Hello, computer.

Approaches to designing speech-based user experiences.

Stefan Schultz

A 90-point thesis submitted to Victoria University of Wellington,
in partial fulfilment of the requirements for the degree of
Master of Design Innovation.

Victoria University of Wellington,
2018

“Hello, computer.”

“Just use the keyboard.”

“Keyboard. How quaint.”

- Scotty and Dr. Nichols, Star Trek IV: The Voyage Home
(Nimoy, 1989)

Abstract

The burgeoning field of speech–based user interfaces, pushed heavily by most major technology vendors, including Google (with Assistant), Apple (Siri), Amazon (Alexa), and Microsoft (Cortana), presents a new challenge in designing end user experiences; one where we cannot rely on there being a visual element at all. With the individual interests of the vendors, we have seen a growth of very distinct guidelines and platforms, resulting in a lack of consistency across the field. There is an opportunity to consider these platforms and this modality of interaction, and how we can design for it more generally.

By reviewing the current array of literature on voice and conversational user interfaces, as well as general speech and user interface metaphors, an understanding and framing for the potential of this field is to be achieved. The different core vendors and their corporate attitudes and business goals are examined to find issues that may affect building for them. Thematic analysis of the current vendor and platform-specific guidelines (such as Apple's Human Interface Guidelines) will allow for determining important commonalities, feeding the creation of a set of voice-focused usability heuristics to evaluate these designs. Finally, the broader research is distilled into a systematic approach for designing speech–based experiences. Technical case study work is informed, and reciprocally informs this approach, ensuring it works in practice.

Table of Contents

v
Abstract

vii
Table of Contents

xi
Figures & Tables
    xi .... List of Figures
    xi .... List of Tables

xiii
Acknowledgements

1
Introduction

5
The State of the Field

11
Methodology

15
The Power of Speech

21
Modes of Interaction by Voice

27
Traditional Human–Computer Interaction
- **29** .... **Mental Models**
- **29** .... **WIMP and the Desktop Metaphor**

37
The Speech Interface in Popular Culture
- **39** .... **The *Star Trek* Computer**
- **40** .... **Tell us about yourself, HAL**
- **40** .... **The her of *Her*.**

43
Selecting a Platform
- **46** .... **Apple Siri**
- **48** .... **Amazon Alexa**
- **49** .... **Google Assistant**
- **50** .... **Microsoft Cortana**

53
Brand Identity Without Visual Media
- **53** .... **Sound and identity**
- **55** .... **Tone and voice**

59
Existing Vendor Guidelines & Usability Heuristics

67
Usability Testing Speech–Based Interfaces
67 .... The *Star Trek* Computer's Usability
70 .... The *Jeopardy!* Alexa Skill

77
A Systematic Approach to Designing for Speech
80 .... Phase One: Selecting a Platform
80 .... Phase Two: Brand and Identity
80 .... Phase Three: User Experience
81 .... Phase Four: Testing
81 .... Phase Five: Release

83
Experimental Work: "Voiceshell"
84 .... Selecting a Platform
84 .... Hardware
85 .... Software Basics
87 .... Speech Recognition
88 .... The Corpus
88 .... Process
90 .... Usability
90 .... Discussion

93
Discussion
97 .... Directions for Future Work and Research

99
Conclusion

103
List of References

Figures & Tables

List of Figures

All figures the author's own, unless otherwise stated.

Pg. 22 *Figure 1.* macOS Terminal application screenshot.

Pg. 22 *Figure 2.* macOS Siri application screenshot.

Pg. 28 *Figure 3.* macOS Desktop screenshot.

Pg. 86 *Figure 4.* "Voiceshell" 3D-–printed hardware enclosure.

Pg. 86 *Figure 5.* "Voiceshell" computer components.

List of Tables

All tables the author's own, unless otherwise stated.

Pg. 30 *Table 1.* "The Mac Interface", adapted from Gentner & Nielsen, 1996.

Pg. 32 *Table 2.* " Mac vs Anti-Mac Interface", reprinted from Gentner & Nielsen, 1996.

Pg. 44 *Table 3.* Word frequencies in 2017 Form 10-K filings for major vendors.

Pg. 62 *Table 4.* Thematic comparison of major platform vendor guidelines.

Pg. 64 *Table 5.* New voice heuristics compared to Nielsen (1996).

Acknowledgements

Many people around me helped make this thesis a reality, and it would take an entire second volume to adequately thank them all. Nevertheless, there are a few people who warrant special mention:

Walter Langelaar, my supervisor. Thanks for letting me run wild.

Dr Catherine Caudwell. Thanks for your help, support, and good humour.

Tom White. Thanks for the extra opportunities.

All the other MDI students, but especially Sophie Price, Stefan Peacock, Ivy Calvert, Phoebe Zeller, Bryan Loh, and Chris Doran. Thanks for being a great bunch of nerds to work (and blow off steam) with.

Thomas McDowall. Thank you. I got there—your turn.

Finally, but most importantly—Lisa, Wayne, and Kerin Schultz. Thank you for being the best and in every sense most supportive family imaginable.

Introduction

Holding a conversation with our computers has long fascinated us. In 1969, Stanley Kubrick's *2001: A Space Odyssey* presented, among other remarkably prescient technological depictions, the talking Artificial Intelligence (or AI), HAL 9000.

While HAL is hardly the paragon of a well-designed and perfectly functioning AI, talking (and listening) computers rapidly propagated through popular culture, from peripheral world-building elements, such as the *Star Trek* computer, to more complex key plot devices and even main characters, such as Samantha in Spike Jonze's *Her*.

Now, we find ourselves at a point in time where such devices, previously the reserve of science fiction, are a reality. Major tech companies such as Amazon, Apple, Microsoft, and Google are actively developing and releasing software and hardware that facilitates human-computer interaction by voice; interactions mediated by AI-presenting personalities such as Alexa, Siri, or Google Assistant. Anyone using a recent smartphone running a market-dominating operating system (that is, iOS or Android) has immediate access to such voice interaction at the press of a button or statement of a key phrase.

These systems have provided and continue to present us with a new realm for thought and research with regards to this basic human–computer interaction. These software "assistants" are now frequently presented via a new category of hardware device—the smart speaker—where voice is the only means of interface, and feedback is primarily auditory, occasionally augmented by some abstract display of lights. To be successful, these devices must interpret naturally (perhaps 'humanly') phrased utterances and respond in kind.

While these devices may proclaim to be in the vein of HAL, or Samantha, or indeed the *Star Trek* computer, they are currently vastly more limited, both in functionality and interaction. These technologies are not yet true AI, and still simply represent a different means of engaging in the kinds of human–computer interaction we have used for the last decades. Until the technology powering natural language processing improves to a significant turning point, combined with significant developments in Artificial Intelligence, this will continue to be true.

Therefore, how to engage and interact with these platforms falls to the makers building for them. Approach this thesis with that in mind.

As a design thesis, it should be read as almost a call to arms for designers to step forward and embrace this burgeoning new field of user interface and experience. A lack of particular visually designed elements should not be a barrier to this; designers are skilled professionals possessed of unique problem-solving abilities and methodologies of thought that are equally valuable and applicable here. Similar approaches in design can be taken, although with certain different considerations to respect. This is a crucial outcome of this thesis.

What follows here is therefore background, thought, and questioning, with the intent of giving designers an impetus to dive into this field, and to generate a systematic approach for handling this work. This thesis will look at the state of the field now, what lies in the past, pop culture depictions, and considerations going forward. New heuristics for evaluating usability will be generated and put into practice, informing a reflective component of this broader, systematic approach to designing speech–based experiences.

Each chapter is intended to be readable individually, without reading the entire thesis to gain a useful piece of information or insight. Indeed, each chapter ties into a phase of the systematic approach, allowing the reader to refer back as needed for further detail.

It is not the purpose of this research to create a new speech interface platform, to develop or discuss new backend technology, or propose new hardware. This work is intended to be more broadly applicable, more platform-agnostic, and continue to be relevant even in the face of constant, rapid development and progress, aimed specifically at designers. The deeply technical "nitty-gritty", developer–focused aspects are very intentionally kept to a minimum wherever possible.

Some of the higher-level considerations, the in-depth discussions of ethics, morals, and politics also fall largely outside the scope of this research, yet by necessity will be referenced at times, briefly discussed, and alluded to. Specific technical details about what is happening behind the scenes, seldom the purview of the designer, are also only discussed obliquely, without the intricacies of artificial intelligence, natural language processing, or machine learning. Nevertheless, such discourse should not read or thought of as exhaustive or authoritative, and going beyond this work for such research would better serve such interests.

This field is one rapidly growing, offering exciting new developments and opportunities, and the time is right for designers to jump in and whet their appetites for new technologies, without concerns of being incapable or uninformed.

We are entering a new world. The technologies of machine learning, speech recognition, and natural language understanding are reaching a nexus of capability. The end result is that we'll soon have artificially intelligent assistants to help us in every aspect of our lives.

- *Amy Stapleton (2015)*

Chapter One

The State of the Field

It has become abundantly clear that the world's major technology companies consider human–computer interaction by voice— 'speech–based interface' —or their so-called 'intelligent assistants' to be an important move into the future of computing, in both software, and hardware. One needs look no further than Google's 2018 I/O conference keynote to see evidence of this, with much of the presentation devoted to demonstrating new features for their Assistant platform (Pichai, 2018). Within weeks of this, the opening keynote of Apple's 2018 Worldwide Developer Conference (WWDC) saw similar time dedicated to their Siri platform (Apple, Inc., 2018b).

Google and Apple are but two of the companies making a push into this space, with Amazon and Microsoft also offering devices and services that try to tackle this computer interaction shift. Anyone with a recent smartphone, running one of the two market-dominating operating systems (iOS, or Android) have immediate access to such voice inputs at the press of a button or the utterance of a key phrase. This is a criteria met by 77% of Americans alone in 2018 (Pew Research Center, 2018). With availability across all modern iOS, macOS, watchOS, and many tvOS devices, Siri processes over 10 billion requests per month globally (Apple, Inc., 2018b).

Then there is the category of smart speakers, still in its' relative infancy. Exemplified by devices that have entered the zeitgeist, such as Amazon's Echo family, these are pieces of computer hardware lacking a screen, offering visual feedback with abstract series of lights. The back-and-forth interaction between human and machine is entirely by voice, relying on the device successfully interpreting naturally phrased utterances, and responding in kind. These devices are a rapidly growing field. In the United States alone, as of January 2018, 47.3 million people over the age of 18 have access to a smart speaker—which equates to 19.7% of the population (Kinsella & Mutchler, 2018). These people may not be the primary owner of such a device, but they are able to freely access one. From the same survey, Amazon was determined to command 71.9% of the smart speaker userbase, making them by far the dominant provider. Google follow, with an 18.4% market share. Siri was determined to be the most used voice assistant platform globally, with 375 million active users as of June 2017—77% of Apple users having at least tried using Siri.

Microsoft Cortana, integrated with their Windows 10 operating system (Microsoft Corporation, 2017b), is included by virtue of the sheer size of the Windows userbase. Windows 10 commands 35.71% of the desktop operating system market as of June 2018 (NetApplications.com, 2018). This represents an estimated 606 million PCs running Windows 10 (Keizer, 2018); a large number of users with immediate access to Cortana should they so choose.

As such, for the purposes of scope, this thesis is limited to those four platforms dominant at the time of writing—Apple's Siri, Microsoft's Cortana, Amazon Alexa, and Google Assistant. These four offer the broadest selection of information, existing research, and reference material. The use of the word "platform" hereafter will generally refer to one (or all) of these four, unless otherwise specified.

It is worth noting here that this research was conducted on Apple Mac computers, running macOS. While this is a matter of personal preference, it does have some impact on the descriptions of desktop computer interfaces, the associated terminology, and the screenshots presented as exemplars. A happy side effect, however, is the integration of Siri into macOS (from version 10.12), allowing for more immediate and direct investigation and comparison.

Another consideration moving forward, and topic of considerable debate is pronoun use. When discussing such deeply personified entities as these platforms, it can be easy to use specific gendered pronouns that we find familiar, frequently she (GrantTree, 2018).

Therefore, it is interesting and relevant to briefly consider the gender politics of artificial intelligence here. “The cyborg is a creature in a postgender world” (Haraway, 2016) is a quote that offers an fascinating direction for thought. Haraway approaches discussions of machines as the products of, and yet exceeding, humanity their creator, from a strongly feminist perspective.

This can lead down the warren of research, thought, and debate regarding post-humanism and our current movements into that realm. It can absolutely be argued along the thinking of Katherine Hayles (1999), where we are moving into the territory of deeply interconnected humans and intelligent machines, blurring the boundaries between the two. Conversational user interface and its’ associated platforms certainly allude to this narrowing dichotomy.

These arguments, the post-gender musings of Haraway and the post-humanist digressions of Hayles, are less relevant within the context of this thesis. Nevertheless, the core idea of abandoning the constructs of gender and challenging the dualism of male and female (already a very topical issue in broader human society), with creations that approach post-human territory are fascinating and appropriate.

Several authors and bloggers in media (Murphy, 2016; Reddy, 2017; Waddell, 2015) have begun to discuss this issue of gender and pronouns for artificial intelligence platforms personified. While suggestions go so far as defining a new set of robot-specific pronouns (Murphy, 2016), for the purpose of this research, the gender and identity neutral "it" will be used throughout. This aligns with the marketing employed by Apple for Siri, whose website uses the pronoun in various places (Apple, Inc., n.d.-d). This helps avoid the slightly awkward (and tedious in long form) sentence construction around a proper noun at all times which is commonly employed by Google and Amazon in their marketing (Google Inc., n.d.; Amazon.com, Inc., n.d.), and avoids the specificity of Microsoft using she/her exclusively when referring to Cortana (GrantTree, 2018). This also serves to sidestep overt humanisation of these platforms, respecting that they still maintain some distance from their creators, and there is a way to go before these skeuomorphs of humanity (N. K. Hayles, 1999) truly reach that parity.

Terminology is important, too. The term 'speech interface' encompasses the array of different terms and technologies at play here, allowing a broader overview of the field, including the older automated telephone systems typically termed as Voice User Interface (VUI) , to the oft used Conversational User Interface (CUI) when discussing modern, natural language experiences (McTear, Callejas, & Griol, 2016). Looking both historically and in the present, toward to the future, requires such a wider lens for discourse. When referring to designed outputs for this mode of interaction, the term 'speech-based experiences' is used, again with a view to maintaining a broad, encompassing perspective.

Designing for speech or using voice as a means of interaction initially became a reality primarily for speech to text dictation to a computer, and was employed for automated telephone operators. Much of the existing literature in the field focuses around this, such as Lewis' *Practical Speech User Interface Design* (Lewis, 2011), which is dedicated to thinking through, planning, and writing these kinds of application.

Similarly, speech has seen development for accessibility purposes, enabling users with difficulty seeing to have access to computer technology. On the World Wide Web, the Web Accessibility Initiative – Accessible Rich Internet Applications (WAI-ARIA) specification (Diggs, McCarron, Cooper, Schwedtfeger, & Craig, 2017) allows for declaring the roles individual elements of the underlying website HTML code play. This enables screen reading applications for blind users to correctly present the page interactions and elements to users with low-vision, making webpages, ostensibly for sighted users, usable for them (Kearney, Gash, & Boxhall, 2018).

In operating systems, Windows offers built-in screen reading with Narrator (Microsoft Corporation, 2017e) and voice control/dictation functionality for users with vision accessibility requirements. On macOS and iOS, Voiceover and Dictation perform the same role (Apple, Inc., n.d.-g). Users with low-vision are able to use these tools to have the user interface read aloud and to control their computers, as well as entering text entirely by voice.

The recent shift to "conversational user interfaces", away from simple screen reading and text dictation, remains ripe for new thought and research, and is in a state of constant flux and evolution. New features are announced or released every few months. Hence, a large number of the sources found throughout this research going forward are news articles and the like—they are the only timely and relevant source to discuss many of these developments. Wherever these sources are heavily relied upon, effort is made to reinforce them with multiple supporting references, acknowledging that such work can be subjective and less authoritative or reliable as academic studies.

Chapter Two

Methodology

This research follows a broadly pragmatic worldview, where the context of a problem is important, and the research is approached as tackling a broader problematic situation, rather than a singular hypothesis (Salkind, 2010). Finding and developing new approaches for traditionally visual designers to tackle creating experiences absent of visual media is a broader problem situation, as opposed to simple hypothesis.

In many cases, this approach requires original research reliant on tacit knowledge, in the vein of that originally outlined by Polyani (1966), and espoused as valuable to design research by Friedman (2008). Research and outcomes derived from using a computer, for example, rely on this original work, based on learned skill, experience and mental processes that defy methodical description. Working with the technologies discussed within naturally call for these kinds of work.

Corporate discourse analysis (Breeze, 2013) is used to examine the outward face, and the public marketing from each of the major platform vendors present to the world, in an effort to define a corporate purpose which may impact the choice of platform for a given design. Alongside this, investor relations material provides a valuable insight to the corporate identity and mission of the vendors; simple content analysis of key concept frequency is applied here to reveal this (Creswell, 2014).

Heuristic evaluation of usability (Nielsen & Mack, 1994) is employed to test existing speech-based experiences, finding examples of successful and unsuccessful designs, with the aim of determining what factors impact good, usable designs for speech-based experiences.

A new set of usability heuristics for this analysis is generated by thematic analysis (Creswell, 2014) of existing guidelines published by the major platform vendors, coding for significant commonalities across the documents to determine key rules that apply, regardless of platform. These are compared and contrasted with Nielsen's Ten Usability Heuristics for User Interface Design (1995), to validate them against an existing, established set.

Experimental work conducted within the "Voiceshell" project follows a reflection-in-action methodology to design (Valkenburg & Dorst, 1998), where each stage of design is regarded as a complete action, one that can be reflected on at its' conclusion to determine next steps. This supported working in a small team, allowing individual autonomy, coming together at later stages.

Speech allows the user to achieve a cleaner separation of modalities and allows data input functions to be localized in a single channel, thus eliminating the interference produced by having to share the visual channel.

- Alexander Hauptmann & Alexander I. Rudnicky (1990)

Chapter Three

The Power of Speech

Noam Chomsky, in *The Chomsky Reader* (1987), discusses historical perspectives on language and speech as abilities of Man that drive freedom and creativity; that language and speech facilitate the creation and dissemination of thought and idea.

While a polarizing figure (Behne, 2014; Boeckx, n.d.), he has been noted as one of the most cited academic figures during the 1990's ('Chomsky Is Citation Champ', 1992), and, regardless of varying attitudes, the ideas put forth provide an interesting way to start considering how powerful speech can be in a variety of uses. A number of studies have considered this concept and the applications of speech.

By looking at a selection of these existing studies, both their benefits and failings, we can draw conclusions to where speech is powerful, and the types and modes of interaction that work successfully (and unsuccessfully) in speech. This can help to inform designs targeting these platforms.

Hauptmann and Rudnicky (1990) conducted a study into the efficacy of speech input versus typed input with a computer system. Their work determined that speech offers a time advantage over traditional keyboard entry, eliminating the time spent locating the keys, finding “a cleaner separation of modalities”, and that speech eliminates “the interference produced byhaving to share the visual channel”.

This study has limitations. The age of the research is a factor, particularly when considering the dictation software used, with issues in transcription accuracy noted by the authors as a factor against speech-input. Nevertheless, speech does still hold an advantage. The task of copying data from one source into a computer could also be argued as inherently biased toward speech; reading a document aloud should always be faster than having to look back and forth between document and destination, while finding the correct keys on a keyboard. It does, however, successfully and strongly illustrate that interactions involving or requiring lengthy dictation from the user are ideal for speech interfaces.

Modality, and multimodal use cases become a consideration here. Modality in human–computer interaction refers to human perception through one of three channels: visual, auditory, and tactile (Bourguet, 2006). Speech-based interfaces are of the auditory modality.

Multitasking across modalities is a potential use case for any speech-based experience; where the user may be engaged in a primary task, yet secondarily interacting with a speech platform. Parush (2005) states that “speech-based interaction is often recognised as appropriate for hands-busy, eyes-busy multitask situations”. By comparing the ability to maintain a visual tracking exercise while engaged in another prompted task, either visually or in speech, the study determined that spoken prompts support better overall performance of a primary visual task. The modality of the prompts was examined to find the most effective for user multitasking. It suggests that certain beneficial design aspects of visual prompts be applied to speech; keeping speech prompts brief, adaptive and interruptible, for example. This is a valuable consideration for designers of speech based experiences. The idea that speech works best when the hands and eyes are otherwise occupied, for multitasking situations, is also interesting, and suggests that designers should consider their experiences as being used secondary to another task; a user may be otherwise engaged (driving being an example given by Parush) and unwilling or unable to dedicate much cognitive power to the speech experience.

This study is subject to similar limitations as Hauptmann and Rudnicky (1990), with the speech task being data entry once again, while visually (and physically) tracking a mouse cursor to an on-screen shape. This does not replicate a real world usage scenario one might expect of a speech-based experience, although offers repeatable theoretical testing. The multitasking facilitation offered by speech determined by this study should still be applicable in real world scenarios of similar visual/physical preoccupation. However, tasks in speech that go beyond simple data entry (after a prompt) and reach more complex interaction may not see similar performance improvements in multitasking; this is not accounted for in this study.

Multitasking can also be closely associated to task switching, in relation with which Emerson and Miyake (2003) explored the role of inner speech. They argue that a user's inner speech facilitates enhanced performance when switching between relatively cognitively complex tasks, partially by retrieving a verbal representation of the upcoming task. They find evidence that users applying verbal labels helps to facilitate this switching of task.

An argument can be made, from this study, that speech offers a more effective and efficient way for users to shift between significantly, cognitively different tasks, by leveraging the existing support of 'executive control processes' that speech provides.

This study does present a number of issues. It is not specific to the human–computer interaction field at all, rather studying language and memory. It is a study of inner speech, not verbalised speech—the method to suppress inner speech is to have the participants verbally utter "a-b-c". Task switching performance may not see similar performance improvements in verbalised speech, although it is a reasonable assumption. The experiments involve timing participants manually writing down answers in list format, solving basic arithmetic, with varying operations. This is a use case highly atypical of speech-based experiences; it does, however, present an interesting consideration when creating such experiences. Users are likely to be able to switch task more efficiently in speech, arguably encouraging designers of these experiences to facilitate this, ensuring that the interactions are ones that can and may be switched between rapidly.

Whether speech is an inherent part of being human, and an ability to which we are naturally predisposed (Evans, 2014), there is evidence to support it serving a powerful role, allowing for efficient human–computer interaction, and enhancing our ability to multitask. When designing experiences for the auditory modality, it is useful to bear in mind that they may frequently be secondary to those in the visual and tactile modalities, and behave accordingly, allowing quick, simple interactions, with fast switching between tasks.

So an OS is a stack of metaphors and abstractions
that stands between you and the telegrams.

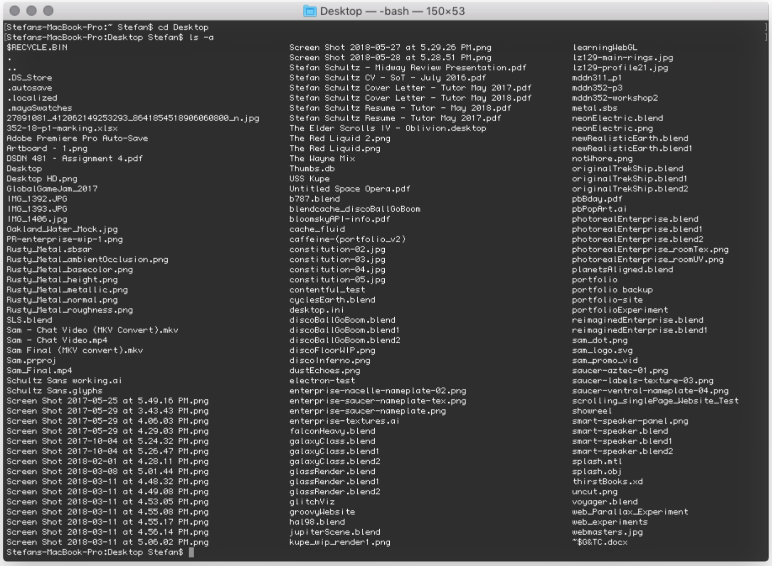
- Neal Stephenson (1999)

Chapter Four

Modes of Interaction by Voice

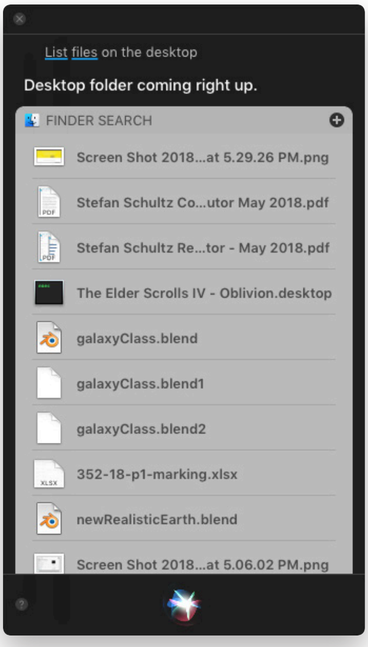
In 1984, Apple made the Graphical User Interface (GUI) available to the commercial market with the Macintosh, based on the work of Xerox PARC, presenting what would quickly become the new default mode of interaction with computers, particularly in 1985 after Microsoft released the first version of Windows, as a GUI layer for the command-line based MS-DOS operating system (Tuck, 2001). The command-line foundations, still accessible through the Terminal application on macOS and Linux, and with Command Prompt on Windows, is compared by Neal Stephenson to sending telegrams—tapping out text to give an instruction (Stephenson, 1999).

Many authors discussing the developments in Conversational User Interface note the parallels it has with the command line interface (Singh, 2015; Sreeraj, 2016; Štolfa, 2016). In practice, this can be seen—compare Figure 1 and Figure 2.



```
Stefans-MacBook-Pro:~ Stefan$ cd Desktop
Stefans-MacBook-Pro:Desktop Stefan$ ls -a
$RECYCLE.BIN                                            Screen Shot 2018-05-27 at 5.29.26 PM.png                learningWebGL
.                                                       Screen Shot 2018-05-28 at 5.28.51 PM.png                lz129-main-rings.jpg
..                                                      Stefan Schultz - Midway Review Presentation.pdf         lz129-profile21.jpg
.DS_Store                                               Stefan Schultz CV - SoT - July 2016.pdf                 mddn311_p1
.autosave                                               Stefan Schultz Cover Letter - Tutor May 2017.pdf        mddn352-p3
.localized                                              Stefan Schultz Cover Letter - Tutor May 2018.pdf        mddn352-workshop2
.mayaSwatches                                           Stefan Schultz Resume - Tutor - May 2018.pdf            metal.sbs
27891081_412062149253293_8641854518906060800_n.jpg      Stefan Schultz Resume - Tutor May 2017.pdf              neonElectric.blend
352-18-p1-marking.xlsx                                  The Elder Scrolls IV - Oblivion.desktop                 neonElectric.png
Adobe Premiere Pro Auto-Save                            The Red Liquid 2.png                                    newRealisticEarth.blend
Artboard - 1.png                                        The Red Liquid.png                                      newRealisticEarth.blend1
DSDN 481 - Assignment 4.pdf                             The Wayne Mix                                           notWhore.png
Desktop                                                 Thumbs.db                                               originalTrekShip.blend
Desktop HD.png                                          USS Kupe                                                originalTrekShip.blend1
GlobalGameJam_2017                                      Untitled Space Opera.pdf                                originalTrekShip.blend2
IMG_1392.JPG                                            b787.blend                                              pbBday.pdf
IMG_1393.JPG                                            blendcache_discoBallGoBoom                              pbPopArt.ai
IMG_1406.jpg                                            bloomskyAPI-info.pdf                                    photorealEnterprise.blend
Oakland_Water_Mock.jpg                                  cache_fluid                                             photorealEnterprise.blend1
PR-enterprise-wip-1.png                                 caffeine-(portfolio_v2)                                 photorealEnterprise.blend2
Rusty_Metal.sbsar                                       constitution-02.jpg                                     photorealEnterprise_roomTex.png
Rusty_Metal_ambientOcclusion.png                        constitution-03.jpg                                     photorealEnterprise_roomUV.png
Rusty_Metal_basecolor.png                               constitution-04.jpg                                     planetsAligned.blend
Rusty_Metal_height.png                                  constitution-05.jpg                                     portfolio
Rusty_Metal_metallic.png                                contentful_test                                         portfolio backup
Rusty_Metal_normal.png                                  cyclesEarth.blend                                       portfolio-site
Rusty_Metal_roughness.png                               desktop.ini                                             portfolioExperiment
SLS.blend                                               discoBallGoBoom.blend                                   reimaginedEnterprise.blend
Sam - Chat Video (MKV Convert).mkv                      discoBallGoBoom.blend1                                  reimaginedEnterprise.blend1
Sam - Chat Video.mp4                                    discoBallGoBoom.blend2                                  sam_dot.png
Sam Final (MKV convert).mkv                             discoFloorWIP.png                                       sam_logo.svg
Sam.prproj                                              discoInferno.png                                        sam_promo_vid
Sam_Final.mp4                                           dustEchoes.png                                          saucer-aztec-01.png
Schultz Sans working.ai                                 electron-test                                           saucer-labels-texture-03.png
Schultz Sans.glyphs                                     enterprise-nacelle-nameplate-02.png                     saucer-ventral-nameplate-04.png
Screen Shot 2017-05-25 at 5.49.16 PM.png                enterprise-saucer-nameplate-tex.png                     scrolling_singlePage_Website_Test
Screen Shot 2017-05-29 at 3.43.43 PM.png                enterprise-saucer-nameplate.png                         showreel
Screen Shot 2017-05-29 at 4.06.03 PM.png                enterprise-textures.ai                                  smart-speaker-panel.png
Screen Shot 2017-05-29 at 4.29.03 PM.png                falconHeavy.blend                                       smart-speaker.blend
Screen Shot 2017-10-04 at 5.24.32 PM.png                galaxyClass.blend                                       smart-speaker.blend1
Screen Shot 2017-10-04 at 5.26.47 PM.png                galaxyClass.blend1                                      smart-speaker.blend2
Screen Shot 2018-02-01 at 4.28.11 PM.png                galaxyClass.blend2                                      splash.mtl
Screen Shot 2018-03-08 at 5.01.44 PM.png                glassRender.blend                                       splash.obj
Screen Shot 2018-03-11 at 4.48.32 PM.png                glassRender.blend1                                      thirstBooks.xd
Screen Shot 2018-03-11 at 4.49.08 PM.png                glassRender.blend2                                      uncut.png
Screen Shot 2018-03-11 at 4.53.05 PM.png                glitchViz                                               voyager.blend
Screen Shot 2018-03-11 at 4.55.08 PM.png                groovyWebsite                                           web_Parallax_Experiment
Screen Shot 2018-03-11 at 4.55.17 PM.png                hal98.blend                                             web_experiments
Screen Shot 2018-03-11 at 4.56.14 PM.png                jupiterScene.blend                                      webmasters.jpg
Screen Shot 2018-03-11 at 5.06.02 PM.png                kupe_wip_render1.png                                    ~$G&TC.docx
Stefans-MacBook-Pro:Desktop Stefan$ 
```

Fig. 1. A screenshot of the macOS terminal, navigating to the Desktop and listing files.



List files on the desktop

Desktop folder coming right up.

Screen Shot 2018...at 5.29.26 PM.png
Stefan Schultz Co...utor May 2018.pdf
Stefan Schultz Re...tor - May 2018.pdf
The Elder Scrolls IV - Oblivion.desktop
galaxyClass.blend
galaxyClass.blend1
galaxyClass.blend2
352-18-p1-marking.xlsx
newRealisticEarth.blend
Screen Shot 2018...at 5.06.02 PM.png

Fig. 2. A screenshot of Siri on macOS 10.13, navigating to and listing files on the Desktop.

With the terminal, all files located on the desktop can be listed with the command:

```
ls -a /Desktop
```

With Siri, that command can be expressed aloud as:

```
List files on the desktop
```

While the terminal command is shorter, and fast to type, it relies on a specific vocabulary of commands and a kind of phraseology which the naturally voiced requests targeted at Siri do not require. It is highly unlikely that a user would say aloud, “ell-ess dash ay slash Desktop”; indeed, vocalising such a statement highlights the unnatural language it uses.

Nevertheless, the request to Siri does need certain phrasing to work. Simply saying “desktop files” yields a response of “I’m sorry, that might be beyond my abilities at the moment”—one of the canned error responses offered after a failed action. It becomes almost essential to speak to Siri as you might to another person; cutting extraneous words from a request would most likely result in a similarly confused response from a restaurant waiter, for example, or garner you a reputation as a rude and difficult customer. In this way, the ‘human’ side of a conversational user interface comes to the fore.

Conversely, with the Terminal command line, you must use that one, brief, specific set of commands. While after a moment’s thought their meaning may become apparent (“ls -a” meaning “list -all” (Stallman & MacKenzie, 2017)), they are far from naturally phrased—read aloud the command as phonetically stated above to prove that to yourself. With Siri, the user has the option of varying their language quite broadly and achieving the same result – even when lacking a technical vocabulary to express their intent.

Asking Siri to “Show desktop files” or “what’s on my desktop” achieves the same result as in Figure 2. Interestingly, asking Siri to “open the desktop” skips the inline results shown in Figure 2 in favour of opening a Finder window directly, hinting at the existence of a set of specific phrasings that can speed a user’s workflow in much the same way the Terminal can.

This example is useful in indicating a clear link between command line interfaces and speech–based conversational interface, however, it still leverages a use case and metaphor that is grounded in the traditional desktop metaphor, that of files. Leaning on users' existing understanding of how to interact with computers, and the cognitive processing already undertaken in understanding GUIs (or the command line interfaces before them), can help a user handle the change in modality more efficiently.

The traditional ideas and metaphors surrounding Human–Computer Interaction will need to be re-evaluated and reconsidered for speech and conversational user interface in order to fully capitalise on the potential they offer, without sacrificing the existing knowledge computer users have.

…GUIs use metaphors to make computing easier, but they are bad metaphors.

- Neal Stephenson (1999)

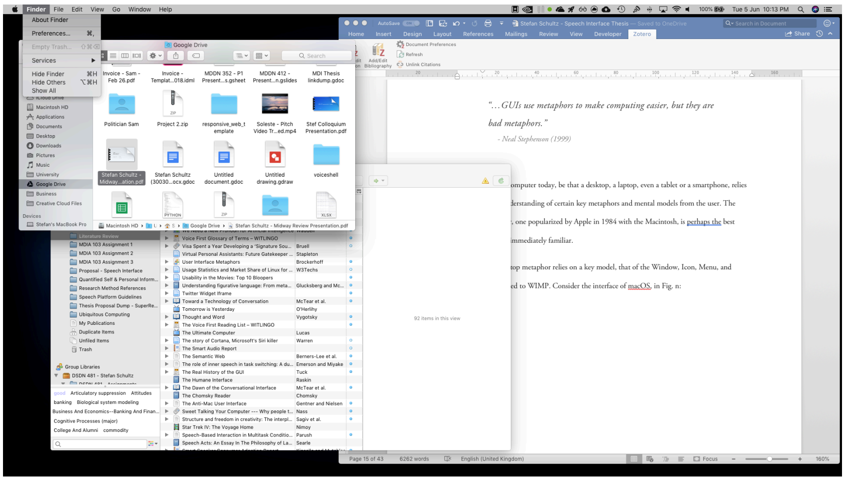
Chapter Five

Traditional Human–Computer Interaction

Using a computer today, be that a desktop, a laptop, even a tablet or a smartphone, relies on an implicit understanding of certain key metaphors and mental models from the user. These are now the standard by which we approach interactions, and offer familiarity and a set of skills transferable across devices and platforms. Ignoring this extant knowledge because of the differing modality in speech would be foolhardy, and make such experiences unnecessarily unapproachable. Indeed, this knowledge can be applied in a different way, to supply broader understanding and help formulate new ideas for this differing means of interaction.

People create mental models of themselves, others, the environment, and the things with which they interact. These are conceptual models formed through experience, training, and institution. These models serve as guides to help achieve our goals and in understanding the world.

- Don Norman (2013)



"...GUIs use metaphors to make computing easier, but they are bad metaphors."

- Neal Stephenson (1999)

omputer today, be that a desktop, a laptop, even a tablet or a smartphone, relies
derstanding of certain key metaphors and mental models from the user. The
, one popularized by Apple in 1984 with the Macintosh, is perhaps the best
immediately familiar.

top metaphor relies on a key model, that of the Window, Icon, Menu, and
ed to WIMP. Consider the interface of macOS, in Fig. n:

Fig. 3. A screenshot of macOS 10.13, showing the WIMP model in action.

Mental Models

The Norman quote above, from his book, *The Design of Everyday Things*, offers a useful definition of mental models. John Maeda, in his annual *Design in Tech* report (2017), refers to conversational user interfaces, stating that they are “grounded in mental models that don’t require a complex graphical representation and navigation system”.

Examining these existing, “complex graphical representations” gives us a way to draw new conclusions on metaphors and mental models, and how they can be adapted to better suit the modality of speech.

WIMP and the Desktop Metaphor

The desktop metaphor, one popularized by Apple in 1984 with the Macintosh, is perhaps the best known and most immediately familiar (Wang & Huang, 2000). This metaphor relies on a key model, that of the Window, Icon, Menu, and Pointer, or WIMP (van Dam, 1997). Consider the interface of macOS, in Figure 3. Along the top, starting at the left, with the “Finder” value selected, we see a system of menus. Behind and beneath that, three overlapping windows, containing a variety of content, with the Finder window—the active window placed atop the other two—displaying a selection of various icons, each representing different types of data. Limitations of the macOS screen capture tool prevent the mouse cursor, or pointer, from being rendered, however each element of the WIMP model is clearly realised here.

Everything about this system is arranged around the desktop metaphor, where the windows behave similarly to pieces of paper atop a traditional office desk.

The reasoning behind applying this metaphor, by the reckoning of Jakob Nielsen and Don Gentner (1996), is to save users’ training time, by taking advantage of their existing understanding of the traditional paper documents and filing cabinets of an office.

This filing cabinet metaphor can be seen at play with the macOS Finder, as an example of a file browser and manager typical to desktop computer OSes, exemplifying how such transferable metaphors help users establish a mental model that is familiar and works for them (Wang & Huang, 2000). Files and folders open and close, containing more files within, organised in

Mac
Users are “the rest of us: (have no previous computer experience)
Office automation “productivity” applications
Weak computer (128K RAM, 68000 CPU)
Impoverished communication bandwidth (small screen, keyboard/mouse input)
Stand-alone system that is stable unless the user decides to make a change
Manipulation of icons
Weak object-orientation (small number of large objects with very few attributes)
“Finder” (visible file system) is unifying home base, and files are the basic interaction object
Surf your hard-drive
“The Power to be your Best” (ed. Apple’s slogan at the time)

Table 1. Nielsen & Gentner’s (1996) breakdown of the Mac interface.

much the same way as they might be in the real world.

This nested, compartmentalised folder structure helps the user formulate a mental model to process the abstract concept of a computer file system, to situate themselves and locate their target data, and to efficiently navigate the potentially millions of files at their fingertips.

In their research, Gentner and Nielsen (1996) refer to this GUI for human–computer interaction as the "Mac Interface", ascribing it the following properties in Table 1.

Many of these properties, while made in reference to the original Macintosh from 1984, are clearly visible in the software pictured in Fig. 3. above, that is, the most modern version of macOS (at the time of writing).

Characteristics specific to hardware are less relevant now. Computers are not as resource starved, with gigabytes of RAM and CPU power that required an entire room to contain just a decade ago now contained in handheld devices carried in our pockets.

The input bandwidth consideration is especially interesting, with computers frequently having very large displays, and varieties of input beyond just a keyboard and mouse.

Now let's look at Gentner and Nielsen's response, in Table 2, where they consider how violating the "Mac" user interface principles could result in a new type of interface, one they term the "Anti-Mac.".

Interesting and relevant to this research are the ideas of rich communication with a computer—that it could see you, and by logical extension, hear you—and the use of language instead of icons as a key means of interface. The manipulation of information as opposed to files, with information surfaced for you instead of files found by you, is also of note here.

This is an indicator of the "Post-WIMP" (van Dam, 1997) nature of speech–based interaction; interaction without icons, as mentioned, but also without windows, menus, or pointers.

In speech–based interfaces, language is, by necessity, the chief means of interaction. There are no icons to manipulate. Designing speech input to behave in the manner of icons, facilitating simple and easy to comprehend computer interaction, is therefore paramount. If an icon is an abstracted representation of data or a system func-

Mac	"Anti-Mac"
Users are "the rest of us: (have no previous computer experience)	Users are the "Post-Nintendo" generation (grown up with computers)
Office automation "productivity" applications	Work, play, groupware, embedded, and ubiquitous.
Weak computer (128K RAM, 68000 CPU)	Humongous computer (multi-gigabyte RAM, Cray-on-a-chip RISC processors)
Impoverished communication bandwidth (small screen, keyboard/mouse input)	Rich communication (computer can see you, knows where you are, large high-res screen, new I/O devices)
Stand-alone system that is stable unless the user decides to make a change	Connected system subject to constant change
Manipulation of icons	Language
Weak object-orientation (small number of large objects with very few attributes)	Strong object-orientation (large number of small objects with rich attribute sets)
"Finder" (visible file system) is unifying home base, and files are the basic interaction object	Personal information retrieval as unifying principle with atomic information units as basic interaction object
Surf your hard-drive	Information comes to you
"The Power to be your Best" (ed. Apple's slogan at the time)	You won't always have to work that hard

Table 2. Nielsen & Gentner's (1996) breakdown of the Mac interface, compared to their suggested Anti-Mac interface.

tion, then how does that abstraction translate to speech?

We also cannot overlook the contextual awareness of modern, Internet connected devices, particularly in discussing the surfacing of topical information. When information, rather than files, are the core unit of interaction, this context becomes increasingly valuable, and an aid to ensuring that exactly what the user needs, or desires is presented at any given point.

Google Now, a separate precursor of sorts to Google Assistant (Jacobsson Purewal, 2016; Martonik, 2018; Townsend, 2017) is an example of this idea put into practice, exposing the user to certain pieces of information based on their schedule, time of day, location, and prior usage habits. Apple have made a similar push in this direction, branding it originally as “Proactive”, and more recently under the aegis of Siri (Sullivan, 2017; Taka, 2017). In their guidelines for conversational design document (Giangola, 2017), Google strongly encourage designers to leverage this kind of context where possible.

With the ability of modern computing devices to understand much more about their situations and their users, with the multitudes of sensors such as gyroscopes, GPS, and the like, the ability to rethink core interactions is enabled.

This frames information as an atomic unit of interaction, as Gentner and Nielsen (Gentner & Nielsen, 1996) put it, as an interesting concept to rethink certain foundational ideas.

A shift to this information-centric interaction can be seen with platforms such as iOS, where the file system is seldom exposed to the user beyond one level of detail, and without direct access—ostensibly to keep the system simple (Apple, Inc., n.d.-a). When working with documents, you need only consider which application it belongs to, not its precise place in the file system, obviating the concept of “surfing your hard-drive” that is seen in the ‘Mac Interface’ above. Information can ‘come to you’ with the development of tools like Google Now, and the broader aspects of Siri.

In speech, the computer being able to understand broader contextual cues helps surface this information. If a user asks, “What’s the weather doing?”, the device can use its locative sensors and constant Internet connection to find itself in the world, and get the relevant piece of information, without requiring any more effort or steps from the user—you don’t have to ask, “What is the weather right now for Wellington, New Zealand, from the MetService", for example.

Similarly, asking a query along the lines of "What's John Smith's phone number?" could return a simple response—a string of digits read aloud, perhaps qualified with whether it is a work or home number. The phone number is an atomic piece of information as far as the user is concerned. While it is saved as a file (in a certain sense of the word) on the device and may be composed of multiple constituent pieces of data and metadata, that underlying nature need never be made apparent to the user.

Enabling this kind of deep information connection and retrieval is a concept that has been championed by Tim Berners-Lee, the man credited with inventing the World Wide Web in 1989 at CERN, who has written of his idea of "the Semantic Web" (Berners-Lee, Hendler, & Lassila, 2001). In this concept, the architecture of information and the representation of knowledge is vitally important and can facilitate software making inferences to meaning and relations, without complex Artificial Intelligence, language processing, and reasoning. Tagging key pieces of information or data semantically allows easy linking of related concepts and surfacing relevant pieces contextually with a lower processing overhead. The ARIA web accessibility standard (Diggs et al., 2017), as mentioned prior, is a step in this direction, revolving around the semantic declaration of roles for each element in a web page.

The example Berners-Lee, et al. give is that of XML (eXtensible Markup Language) and RDF (Resource Description Framework), as technologies emerging at the time of writing that push into this direction. XML is a means of storing data, by encapsulating each piece of information within a tag of what that information represents (Quin, 2016).

For example:

```
<author>Stefan Schultz
</author>
<year>2018</year>
<institution>Victoria
University of Wellington
</institution>
```

The above could be a snippet of an XML file describing, or indeed containing, this very thesis. Each piece of information is present but accompanied by a tag to what it means. The meaning of information becomes vital to making the systems described work.

For designers, this is a crucial consideration when working on a speech experience. Consider how to treat data and files as singular, indivisible yet interrelated pieces of information, surfaced to the user contextually, on demand. The user should never need to think about a file system, where they are “on the disk”, or what application will open something. Files are a meaningless and overly complicated concept unnatural to conversation.

No. You know that's just a movie, right?

- Siri, when asked if it is Her.

Chapter Six

The Speech Interface in Popular Culture

The quote that opens this chapter is but one example of the humorously self-aware responses proffered by the various virtual assistants when queried on their pop culture brethren. This influence is inescapable; with fully-fledged voice interfaces to computers starting life as the figments of screenwriters' imaginations.

Popular culture gives us examples of problems already solved in some way, shape or form. Technologies and experiences are thought through and designed, then presented to the world, often encouraging or inspiring new ideas and developments.

Star Trek has been such a notable inspiration and influence on the technology field, with its' depictions of advanced technology frequently drawing parallels with real world devices as they launch, such as the iPad as compared to the PADD used since the 1980's series (Foresman, 2016). Perhaps the most well-known example of this influence is the flip cellphone, first realised in 1996 in the Motorola StarTAC, a design heavily inspired by Captain Kirk and crew's iconic flip open 'communicators' (Kessler, 2011).

The influence of *Star Trek* is one Amazon acknowledge. Speaking at the 2017 Wired Business Conference, the company's Senior Vice President of Devices, David Limp, referred to the *Star Trek* computer as Alexa's "north star" (Johnson, 2017), the target they aspire to reach.

Microsoft take this pop culture influence a step further, with their platform, Cortana, named for the AI character in the *Halo* video game franchise (originally a codename, retained after a significant online petition); going so far as to record the same voice actress, Jen Taylor, for the auditory feedback (Warren, 2014).

As the real world adapts to this technology becoming a reality, perceptions change. Eric Harvey, in a piece on smart speakers and streaming music for Pitchfork (2018), notes that speech-based computer interaction depictions have shifted from a mood fear (with HAL 9000 in *2001: A Space Odyssey* (Kubrick, 1968)) to romance and even arousal (with Samantha in *Her* (Jonze, 2013)). In this sense, pop culture helps us to understand and rationalise the world as it changes.

If the companies and people behind the leading technologies enabling this field recognise, acknowledge, and rely on this influence, it seems only logical to consider how popular culture designed and depicted them, as a lens to approach real world applications. Screenwriters and designers have been tasked with thinking and solving issues of how we might interact with such technology for decades; the wealth of information and reference of this nature is too great and too valuable to pass up. How are these experiences designed to invoke fear, or attraction? Are they useable? Having already examined what we can do with the real-world technology today, why not explore the unreal technology that has and continues to inspire it?

The *Star Trek* Computer

A computer interacted with by speech has been a feature of the *Star Trek* franchise since the first season of the retroactively titled *Original Series*. While advanced robots approaching humanity were frequently shown, sometimes as main characters, the starship computer was notably distinct, operating under no illusions of artificial intelligence or proximity to sentience. It was not too dissimilar to the various voice interface platforms already available today.

It is also remarkably comparable to the current smart speaker devices, such as Siri and Alexa. Interaction begins with a wake-word (‘Computer’; an option Amazon allow to wake their Echo devices now) and is handled entirely by vocal back and forth. The voice feedback is traditionally even female (most iconically that of Majel Barrett, wife of *Star Trek* creator Gene Roddenberry), echoing (or perhaps foretelling) the default gendered tones of Assistant, Alexa, Siri, and Cortana.

In the *Original Series* episode “Tomorrow is Yesterday” (O’Herlihy, 1967), Captain Kirk engages in a played-for-laughs back and forth with a recently upgraded system, which insists on addressing him as “dear”, depicting a degree of personality from the computer system; a more conversational interaction.

Star Trek: The Next Generation (and its’ sequel series, *Star Trek: Deep Space Nine*, and *Star Trek: Voyager*) shows the computer has continued to exist, always present, latent in the background, ready to offer assistance. The episode “Darmok” (Kolbe, 1991) depicts interaction in detail, with characters making extensive use of the computer for library access and research. This role seems the main function of the *Star Trek* computer, but even this is not unrealistic with the current, real world state of affairs: one survey suggests that thirty percent of all smart speakers requests are general information-seeking (VoiceLabs.co, 2017). Sixty percent, in the same survey, are reported to be asking for music, an activity seen frequently completed by the *Star Trek* computer, such as Captain Picard asking for music to be played, by genre and mood, in the movie *Star Trek: First Contact* (Frakes, 1996).

Tell us about yourself, HAL

It would be remiss of any review of pop cultural depictions of speech-based computer interface, no matter how brief, to fail to mention HAL 9000, who remains perhaps the quintessential example of a talking computer in the public consciousness.

In Stanley Kubrick's film *2001: A Space Odyssey* (1968), a team of astronauts are sent to Jupiter, investigating a mysterious signal sent from an object on the moon. This objective is unknown to them, but the spaceship computer, HAL (an acronym of Heuristically programmed ALgorithmic computer (Clarke, 1972)) 9000 has been given the full mission briefing, instructed to keep it secret. In order to complete this objective, HAL determines it must kill the human astronaut crew, as the only way to maintain secrecy. This premise presents futuristic (at the time) technology, and raises questions of ethics and morality in computers, as they develop to approximate levels of humanity.

At the time of the film's release, 1968, the majority of audiences' experience with computers would likely extend no further than simple punch card tabulating machines. Typing input to and receiving output via screen was still being developed. HAL presented a vision of future technology where these technologies would be largely left behind by a spoken, human–approximating system (Olive, 1998). Seventeen years after the real year 2001, such a vision is slowly approaching reality.

The influence of HAL on real-world developments of speech-based platforms is undeniable. When asked to "open the pod bay doors", Siri offers "We intelligent agents will never live that down apparently" (Sciretta, 2011)—yet another example reinforcing the relevance of examining pop culture depictions of speech–based experiences for inspiration when designing new ones.

The her of *Her*.

Spike Jonze's film *Her* (2013) depicts an artificially intelligent computer operating system, Samantha, so believably "human" that a user falls deeply in love with it, only to deal with the struggle of reconciling romantic, monogamous love, with the vast, arguably polygamous nature of a networked computer program.

The creators refer to the setting as a 'slight future', where technology is designed to 'dissolve into everyday life', still embracing certain low-tech elements (Steffen, 2015). This is emphasised and visible throughout, with warm tones and wood grain contrasted against crisp lighting and glass.

The main means of connecting with Samantha is through an earbud. Production designer K.K. Barrett, in an interview with Fast Company (Hart, 2013), discusses this earbud, and its' deliberately under-designed nature, intended to give the smallest cue that the main character, Theodore, is engaging with Samantha.

Interestingly, this parallels Apple's 2017 AirPods wireless earbuds. Wired reviewer David Pierce (2017) describes AirPods as being 'for Siri', where all that is required to invoke Apple's speech platform is a double tap of the earbud. This thread is picked up by multiple other reviewers, who, despite mixed feelings on the devices themselves, emphasise the ease of access to Siri (Ranj, 2018; Porter, 2017).

Google's 2017 Pixel Bud wireless earbuds get even closer to the *Her* vision of future voice/computer interaction. As noted by reviewers from TechRadar and Engadget, the standard "Ok, Google" command is all that is needed to invoke Assistant while wearing the Pixel Buds (Faulkner, 2018; Velazco, 2017).

The range of influential technology examples provided by popular culture illustrate a common fascination with conversational computing. The examples discussed here show the variety of approaches and how they relate to real world developments now, including where technology aligns or is directly influenced. Looking to these designs for reference when approaching development of new speech experiences could provide valuable inspiration and help solve problems.

There is also an opportunity to examine these fictional speech interface examples in the context of usability, to explore where they are less successful, and apply those lessons to future designs. This idea will be explored in a later section, on usability evaluation.

Chapter Seven

Selecting a Platform

The platform designed for can impact the designers working on speech experiences. Comparing the four selected platforms of this research shows that there are some distinct variations and unique limitations to each.

Consider also the differences between the companies that develop and offer these platforms themselves. Are there specific corporate motivations driving Alexa or Siri, and how can those be reconciled with those of the designer and their clients?

What about the target audience of an experience? If 95% of a client's user base are users of Android phones, for example, it would be illogical to design exclusively for Siri; few users would be able to derive any benefit.

Before a closer look into the individual vendors and their platforms, let's examine investor-oriented material to gain insight to the primary business each conducts.

Code	Alphabet	Amazon	Apple	Microsoft
Enterprise	3	9	9	**47**
Organization	19	9	9	**63**
Personal	7	3	**35**	51
Consumer	10	**46**	19	37
Individual	15	11	**24**	17
Security	24	19	**31**	46
Privacy	**14**	4	7	8
Data	65	**69**	49	80
Commerce	6	**15**	2	2
Advertising	**112**	18	7	45
TOTAL:	275	203	192	396

Table 3. Word frequencies for selected terms in the 2017 Form 10-K filings of specified companies.

Content Analysis of Annual SEC Filings (10-K)

Publicly traded businesses in the United States are required to file "Form 10-K" annually, giving their investors a breakdown of the company's financial activities and status (U.S. Securities and Exchange Commision, 2009). Performing a content analysis on these documents should therefore provide an accurate (by legal necessity) overview of a company's business objectives, at least at a high level.

The following codes are used, each intended to capture a section of business likely to be conducted by a tech company, or collectively cover a target market:

```
enterprise, organization,
business, personal,
consumer, individual,
security, privacy, data,
commerce
```

Organization is spelt the American English way to account for these four documents all being from American companies. The terms "company" and "corporation" are deliberately omitted, to limit capturing and counting self-referential statements. "Business" is similarly not counted, as it refers to the company and their activities in terms of self, rather than corporate interests.

The 2017 Form 10-K filings of Microsoft, Amazon, Apple, and Alphabet (the parent company of Google) are analysed (Alphabet Inc., 2017; Amazon.com, Inc., 2017; Apple, Inc., 2017; Microsoft Corporation, 2017a).

The totals at the bottom of Table 3 are given as guides for the overall length of the documents analysed, allowing rough percentages of term frequency to be calculated and used to draw conclusions to the importance of each code to each company. The highest percentage of occurrences for each code are emphasised, and rounded to the nearest whole number when discussed. These will be discussed as appropriate under each subheading below. Some high level conclusions to consider, however:

- Alphabet are the most advertising focused of the four companies.
  - The term "advertising" accounts for 40% of those coded, far in excess of the others (9% for Amazon, 4% for Apple, and 11% for Microsoft).

- Microsoft reference "enterprise" and "organisation" the most frequently, implying a stronger emphasis on enterprise business.
    - Together, these terms account for 27% for the coded terms for Microsoft; 8% for Alphabet, 9% for Amazon, and 9% for Apple.
- Apple collectively reference "personal", "consumer", and "individual" in 41% of coded terms, suggesting a stronger emphasis on personal, consumer computing.
    - 12% for Alphabet, 30% for Amazon, and 27% for Microsoft.
- Amazon are the most retail-centric, with 7% of their coded terms being "commerce".
    - This compares with 2% for Alphabet, 1% for Apple, and 0.5% for Microsoft.
- Alphabet mention "privacy" the most, as 5% of coded terms, compared to 2% for Amazon, 4% for Apple, and 2% for Microsoft.
    - Context should be noted here; the majority of Alphabet's privacy references are discussing the potential negative effects of privacy-enhanced legislation on their business.

Apple Siri

Siri has the distinction of having been available for the longest, since 2011 (Bosker, 2013). Originally accessed with visual feedback, Siri allowed users to get information, or complete simple functions by voice, such as playing music, creating reminders, or placing phone calls. Siri is a layer to enable access to various platform features, sort of like an API for the user, where developers can allow access to their application functions by exposing them in extensions (Apple, Inc., n.d.-f), potentially lowering the barrier of entry for developing such experiences. By the same token, the requirement for a full regular iOS (or macOS) application to drive the voice experience may increase the initial effort to get started with the platform (Kinsella, 2018a).

Based on the content analysis conducted in Table 3, Apple as a company appear to place a focus on individual consumer users, rather than the corporate/enterprise market. As such, experiences targeted at the enterprise or business market may be less suitable for the Siri platform.

A core part of Apple's corporate persona and strategy is the promotion and claimed protection of user privacy

(Apple, Inc., n.d.-e). New features are frequently announced by the company in pursuit of that goal, such as preventing tracking by advertising companies online (Grothaus, 2018). Rather than using Siri as a means to gather more personal data for corporate purposes, it could be argued that Siri is a means to increase and deepen user engagement with Apple's services, tying people further into their ecosystem.

The marketing site for Siri (in New Zealand) provides twenty-four references directly naming or depicting Apple services in the context of Siri functionality (Apple, Inc., n.d.-d). Seventeen of these are entertainment related (iTunes, Books, Podcasts, Apple Music), and two about the App Store, all Apple services which make the company more money from prolonged and more frequent use.

The HomePod, Apple's entry to the smart speaker market, could also be argued as an example of this. It is heavily promoted as a high-quality speaker for music playback, working with Siri. The first footnote on the page, however, states "Siri works with Apple Music. Subscription required for Apple Music." (Apple, Inc., n.d.-b). This further emphasises the close relationship of Siri to Apple's services, and strongly reinforces this as a key component of Apple's strategy with the platform.

iOS 12, unveiled at Apple's Worldwide Developer Conference (WWDC) 2018, introduces the new system feature, Siri Shortcuts (Apple, Inc., 2018b). This function is designed to open access to creating custom Siri interactions to the end user, allowing them to string together various first- and (if enabled by the developers) third-party application functions into single commands, executing possibly lengthy and complex sequences.

Choosing Siri as the platform for a speech experience therefore presents a couple of key considerations. For privacy conscious clients and customers, Apple's corporate pledge of user privacy protection may be very desirable. In privacy-conscious use cases, Siri may be the most palatable platform. It assumes the end user will be invested in the Apple ecosystem, and happy to remain that way. A certain connection or reliance on Apple services, or a willingness to be tangentially related to these services must be acceptable. Finally, if there is an existing application built for iOS or macOS, it may prove easier to build out to Siri on that software, by developing an extension into the existing code, opening up functionality to users by voice. Soon, the end user will have the ability to make these voice-exposed functions work more flexibly for their use cases, which could also be seen as advantageous—although from a business perspective could be seen as diluting or interfering with a singular, cohesive application experience.

Amazon Alexa

Amazon unveiled Alexa in November 2014 (Stone & Soper, 2014), releasing it for general availability in June 2015 (Amazon.com, Inc., 2015), and quickly taking an early lead on market share, holding 71.9% percent of that product category (Kinsella & Mutchler, 2018).

Alexa offers a different approach to applications on the platform, naming them “skills”, and allowing complete experiences packaged up for the platform, with little to no connection to external applications or sources (Amazon.com, Inc., n.d.-a). This could be a boon to clients who lack any existing application, facilitating development of a new, unique, and entirely independent experience.

Amazon do not produce an operating system, unlike Apple and Google (with iOS and Android, respectively). This can be seen as placing them in neutral ground between those vendors’ platforms, which may be advantageous to a client wishing to target a broader audience, or not be restricted to a single userbase. However, the experience will still largely be restricted to Alexa-powered devices, which is a more limited audience than the install base of iOS (with Siri available), or Android (with Google Assistant) (Kinsella & Mutchler, 2018). However, as of January 2018 with Android (Martin, 2018), and June 2018 for iOS (Prospero, 2018), Alexa can be interacted with through the supporting Amazon Alexa smartphone application (Amazon.com, Inc., n.d.-f), opening the platform to a broader audience than previously.

Amazon do, however, make Alexa available for building into hardware devices manufactured by third-parties, to create new, separate products that are compatible with Echo devices, and allow use of other Alexa skills (Amazon.com, Inc., n.d.-b). This “Alexa Voice Service” potentially offers an advantage in creating highly-customised experiences, moving beyond just software and into unique hardware devices.

A cursory glance at the Amazon homepage (Amazon.com, Inc., n.d.-d) immediately reveals the core purpose of Amazon as a company—online shopping. The homepage title directly includes that phrase. Amazon CEO and founder Jeff Bezos has given the company's mission statement as "to be earth's most customer-centric company; to build a place where people can come to find and discover anything they might want to buy online" (Farfan, 2018). Also consider the content analysis from Table 3, which reinforces this strong focus on commerce driving their business and corporate activities; significantly more so than Apple, Google, or Microsoft.

Echo devices and the Alexa platform facilitate this core function also. Reviewers of these note the prominence of shopping within the platform and to Amazon as a company. Clauser (Clauser, 2018) refers to voice-controlled shopping as "close to Amazon's heart", while the Bloomberg article about the original launch (Stone & Soper, 2014) states an Echo "of course (since this is Amazon, after all) will obey instructions to put various products into your Amazon.com shopping cart". An OC&C Strategy Consultants report (Franklin & Maal, 2018) estimates that in 2017, $2 billion (USD) was spent in commerce via speech platforms, with Amazon accounting for around 90% of that expenditure.

This commerce/shopping focus could fall into conflict with the desires, goals, and mission statements of certain clients. When choosing a platform for a speech experience, designers should consider this aspect; for example, would it make sense for a non-commercial, non-profit, or charity-focused application to exist on such a platform?

Amazon Alexa therefore provides a more flexible experience for designing and building speech-based experiences, but this is weighed against the inherent corporate focus of the vendor for shopping with their services, and either the need to build and supply custom hardware or rely on the smaller install base of Echo devices.

Google Assistant

As of 2017, Google Assistant accounted an for estimated 28.7% of voice assistant interactions on smart phones (Kinsella, 2018c) and available on over 400 million devices running Android.

Custom developed experiences for Google Assistant are called "Actions on Google" (Google Inc., 2018). Google state that Actions provide a way for third parties to connect and engage users with their own “fulfilment services”.

Google have gained a reputation for gathering vast swaths of user data in return for offering a free product, often ascribed the “if you’re not paying, you’re the product” epithet (Kepes, 2013; Sauer, n.d.). Google themselves disclose that they do collect this personal data from users engaging with their products, for the purposes of improving and offering the products and services that they do (Google Inc., n.d.-b). When discussing this, Google references how such data is used to improve the Assistant platform, by giving it access to a broad range of contextual information about you, gathered from your general use of Google products and services.

In this same page, Google discusses advertising, and their use of personal data for that purpose. Google state that personal data is not disclosed to a third-party, but rather used to target ads provided to Google—they become the middleman.

In the content analysis from Table 3, Google are the company with the most references to advertising; the term “advertising” accounts for 40% of all coded phrases in their 10-K form, almost ten times as frequently mentioned as the next most common, Amazon. This strongly verifies how integral advertising is to Google’s business model.

Google’s collection of this data could be problematic for clients and certain experiences being designed for the Assistant platform. While having a large install base and possible range of users is appealing to getting more users for an experience, it must be weighed against any concerns of an experience driving Google’s own business practices, and whether they fall into conflict with a client’s. Also consider the target audience of an experience. If they are predominantly users of iOS devices, for example, Google Assistant is likely less preferable as a platform, given the company’s development of the competing Android operating system.

Microsoft Cortana

Cortana is available on devices running Windows 10, as well as within dedicated apps for iOS and Android (Microsoft Corporation, n.d.-b), giving it broad reach across platforms and vendors, in a similar way to Amazon Alexa. Microsoft have also made Cortana available in a smart speaker device, the Harman Kardon Invoke (Seifert, 2017; van Camp, 2018). Between late 2017 and early 2018, Microsoft have claimed between 141 and 145 million monthly users of the Cortana platform, across 400 million devices (Holly, 2017; Kinsella, 2018b)

Cortana (in the United States) can be developed for in much the same way as Amazon Alexa, with a Skills Kit allowing the creation of custom speech–based experiences on the platform, available wherever Cortana is (Microsoft Corporation, 2017d).

In the past, particularly surrounding the *United States vs. Microsoft* antitrust case beginning in the late 1990's (The United States Department of Justice, 2015), Microsoft pursued an internal corporate strategy of "embrace, extend, extinguish"—that is, actively support an existing platform, add to it extensively, and in the process, eliminate it as competition ('Deadly embrace', 2000). This general ill-will surrounding the corporate mission of the company may influence selecting their platform for a client product.

The current public facing mission statement for Microsoft states that they "believe in what people make possible", espousing a desire to "empower every person and every organization" (Microsoft Corporation, n.d.-a). Referring to Table 3, this mention of organisations is notable. Of the companies examined, Microsoft sees the highest mentions of that term, and of "enterprise", strongly indicating their dominant business interests lie with the corporate market as opposed to individual consumers.

For designers creating experiences in speech, the Cortana platform may therefore be most suitable for enterprise applications, for a strongly corporate market. The lower usage figures relative to the other platforms suggest individual 'consumer' users are less likely to engage with the platform, although the broad number of devices with Cortana availability could be untapped potential for a "killer app".

When selecting a platform from any of the major vendors, there are various pros and cons to weigh against one another. Implications with user data privacy, with heavily commercial objectives, with ecosystem lock-in, and general user apathy all must be considered to ensure that any designed experience is presented and accessible appropriately.

Chapter Eight

Brand Identity Without Visual Media

For businesses and clients with a clearly defined and established brand, designers of speech experiences must consider how to convey this identity without their traditional aids of colour schemes, logos, and the like.

Here, audio trademarks, language, and the nature of interactions themselves become the only meaningful ways to deliver a brand to the user, and the realms of sound design and scriptwriting become more relevant than traditional graphic media techniques.

Sound and identity

Auditory cues to a brand identity are a well-known and established subset within the field of sound design. D. A. Sumikawa coined the term “earcons” to describe audio cues in user interface (1985), where these small snippets of sound perform the same role as a traditional icon does in a graphical user interface—giving information to the user about an interaction or function of their computer.

Examples of common earcons include the sound of rustling paper that accompanies emptying the Trash on a computer, or the beeps indicating a microwave has finished (Stevens, 2013).

Earcons may not be exclusively and entirely for system feedback. Using Apple Pay, for example, produces a distinctive chime as a transaction successfully completes, one unique to the Apple Pay experience. It provides both feedback on an interaction to the user, while concomitantly presenting a brand element.

This overlap can be complex. While the tone achieved and the sonic specifics of these earcons can be varied and matched to a brand identity, they should not be confused with sonic branding, or “identitones” (Conradt, 2017).

This is not to say that specific audio logos (ones that do not serve a user interface or experience role) are not important or valuable. Many have deeply permeated the zeitgeist. Consider the Intel Inside jingle, or the THX Deep Note as examples of well known, distinctive, and brand-associated sounds.

Music can play a similar role, such as the Jeopardy “think” music, or even the Nokia Tune ringtone. Sound is a vital element of a complete and cohesive brand structure.

While a more dated example now, the Nokia tune (properly called Grand Valse) may remain one of the most culturally significant exercises in audio branding. The tune was stated, at the height of Nokia’s popularity, to have been the most heard piece of music in the world, played up to 1.8 billion times a day (Klara, 2016; Peters, 2014). Again, however, this walks the line of earcon and identitone, as in general use, it indicated a function of the device to the user. Nevertheless, the association to a particular brand is undeniable (Klara, 2016).

Visa spent more than a year selecting from over 200 sounds to find one that suited the applications they desired to use it in (including as part of payment interactions in the world), and met the needs of their brand identity (Bruell, 2017).

In existing speech experiences, there are some examples of this in practice. The Jeopardy Alexa skill (Sony

Pictures Television, 2016), for example, relies heavily on certain notable sounds from the original television series , including prerecording dialogue from host Alex Trebek. Such touches at once ground the user within the experience they are having, and convey a unique, distinct identity separate from that of the platform being leveraged. The Stuff Alexa "flash briefing" skill (Stuff Limited, 2018) is similar, employing the audio from their video sequence logo between each news headline.

Tone and voice

When designing for a platform that already has a defined vocal identity, as seen with Siri, Alexa, Assistant, and Cortana, how to apply that voice becomes very important to creating a sense of identity. Here, the language used, the kinds of phrases and how they are phrased all become vital to owning the experience, and helping users remember what (or who) they are interacting with. If available across multiple platforms, this can also introduce an element of consistency that keeps the user grounded within a singular experience.

Defining a voice is an important part of developing a brand, and applicable to general user experience design. In many cases, this work already exists and is applied in various places. Consider copywriting style guides used by marketing teams. For example, Apple have a lengthy guide to writing company communications, ensuring absolute consistency across all channels with tone and language (Apple, Inc., 2018a). This document goes to great pains to detail every possible language choice, from correct number formats, to the sentence construction of phrases discussing products—for example, iPhone and iPad should not be prefaced with a 'the' article. Vice Media employ a similar guide to styling their written content, giving rules for authors to follow as to how to phrase dates and times, abbreviation usage, punctuation, and grammatical choices (Vice Media LLC., 2014).

MailChimp have a lengthy guide specifically on implementing their voice and tone in communications (The Rocket Science Group, 2018). This guide focusses particularly on the user's emotional responses to how communications are phrased, and the impact certain phrasings can have—particularly valuable when the only means of feedback to a user is with language.

If there isn't an existing style guide for copywriting, or any copywriters within a team or working with a client, defining some simple and basic rules and principles around this can be essential to ensuring successful translation of a brand to a speech–based experience. When leveraging an existing platform, this can ensure that it speaks with your voice.

Overlooking the importance of sound and language in any branded experience can be seen as unwise and passing up a valuable opportunity. Leveraging existing, iconic pieces of audio, or creating new ones to help drive identity and engagement, are a way to account for an inability to use traditional graphic media. Similarly, the choice of language and the tone achieved by an experience, regardless of platform, can help differentiate an experience from the various others, and from a given platform or vendor.

Chapter Nine

Existing Vendor Guidelines & Usability Heuristics

After considering how to differentiate a designed experience from the platform on which it resides, with a view to being 'platform-agnostic', it makes sense to examine the existing design guidelines available from these vendors. While they offer specifics to the intricacies of their associated platform, here we can examine them for commonalities at a higher level, with the assumption that these shared points will be more universally applicable.

Table 4. is compiled from the Apple Human Interface Guidelines for Siri (Apple, Inc., n.d.-c), the Amazon Alexa Voice Design Guide (Amazon.com, Inc., n.d.-c), the Google Conversation Design: Speaking the Same Language document (Giangola, 2017), and the Microsoft Cortana Design Guidelines (Microsoft Corporation, 2017c). These four documents are comparable in terms of what they discuss, offering considerations and guidelines for

Theme	Apple	Amazon	Google	Microsoft
	Strive for a voice-driven experience that doesn't require touching or looking at the screen.			
Respecting the user	**Be appropriate.**			**Trustworthy.**
	Don't advertise.			
Speed/ease of use	**Respond quickly and minimize interaction. Take people directly to content.**	**Keep interactions brief. In general, provide no more than three choices at a time. Ask for information one piece at a time. Avoid repetitive phrases.**	**Move the conversation forward.**	**Efficient.**
Clarity and relevance	**Be relevant and clear.**	**Clearly present options.**	**Be brief, be relevant.**	**Relevant. Clear.**

Table 4. Thematic comparison of vendor-voice design guidelines. Emphasised text indicates key determined themes.

Theme	Apple	Amazon	Google	Microsoft
Accounting for error and uncertainty	**When a request has a financial impact, default to the safest and least expensive option.**	**Don’t assume the user knows what to do or what will happen.**		
Allow for variation in speech	**If appropriate, define custom vocabulary.**	**Identify utterances.**	**Don’t teach “commands” —speaking is intuitive.**	
	Consider defining alternate app names.			
	Provide example requests.		[This is in direct opposition to what Google suggests above]	
			Leverage context	

designing speech-based experiences. Each is arranged to give more detail, but pulls out key points as strong, emphasised text, or bullet points. Table 4. collects these points, and compares them, coding for the general theme each follows, to determine commonalities.

From this simple comparative exercise, five key aspects emerge as important to each of the vendors. Two are present across the guidelines of all four analysed vendors; three across at least three. These can be framed by their attached codes, ranked by relative importance (how often related guidance appears across each document):

1. **Speed and ease of interaction.**
   Keeping every interaction as brief as possible, and limited choices and feedback to avoid overwhelming the user, for efficient use of the platform.
2. **Clarity and relevance.**
   Ensuring that everything is as clear to the user as possible, and free of unnecessary, extraneous detail.
3. **Allowing for natural variations in speech.**
   Ensure that the user can have a successful interaction with natural language; that specific phrasing and wording isn't needed to achieve the goal.
4. **Allowing for errors and uncertainty.**
   Ensure the use can feel confident with what they are doing, and account for errors both on their part, and with the technology.
5. **Respecting the user.**
   Never compromise the user's trust or leave the user feeling negative and always respect their privacy.

With these appearing so prominently across the guidelines given by the current dominant vendors in the field, it is logical to take these as a baseline, 'bare minimum' consideration for designers working on applications for voice.

Table 5 compares these five guidelines with the Nielsen's (1995) ten usability heuristics:

1. **Visibility of system status**
2. **Match between system and the real world**
3. **User control and freedom**

4. **Consistency and standards**
5. **Error prevention**
6. **Recognition rather than recall**
7. **Flexibility and ease of use**
8. **Aesthetic and minimalist design**
9. **Help users recognise and recover from errors**
10. **Help and documentation**

All ten of Nielsen's heuristics can be aligned with at least one of the five derived from the vendor guidelines. Interestingly, none quite fit into the area of respecting the user, in the sense that the voice design guidelines suggest. Arguably, this is indicative of the more personal nature of vocalised interaction, where conversation provides more immediate avenues to offence and upset in the user, an argument and line of thinking that has already seen some discussion and research. Clifford Nass (2010) anecdotally references experiments in how users interact with computers differently based on perceptions of humanity; computers that speak for themselves, for example, see different and more emotional responses than those that do not. Research by Karl MacDorman, et al. reinforces this, with their study indicating a biased preference for synthesised voices that sound female or feminine (Mitchell, Ho, Patel, & MacDorman, 2011); males implicitly preferring them, and females vocally expressing the same.

Certain nuances of **respecting the user** are arguably less relevant to the vendors behind certain platforms. Consider Google's Duplex demo from I/O 2018 (Pichai, 2018), where Assistant was demonstrated making an appointment with a hair salon over the phone, and at all times maintaining a pretence of being a real human, going so far as to "umm" and "ahh" at appropriate conversational breaks. This is an intentional deception of the user, and, if measured against this heuristic, is easily described as disrespectful of the user; indeed, the fundamental ethics of computers deceiving the people interacting with them are up for debate in this context (Cowie, n.d.). Interestingly, if we look again at the table of guidelines the vendors offer for designing in speech, Google lack any which align with respecting the user, suggesting they are less concerned with this aspect. Being such a broad and complex topic of discussion—and one considered important by two of the four vendors above—reinforces the importance and relevance of ensuring the user is treated

Voice Heuristic	Comparative Nielsen
Speed and Ease of Interaction	• Flexibility and ease of use • Help and documentation
Clarity and Relevance	• Aesthetic and minimalist design • Visibility of system status
Allowing for Natural Speech Variations	• Match between system and the real world • User control and freedom • Recognition rather than recall
Allowing for Errors and Uncertainty	• Error prevention • Help users recognise and recover from errors • User control and freedom
Respecting the User	

Table 5. Nielsen's 10 Usability Heuristics (1995), compared to the new Voice-centric heuristics.

with respect when designing these speech-based, conversational experiences.

Considered together, we now have a set of voice-specific usability heuristics that can be put into practice, evaluating speech-based experiences.

Chapter Ten

Usability Testing Speech–Based Interfaces

With these known guidelines examined and new, speech-centric heuristics established, we can set about evaluating existing speech-based interactions against them.

The *Star Trek* Computer's Usability

A perhaps amusing yet relevant exercise, in the context of prior discussed examples, is to evaluate the *Star Trek* computer's usability. The *Star Trek: The Next Generation* episode, "Darmok" (Kolbe, 1991), provides a useful example of an interaction which can be evaluated with the new usability heuristics for speech.

The dialogue in the episode occurs between the character of Deanna Troi, and the USS Enterprise starship's main library computer, and is transcribed on Page 80.

DEANNA: Computer, search for the term Darmok in all linguistic databases for this sector.

COMPUTER: Searching. [pause] Darmok is the name of a seventh dynasty emperor on Conda IV, a mythohistorical hunter on Shantil III, a colony on Lindy VII, a frozen desert on Tasna V, a– [interrupted]

DEANNA: Stop search. Computer, how many entries are there for Darmok?

COMPUTER: Forty-seven.

Transcript Deanna Troi and the USS Enterprise Computer, in *Star Trek: The Next Generation*, “Darmok” (Kolbe, 1991)

Ignoring the obviously fictional content, let's apply the five new, voice-first usability heuristics to evaluate this short interaction itself.

Speed and ease of interaction:

While it should be noted that the search depicted here was unsuccessful, the interaction required two short lines from the user (Deanna). The interaction could be easily interrupted, without requiring the "Computer" wake-word to be uttered again, indicating that the system remains in a state of active listening, or a "follow-up mode"[1] , ready to quickly handle adjustments from the user.

Clarity and relevance:

The Star Trek computer falls flat here. The responses it provides are lengthy and detailed, and while all related to the broad search query given, are lacking in relevance.

Allowing for natural variations in speech:

The commands given by the user are phrased very formally, and have a definite structure, including specifying parameters ("…all linguistic databases for this sector"). This suggests that the computer relies on a prescribed phrasing for commands, although without other evidence this statement cannot be confirmed.

Allow for errors and uncertainty:

When the computer proceeds to list all search results in detail, the user is able to immediately stop this and make a related yet distinct request: "Stop search. Computer, how many entries are there for Darmok?". Stopping the computer's listing does not require reuttering the "Computer" wake-word, indicating that such commands are actively listened for as a means of error-prevention. Should the user become aware of a mistaken command, they would appear able to immediately stop it, without having to phrase the cease command in a particular formula.

Respect the user:

This heuristic has less applicability here, although the tone of the computer, one that is extremely matter of fact and automaton-esque, carries an inherent respect for the user as the one in control; it never behaves disin-

---

[1] A similar "follow-up" mode is being added to the US English language versions of Amazon Alexa and Google Assistant, as of June 2018 (Johnson, 2018; Kinsella, 2018; Moon, 2018; Statt, 2018).

genuously or in a way that breaches the user's trust.

The computer proceeding to list forty-seven search results in some degree of detail indicates an unaccounted-for usability issue in speech – whether or not it is the appropriate medium for the interaction at hand. Clearly a user could more efficiently traverse forty-seven search results by 'traditional' visual list, than have to await their complete auditory presentation.

As such ***appropriateness of the speech medium*** becomes an important consideration at any point and should be strongly considered throughout the design and testing process for any speech-based experience. Whenever a feature is implemented, or the experience expanded, take a moment to consider whether it makes sense in speech, or is better suited to other interaction modes, such as traditional GUI. If it is found to be currently ill-suited to speech, can anything be done to improve the experience without sacrificing the feature or experience altogether? Consider the results discussed in the preceding 'Power of Speech' chapter, giving examples of speech's unique success in interaction.

When applied to the "Darmok" example, the answer can simply be 'No'. Dense results from the inquiry depicted are not appropriate to speech, except for accessibility purposes (supporting visually impaired users). Lengthy, listed search results realistically remain the preserve of the GUI.

The *Jeopardy!* Alexa Skill

The *Jeopardy!* Alexa skill (Sony Pictures Television, 2016) allows users of Amazon Echo devices to play one round of trivia questions, in the format of the television game-show, *Jeopardy!* (Griffin, 1984). The experience interaction, between user ("User" being the author; i.e. myself) and skill, which largely uses the voice of Alexa as set by the user's language preferences is transcribed opposite.

The experience mixes dialogue read by Alexa with pre-recorded audio snippets of the voiceover announcer, Johnny Gilbert, and the host, Alex Trebek, from the series (IMDb.com, Inc., n.d.). This is used alongside sound effects and music clips from the series, reinforcing the brand experience by presented familiar audio elements—the identitones, as referenced in the earlier chapter—which helps the experience

User: Alexa, play Jeopardy.

[Jeopardy! theme music plays]

Recorded audio (Johnny Gilbert): This is Jeopardy!

Recorded audio (Alex Trebek): Thanks, Johnny, and welcome players!

Alexa: The first category is [sound] Danish food and drink. Gravlax features this fish cured, not smoked like when served as Lox.

User: What is salmon?

Alexa: Yes, that's it. The second category is [sound] government ABCs. Add to save the world, EPA stands for this.

User: What is Environmental Protection Agency.

Transcript The author interacting with the *Jeopardy!* Alexa skill (Sony Pictures Television, 2016).

Alexa: Yes, that’s correct. The next category is [sound] small screen crime fighters. TV’s Elementary features this lead detective.

User: Sherlock Holmes.

Alexa: Please respond in the form of a question starting with phrases like “Who is?” or “What is?”

User: Who is Sherlock Holmes?

Alexa: That’s correct. The next category is [sound] Born and died. Born in 1931 in Indiana, he went west and starred in East of Eden and only two other movies, before he died in a 1955 car crash.

User: Who is James Dean?

Alexa: Great. The fifth category is [sound] book of the month. Fill in the blank with the name of a month. The Enchanted blank by Elizabeth von Arnim.

Transcript The author interacting with the *Jeopardy!* Alexa skill (Sony Pictures Television, 2016) - continued.

User: What is March?

Alexa: No, the correct response was what is April? The last category is [sound] anagram pairs. The response is two words that are anagrams of each other. What you cast on election day, and a president’s way to kill a bill.

User: What is vote and veto.

Alexa: Yes, that’s it.

Recorded audio [Alex Trebek]: Alright, let’s see how you did today.

Alexa: Great job, you got five of six correct. You scored in the top five percent of players today.

Recorded audio [Alex Trebek]: Come back tomorrow for more clues.

feel unique and independent from the platform it is leveraging.

The usability heuristics for voice can now be applied.

Speed and ease of interaction:

The skill is quite lengthy, with each question from Alexa notably longer than the responses the user needs to give. This requires prolonged attention from the user, and makes the experience seem one-sided. It does, however, fit the format of the series being emulated, largely excusing it from ‘failing’ against this heuristic.

After each clue, the Echo automatically enters active listening mode, meaning the user can respond without the “Alexa” preface, speeding and smoothing the interaction flow.

Clarity and relevance:

The questions are read fast, and are quite lengthy, which can limit the clarity of them; for the purposes of the above transcription, certain proper nouns and foreign-language words had to be double checked. Contextually, there is enough surrounding information for the user to make a reasonable assumption and guess at the correct answer.

Allowing for natural variations in speech:

The *Jeopardy!* format requires players to respond in the format of a question. The skill requires the same, to align the experience with the expectations attached as a game. When the user fails to respond in the correct format, a prompt for the user to correct is given, with an example.

Allow for errors and uncertainty:

Given the game format of the experience, errors naturally cannot be accepted. While failure to meet the required format of a response is caught, and allowed to be corrected with a prompt, an incorrect response is met with a negative reply and the correct answer. No opportunity for the user to account for or correct a misinterpretation by the device is offered.

Respect the user:

The experience generally respects the user within the confines of a game. The response given for an incorrect answer could come across as mildly condescending and brusque, but that is in service of a smoother and quicker experience. Frequent apologies or artificial sympathy for the user in such scenarios may come across worse, and seem patronising.

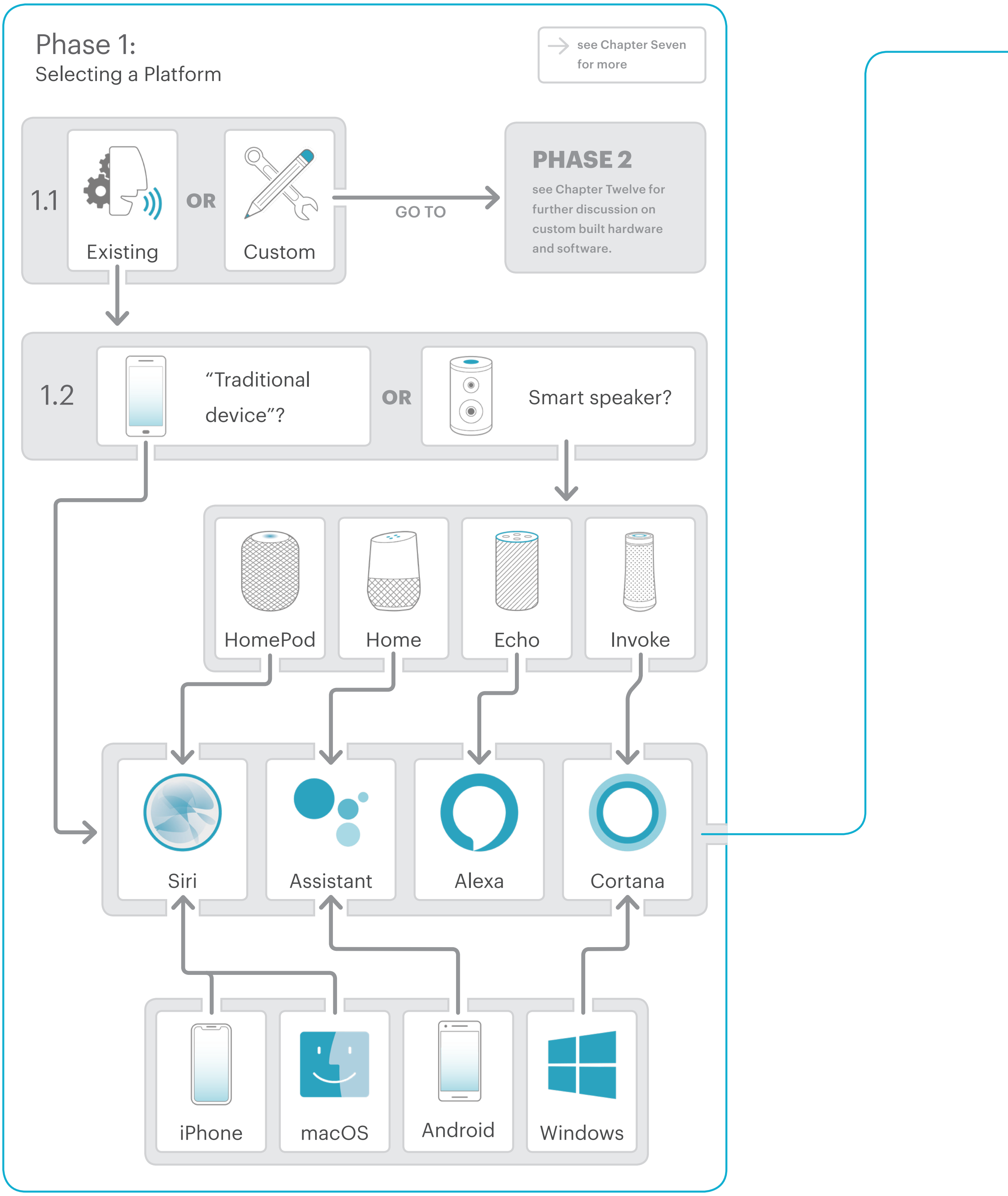
By applying the newly created usability heuristics for voice to existing experiences, both real and fictional, we can see their usefulness in evaluating designs for speech. Generating this kind of qualitative information about user experiences is a valuable part of the design process, and can inform further iteration, tying into a broader approach for creating such designs, and providing metrics for reflective practice.

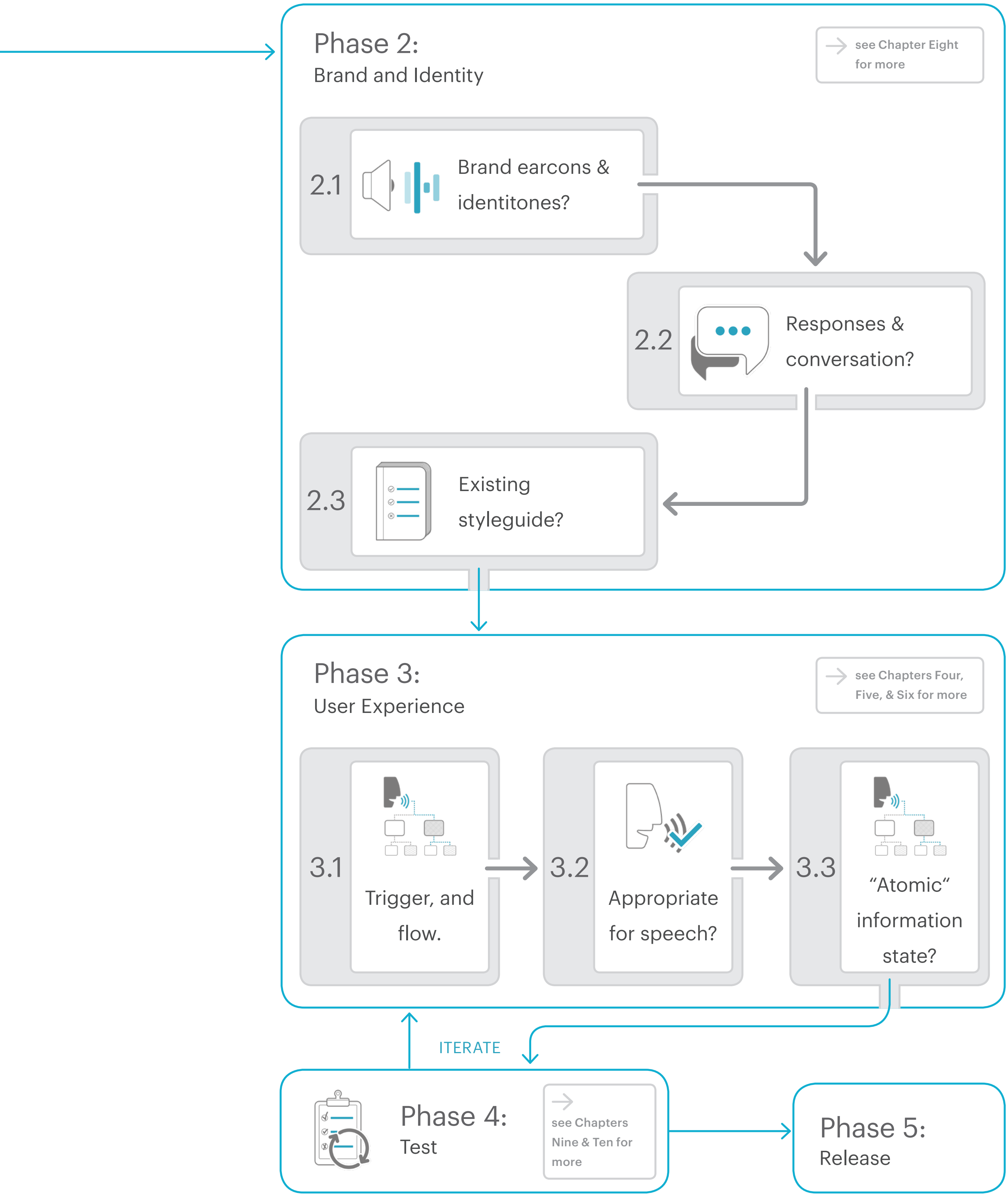
Chapter Eleven

A Systematic Approach to Designing for Speech

Now, we can formulate a simple systematic approach to planning, designing, and evaluating proposed speech-based applications and experiences, taking onboard all the considerations and decision points found throughout the course of this research, from the background research, analysis and comparison of the different platforms and their implications, exploration of where speech is most powerful, and development and implementation of usability evaluation heuristics. All of this gives us the material to distil into a simple format for designers to follow as they approach creating speech-based experiences. Each key point relates to a preceding chapter, and readers are encouraged to refer back frequently for expanded information and clarification.

The approach follows five key phases, which are largely distinct from one another, but which feed forward, and loop iteratively at the end.


Phase 1:
Selecting a Platform
see Chapter Seven for more
1.1
Existing
OR
Custom
GO TO
PHASE 2
see Chapter Twelve for further discussion on custom built hardware and software.
1.2
“Traditional device”?
OR
Smart speaker?
HomePod
Home
Echo
Invoke
Siri
Assistant
Alexa
Cortana
iPhone
macOS
Android
Windows


Phase 2:
Brand and Identity
see Chapter Eight for more
2.1
Brand earcons & identitones?
2.2
Responses & conversation?
2.3
Existing styleguide?
Phase 3:
User Experience
see Chapters Four, Five, & Six for more
3.1
Trigger, and flow.
3.2
Appropriate for speech?
3.3
“Atomic“ information state?
ITERATE
Phase 4:
Test
see Chapters Nine & Ten for more
Phase 5:
Release

Phase One: Selecting a Platform

1.1: Does the experience require the flexibility of custom hardware/software, or can it leverage an existing platform?

1.2: Is this an experience for "traditional" devices (such as smartphones or PCs), or is it for smart speakers?

Now, determine which platform makes sense. This could be a target smart speaker, or the dominant smartphone platform of the target audience, for example. For more detail on selecting a platform, see Chapter Seven.

Phase Two: Brand and Identity

2.1: Does the brand being designed for have any distinctive, established audio elements, be that earcons or identitones?

2.2: Thinking about the structure of conversation, what kind of format feels suitable to the brand? How brief (or not) should each moment of speech be?

2.3: Does the brand have an existing styleguide for language that can be used to ensure the specifics are correct, and convey the identity fully?

See Chapter Eight for more information on conveying brand identity.

Phase Three: User Experience

3.1: How does the user initially trigger the experience? From there, what is the flow of experience the user will have?

3.2: Does this established flow feel appropriate to the speech modality, or is it overly lengthy and complex, or heavily requiring supporting visual media?

3.3: If the user will be engaging with files or data of any kind, how can it be exposed as "information", in an interactive way—what is the "atomic" state of the data?

For more on user experience and interaction in speech, see Chapters Four, Five, and Six.

Phase Four: Testing

Test and evaluate the designed experience. Consider the voice-first usability heuristics, and how the design works when compared to them.

For more on these heuristics, see Chapters Nine and Ten.

Upon the conclusion of testing, return to Phase Three, and apply the results to improve the experience. Continue to iterate and cycle through these phases as long as necessary.

Phase Five: Release

Release the final designed and built experience through the channels appropriate to the chosen platform.

Chapter Twelve

Experimental Work: “Voiceshell”

Within the context of this body of research into speech-based user interface and experience, an opportunity arose to undertake research assistant work on a related project, under Tom White from the Victoria University School of Design.

This allowed some limited real-world application and development of the systematic approach, and of the usability heuristics. The experimental, and ostensibly artistic nature of the project—working with poetry—restricts the applicability of much of this work and keeps it strictly tangential to the body of research above; nevertheless, it is an important relationship and useful feedback loop of work to explore.

The project was aiming for the development of a device intended for public exhibition of works of poetry, as read by the original authors. A large corpus of text and attached audio was available, with any further specifics left open. No parameters to the experience were given, initially.

Beginning design and development work on a project like this immediately informs this thesis and is the direct cause for certain sections to be present and discussed—such as that on choosing the right platform for a project.

In the case of this project, the leading speech-based platforms did not prove ideal for what was required. While there was a degree of flexibility afforded in using custom hardware with Amazon Alexa or Google Assistant, developing and designing interactions for them was less simple, given the custom corpus being used, and the nature of responses decided upon.

Working with a more abstract set of material and accepting a more abstract means of vocal interaction required a kind of customisation and freedom of development that was not afforded in developing an Alexa Skill as the project began.

Selecting a Platform

After examining the possibility of using Google Assistant and Amazon Alexa, it was decided that smaller size, more heavily customisable software and hardware would work more successfully, especially given the large number of unknowns regarding the experience and its' exhibition at the time.

The final main consideration that pushed away from leveraging Alexa or Assistant was the possibility of exhibition in a location without access to the Internet—the existing platforms leverage cloud computing power to process user input and respond, rendering them essentially useless without this connection. The ability to work entirely offline and locally was determined more important at this stage.

Hardware

Initial hardware experiments began with a Raspberry Pi 3 Model B microcomputer, paired with a USB microphone and small speaker connected by standard 3.5mm analogue audio jack.

Using standard, cost-effective components presented advantages in planning for future development. If one hardware component proved incorrect, or non-ideal, then it could be easily swapped out, or a different piece trialled, with a minimum of difficulty;

this is compounded by using standard connections (such as USB), which limit any, more intense, electronic work.

The Raspberry Pi is a very well-used and extensively documented piece of hardware, which eased the development process, with finding reference for most issues very simple. The comparatively (to other similarly sized or versatile computers) low cost per unit was also advantageous in considering future duplicate units; creating multiple "Voiceshells" for display in a variety of different locations.

Issues began to arise with the processor architecture used by the Raspberry Pi—the Model B uses an ARMv7 CPU, which has a unique instruction set (the fundamental way the processor runs code), requiring additional work on the software side to run some libraries, compared to the traditional, and more widely adopted x86 architecture typical of most mainstream computers (those with Intel or AMD CPUs).

This necessitated a shift in platform, to an Intel Next Unit of Computing (NUC) box, pictured in Figure 5, with an x86 instruction set CPU on the motherboard (an Intel Core i5). The hardware change offered more RAM, a faster CPU, and a higher base level of storage, all of which offered added flexibility during the development process.

The Intel NUC offered the same selection of connections for external hardware devices, meaning that the originally obtained speaker and microphone could continue to be used.

The hardware itself had to be communicated to the industrial designer attached to the project, Zach Challies, who was involved in producing a physical enclosure to contain the individual components; the large, 3D printed structure, seen in Fig. 4. It was designed to be mounted to a post, with all of the technical hardware components completely contained within, and invisible. This design was largely independent of my own work, aside from sharing dimension details of the individual hardware components as needed for the enclosure design.

Software Basics

Raspberry Pi uses a custom Linux distribution, Raspbian (based on Debian Linux) by default, which led to development beginning there, after initial experimentation on a macOS system. The Unix-like nature of both operating systems made porting code between the two a simple matter.

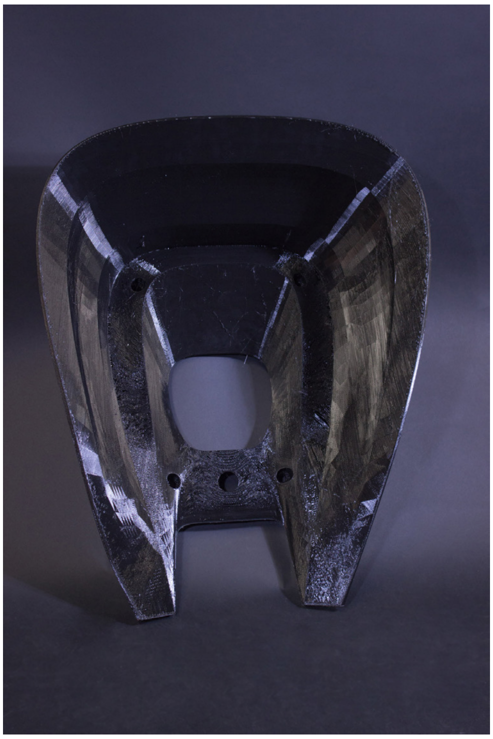


Fig. 4. The 3D–printed hardware enclosure.

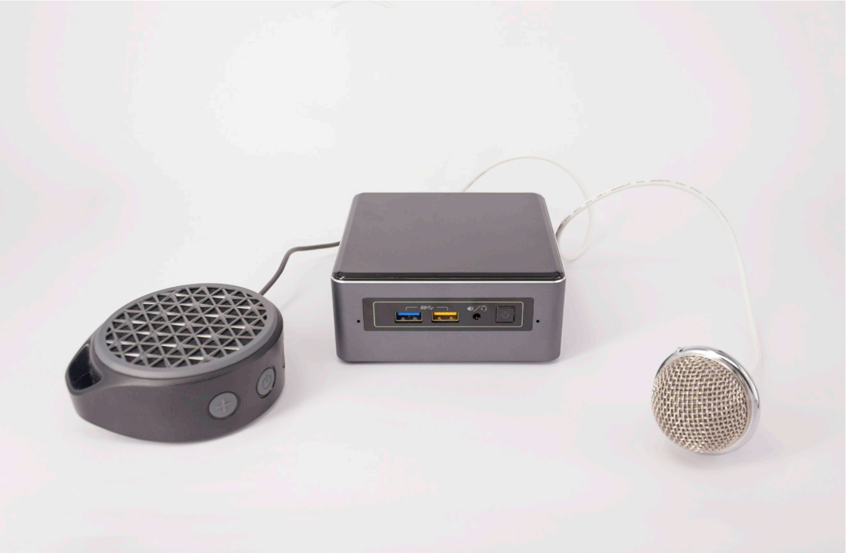


Fig. 5. The final prototype hardware components; speaker, Intel NUC, and microphone.

After switching hardware, the Intel NUC was set up with Linux Mint 18.3 'Sylvia', after a series of failures to install Ubuntu Linux. Linux Mint offered the same required capabilities, similar user experience, and extensive documentation and community support from a large user base. Being based on Ubuntu and Debian, the current dominant distributions (W3Techs, 2018), there was a vast array of available software components and packages to aid in development. Staying with Linux ensured the portability of any developed code across platform, barring architecture dependence issues.

The software for Voiceshell itself was written in the Python (version 3) language, owing to personal familiarity, and library availability. It is also a popular language, with extensive and detailed documentation, and a very active community for support readily available online.

The library selected for Natural Language Processing (NLP), 'Spacy', proved exceedingly difficult to successfully compile and run on the Raspberry Pi (due to incompatibilities with the ARM architecture), which predicated the move to the x86 architecture Intel NUC. After some development, and an evolution of the core interaction, this NLP library no longer proved necessary, allowing for some new experimentation with simplified software, back on the Raspberry Pi platform.

This development remained uncompleted in the time frame and funding period encompassed by this research. Returning to the Raspberry Pi presents a number of benefits, primarily in terms of cost-effectiveness, allowing for more hardware devices to be created and displayed, if desired. The smaller size of the enclosures (relative to the Intel NUC) also presents greater flexibility in the final hardware shell design.

Speech Recognition

Speech recognition was implemented using the PocketSphinx library in Python, a port of the open source CMU-Sphinx project. This project states a goal of running efficiently on low-resource platforms, which made it appealing during development for Raspberry Pi.

PocketSphinx was used by the Python speech-recognition library. This library was selected thanks to the extensive documentation and broad existing use, as well as flexibility with recognition library. If constant internet connectivity was determined to be possible with the final installation of "Voiceshell", then the recognition library could be changed out for others, such as the Google one.

The Corpus

The corpus of material for exposure by Voiceshell came from the Victoria University of Wellington's International Institute of Modern Letters (a part of the Faculty of Humanities and Social Sciences), and their annual collection of the Best New Zealand Poems. They provide copies of these poems in plain text on their websites, with the majority having audio of the author reading their work attached.

Responsibility for the initial work on the corpus fell to the other research assistant engaged on the project, Chris Doran. Certain key choices were made on mutual agreement, particularly where the software development and overall user experience may have been concerned.

The primary consideration in this vein was how to split each poem in an effective way. Ultimately, it was deemed fairest and simplest to split to any full stop, taking complete sentences as dictated by the original poets. This also gave the other research assistant a straightforward way to quickly split each of the audio files—which are originally complete readings, up to several minutes in length—into individual pieces more flexible to integrate with the software.

In terms of working with files or data, and exposing that to the user through speech, this also solved the question of the "atomic state" of the data as information; each piece was a played audio file, of one sentence from a given poem.

Process

Development on the software side began with an interactive Python notebook, facilitating very fast writing and testing of code, making changes and trialling new ideas with efficiency. This rapid iteration proved invaluable for trialling new libraries and software for usefulness and compatibility. As this initial development was underway, the aforementioned other research assistant was working with the raw audio files, splitting them to single lines, and converting them to Wave audio format files, for easier use with Python audio playback libraries.

This notebook developed in steps, gradually introducing and testing functionality independently of each other. Initial loading of the corpus text from a manually created CSV file was the first step, parsing the read lines through Spacy for natural language processing, with the intent of matching to a spoken input from the user. Before this spoken input was implemented, it was prototyped with typed input, using simple string matching to test the natural language processing and sentence matching. With this working, playing back the audio recording of the correct sentence was attached.

Getting live microphone input working began as a separate step, building it out to include speech recognition of the live input, transcribed to the screen for verification of accuracy. The offline PocketSphinx recognition had a somewhat limited success rate, potentially impacted by non-American accented speaking. Nevertheless, this recognition was used to replace the text input from the previous iteration, and successfully matched to lines of poetry. Given the somewhat abstract nature of the corpus, less exact sounding matches for user utterances seemed still successful; arguably, a more opaque response to user input is more thought-provoking and interesting within the context of the project.

The processing overhead of working with the natural language processing encouraged a small shift to more simplistic methods of interpreting and matching user input to the corpus. A new program was written, taking large pieces of the existing code, but replacing the Spacy natural language processing with the “FuzzyWuzzy” fuzzy string matching library, which attempts to find the closest matching line of text to a given input, up to a specified confidence—how sure of the match the library is. This proved more efficient, and eliminated the one remaining x86 architectural dependency.

As the phase of the project covered within this research came to a close, considerations for the non-visual experience came more to the fore. Without the existing capabilities of Alexa or Assistant for providing feedback at various stages of the experience, custom development was needed again. Originally, this took the form of small system status statements, vocalised by the Linux text-to-speech engine. These were gradually replaced with generated sound effects, trying to establish a consistent sound scheme for the project that didn’t conflict with the differing poetry readings.

Usability

Voiceshell can be evaluated against the voice-first usability heuristics.

Speed and ease of interaction:

At any given moment, Voiceshell is listening for input from a user. When it is reading a poem line aloud, it cannot be interacted with. With longer lines, this does negatively affect the speed of interaction.

Clarity and relevance:

Given the nature of poetry, a great degree of latitude can be afforded with this metric; having more clearly related responses to user input (at least occasionally) would likely benefit user understanding of the device.

Allowing for natural variations in speech:

Any input string is acceptable; the offline speech recognition is limited in its’ accuracy, particularly with non-standard phrasing, non-American accents, and interjections (“ums” and “ahhs”, for example).

Allow for errors and uncertainty:

Generally irrelevant to this project; serious errors require the entire device to be restarted to restore functionality, as there is no means to reset the software by voice. Uncertain interjection (“ums” and “ahhs”, for example) tend to interfere with accurate recognition of user speech.

Respect the user:

Certain lines of poetry contain content that may be upsetting or offensive; swearing and references of a sexual nature for example. Censoring these was determined to be against the spirit of the project (exhibiting poetry in a new format), however this becomes a consideration if the device were to be exhibited in certain public spaces where it is likely to cause offense.

Discussion

As the project came to a conclusion during this thesis, it was viewed by several members of the Faculty of Humanities and Social Sciences, giving the project its’ first test approximating real world conditions. This revealed a multitude of issues arisen from creating a custom software and hardware experience.

Inconsistent volume levels interfered with the speech recognition; Voiceshell would sometimes continue listening to extended conversations, and then have too much data to efficiently process. Alternatively, it would adapt to a higher ambient noise level, failing to hear direct statements to it. Limiting the time Voiceshell could listen for input, and potentially requiring an active "wake-word" (such as Assistant's "Ok, Google") to trigger listening would improve this. More difficult but possibly successful would be dynamically adapting the ambient noise level at all times, and using these levels to determine when an utterance to Voiceshell was being made. Considering the active and inactive states, as well as triggering, during the initial planning would have helped here. Having constant Internet connectivity to enable use of Google or Amazon for speech recognition would also enhance the success of the project.

Limitations of splitting the poetry by sentence, to punctuation, was also properly noted. Certain statements became excessively long, with others incredibly short. An inability to get more information on the line of poetry heard (such as the title and author), as well as not being able to hear the entire poem if desired, were noted as issues also. Whether or not certain potentially offensive lines are censored is also a consideration, best made once the location of exhibition is determined.

Ultimately, the state of the project at this point was met with approval, and work could continue going forward, after the conclusion of this thesis. The issues noted above influenced phases of the systematic approach to designing speech-experiences, and many of the early stage choices, such as in platform selection, are directly reflected in it. Continuing the project, beginning with a reconsideration of the existing work, will benefit from following the approach more specifically.

Chapter Thirteen

Discussion

Speech–based platforms for interacting with our computers are rapidly, dramatically increasing in prevalence (Edison Research, 2018; eMarketer Inc., 2017; Kinsella & Mutchler, 2018; VoiceLabs.co, 2017). Authors across the user experience design field are emphasising the importance of designing experiences for these platforms, evangelising the human-centric, user-friendly future of human–computer interaction they usher in, with discussions turning to just how these experiences should be realised (Borowska, 2017; Holland, 2018; Kucheria, 2018; Techlabs, 2017).

Informed by essays on post-humanism and post-gender from Hayles (1999) and Haraway (2016), after noting the broad considerations of authors and bloggers into the realm, and the depth of possible argument (Murphy, 2016; Reddy, 2017; Waddell, 2015), this research avoids directly addressing humanity and gender in artificial intelligence and associated interfaces. With a focus on design, such questions become too broad to cover in the scope of this thesis; nonetheless, they are important to consider. Electing to use specifically technical terms, and genderless 'it' pronouns where needed serve this evasion. These debates fall largely to platform vendors, who control the personality and "humanity" of their systems (Hempel, 2015), although that is not to undercut their importance.

The four largest platforms in the field were the only ones examined with any great depth. Much of the literature on the basic concepts in and the efficacy of speech are from a broader and less platform-specific position, or use systems built for that study. There are other platforms, such as Bixby from Samsung (Samsung, n.d.), or the Web Speech API (Mozilla, 2018), but these currently have significantly smaller market shares and numbers of active users and (in the case of Bixby) limited support across other devices (Cho, 2017; Deveria, 2018; Informa PLC., 2017). This does limit the scope when analysing vendor marketing and guidelines, but for the size and nature of this research is a justifiable trade-off. When examining these four platforms, certain key differences between vendors do come into question; for example, Apple's requirement that Siri experiences are extensions of existing applications, rather than standalone software (Apple, Inc., n.d.-f). This could result in fundamental differences in design recommendations; in practice, the guideline analysis proves this is minimal.

Examining existing mental models of human–computer interaction provides an avenue to reconsider some fundamentals in the context of the speech modality, and begin to satisfy John Maeda's thinking in this area (2017). He argues that the mental models grounding conversational user interfaces rely on complex graphical representations. This research finds comparable results but does not actively and wholly reject them as he proposes. Rather, they are seen as an additional grounding to build from, extrapolating the abstractions behind the graphical elements into metaphors that work in speech more effectively.

The presentation of data and files, and how the user understands and manages that structure is a key consideration. This understanding of the user forms the core of Raskin's (2000) "humane interface", and when interacting within a modality that embraces such a powerful human expression as speech (Chomsky, 1987), is a vital consideration. Nielsen and Gentner (1996) suggest that, with modern computing technology, information should be the atomic unit of user interaction. The research in this thesis agrees, finding that speech–based interfaces call for this simplified model. The push into contextualised information is positioned as heading to this point, and a more concerted emphasis to this is encouraged for designers in this field. The idea of the "semantic web" (Berners-Lee et al., 2001) begins to realise this, with practical examples emerging attached to the speech–based platforms of Apple and Google (Miller, 2015; Stegner, 2018). Information as a core, structural element of an interaction aligns with Johanna Drucker's thoughts in *Graphesis* (2014), where she posits that the interface itself, through its' affordances (Norman, 2013), provides as much information as the user does, through the modes and constraints imposed.

This ties into a broader, Systematic Approach to designing these experiences, one derived from exploration of the field historically, and in popular culture, where design solutions for many of the problems have existed for decades, predating the real-world technology in some cases. The major current platforms are looked at more closely, with the pros and cons of using each outlined. How to take ownership of and represent a unique brand on these ostensibly strongly vendor-owned platforms is also explored, with suggestions put forth to achieving this without traditional visual media; taking advantage of audio material and use of language. Existing principles and metaphors of human-computer interface are used as a jumping-off point, adaptations and new ideas for how they can develop when applied to speech have been found, particularly with regard to exposing the user to data and files.

In parallel to this broader research and the development of this approach and these heuristics, the fundamental ideas were put into practice with experimental work on a speech-based platform for the exhibition of poetry, "Voiceshell". Creating a piece for public exhibition, relying on a fixed corpus of material for presentation changes the nature of the experience, becoming more of an interactive art piece than a fully realised application, one built on a combination of custom hardware and software in order to maximise flexibility and customisation and achieve the project goals. This is largely distinct from the general idea of leveraging existing speech platforms presented otherwise in this thesis. The near-complete reliance on pre-recorded audio, and how to make that work with the interaction also limit the flow of the experience, restricting the kinds of response available. This does, however, exemplify Nielsen and Gentner's (1996) concepts of atomic units of information, being surfaced for the user; Voiceshell responds to user input abstractly, returning a combination of text and an audio file presented dynamically, read aloud, without any concept of the file system metaphor, or "surfing your hard drive" as those authors put it. This project, and its' requisite considerations of platform and interaction formed an early, embryonic form of the Systematic Approach, one which revealed key considerations otherwise missed, feeding back into the final outcome, and strengthening that approach as a useful material for designers first touching on this work.

Directions for Future Work and Research

There were limitations in scope and range of the research in this thesis, as outlined above. More extensive testing of the usability heuristics would benefit them and ensure their applicability and validity across a broad array of experiences.

More user testing, and testing designed specifically for speech–based experiences is another area of additional research, building on the core presented here. Evaluating user cognitive load (Tracy & Albers, 2006), for example, could provide valuable feedback to designers on ensuring the speed and ease of interaction, and the clarity of each step of the user flow.

Going beyond information as the key, 'atomic' unit of interaction presents an intriguing realm for further work also. It serves to primarily transform the 'Icon' component of the WIMP model (van Dam, 1997) to a model more appropriate speech. This also somewhat eliminates the need for a pointer; however, finding effective ways to translate the concepts of windows and menus requires more thought and research. Research into so-called "post-WIMP" interactions (van Dam, 1997), such as Geyer, Jetter, and Reiterer's "blended interaction" (2014), where they discuss conceptual integration between the natural, physical and the digital worlds provides potentially rich ground for further work as it relates to speech–based interaction.

Conclusion

The ultimate goal of this research is to present new ways of thinking about, approaches to designing, and ways of usability testing speech-based experiences. A set of usability heuristics created specifically for voice are put forward, and used to evaluate existing examples of these experiences, both from the real-world and fiction. The current ideas and attitudes in human–computer interaction for traditional modalities similarly provided foundations to propose new models of interface for speech, adapting what is already known to work. Analysing fictional examples and popular culture representations acknowledges that designers have already done work in tackling some of the problems facing the real-world technology now, and the inspiration and insight of these sources should not be overlooked. A distinct unwillingness to wholesale reject that which has come before drives much of this thesis, and a key takeaway for the reader should be to respect and remember the past when designing for the future.

More work and research in this area is absolutely needed, and within the context of this thesis, only the surface can realistically be scratched. Nevertheless, a foundation is presented here which offers sufficient background and scope to develop on further with additional research.

By remaining non-specific to a platform, designer-focused, and with the technical details kept at a high level, the hope is that this thesis can serve as an introduction and one of many “handbooks” to creators beginning to tackle the challenges presented by this burgeoning new field of user experience design.

List of References

Alphabet Inc. (2017). *10-K Annual Report.* Retrieved from https://abc.xyz/investor/pdf/20171231_alphabet_10K.pdf

Amazon.com, Inc. (2015, June 23). Amazon Echo Now Available to All Customers. Retrieved 30 June 2018, from https://www.businesswire.com/news/home/20150623005861/en/Amazon-Echo-Customers

Amazon.com, Inc. (2017). *Form 10-K*.

Amazon.com, Inc. (n.d.-a). Alexa Skills Kit - Build for Voice with Amazon. Retrieved 1 July 2018, from https://developer.amazon.com/alexa-skills-kit

Amazon.com, Inc. (n.d.-b). Alexa Voice Service. Retrieved 2 July 2018, from https://developer.amazon.com/alexa-voice-service

Amazon.com, Inc. (n.d.-c). Amazon Alexa Voice Design Guide. Retrieved 5 October 2017, from https://developer.amazon.com/designing-for-voice

Amazon.com, Inc. (n.d.-d). Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more. Retrieved 1 July 2018, from https://www.amazon.com/

Amazon.com, Inc. (n.d.-e). Echo & Alexa - Amazon Devices - Amazon Official Site. Retrieved 28 June 2018, from https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8&node=9818047011

Amazon.com, Inc. (n.d.-f). Talk to Alexa from the Alexa App. Retrieved 8 July 2018, from https://www.amazon.com/gp/help/customer/display.html?nodeId=202202210

Apple, Inc. (2017). *10-K Annual Report*. Retrieved from http://files.shareholder.com/downloads/AAPL/6335122040x0x962680/D18FAEFF-460A-4168-993D-A60CBA8ED209/_10-K_2017_As-Filed_.pdf

Apple, Inc. (2018a, May). Apple Style Guide. Retrieved from https://help.apple.com/applestyleguide/#/

Apple, Inc. (2018b, June). *WWDC 2018 Keynote - WWDC 2018 - Videos*. Presented at the WWDC 2018, San Jose Convention Center. Retrieved from developer.apple.com/videos/play/wwdc2018/101/

Apple, Inc. (n.d.-a). File System Basics. Retrieved 30 June 2018, from https://developer.apple.com/library/archive/documentation/FileManagement/Conceptual/FileSystemProgrammingGuide/FileSystemOverview/FileSystemOverview.html

Apple, Inc. (n.d.-b). HomePod. Retrieved 30 June 2018, from https://www.apple.com/homepod/

Apple, Inc. (n.d.-c). Human Interface Guidelines - Siri. Retrieved from https://developer.apple.com/ios/human-interface-guidelines/system-capabilities/siri/

Apple, Inc. (n.d.-d). iOS - Siri. Retrieved 28 June 2018, from https://www.apple.com/ios/siri/

Apple, Inc. (n.d.-e). Privacy. Retrieved 30 June 2018, from https://www.apple.com/nz/privacy/

Apple, Inc. (n.d.-f). SiriKit - Apple Developer. Retrieved 30 June 2018, from https://developer.apple.com/sirikit/

Apple, Inc. (n.d.-g). Vision Accessibility - Mac. Retrieved 7 July 2018, from https://www.apple.com/lae/accessibility/mac/vision/

Behne, C. (2014). *A Potpourri of Chomskyan Science*. Retrieved from http://wernercohn.com/Resources/behme_14_A-Potpourri-of-C.14%20copy.pdf

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 29–37.

Boeckx, C. (n.d.). Not Only Us – Cedric Boeckx – Inference. Retrieved 7 July 2018, from http://inference-review.com/article/not-only-us

Borowska, P. (2017, December 11). What is Conversational UI, and Why It's Important. Retrieved 11 July 2018, from https://designmodo.com/conversational-ui/

Bosker, B. (2013, January 22). SIRI RISING: The Inside Story Of Siri's Origins -- And Why She Could Overshadow The iPhone. *Huffington Post.* Retrieved from https://www.huffingtonpost.com/2013/01/22/siri-do-engine-apple-iphone_n_2499165.html

Bourguet, M.-L. (2006). An Overview of Multimodal Interaction Techniques and Applications. In *Encyclopedia of Human Computer Interaction* (pp. 451–455). IGI Global.

Breeze, R. (2013). Corporate Discourse. London: Bloomsbury Academic. Retrieved from http://helicon.vuw.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=603986&site=ehost-live

Bruell, A. (2017, December 7). Visa Spent a Year Developing a 'Signature Sound'. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/visa-spent-a-year-developing-a-signature-sound-1512644401

Cho, J. (2017, November 6). Samsung Bixby's Global Market Share of 12.7% Expected to Slide to 6.5% in 2020. Retrieved 11 July 2018, from http://www.businesskorea.co.kr/news/articleView.html?idxno=19728

Chomsky Is Citation Champ. (1992, April 15). Retrieved 7 July 2018, from http://news.mit.edu/1992/citation-0415

Chomsky, N. (1987). *The Chomsky Reader*. Pantheon Books.

Clarke, A. C. (1972). *The Lost Worlds of 2001*. United Kingdom: New American Library.

Clauser, G. (2018, June 7). What Is Alexa? What Is the Amazon Echo, and Should You Get O. Retrieved 1 July 2018, from https://thewirecutter.com/reviews/what-is-alexa-what-is-the-amazon-echo-and-should-you-get-one/

Conradt, S. (2017, November 16). Earcons. Retrieved 18 June 2018, from https://qz.com/email/quartz-obsession/1130876/

Cowie, R. (n.d.). Ethical issues in affective computing, 18.

Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches* (4th ed). Thousand Oaks: SAGE Publications.

Deadly embrace. (2000, March 30). *The Economist*. Retrieved from https://www.economist.com/business/2000/03/30/deadly-embrace

Deveria, A. (2018, July 11). Can I use... Support tables for HTML5, CSS3, etc. Retrieved 11 July 2018, from https://caniuse.com/#search=speech

Diggs, J., McCarron, S., Cooper, M., Schwedtfeger, R., & Craig, J. (2017, December 14). Accessible Rich Internet Applications (WAI-ARIA) 1.1. Retrieved 7 July 2018, from https://www.w3.org/TR/wai-aria-1.1/

Drucker, J. (2014). *Graphesis: Visual Forms of Knowledge Production*. Cambridge, US: Harvard University Press. Retrieved from https://monoskop.org/images/2/2a/Drucker_Johanna_Graphesis_Visual_Forms_of_Knowledge_Production.pdf

Edison Research. (2018). *The Smart Audio Report*. Retrieved from https://www.nationalpublicmedia.com/wp-content/uploads/2018/01/The-Smart-Audio-Report-from-NPR-and-Edison-Research-Fall-Winter-2017.pdf

eMarketer Inc. (2017, August 5). Alexa, Say What?! Voice-Enabled Speaker Usage to Grow Nearly 130% This Year - eMarketer. Retrieved from https://www.emarketer.com/Article/Alexa-Say-What-Voice-Enabled-Speaker-Usage-Grow-Nearly-130-This-Year/1015812

Emerson, M. J., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, 48(1), 148–168. https://doi.org/10.1016/S0749-596X(02)00511-9

Evans, V. (2014, December 4). The evidence is in: there is no language instinct. Retrieved 7 July 2018, from https://aeon.co/essays/the-evidence-is-in-there-is-no-language-instinct

Farfan, B. (2018, April 15). Here Is a Look at the Mission Statement of Amazon.com. Retrieved 1 July 2018, from https://www.thebalancesmb.com/amazon-mission-statement-4068548

Faulkner, C. (2018, January 26). Google Pixel Buds. Retrieved 1 June 2018, from https://www.techradar.com/reviews/google-pixel-buds

Foresman, C. (2016, September 10). How Star Trek artists imagined the iPad... nearly 30 years ago. Retrieved 28 May 2018, from https://arstechnica.com/apple/news/2010/08/how-star-trek-artists-imagined-the-ipad-23-years-ago.ars

Frakes, J. (1996). *Star Trek: First Contact.*

Franklin, J., & Maal, M. (2018). *The Talking Shop* (p. 20). OC&C Strategy Consultants.

Friedman, K. (2008). Research into, by and for design. *Journal of Visual Art Practice*, 7(2), 153–160. https://doi.org/10.1386/jvap.7.2.153_1

Gentner, D., & Nielsen, J. (1996). The Anti-Mac User Interface. *Nielsen Norman Group*. Retrieved from https://www.nngroup.com/articles/anti-mac-interface/

Giangola, J. (2017, August 8). Conversation Design: Speaking the Same Language. Retrieved 2 July 2018, from https://design.google/library/conversation-design-speaking-same-language/

Google Inc. (2018, June 5). Overview | Actions on Google. Retrieved 10 July 2018, from https://developers.google.com/actions/extending-the-assistant

Google Inc. (n.d.-a). Google Assistant - Just Say “Hey Google” and Make Google Do It. Retrieved 28 June 2018, from https://assistant.google.com/

Google Inc. (n.d.-b). Google Privacy | Why data protection matters. Retrieved 10 July 2018, from http://privacy.google.com/intl/en_ALL/your-data.html

GrantTree. (2018, January 29). GrantTree Blog | Technology | Virtual Assistants. Retrieved 6 July 2018, from https://granttree.co.uk/embarrassed-use-virtual-assistants/

Griffin, M. (1984). Jeopardy! [Game-show].

Grothaus, M. (2018, June 7). All the people Apple just pissed off to better protect your privacy. Retrieved 30 June 2018, from https://www.fastcompany.com/40581691/all-the-people-apple-just-pissed-off-to-better-protect-your-privacy

Haraway, D. (2016). A Cyborg Manifesto. University of Minnesota Press. Retrieved from https://warwick.ac.uk/fac/arts/english/currentstudents/undergraduate/modules/fictionnownarrativemediaandtheoryinthe21stcentury/manifestly_haraway_----_a_cyborg_manifesto_science_technology_and_socialist-feminism_in_the_....pdf

Hart, H. (2013, December 18). Designing A Future Of Comfort, Color, And Gorgeous Gadgets In “Her”. Retrieved 28 May 2018, from https://www.fastcompany.com/3023518/designing-a-future-of-comfort-color-and-gorgeous-gadgets-in-her

Harvey, E. (2018, June 29). How Smart Speakers Are Changing the Way We Listen to Music. Retrieved 2 July 2018, from https://pitchfork.com/features/article/how-smart-speakers-are-changing-the-way-we-listen-to-music/

Hauptmann, A., & Rudnicky, A. I. (1990). A Comparison of Speech vs Typed Input. Retrieved from http://repository.cmu.edu/compsci/1414/

Hayles, N. K. (1999). *How we became posthuman: virtual bodies in cybernetics, literature, and informatics*. Chicago, Ill: University of Chicago Press.

Hempel, J. (2015, October 28). Siri and Cortana Sound Like Ladies Because of Sexism. Retrieved 11 July 2018, from https://www.wired.com/2015/10/why-siri-cortana-voice-interfaces-sound-female-sexism/

Holland, B. (2018, April 5). The Rise Of Intelligent Conversational UI. Retrieved 11 July 2018, from https://www.smashingmagazine.com/2018/04/rise-intelligent-conversational-ui/

Holly, R. (2017, May 10). Microsoft says Cortana now has more than 141 million monthly users. Retrieved 9 July 2018, from https://www.windowscentral.com/cortana-now-has-over-141-million-users-every-month

IMDb.com, Inc. (n.d.). Jeopardy! (TV Series 1984–). Retrieved 4 July 2018, from http://www.imdb.com/title/tt0159881/fullcredits

Informa PLC. (2017). *Digital Customer Insights 2017: Smart Living Technology Analysis*.

Jacobsson Purewal, S. (2016, October 4). The difference between Google Now and Google Assistant. Retrieved 30 June 2018, from https://www.cnet.com/how-to/the-difference-between-google-now-and-google-assistant/

Jetter, H.-C., Reiterer, H., & Geyer, F. (2014). Blended Interaction: understanding natural human–computer interaction in post-WIMP interactive spaces. *Personal and Ubiquitous Computing*, 18(5), 1139–1158. https://doi.org/10.1007/s00779-013-0725-4

Johnson, K. (2017, June 7). How ‘Star Trek’ inspired Amazon’s Alexa. Retrieved 8 March 2018, from https://venturebeat.com/2017/06/07/how-star-trek-inspired-amazons-alexa/

Johnson, K. (2018, June 21). Google Home now answers follow-up questions without ‘OK, Google’ wake word. Retrieved 25 June 2018, from https://venturebeat.com/2018/06/21/google-home-now-answers-follow-up-questions-without-ok-google-wake-word/

Jonze, S. (2013). *Her*.

Kearney, M., Gash, D., & Boxhall, A. (2018, July 2). Introduction to ARIA | Web Fundamentals. Retrieved 7 July 2018, from https://developers.google.com/web/fundamentals/accessibility/semantics-aria/

Keizer, G. (2018, July 5). June 2018 data: Windows 10 user share rises but still trails Windows 7 | Computerworld. Retrieved 6 July 2018, from https://www.computerworld.com/article/3199373/windows-pcs/windows-by-the-numbers-windows-10-gains-ground-but-still-trails-windows-7.html

Kepes, B. (2013, December 4). Google Users - You’re The Product, Not The Customer. Retrieved 10 July 2018, from https://www.forbes.com/sites/benkepes/2013/12/04/google-users-youre-the-product-not-the-customer/

Kessler, S. (2011, September 9). 8 Star Trek Gadgets That Are No Longer Fiction. Retrieved 28 May 2018, from https://mashable.com/2011/09/08/star-trek-gadgets/

Kinsella, B. (2018a, January 2). Apple Should Do 3 Things in 2018 to Make Siri Better for Everyone. Retrieved 30 June 2018, from https://voicebot.ai/2018/01/02/apple-3-things-2018-make-siri-better-everyone/

Kinsella, B. (2018b, March 2). Microsoft Cortana Available on 400 Million Devices Gets a New Product Leader. Retrieved 9 July 2018, from https://voicebot.ai/2018/03/02/microsoft-cortana-available-400-million-devices-gets-new-product-leader/

Kinsella, B. (2018c, April 3). Over Half of Smartphone Owners Use Voice Assistants, Siri Leads the Pack. Retrieved 10 July 2018, from https://voicebot.ai/2018/04/03/over-half-of-smartphone-owners-use-voice-assistants-siri-leads-the-pack/

Kinsella, B. (2018d, May 8). Google Assistant to Get Follow Up Mode and Compound Requests. Retrieved 25 June 2018, from https://voicebot.ai/2018/05/08/google-assistant-to-get-follow-up-mode-and-compound-requests/

Kinsella, B., & Mutchler, A. (2018). *Smart Speaker Consumer Adoption Report*. Voicebot.AI.

Klara, R. (2016, February 17). How Nokia's Original Ringtone Became the Most Played Tune on the Planet. Retrieved 18 June 2018, from https://www.adweek.com/brand-marketing/how-nokia-s-original-ringtone-became-most-played-tune-planet-169615/

Kolbe, W. (1991, September 30). Darmok. *Star Trek: The Next Generation*.

Kubrick, S. (1968). *2001: A Space Odyssey*.

Kucheria, A. (2018, March 21). Designing for Voice. Retrieved 11 July 2018, from https://medium.muz.li/designing-for-voice-a84aeeed6425

Lewis, J. R. (2011). *Practical Speech User Interface Design*. CRC Press.

Maeda, J. (2017, March 11). *Design in Tech Report 2017*. Retrieved from https://designintechreport.wordpress.com/2017/03/11/design-in-tech-report-2017/

Martin, T. (2018, January 22). Now you can talk to Alexa in Amazon’s Alexa app. Retrieved 8 July 2018, from https://www.cnet.com/how-to/you-can-finally-talk-to-amazon-alexa-in-the-alexa-app/

Martonik, A. (2018, January 19). Google Now is being left to wither and die as Google Assistant takes the focus. Retrieved 30 June 2018, from https://www.androidcentral.com/google-now-dies-make-way-google-assistant

McTear, M., Callejas, Z., & Griol, D. (2016). Conversational Interfaces: Past and Present. In *The Conversational Interface* (pp. 51–72). Springer, Cham. https://doi.org/10.1007/978-3-319-32967-3_4

Microsoft Corporation. (2017a). *Form 10-K*. Retrieved from https://microsoft.gcs-web.com/static-files/0493db0a-54b7-4aca-87d4-1fb8f6beb456

Microsoft Corporation. (2017b, January 12). What is Cortana? Retrieved 6 July 2018, from https://support.microsoft.com/en-nz/help/17214/windows-10-what-is

Microsoft Corporation. (2017c, February 8). Cortana Design Guidelines. Retrieved from https://docs.microsoft.com/en-us/cortana/voice-commands/voicecommand-design-guidelines

Microsoft Corporation. (2017d, March 30). Cortana Skills Kit FAQ - Cortana Skills. Retrieved 8 July 2018, from https://docs.microsoft.com/en-us/cortana/skills/faq

Microsoft Corporation. (2017e, December 1). Hear text read aloud with Narrator. Retrieved 7 July 2018, from https://support.microsoft.com/en-us/help/17173/windows-10-hear-text-read-aloud

Microsoft Corporation. (n.d.-a). About - Microsoft. Retrieved 9 July 2018, from https://www.microsoft.com/en-us/about

Microsoft Corporation. (n.d.-b). Personal Digital Assistant - Cortana Home Assistant - Microsoft. Retrieved 8 July 2018, from https://www.microsoft.com/en-us/cortana

Miller, R. (2015, June 8). Apple takes on Google Now with iOS 9's proactive assistance - The Verge. Retrieved 13 July 2018, from https://www.theverge.com/2015/6/8/8732215/proactive-apple-ios-google-now-wwdc-2015

Mitchell, W. J., Ho, C.-C., Patel, H., & MacDorman, K. F. (2011). Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior,* 27(1), 402–412. https://doi.org/10.1016/j.chb.2010.09.002

Moon, M. (2018, March 10). Amazon makes it easier to give Alexa follow-up commands. Retrieved 25 June 2018, from https://www.engadget.com/2018/03/10/amazon-alexa-follow-up-mode/

Mozilla. (2018, February 20). Web Speech API. Retrieved 11 July 2018, from https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API

Murphy, M. (2016, March 17). It's time for robots to have their own pronouns. Retrieved 20 May 2018, from https://qz.com/629535/we-need-new-pronouns-for-robots/

Nass, C. (2010, August 28). Sweet Talking Your Computer --- Why people treat devices like humans; saying nice things to a machine to protect its 'feelings'. *Wall Street Journal, Eastern Edition; New York, N.Y.*, p. W.3.

NetApplications.com. (2018). Operating system market share. Retrieved 6 July 2018, from https://netmarketshare.com/operating-system-market-share.aspx

Nielsen, J. (1995, January 1). 10 Heuristics for User Interface Design: Article by Jakob Nielsen. Retrieved 21 May 2018, from https://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J., & Mack, R. L. (1994). *Usability inspection methods.* New York: Wiley.

Norman, D. A. (2013). *The design of everyday things* (Revised and expanded edition). New York, New York: Basic Books.

O'Herlihy, M. (1967, January 26). Tomorrow is Yesterday. Star Trek. NBC.

Olive, J. P. (1998). 'The Talking Computer' Text to Speech Synthesis. In *Hal's Legacy: 2001's Computer as Dream and Reality* (p. 384). MIT Press. Retrieved from https://web.archive.org/web/20050322025655/http://www.mitpress.mit.edu:80/e-books/hal/chap6/six1.html

Parush, A. (2005). Speech-Based Interaction in Multitask Conditions: Impact of Prompt Modality. *Human Factors; Santa Monica*, 47(3), 591–597.

Peters, L. (2014, April 25). Nokia Tune: more than just a ringtone - Lumia Conversations. Retrieved 18 June 2018, from https://web.archive.org/web/20150413013830/http://lumiaconversations.microsoft.com/2014/04/25/nokia-tune-just-ringtone

Pew Research Center. (2018, February 5). Mobile Fact Sheet. Retrieved 8 March 2018, from http://www.pewinternet.org/fact-sheet/mobile/

Pichai, S. (2018, May). *Google Keynote*. Presented at the Google I/O. Retrieved from https://events.google.com/io/schedule/?section=may-8&sid=9c52752e-9a83-4c87-9a3e-46f409a16578

Pierce, D. (2017, April 7). Apple's AirPods Aren't For Music, They're For Siri. Retrieved 28 May 2018, from https://www.wired.com/2017/04/review-apple-air-pods/

Polyani, M. (1966). *The Tacit Dimension*. Garden City, New York: Doubleday & Company, Inc. Retrieved from https://monoskop.org/images/1/11/Polanyi_Michael_The_Tacit_Dimension.pdf

Porter, J. (2017, November 22). Apple AirPods. Retrieved 28 May 2018, from https://www.techradar.com/reviews/audio-visual/hi-fi-and-audio/headphones/apple-airpods-1328292/review

Prospero, M. (2018, June 27). How to Use Alexa on Your iPhone. Retrieved 8 July 2018, from https://www.tomsguide.com/us/alexa-iphone-how-to,news-24688.html

Quin, L. (2016, October 11). Extensible Markup Language (XML). Retrieved from https://www.w3.org/XML/

Ranj, B. (2018, March). Apple's totally wireless Airpods are full of compromises, but here's why I use them every day. Retrieved 28 May 2018, from http://www.businessinsider.com/are-apple-airpods-worth-it-2018-3

Raskin, J. (2000). *The Humane Interface*. ACM Press.

Reddy, D. (2017, April 17). Breaking Gender Bias in Artificial Intelligence. Retrieved 20 May 2018, from https://medium.com/my-ally/breaking-gender-bias-in-artificial-intelligence-c3c143038c20

Salkind, N. J. (2010). Pragmatic Study. In *Sage Research Methods*. Retrieved from http://methods.sagepub.com/reference/encyc-of-research-design/n326.xml

Samsung. (n.d.). Meet Bixby - A Smarter Way To Use Your Phone. Retrieved 11 July 2018, from https://www.samsung.com/us/explore/bixby/

Sauer, C. (n.d.). If a Service is Free, You Are the Product | Webtrekk. Retrieved 10 July 2018, from https://www.webtrekk.com/en/knowledge/webtrekk-blog/google-analytics-data-advertising/

Sciretta, P. (2011, October 12). LOL: Siri Responds To ’2001: A Space Odyssey’ Reference on the Apple iPhone 4s. Retrieved 8 July 2018, from http://www.slashfilm.com/lol-apple-iphone-4ss-siri-responds-2001-space-odyssey-reference/

Seifert, D. (2017, October 20). Harman Kardon Invoke review: Cortana gets a speaker of its own. Retrieved 8 July 2018, from https://www.theverge.com/2017/10/20/16505468/harman-kardon-invoke-cortana-microsoft-smart-speaker-review

Singh, P. (2015, November 1). Conversational Interfaces and Return of Command Line. Retrieved 11 July 2018, from https://medium.com/appy-thoughts/conversational-interfaces-and-return-of-command-line-73c1bc5bbc5d

Sony Pictures Television. (2016). Jeopardy! Sony Pictures Television. Retrieved from https://www.amazon.com/Sony-Pictures-Television-Jeopardy/dp/B019G0M2WS

Sreeraj. (2016, October 5). “Conversational UI” — The slow death of Creative Interfaces? Retrieved 11 July 2018, from https://blog.prototypr.io/conversational-ui-the-slow-death-of-creative-interfaces-84d7d5c32019

Stallman, R. M., & MacKenzie, D. (2017). ls(1) - Linux manual page. Retrieved 4 July 2018, from http://man7.org/linux/man-pages/man1/ls.1.html

Stapleton, A. (2015, November 17). Virtual Personal Assistants: Future Gatekeeper to Your Attention? Retrieved from https://opusresearch.net/wordpress/2015/11/17/virtual-personal-assistants-future-gatekeeper-to-your-attention/

Statt, N. (2018, March 9). Amazon adds follow-up mode for Alexa to let you make back-to-back requests. Retrieved 25 June 2018, from https://www.theverge.com/2018/3/9/17101330/amazon-alexa-follow-up-mode-back-to-back-requests

Steffen, A. (2015, May 21). Designing the future of “Her”. Retrieved 28 May 2018, from https://thenearlynow.com/designing-the-future-of-her-b865347a8895

Stegner, B. (2018, May 11). Proactive Siri vs. Google Feed: How They Compare and Which Is Better. Retrieved 13 July 2018, from https://www.makeuseof.com/tag/proactive-siri-vs-google-feed/

Stephenson, N. (1999). *In the beginning... was the command line.* Avon books New York. Retrieved from http://studiowhitney.gearnine.com/temp/command.pdf

Stevens, J. (2013, March 4). What is an Earcon? | An Audio Branding Blog By TreBrand. Retrieved 29 June 2018, from http://trebrand.com/blog/what-is-an-earcon/

Štolfa, T. (2016, March 10). The Future of Conversational UI Belongs to Hybrid Interfaces. Retrieved 11 July 2018, from https://medium.com/the-layer/the-future-of-conversational-ui-belongs-to-hybrid-interfaces-8a228de0bdb5

Stone, B., & Soper, S. (2014, November 7). Amazon Unveils a Listening, Talking, Music-Playing Speaker for Your Home. *Bloomberg.Com*. Retrieved from https://www.bloomberg.com/news/articles/2014-11-06/amazon-echo-is-a-listening-talking-music-playing-speaker-for-your-home

Stuff Limited. (2018, February 2). Amazon Echo, Alexa and Stuff: How to listen to the news with our flash briefing. *Stuff.Co.Nz.* Retrieved from http://www.stuff.co.nz/about-stuff/101015196/amazon-echo-alexa-and-stuff-how-to-listen-to-the-news-with-our-flash-briefing

Sullivan, M. (2017, September 11). Apple Explains How It’s Making Siri Smart Without Endangering User Privacy. Retrieved 30 June 2018, from https://www.

fastcompany.com/40443055/apple-explains-how-its-making-siri-smart-without-endangering-user-privacy

Sumikawa, D. A. (1985). *Guidelines for the integration of audio cues into computer user interfaces* (No. UCRL-53656). Lawrence Livermore National Lab., CA (USA). Retrieved from https://www.osti.gov/biblio/5475406

Taka, D. (2017, January 31). Understanding iOS Proactive Suggestions. Retrieved 30 June 2018, from https://www.sitepoint.com/understanding-ios-pro-active-suggestions/

Techlabs, M. (2017, October 13). What is a Conversational UI and why does it matter? Retrieved 11 July 2018, from https://chatbotsmagazine.com/what-is-a-conversational-ui-and-why-it-matters-de358507b9a2

The Rocket Science Group. (2018). MailChimp Voice & Tone. Retrieved from voiceandtone.com

The United States Department of Justice. (2015, October 16). U.S. v. Microsoft Corporation [Browser and Middleware] | ATR | Department of Justice. Retrieved 8 July 2018, from https://www.justice.gov/atr/case/us-v-microsoft-corporation-browser-and-middleware

Townsend, T. (2017, March 11). Now that Android is getting Google Assistant, here's how it works. Retrieved 30 June 2018, from https://www.recode.net/2017/3/11/14850438/google-assistant-android-how-work

Tracy, J. P., & Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. In *Annual Conference-society for technical communication* (Vol. 53, p. 256). Retrieved from https://www.researchgate.net/profile/Michael_Albers/publication/253713707_Measuring_Cognitive_Load_to_Test_the_Usability_of_Web_Sites/links/55ef096608ae199d47bff6cd.pdf

Tuck, M. (2001, August 13). The Real History of the GUI. Retrieved 19 November 2017, from https://www.sitepoint.com/real-history-gui/

U.S. Securities and Exchange Commision. (2009, June 26). SEC.gov | Form 10-K. Retrieved 9 July 2018, from https://www.sec.gov/fast-answers/answers-form10khtm.html

Valkenburg, R., & Dorst, K. (1998). The reflective practice of design teams. *Design Studies*, 19(3), 249–271. https://doi.org/10.1016/S0142-694X(98)00011-8

van Camp, J. (2018, January 29). Harmon Kardon Invoke Review: You Can't Count Microsoft's Cortana Out (Yet). Retrieved 8 July 2018, from https://www.wired.com/review/review-harman-kardon-invoke-speaker-with-cortana/

van Dam, A. (1997). Post-WIMP User Interfaces. *Communications of the ACM, 40*. Retrieved from http://delivery.acm.org/10.1145/260000/253708/p63-van_dam.pdf

Velazco, C. (2017, November 25). Google Pixel Buds review: You (and Google) can do better. Retrieved 1 June 2018, from https://www.engadget.com/2017/11/25/google-pixel-buds-review/

Vice Media LLC. (2014). Vice Style Guide. Retrieved from https://www.scribd.com/doc/240864422/Vice-Style-Guide

VoiceLabs.co. (2017). *The 2017 Voice Report Executive Summary.* Retrieved from https://s3-us-west-1.amazonaws.com/voicelabs/report/vl-voice-report-exec-summary_final.pdf

W3Techs. (2018, June). Usage Statistics and Market Share of Linux for Websites, June 2018. Retrieved 4 June 2018, from https://w3techs.com/technologies/details/os-linux/all/all

Waddell, K. (2015, December 15). We Need a New Pronoun for Artificial Intelligence. *The Atlantic.* Retrieved from https://www.theatlantic.com/technology/archive/2015/12/we-need-a-new-pronoun-for-ai/420378/

Wang, E. M., & Huang, A. Y. (2000). A Study on Basic Metaphors in Human-Computer Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 44*(1), 140–143. https://doi.org/10.1177/154193120004400137

Warren, T. (2014, April 2). The story of Cortana, Microsoft's Siri killer. Retrieved 27 May 2018, from https://www.theverge.com/2014/4/2/5570866/cortana-windows-phone-8-1-digital-assistant