazon echo, 2019. URL: https://www.theverge.com/2019/8/7/20757749/amazon-alexa-talk-faster-slower-speed-echo.

[175] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. Style and alignment in information-seeking conversation. CHIIR '18, pages 42–51. ACM, 2018.

[176] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. Informing the design of spoken conversational search: Perspective paper.

In *Proceedings of the 2018 Conference on Conference Human Information Interaction and Retrieval*, pages 32–41. ACM, 2018.

[177] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 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*, pages 325–328. ACM, 2017.

[178] Johanne R Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 991–994. ACM, 2015.

[179] Janice Y. Tsai, Tawfiq Ammari, Abraham Wallin, and Jofish Kaye. Alexa, play some music: Categorization of alexa commands. In *Voice-based Conversational UX Studies and Design Wokrshop at CHI*. ACM, 2018.

[180] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464, 2019.

[181] Antal Van den Bosch, Toine Bogers, and Maurice De Kunder. Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics*, 107(2):839–856, 2016.

[182] Marilyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL'90, page 70–78, USA, 1990. Association for Computational Linguistics.

[183] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[184] Ryen W White, Joemon M Jose, and Ian Ruthven. Using top-ranking sentences for web search result presentation. In *WWW (Posters)*, 2003.

[185] Ryen W White, Matthew Richardson, and Wen-tau Yih. Questions vs. queries in informational search tasks. In *Proceedings of the 24th International Conference on World Wide Web*, pages 135–136. ACM, 2015.

[186] Ryen W White and Resa A Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–98, 2009.

[187] Mark E Whiting, Grant Hugh, and Michael S Bernstein. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206, 2019.

[188] Alex C Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28, 2019.

[189] Max L Wilson. Search user interface design. *Synthesis lectures on information concepts, retrieval, and services*, 3(3):1–143, 2011.

[190] Arthur Wingfield and Patricia A Tun. Working memory and spoken language comprehension: the case for age stability in conceptual short-term memory. In *Constraints on language: Aging, grammar, and memory*, pages 29–52. Springer, 2002.

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

[192] Wan Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. *IIiX 2012 - Proceedings 4th Information Interaction in Context Symposium: Behaviors, Interactions, Interfaces, Systems*, pages 254–257, 2012.

[193] Linda Wulf, Markus Garschall, Julia Himmelsbach, and Manfred Tscheligi. Hands free-care free: elderly people taking advantage of speech-only interaction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, pages 203–206. ACM, 2014.

[194] Yandex. Alice ai assistant coming to more smart speakers., 2018. URL: https://yandex.com/company/blog/small-speakers.

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

[196] Svetlana Yarosh, Stryker Thompson, Kathleen Watson, Alice Chase, Ashwin Senthilkumar, Ye Yuan, and AJ Brush. Children asking questions: speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*, pages 300–312. ACM, 2018.

[197] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*, 2014.

[198] Jiahong Yuan, Mark Liberman, and Christopher Cieri. Towards an integrated understanding of speaking rate in conversation. In *Ninth International Conference on Spoken Language Processing*, 2006.

[199] Haotian Zhang, Mustafa Abualsaud, and Mark D Smucker. A study of immediate requery behavior in search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 181–190, 2018.

[200] Yu Zhong, TV Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P Bigham. Justspeak: enabling universal voice control on android. In *Proceedings of the 11th Web for All Conference*, page 36. ACM, 2014.

[201] Shaojian Zhu, Daisuke Sato, Hironobu Takagi, and Chieko Asakawa. Sasayaki: an augmented voice-based web browsing experience. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, pages 279–280. ACM, 2010.